

**PREDICTING MELANOMA RISK FROM ELECTRONIC HEALTH
RECORDS WITH MACHINE LEARNING TECHNIQUES**

by

Aaron N. Richter

A Dissertation Submitted to the Faculty of
The College of Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

Florida Atlantic University

Boca Raton, FL

August 2019

Copyright 2019 by Aaron N. Richter

PREDICTING MELANOMA RISK FROM ELECTRONIC HEALTH
RECORDS WITH MACHINE LEARNING TECHNIQUES

by

Aaron N. Richter

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Taghi M. Khoshgoftaar, Department of Computer and Electrical Engineering and Computer Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

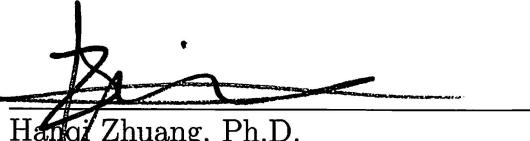
SUPERVISORY COMMITTEE:



Taghi M. Khoshgoftaar, Ph.D.
Dissertation Advisor



Mehrdad Nojoumian, Ph.D.



Hanchi Zhuang, Ph.D.
Interim Chair, Department of Computer
and Electrical Engineering and Com-
puter Science



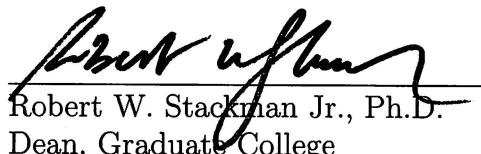
Dingding Wang, Ph.D.



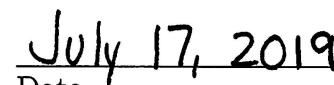
Xingquan Zhu, Ph.D.



Stella N. Batalama, Ph.D.
Dean, The College of Engineering and
Computer Science



Robert W. Stackman Jr., Ph.D.
Dean, Graduate College



July 17, 2019

Date

ACKNOWLEDGEMENTS

First, I want to thank my Creator, for the vision, wisdom, and strength to complete this dissertation. I thank my parents for instilling in me the value of education and hard work, and continually encouraging me throughout my studies.

I am very grateful for my Ph.D. advisor, Dr. Taghi M. Khoshgoftaar, who saw potential in me, guided me along the way, and introduced me to many academic and professional opportunities. I would also like to thank Dr. Nojoumian, Dr. Wang, and Dr. Zhu for serving on my supervisory committee. Thank you to various members of the FAU Data Mining and Machine Learning Laboratory for their collaboration and reviews of my work over the years.

To those that have mentored and taught me: Jeff Greene, Rich Kroll, Michael Crawford. Thank you for your guidance and inspiration. To the folks at Modernizing Medicine: Daniel Cane, Dr. Michael Sherling, Alice Rose, Robert Hasty. Thank you for the opportunity for, and support of, this research.

A special thanks to the places where most of my work was performed: Switchbox Coffee, Mane Coffee, and the Seed. The copious amounts of tea and kombucha powered me along the way.

ABSTRACT

Author: Aaron N. Richter
Title: Predicting Melanoma Risk from Electronic Health Records with Machine Learning Techniques
Institution: Florida Atlantic University
Dissertation Advisor: Dr. Taghi M. Khoshgoftaar
Degree: Doctor of Philosophy
Year: 2019

Melanoma is one of the fastest growing cancers in the world, and can affect patients earlier in life than most other cancers. Therefore, it is imperative to be able to identify patients at high risk for melanoma and enroll them in screening programs to detect the cancer early. Electronic health records collect an enormous amount of data about real-world patient encounters, treatments, and outcomes. This data can be mined to increase our understanding of melanoma as well as build personalized models to predict risk of developing the cancer. Cancer risk models built from structured clinical data are limited in current research, with most studies involving just a few variables from institutional databases or registries. This dissertation presents data processing and machine learning approaches to build melanoma risk models from a large database of de-identified electronic health records. The database contains consistently captured structured data, enabling the extraction of hundreds of thousands of data points each from millions of patient records. Several experiments are performed to build effective models, particularly to predict sentinel lymph node metastasis in known melanoma patients and to predict individual risk of developing melanoma. Data for these models suffer from high dimensionality and class imbalance.

Thus, classifiers such as logistic regression, support vector machines, random forest, and XGBoost are combined with advanced modeling techniques such as feature selection and data sampling. Risk factors are evaluated using regression model weights and decision trees, while personalized predictions are provided through random forest decomposition and Shapley additive explanations. Random undersampling on the melanoma risk dataset shows that many majority samples can be removed without a decrease in model performance. To determine how much data is truly needed, we explore learning curve approximation methods on the melanoma data and three publicly-available large-scale biomedical datasets. We apply an inverse power law model as well as introduce a novel semi-supervised curve creation method that utilizes a small amount of labeled data.

PREDICTING MELANOMA RISK FROM ELECTRONIC HEALTH RECORDS WITH MACHINE LEARNING TECHNIQUES

List of Tables	xi
List of Figures	xiv
List of Equations	xvii
1 Introduction	1
1.1 Background and Motivation	1
1.1.1 Melanoma	1
1.1.2 Electronic Health Records	2
1.1.3 Machine Learning and Big Data	3
1.1.4 Limited Data and Limited Labels	5
1.2 Contributions	6
1.3 Dissertation Structure	7
2 Theory and Methodology	8
2.1 Machine Learning	8
2.1.1 Algorithms	9
2.1.2 Dimensionality Reduction	15
2.1.3 Data Sampling	17
2.1.4 Interpretability	18
2.2 Experimental Design	19
2.2.1 Metrics and Evaluation	19
2.2.2 Statistical Tests	23

2.3	Data Engineering and Infrastructure	23
2.3.1	Hadoop and Spark	23
2.3.2	Cloud Computing	26
2.4	Software	28
3	Clinical Data Preparation	29
3.1	Background and Motivation	29
3.1.1	Data Sources and Features	29
3.2	MAMEL Dataset	33
3.2.1	Modernizing Medicine EHR	33
3.2.2	Data Processing Architecture	34
3.3	Exploratory Data Analysis	36
3.4	Discussion	47
3.4.1	Related Datasets	47
3.4.2	Necessity of Structured and Available Clinical Data	48
3.5	Chapter Summary	50
4	Predicting Sentinel Lymph Node Metastasis in Melanoma	51
4.1	Background and Motivation	51
4.2	Related Works	53
4.3	Materials and Methods	55
4.4	Results and Discussion	60
4.5	Chapter Summary	66
5	Predicting Melanoma Risk	68
5.1	Background and Motivation	68
5.2	Literature Review	70
5.2.1	Methodology	70

5.2.2	Models in Practice	71
5.2.3	Cancer Risk Models	74
5.2.4	Melanoma Risk Models	80
5.2.5	Section Summary	81
5.3	Clinical Risk Model	82
5.3.1	Materials and Methods	82
5.3.2	Results	91
5.3.3	Discussion	93
5.3.4	Section Summary	95
5.4	Advanced Machine Learning Techniques	96
5.4.1	Methods	96
5.4.2	Results	98
5.4.3	Section Summary	104
5.5	Interpretability	105
5.5.1	Global Interpretability	105
5.5.2	Local Interpretability	108
5.5.3	Section Summary	111
5.6	Chapter Summary	113
6	Learning from Limited Data	116
6.1	Background and Motivation	116
6.2	Limited Positive Samples	117
6.2.1	Related Works	118
6.2.2	Materials and Methods	119
6.2.3	Results	125
6.2.4	Section Summary	134
6.3	Learning Curve Approximation with Limited Labels	136

6.3.1	Related Works	137
6.3.2	Data	139
6.3.3	Methods	142
6.3.4	Results	145
6.3.5	Section Summary	154
6.4	Chapter Summary	155
7	Conclusions and Future Work	156
7.1	Structured Clinical Data	156
7.2	Sentinel Lymph Node Metastasis	156
7.3	Melanoma Risk	157
7.4	Learning from Limited Data	158
	Bibliography	159

LIST OF TABLES

2.1	Confusion matrix	20
3.1	High-level data elements	34
3.2	Melanoma subtypes	38
3.3	Demographics	38
3.4	Top chief complaints	43
3.5	Top ROS questions	44
3.6	Top exam elements	44
4.1	Demographics	56
4.2	Tumor characteristics	57
4.3	Class distributions	57
4.4	Model configurations	62
4.5	Validation results	63
4.6	LR model coefficients: Full dataset	64
4.7	LR model coefficients: $\leq 1\text{mm}$	65
5.1	Data elements	85
5.2	EC2 instance types	90
5.3	Patient population	91
5.4	Average results for each dataset and classifier	92
5.5	Additional metrics	92
5.6	Dataset statistics	98

5.7	Demographics and clinical characteristics	99
5.8	ANOVA: All models	100
5.9	ANOVA: RF	100
5.10	HSD: RF	101
5.11	Classification results	103
5.12	New vs. established patient results	104
5.13	LR model weights: No history	106
5.14	LR model weights: History	106
5.15	LR model weights: 10 features	107
5.16	RF: Top 15 feature importances	109
5.17	Example prediction: Positive class	110
5.18	Example prediction: Negative class	110
6.1	Hyperparameter grids	121
6.2	Negative samples by dataset and RUS Ratio	122
6.3	ANOVA: Dataset/Classifier/RUS Ratio	127
6.4	HSD: RUS Ratio	129
6.5	HSD: Classifier	129
6.6	ANOVA: Dataset/Classifier/RUS	130
6.7	HSD: Classifier/RUS interaction	130
6.8	HSD: Dataset/RUS interaction	131
6.9	EC2 instance types	133
6.10	Datasets	139
6.11	Sampling schedules	146
6.12	Best fitting inverse power law curves	150

6.13 Curve slopes by region	153
---------------------------------------	-----

LIST OF FIGURES

2.1	Example Kaplan-Meier curve	11
2.2	Example decision tree	13
2.3	Basic framework of an SVM	14
2.4	Basic framework of an ANN	16
2.5	Example ROC curve	22
2.6	Predictive modeling pipeline	22
2.7	HDFS architecture	24
2.8	MapReduce programming paradigm	25
2.9	Spark architecture	26
3.1	Big data processing architecture	35
3.2	Medical conditions (ICD10)	39
3.3	Family history, procedures, and social history (SNOMED)	40
3.4	Pairwise completion rates of cancer variables	42
3.5	Top morphologies by diagnosis	45
3.6	Top procedures by diagnosis	46
4.1	Nomogram model online form	53
4.2	Probability of SLN metastasis vs. tumor thickness	59
4.3	ROC curves for the benchmark model	60
4.4	Predicted probabilities of each sample using the benchmark model . .	61
4.5	AUC compared to benchmark	62

4.6	Sensitivity and specificity compared to benchmark	63
5.1	Example nomogram	73
5.2	Feature selection and model algorithm methods	75
5.3	Variable missingess	87
5.4	Sparse matrix creation process	88
5.5	Example ML pipeline	90
5.6	Average results for each classifier and dataset	93
5.7	Historical visit distributions	99
5.8	AUC of each model configuration	100
5.9	ROC curves for each classifier	102
5.10	Predicted probabilities	103
5.11	ROC curves for each selected model	104
5.12	Sample nodes from decision tree	108
5.13	Cumulative feature importances for the RF model	108
5.14	XGB SHAP values	112
5.15	Model performance as more features are included	113
6.1	Machine learning pipeline	124
6.2	Average AUC for each classifier by dataset	126
6.3	RUS results	128
6.4	RMSD distributions	132
6.5	Average training costs	135
6.6	Full learning curves for each dataset	147
6.7	Approximated learning curves	148
6.8	MAE as prediction size increases	149

6.9	Inverse power law method with varying fit schedules	151
6.10	LRLS slopes	153

LIST OF EQUATIONS

2.1	Cox hazard function	10
2.2	LR log-odds	11
2.3	L2 cost function	11
2.4	Gini impurity	12
2.5	Naïve Bayes	14
2.6	Multinomial naïve Bayes	15
2.7	χ^2 statistic	16
4.1	SLN metastasis benchmark	59
5.1	EC2 model training cost	90
6.1	Root-mean-squared distance	132

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

1.1.1 Melanoma

Melanoma is the most dangerous form of skin cancer: on average, one person in the U.S. dies every hour from this disease [51]. The rates of incidence are rising, with over 90,000 new cases of melanoma expected in 2019 [4, 13]. Early detection of the cancer is key, as the 5-year survival rate is less than 20% for patients with distant metastases [66]. A key component of early detection is routine screening for the disease by both the patient and dermatologist. Screening guidelines, however, vary across different countries [168]. Melanoma is also one of the most common cancers for young adults [1], people who may not be concerned or motivated enough to enroll in screening programs. While thousands of new cases are expected every year, the number of new melanoma patients is significantly smaller than the full population of the U.S., which would make mass screening too costly [156]. Therefore, we need predictive models that can target high-risk patients for regular screening.

Ultraviolet (UV) light exposure is known to be a risk factor for the cancer, and geographic location and lifestyle habits can affect the amount of exposure a patient has over his or her lifetime [69]. Family history of melanoma and personal history of dysplastic or benign moles are also present in patients that develop the disease. Similar to screening guidelines, different countries have different categorizations of various risk factors for melanoma [168]. While a few known risk factors exist, modern technologies and data collection enable us to uncover many risk factors that have

been previously unknown.

Beyond physicians, there is a need to educate patients about their risk for the disease. Studies have found that some patients with a previous diagnosis of melanoma actually increase their time under the sun [104]. Additionally, a large proportion of patients report never applying sunscreen or seeking shade when outside on a sunny day, and some still utilize tanning beds. While physicians instruct melanoma patients to limit UV exposure, many do not change their behavior. This shows that some patients value physical appearance or lifestyle choices over melanoma prevention. A predictive model is able to show, using the patient's own data, why he or she is at high risk for the disease, and may be able to influence behavior more than a physician alone. For more information about the epidemiology, risk factors, and treatment of melanoma, please refer to [51].

1.1.2 Electronic Health Records

Electronic Health Record (EHR) systems capture large databases of clinical patient data relating to office and hospital visits, medical history, lab and pathology results, prescriptions, and social and demographic information. The biggest promise of EHR systems is being able to quickly collect structured data from medical providers, improve the efficiency and accuracy of care. This also results in consistent and clinically relevant medical datasets that can be utilized for research purposes. Due to electronic record-keeping requirements such as the Health Information Technology for Economic and Clinical Health (HITECH) Act [11], the last few years have seen an immense increase in the use of EHR systems [15]. The 21st Century Cures Act, passed in 2016, provided \$1.8 billion to support cancer research through the Cancer Moonshot [6]. This funding will go to advancing precision medicine initiatives by increasing operability between EHR systems.

Holistic data about a patient is advantageous for modeling and treatment of

melanoma, as early detection is key to effective treatment. EHR data more desirable for personalized medicine applications, because EHRs collect real-world data as opposed to highly curated data from clinical trials and prospective studies. While many retrospective and observational studies are performed at research institutions, the patients in those datasets often do not reflect the diverse population across the country.

There are barriers, however, to fully unlocking the potential of this data. EHR systems are developed independently and often maintain proprietary standards for data collection and storage, and database schemas can be different across hospitals and physician practices. As data is collected from real-world environments, data elements that are important for research may not be collected due to time constraints. There can also be inconsistencies between how different physicians and facilities record the same types of information. Furthermore, many EHRs capture clinical information via free-text notes, making it difficult to extract structured information for use in automated decision support algorithms. While there is a great deal of research involving natural language processing techniques to extract structured elements from free-text data [46], the complexity of clinical information prevents data from multiple systems or doctor's offices to be used together.

1.1.3 Machine Learning and Big Data

Machine learning (ML) is the process of feeding data into an algorithm that can analyze patterns to make predictions for new data. Datasets for machine learning are increasing in both availability and size, especially in the healthcare space. ML models can be extremely useful in the context of melanoma risk prediction, as they can extract complex information from clinical records, and provide predictions for new patients presenting for screening. In addition to accurate predictions, the models must be *interpretable*, meaning the predictions can be explained to physicians and

their patients.

If a large quantity of consistent patient data can be collected for a predictive model, computational challenges arise when transforming the data and training a machine learning algorithm. First, data elements must be extracted from the EHR system and transformed into a tabular format to be passed to a machine learning model. The size of the dataset and variety of data can cause traditional processing and analytic tools to fail [83]. The field of big data has arisen in the last several years to be able to extract insight and build models from vast amounts of data. While big data has been historically defined by the 5 V's (Volume, Velocity, Variety, Veracity, Value), it suffices to consider a dataset to be "big data" when traditional computing techniques and resources are unable to analyze or model the data.

Cloud computing offers access to virtually infinite computing infrastructure, allowing for processing and modeling of big data. This technology can be utilized to evaluate a wide range of algorithms to produce accurate models. When dealing with big data and cloud computing, predictive accuracy is not the only consideration when choosing classifiers and machine learning techniques; computational complexity and cloud computing cost must also be factored in the selection process.

Many datasets for machine learning can suffer from class imbalance, namely, when a particular class of interest is much less represented than other classes in a dataset [85, 166]. A classic example is in binary classification for disease detection. The estimated cancer incidence in the U.S. is 439.2 per 100,000 men and women [109]. Therefore, a predictive model for general cancer risk would need to detect positive instances from a 0.44% class distribution (number of positive cases / number of total cases). Machine learning methods such as data sampling can be used to address this class imbalance.

1.1.4 Limited Data and Limited Labels

The era of big data has enabled vast amounts of data to be processed and analyzed in a cost-efficient manner on a scale like never before. Limited data, however, is still a challenge even when dealing with big data. Just because there is a large amount of data, it is not necessarily the *right* data. Supervised classification algorithms require that the data is labeled, meaning the class membership of each instance in the training data is known. For many applications, this requires expensive and time-consuming human annotation. Therefore, even if there is an infinite amount of computing power for model training, there is still a large cost that must be dedicated to labeling [73]. The question of “How much data is needed?” has been asked many times and explored through numerous studies, especially within the bioinformatics and biomedical community [55, 103]. More recently, the problem of learning from limited labels has been formulated as an active area of research [141], even spawning a new research program funded by DARPA [3]. Generally, the problem of “limited labels” refers to when there is a large amount of unlabeled data available, but only a small amount of labeled data. Class imbalance is even more important when dealing with limited labels, as a theoretical cancer detection dataset with 10,000 labeled instances would only have 44 positive cases (according to the 0.44% class distribution discussed above).

The opinion in the machine learning community is often that more data produces better models, and this assumption has not been made without experimental evidence [164]. With most ML problems, however, there is a point at which the law of diminishing returns takes effect, and the achieved classification performance hits a plateau with respect to dataset size [155]. This phenomenon can be visualized by creating a *learning curve*: training models on increasing sizes of data and plotting the data size versus classification performance on a graph. Approximating learning curves is a useful exercise for scenarios of limited labeled data where more labels can

be gathered at a known cost. The shape of the curve along with the labeling cost can be used to estimate the point of diminishing returns: where it would not be worth it to collect more labeled data.

1.2 CONTRIBUTIONS

In this work, we present the data engineering methods used to create a dataset for melanoma risk prediction along with machine learning approaches to build risk models. The risk problems studied include sentinel lymph node metastasis and individual risk of developing melanoma. Additionally, we explore methods to build effective models from limited data, as well as approaches to estimate how much data is needed for future biomedical studies. The key contributions of this dissertation are as follows:

- Present a large, unique de-identified dataset of dermatology patients for melanoma risk research along with novel data engineering and feature processing methods.
- Review current literature for predicting cancer risk and identify shortcomings in the research.
- Build a model to predict sentinel lymph node metastasis in melanoma from tumor and patient encounter information.
- Build an accurate, clinically relevant, and explainable model to predict melanoma risk from routine office visits.
- Investigate advanced machine learning methods such as data sampling, feature selection, and model interpretation for the melanoma risk problem.
- Apply inverse power law learning curve fitting to four big biomedical datasets.
- Present a novel semi-supervised method for learning curve approximation.

1.3 DISSERTATION STRUCTURE

The remainder of this document is organized as follows:

Chapter 2 describes machine learning algorithm theory and data engineering methods used throughout the dissertation, while Chapter 3 presents the de-identified dataset created for the melanoma risk problems. Chapter 4 discusses an experiment for detecting sentinel lymph node metastasis for patients with a diagnosis of melanoma. Several studies are described in Chapter 5 for building melanoma risk models and machine learning techniques to improve accuracy and interpretability of the models. Also included is a literature review of existing studies that predict cancer risk from clinical data (Section 5.2). Chapter 6 presents two experiments focused on learning from limited data, one for class imbalance in melanoma risk, and one for learning curve approximation with four large biomedical datasets (of which melanoma risk is one). Where necessary, each chapter contains additional related works and methodology information. Finally, the conclusions and future work are presented in Chapter 7.

CHAPTER 2

THEORY AND METHODOLOGY

This chapter presents the necessary theoretical background for the experiments conducted in this dissertation. Sections 2.1-2.2 review machine learning algorithms, techniques, and experimental design, while Section 2.3 presents the infrastructure and tooling used for data engineering. The various software packages used throughout are outlined in Section 2.4.

2.1 MACHINE LEARNING

At the highest level, ML methods can be separated into *unsupervised* and *supervised* learning; namely, whether or not the dependent variable (i.e., class label) for a group of data is known. The most straightforward scenario is supervised learning, where the data has a number of independent variables (i.e., attributes or features) and the class labels are known. An example of this is for email spam detection. Known “real” and “spam” emails are fed into a algorithm using various attributes of the email text and metadata. A trained model can then make future predictions for new emails which are not yet classified into each group [40]. In unsupervised learning, the attributes are available for a set of data but the class labels are unknown. With email data, an unsupervised learning activity might be to divide an inbox into a group based on related categories (e.g., financial, work, personal), when the categories are not known up-front. *Semi-supervised* learning is a combination of the two methods where there is a (typically large) set of unlabeled data and a (typically small) set of labeled data. Semi-supervised learning is useful when assigning labels involves expensive human

annotation. An application of semi-supervised learning for email spam would be to use a small set of emails where the real/spam class was annotated by humans, and a large corpus of unlabeled emails to augment the labeled data.

Within supervised, unsupervised, and semi-supervised learning, there are several families of algorithms that can be used for different applications. Classification is a supervised learning task that seeks to discriminate a class label from a set of inputs. When there are exactly two different values for the class, this is known as binary classification. Examples beyond email spam are predicting patient response to a drug [44] and sentiment analysis of text data [118]. Regression is also a supervised learning task that fits a model to a group of data, but has a continuous dependent variable allowing for numeric predictions rather than a class [77]. Recommendation engines predict meaningful relationships between entities. Recommending books, music, movies, and other products to users based on their relationships with other users is a prime example of these systems. Collaborative filtering methods are widely used to create recommender systems [154]. Clustering is an unsupervised learning task that groups elements in a population together by examining their various features [184]. Additionally, data processing techniques outside of the learning tasks are used to improve performance of the models. Data sampling alters the distributions of instances in the data to allow algorithms to learn from imbalanced data. Feature selection is a dimensionality reduction technique that selects the most informative features for training a model. While complex algorithms can produce ideal results, there must be consideration for interpretability of these models.

2.1.1 Algorithms

This dissertation focuses primarily on binary classification tasks to discriminate between low-risk and high-risk patients from EHR records in the context of melanoma. Clustering is used to understand certain properties of the data, while regression is

used for performance analysis and learning curve predictions. The remainder of this sub-section summarizes various algorithms used in our experiments or reviewed in related literature. Each model has some parameters (also known as hyperparameters) that must be chosen before training the model; we introduce the parameters in this section, while the methods sections in subsequent chapters describe the selection process for each experiment.

Survival Analysis

Survival analysis involves modeling time to a specific event, such as disease development or death. A Kaplan-Meier curve estimates the survival function of different cohorts of patients and plots the probability of survival along a time axis. Traditionally, this allows for comparison of patient cohorts with different characteristics of treatment regimens to determine which treatment to select for a new patient. An example is shown in Figure 2.1, where Kim et al. use a Kaplan-Meier curve to compare the survival rates of high-risk and low-risk patients for breast cancer [78]. Cox proportional hazards [39] is the typical model of choice for survival analysis, as it allows for time censoring and multivariate analysis. It is a regression model that creates a function of time, from baseline covariate values, that model the probability of an event occurring at any future time (disease development, death, etc.), according to the following hazard function:

$$h(t) = h_0(t) \times \exp(b_1x_1 + b_2x_2 + \dots + b_px_p) \quad (2.1)$$

In risk prediction studies, the event is the diagnosis of cancer, and time zero is either the enrollment in a study, or start of the observation period. A patient is censored when follow-up is lost before the event occurs, which is typically the end of the follow-up period, but may be other scenarios such as a patient dropping out of

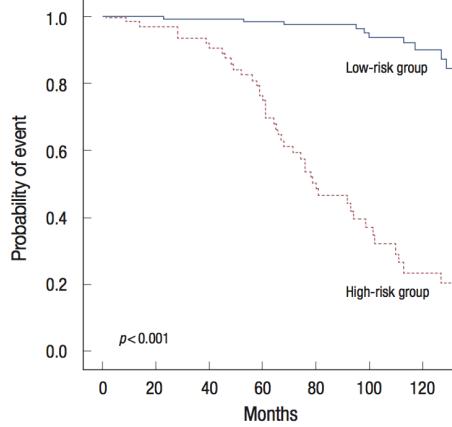


Figure 2.1: Example Kaplan-Meier curve [78]

the study or death.

Logistic regression

Logistic regression (LR) is a widely used linear model for classification. This technique allows for multivariate analysis and modeling of a binary dependent variable [74]. In LR, a linear model is built on the independent features, and then a logistic function is applied to discriminate between the two classes of output [74]. The log-odds for a particular class is:

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n \quad (2.2)$$

Weights (β) for each feature are learned by optimizing a cost function, and these are used to make a prediction on a new case. Linear models are improved by using regularization in the cost function; we use L2 regularization (also known as ridge regression) which minimizes the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log (\exp (-y_i (X_i^T w + c)) + 1) \quad (2.3)$$

We tested various values of the regularization parameter C during the hyperparameter selection process.

Decision trees

Decision trees (DT) are ML models that can be used for regression or classification. They produce an output similar to a flow chart, allowing a path to be traversed based on the value of the instance in question, resulting in a predicted value. The model is trained by selecting a feature that best discriminates between the different outcomes, splitting the tree on this feature (node), and recursively performing this split on each new node that is generated. Splitting candidates are determined by the split that minimizes impurity. We use Gini impurity in our experiments:

$$H(X_m) = \sum_k p_{mk} (1 - p_{mk}) \quad (2.4)$$

The splitting produces a tree-like graph, and new instances can be scored by traversing the path created based on the instance's feature values. Various parameters of the model will determine when this splitting stops (number of iterations, number of nodes, etc.). Since the model is selecting features to split the tree on at each node, there is an inherent feature reduction that occurs, resulting in the most informative features being included in the model. Common algorithms for decision trees are CART [27], C4.5 [119], C5.0 (an improved, commercialized version of C4.5), and Bayesian trees [35]. An example tree is shown in Figure 2.2 [35].

Tree ensembles

An extension of the decision tree model is called random forest (RF). In a random forest, multiple trees are built and predictions are decided by majority voting [26, 75]. Bagging is used to construct the trees so that a random subset of features and a

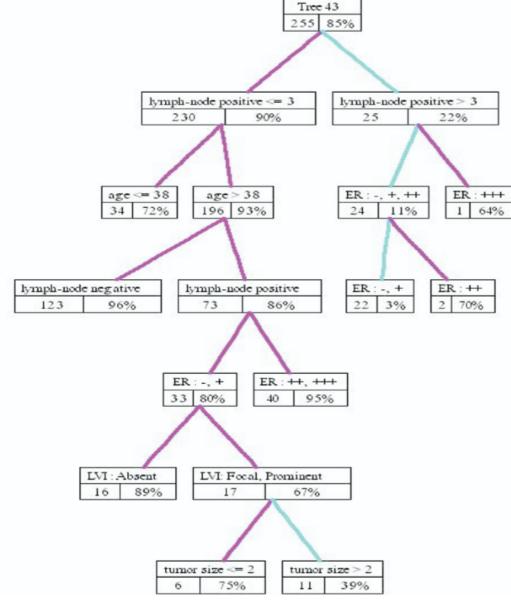


Figure 2.2: Example decision tree [35]

random subset of data are selected to build each tree. While building the trees, a random subset of features are considered at each decision node. After all trees are built, classification takes place by evaluating the instance with respect to all trees and the decision is the one agreed upon by the majority of the trees. We evaluated several values for the maximum depth of each tree and the total number of trees for each forest.

Gradient boosting is a particular boosting process that uses an additive model to select weak learners for the ensemble by optimizing a loss function. XGBoost (XGB) is an enhanced version of a typical gradient boosted machine that uses regularization in model building to improve performance [34]. It utilizes tree models as the weak learners, and thus maximum depth and number of trees must be selected. In addition, we test various values of the learning rate used in the boosting process.

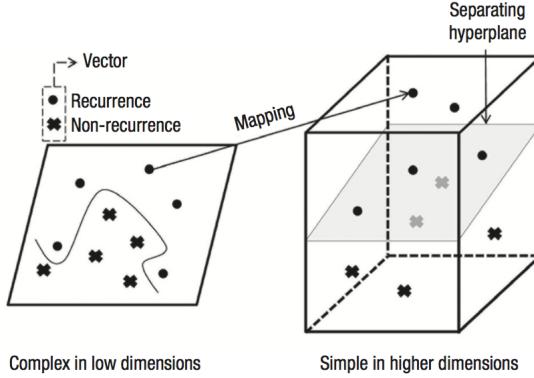


Figure 2.3: Basic framework of an SVM [78]

Support vector machine

Another widely used classification model is the support vector machine (SVM). An SVM creates a set of hyperplanes for each feature in an infinite dimensional space, and fits linear or nonlinear models that most effectively discriminate between the values of a binary output variable. Kim et al. (Figure 2.3) provide a basic description of an SVM model in their paper that discriminates between recurrence and non-recurrence in breast cancer patients [78]. Due to the size of our data we chose a linear kernel and tested various values of the L2 penalty parameter, C . Since SVM does not return class membership probabilities (as opposed to the other models used in this study), we calibrate probabilities with cross-validation using Platt's method [116].

Naïve Bayes

Naïve Bayes (NB) is a classifier that utilizes Bayes' theorem and independence assumptions to make predictions based on the feature probabilities of each instance [183]:

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i|y) \quad (2.5)$$

We used a multinomial NB model, where parameters are estimated by relative frequency counting, and selected from various values of the smoothing parameter α [124, 125]:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (2.6)$$

Deep learning

A popular model in the machine learning community is the artificial neural network (ANN), which can be used for classification, regression, and unsupervised learning. Variations of ANNs have been shown to be highly effective in complex tasks such as image recognition [96]. A neural network is roughly modeled after the way the human brain works, by creating nodes (neurons) that give weights to certain inputs and produce an output value. Multiple layers of nodes are tied together with an input layer taking in the value of the independent variables, and an output layer with nodes representing each of the possible outcome values. The weights at each layer in the network are modified as the model learns through back-propagation. When one node in the output layer is positive, the value at the node is taken as the prediction. When there is a large number of intermediate layers, this is often called “deep learning”, and has shown impressive results for very complex modeling problems [105]. Ahmad et al. provide an illustration of a basic network in Figure 2.4 [9].

2.1.2 Dimensionality Reduction

Dimensionality reduction minimizes data size by combining, transforming, and removing features. Notable algorithms for dimensionality reduction are principal component analysis and singular value decomposition. Feature selection is a technique in dimensionality reduction that chooses specific features based on a statistic or eval-

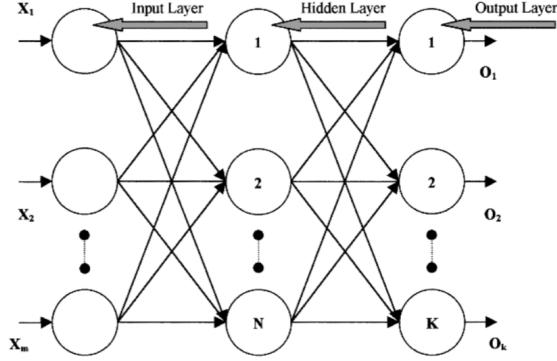


Figure 2.4: Basic framework of an ANN [9]

ation of how useful the feature is to the overall modeling task [45]. Feature rankers order the features according to a certain statistic, leaving the practitioner to determine how many of the top features he or she wants to include in the model. Many of these algorithms utilize univariate analysis methods, such as information gain or mutual information. Several predictive models, such as decision trees, effectively perform feature selection as part of the model building process. The p-values from a statistical model can also be used as a form of feature selection, by only selecting those features that have significant p-values (often <0.05).

For dimensionality reduction in our experiments, we first remove all features with zero variance (i.e. same value in all samples) and utilize the χ^2 feature ranker to select the top K features. Various values of K are tested throughout the experiments. χ^2 tests the independence of two variables; in feature selection these are a particular feature and the class:

$$\chi^2 = \sum_{i=1}^n \frac{(N_i - E_i)^2}{E_i} \quad (2.7)$$

If a feature is independent of the class, then it is not informative for ML purposes and can be discarded.

Feature selection and classification algorithms can depend on assumptions about

distributions of independent variables and therefore, we standardize all features by removing the mean and scaling to unit variance. In the event of sparse data we do not remove the mean, as it would destroy the sparse properties of the data.

2.1.3 Data Sampling

Many supervised learning problems suffer from class imbalance. Specifically, this means that certain cases of interest are less represented in a dataset than normal cases [70, 166]. In healthcare applications, this can mean that healthy patients far outnumber those with a disease of interest. This becomes especially challenging when attempting to train a machine learning model, because the model will tend to focus on the majority cases rather than those of interest (minority cases). The problem of class imbalance can be exacerbated when dealing with big data, as there can be millions of negative (majority) samples, but only hundreds or thousands of positive (minority) samples.

Data sampling methods are one way to address class imbalance for model training. They involve either undersampling majority cases or oversampling minority cases, or combination of both. Undersampling techniques are desired when dealing with large datasets to reduce computational complexity and runtime considerations. Random undersampling (RUS) randomly removes instances from the majority class based on a specific target class ratio. It is not always advantageous to fully balance the classes, especially when the classes are severely imbalanced, because it results in throwing away a large proportion of data. Therefore, we explore various sampling ratios throughout the experiments to see which is most effective. While not used in this dissertation, oversampling techniques include random oversampling (ROS), which randomly duplicates samples from the positive class, and Synthetic Minority Over-sampling Technique (SMOTE), which selects minority samples to duplicate based on a nearest-neighbor approach [33].

2.1.4 Interpretability

In many contexts, it is important to not only to get accurate predictions from an ML model, but also to understand (i.e., interpret) why the model made that prediction. This is especially important for healthcare applications, as the predictions from a model may influence patient care. In this sub-section, we explore various interpretability considerations of the ML algorithms explored in this research.

Selecting the most important features to input to a model is important for both performance and interpretability considerations [167]. Linear models, such as logistic regression, are much easier to interpret when the number of features included in the model is small. This is because each feature has a level of contribution toward the final prediction, and it would be difficult for a person to understand contributions given by thousands of variables. Limiting the number of features, however, may impact performance of the model.

Models have different levels of *global* and *local* interpretation. Global interpretation is when a model can describe how a prediction is made generally, across all samples. Local interpretation is when predictions can be explained for a local region of the data (or a single sample). Linear models and tree models allow for both global and local interpretation, since the coefficients of a linear model and decision path of a tree are the same for all instances that pass through the model. This is beneficial for identifying and studying risk factors for a disease, as important factors in these models can be generalized across large patient populations. Linear models are widely used due to the simplicity of the fitted model. It is easy for a practitioner to see which features contribute toward the prediction [153]. However interpretability of both linear and tree models is limited when a large number of variables are involved, as a person will not be able to grasp the contributions of more than ten or so features.

As ensembles such as RF or XGB contain a combination of potentially hundreds of decision trees, it is difficult to generalize the decision path of all instances. Global

interpretability can be assessed by calculating the average reduction of loss when splitting on each particular feature (Gini importance [101]). This is known as feature importance; feature importance does not describe the directional impact of each feature toward a prediction, but it does provide a sense of how important each feature is for the model. Local interpretability, however, is possible with ensembles. The contributions of each feature can be explored for a specific prediction, by decomposing each prediction into a bias and feature contributions, similar to a regression function [140]. The contributions of each feature can be explored for a specific prediction, showing why a certain instance was classified as positive or negative.

As shown by Lundberg et al., the Gini importance approach for interpreting tree ensembles can be inconsistent when comparing different models [91]. Therefore, we use their proposed method, Shapley additive explanations (SHAP), to achieve a consistent and accurate representation of feature importance for RF and XGB, which can be used for both global and local interpretation [92]. This method provides a promising avenue for explaining model predictions, as the same method can explain an individual prediction as well as generalize to groups of instances.

2.2 EXPERIMENTAL DESIGN

The concepts of model evaluation and a template ML pipeline used throughout the experiments are presented below. The specific experiments in the subsequent chapters then explain in detail the design for each.

2.2.1 Metrics and Evaluation

Performance evaluation is an important step when creating a classification model, as the model must be proven to be accurate before using it to inform decision-making. The most basic form of performance evaluation is predictive accuracy, which gives the percentage of instances that the model correctly labeled. This can be a biased

Table 2.1: Confusion matrix

		Predicted Values	
		Positive	Negative
Actual Values	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

measure if the classes are imbalanced. For example, if 10 out of 1,000 patients in a dataset develop a disease, the model can simply label all 1,000 patients as negative (not developing the disease), and still achieve an accuracy of 99%. Therefore, other metrics based on a confusion matrix (see Table 2.1) are calculated [143]:

- Accuracy: $(TP + TN)/(TP + TN + FP + FN)$
- True positive rate (TPR, sensitivity, recall): $TP/(TP + FN)$
- True negative rate (TNR, specificity): $TN/(TN + FP)$
- False negative rate (FNR): $FN/(TP + FN)$
- False positive rate (FPR, 1 - specificity): $FP/(FP + TN)$
- Positive predictive value (PPV, precision): $TP/(TP + FP)$
- Balanced accuracy (arithmetic mean): $\frac{1}{2}(TPR + TNR)$
- G-mean (geometric mean): $\sqrt{TPR * TNR}$

Confusion matrix-based metrics can also be biased, as most models produce a score, or a probability as the output rather than a concrete class label. A discrimination threshold must be set to determine at which point the score results in a positive or negative class value. To evaluate the accuracy of a prediction, that probability must be converted to a binary decision as the actual class labels are binary. A probability

above the threshold is classified as positive, while probabilities below the threshold are classified as negative. The default threshold is 0.5, but it is advantageous to explore the distribution of predicted probabilities of each model to select a threshold that will result in better TPR and TNR. We do this experimentally by plotting the predicted probabilities of instances versus their actual class membership, and then explore the TPR and TNR of various thresholds. We choose a threshold where the TPR is maximized without a large drop in TNR. Additionally, the threshold may be selected by the one that produces the best value of a certain confusion matrix-based metric (such as balanced accuracy or G-mean).

To handle multiple different discrimination thresholds, a receiver operating characteristic curve (ROC) is generated [185]. The ROC curve plots the TPR (or sensitivity) against the FPR (or 1 - specificity) against a range of discrimination thresholds. An example ROC curve is shown in Figure 2.5, taken from Kim et al. [78]. By taking the area under the curve (AUC), a single metric is produced that is not dependent on the discrimination threshold. This metric is the probability that the model will rank an arbitrary positive instance higher than an arbitrary negative instance (in terms of the probability of an instance being positive).

In addition to reporting performance measures on the training dataset, some sort of validation set must be used to prove that the model can accurately predict on new instances, and is not overfit to the training data. This can be accomplished by splitting the dataset into training and testing sets, using an independent validation set, or by performing bootstrapping or cross-validation. Validation using bootstrapping resamples the training data with replacement to create a training set, and uses the rest of the instances as a test set. This is repeated n number of times and the results combined to produce the final performance score. Cross-validation is similar to bootstrapping, but divides the dataset into n folds, using $n - 1$ folds for training and the final fold for testing. This is then repeated for the rest of the folds and the

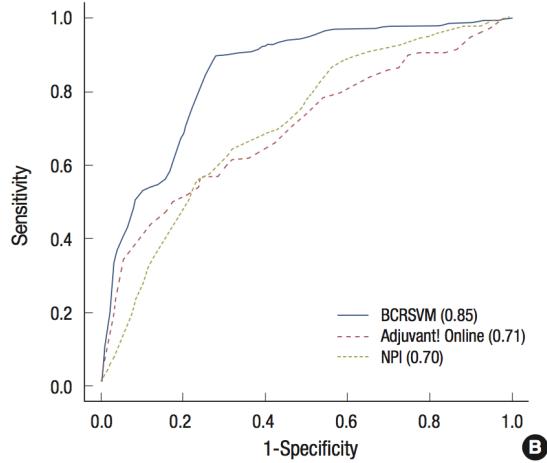


Figure 2.5: Example ROC curve [78]. A larger area under the curve indicates better model performance.

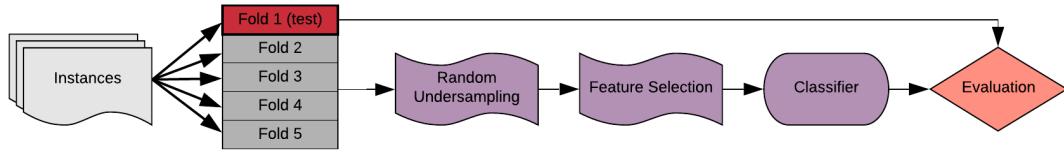


Figure 2.6: Predictive modeling pipeline. This process is repeated for each dataset split from cross-validation.

results are combined to produce the final performance score.

In our experiments we used both train/test splits and cross-validation for model training and evaluation. ML pipelines are created for each model configuration to ensure proper splitting of data in the cross-validation process. Figure 2.6 shows an example of a model pipeline including undersampling, feature preprocessing, and finally classifier training and evaluation. Note that each preprocessing step (both sampling and feature processing) occurs *within* a single fold of cross-validation. This is important because the full dataset cannot be processed before cross-validation splitting. Otherwise, this would bias the fold datasets according to properties of the full dataset.

2.2.2 Statistical Tests

When possible in our experiments, we perform multiple runs of the same model configuration (i.e., dataset, classifier, sampling, feature selection) to enable hypothesis testing of the results. We perform Analysis of Variance (ANOVA) tests to determine whether the means of factors are equal [58]. Additionally, post-hoc analysis is performed using Tukey’s Honestly Significant Difference (HSD) test [160]. This method compares the means of all possible pairs of factor levels and interactions. A letter is assigned to each group of levels that are not statistically different from each other. Welch’s two-sided t-test is used for ad-hoc significance testing [170]. The significance level for all tests is set at 0.95.

2.3 DATA ENGINEERING AND INFRASTRUCTURE

Datasets for machine learning are increasing in both availability and size. The field of big data has arisen in the last several years to be able to extract insight and build models from vast amounts of data. While big data has been historically defined by the 5 V’s (Volume, Velocity, Variety, Veracity, Value), it suffices to consider a dataset to be “big data” when traditional computing techniques and resources are unable to analyze or model the data. In this dissertation, the data sources used are all “big”, so non-traditional data engineering methods were used to process the data and extract features for use in ML models. We discuss the data processing frameworks in general here, while the specifics of data engineering for the EHR data used for modeling is presented in Chapter 3.

2.3.1 Hadoop and Spark

When the research for this dissertation began (circa 2014), Apache Hadoop with MapReduce was the go-to processing engine for big data. Apache Hadoop¹ is a

¹<http://hadoop.apache.org/>

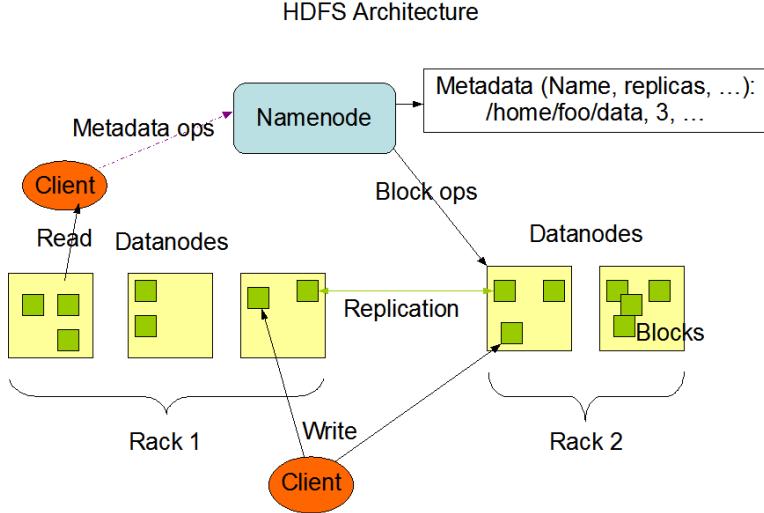


Figure 2.7: HDFS architecture²

distributed storage system for big data that allows data from multiple sources to be ingested and stored in the same cluster of machines [83]. Hadoop introduced the Hadoop Distributed File System (HDFS) to allow for redundant, fault-tolerant storage of large amounts of data at scale. Datanodes are used to distribute storage and computation, while the Namenode orchestrates data pipelining and job scheduling. The HDFS architecture is presented in Figure 2.7.

The original data processing engine for Hadoop was MapReduce, which is named after its theoretical distributed programming paradigm. The general concept of MapReduce is to send computation to the data, rather than sending data to the computation (Figure 2.8). This is accomplished by performing tasks on each chunk of data stored on the individual datanodes (“map” task), then combining the intermediate results and writing them out (“reduce” task). MapReduce writes all intermediate computations to disk, which allows for robust fault-tolerance, but at a tradeoff of speed. This allows processing to scale to very large datasets, but the overhead becomes significant when using smaller datasets. Additionally, writing pure MapReduce programs takes very skilled programming to translate tasks into concepts involving

²<http://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>

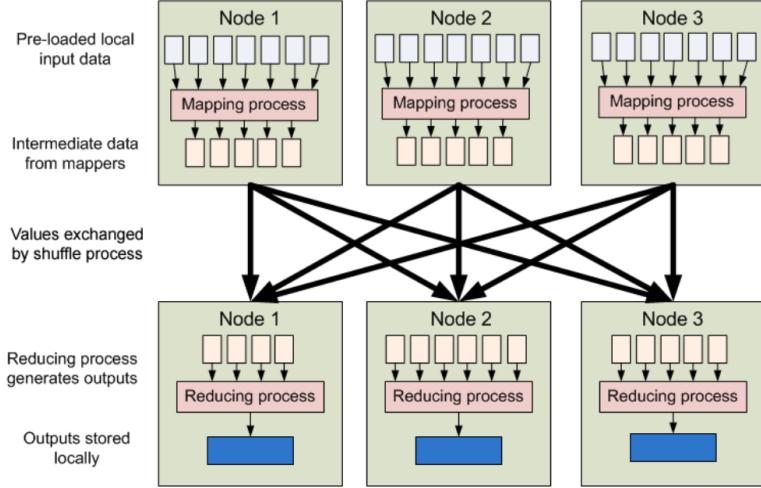


Figure 2.8: MapReduce programming paradigm ³

mappers and reducers.

Apache Spark⁴, a distributed computing engine, is used on top of a Hadoop cluster for fast and fault-tolerant data processing. Spark was created at the University of California, Berkeley, with the goal of solving many problems inherent with the Hadoop MapReduce processing architecture [181]. The project introduced the concept of Resilient Distributed Datasets (RDD) which store and process data in-memory across nodes in a cluster [180]. Fault tolerance is provided by creating a Directed Acyclic Graph (DAG) of operations and evaluating actions in a lazy manner. When a node fails, results are re-computed using the actions stored in the DAG [72]. This significantly reduces the number of read/write operations typically used in MapReduce programs, resulting in greater time efficiency. Spark set the Sort Benchmark⁵ record in October 2014, sorting 100TB of data in 23 minutes using 206 nodes. The previous record, held by Hadoop MapReduce, sorted the data in 72 minutes with 2,100 nodes. Additionally, Spark sorted one petabyte in 234 minutes with 190 nodes⁶, though not part of the official competition.

³<https://developer.yahoo.com/hadoop/tutorial/module4.html>

⁴<http://spark.apache.org/>

⁵<http://sortbenchmark.org/>

⁶<https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html>

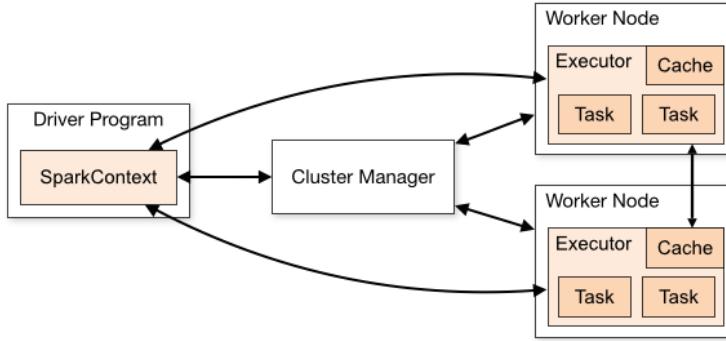


Figure 2.9: Spark architecture ⁷

While Spark can run on Hadoop clusters, it does not contain a storage layer, but connects to various data storage mechanisms such as HDFS. The infrastructure of a Spark cluster is similar to that of a Hadoop cluster, with a driver node to orchestrate operations, and worker nodes that perform the data processing (Figure 2.9). Within just a couple of years, Spark overtook MapReduce as the de-facto big data processing engine due to its immense speed and usability improvements.

Spark is compatible with Hadoop elements such as HDFS, and can run on top of a Hadoop cluster using its built-in job scheduler (YARN). Additionally, the package can be downloaded and run on a single machine, and the same code can be deployed to a Spark standalone or Mesos cluster. Spark is written in Scala and uses the Java Virtual Machine (JVM), but the engine is accessible through Scala, Java, SQL, Python, and R. MLlib is a component within Spark that can be used to train and evaluate ML models using the Spark engine [100, 137].

2.3.2 Cloud Computing

As data size and algorithm complexity increases, so does the need for computing infrastructure. Cloud computing, which is infrastructure accessed through the internet, enables users to launch machines of varying size with pre-built libraries for

⁷<https://spark.apache.org/docs/latest/cluster-overview.html>

data processing and machine learning algorithms. All data processing and ML experimentation for this dissertation was performed in the cloud using Amazon Web Services (AWS)⁸. We used Spark on AWS Hadoop clusters⁹ to process the EHR data and extract features. For the machine learning pipelines, we used instances from the Amazon Elastic Compute Cloud (EC2)¹⁰. EC2 allows users to spin up different types of machines for different workloads with various amounts of CPU cores and memory. This allows for many machines to be run in parallel, executing various components of an ML experiment.

Distributed computing is often required to deal with big data. While packages such as MLlib provide distributed ways to train machine learning algorithms, there are many more libraries and algorithms available on non-distributed infrastructures (i.e., single machines). Additionally, certain techniques for handling class imbalance, such as random undersampling, will actually remove majority cases from a dataset. Therefore, a dataset can start out as “big data”, but if undersampling is performed, the data that the machine learning model is trained on may very well be “small”. Machine learning research and model building is an experimental and iterative process; therefore, the cost and time to train a model can be significant for big data tasks. If a single model takes a long time to train, it can limit the amount of experimentation that can be done to achieve an exemplary model. We found that the best approach for our scenario was to use Spark to perform data collection and transformation and then use non-distributed infrastructure to perform ML experimentation. By doing so, we were able to use multiple machines to train and evaluate multiple machine learning models in parallel rather than using a cluster of machines to train one model at a time.

⁸<https://aws.amazon.com/>

⁹<https://aws.amazon.com/emr/>

¹⁰<https://aws.amazon.com/ec2/>

2.4 SOFTWARE

The Python API for Spark was used to perform all data engineering on the Spark clusters. A data processing library was built on top of Spark as part of this research and is described in Section 3.2.2.

We utilized several libraries in the Python data science ecosystem to execute the ML experiments. We used *scikit-learn* [114] to train and evaluate all classifiers, *numpy* [165] and *scipy* [71] for matrix processing, *imbalanced-learn* [87] for data sampling, and *treeinterpreter* [140] and *shap* [92] for explaining tree ensembles. An open-source ML experimentation framework was built as part of this research to seamlessly launch EC2 instances and execute model training and evaluation. The framework is available on GitHub¹¹.

Results analysis, figure generation, and hypothesis testing were performed in *R* [120]. We used many packages from the *tidyverse*¹², a group of packages designed to produce a consistent and easy-to-use API across tasks, but the notable ones are: *dplyr* [172] for data manipulation and *ggplot2* [171] for visualization.

¹¹<https://github.com/rikturr/aws-ml-experimenter>

¹²<https://www.tidyverse.org/>

CHAPTER 3

CLINICAL DATA PREPARATION

3.1 BACKGROUND AND MOTIVATION

In the era of personalized medicine, detailed patient data is crucial. Clinically-relevant data points and biomarkers must be captured for a wide range of patients to enable large-scale population health modeling. Large scale health models have the ability to transform the way we approach medicine from a reactionary approach to a predictive and personalized one [99]. There are various sources of patient data, such as molecular and genetic markers, medical records, clinical registries, and even non-health related information such as social media and lifestyle data [65]. When used in aggregate, the data are more powerful. However, due to technology limitations and privacy policies, it is often very challenging to collect all sources of data for each individual patient. In this chapter, we review various data sources for clinical data, and outline the data collection and preparation process for the dataset used in this dissertation.

3.1.1 Data Sources and Features

Patient data is collected from a variety of sources, and the availability of each varies based on the ease of collection, cost, and data storage methods [65]. This dissertation focuses on applications that utilize structured clinical information (not free-text or genomic), as this data is widely collected and has the greatest value for efficient modeling in large-scale machine learning applications.

Clinical and Practice Data

There is a large amount of information collected about routine clinical encounters in hospitals and private practices. Billing data, such as insurance claims for procedures and medications, have mature data sharing standards due to their financial impact and need for consistency. Coding standards include Current Procedural Terminology (CPT) [14] for procedures performed by a physician, and International Classification of Diseases (ICD) for specifying which diagnoses warrant the procedure being billed for [110]. While these codes provide a standard for data collection, there is more clinically-relevant information that is not captured through routine billing data. For example, the ICD-10 code C50.111 represents “malignant neoplasm of central portion of right female breast”, but the tumor information, progression of the patient’s health, and the patient’s medical and social history are all unknown.

Electronic Health Record (EHR) systems have the potential to capture large databases of clinical patient data relating to office and hospital visits, medical history, lab and pathology results, prescriptions, and social and demographic information. The biggest promise of EHR systems is being able to collect structured data at the point of care by medical providers, creating consistent and clinically accurate medical datasets. This information is more advantageous for personalized medicine applications, because clinical information is often more reliable than billing information [106]. For example the number of adenomatous polyps, or family history determines the risk profile for colon cancer. With melanoma, family history, proximity to the equator, number of sunburns, and the number of clinically atypical nevi are all factors that lead to developing the cancer. Due to electronic record keeping requirements, the last few years have seen an immense increase in the use of EHR systems [15]. The 21st Century Cures Act, passed in 2016, provided \$1.8 billion to support cancer research through the Cancer Moonshot [6]. This funding will go to advancing precision medicine initiatives by increasing operability between EHR sys-

tems. However, there are barriers to fully unlocking the potential of this data. EHR systems are developed independently and often maintain proprietary standards for data collection and storage. Furthermore, many EHRs capture clinical information via free-text notes, making it difficult to extract structured information for use in automated decision support algorithms. While there is a great deal of research involving Natural Language Processing (NLP) techniques to extract structured elements from free-text data [46], the complexity of clinical information prevents data from multiple systems or doctor’s offices to be used together.

Standards outside of financial applications exist for capturing clinical data that is transferred between multiple parties. ePrescriptions, prescriptions that are sent electronically from the doctor’s office to a pharmacy, use standards such as National Drug Code (NDC) numbers and RxNorm [108] to ensure the correct medications are given to the patient. Logical Observation Identifiers Names and Codes (LOINC) are used to maintain consistency in the ordering and reporting of lab results and other clinical observations [90]. The Systematized Nomenclature of Medicine (SNOMED) maintains coding standards for clinical information such as diagnoses, family history, allergies, social information, and others [148]. The adoption of these standards is not consistent across medical providers, but when used, they provide valuable structured information that can be used to advanced population health research.

Social and Lifestyle Data

Social and lifestyle data can be important to modeling the risk for certain cancers. Smoking has been shown to be associated with lung cancer [151], alcohol consumption with liver cancer [161], and UV light exposure with skin cancer [168]. This data can be captured through routine clinical encounters using EHR systems, or through surveys and questionnaires given to patients.

Clinical Registries

Clinical registries help solve research problems by maintaining a centralized database of clinical information specific to certain patient populations. The data points captured are often based on expert knowledge of the disease being studied, and can be submitted through electronic connections with digital record systems or manual input. Therefore, the data stored in these registries can be from multiple different sources, such as demographic, billing, pathologic, and tumor information. Registries are common for high-profile diseases, such as cancer, and many governments require that all cancers be recorded in a local or national cancer registry [2].

Feature Types

Various features are collected through the above data sources, and can be grouped into the following categories:

- **Demographic:** Patient demographic information, such as age, sex, race, and geographic location.
- **Lab:** Laboratory test results, such as white blood cell count, hemoglobin, glucose, triglycerides, etc.
- **Histopathologic:** Cancer and tumor-related information, such as the location, tumor size, metastasis, stage, margins, etc.
- **Clinical:** Treatments, family history, vitals, and other routinely captured clinical information that does not fit into any of the other categories.
- **Lifestyle:** Social history information such as smoking status and alcohol use.

3.2 MAMEL DATASET

Modernizing Medicine, Inc. is an EHR provider for private-practice surgical specialties that solves many interoperability and data management problems by employing an innovative technology stack. We extracted a dataset from the de-identified EHR data for use in this research, titled “Modernizing Analytics for ME^Lanoma” (MAMEL) [127]. MAMEL includes de-identified patient data from over 20,000,000 dermatology patients and serves as the base dataset for experiments presented in Chapters 4-6. Image or genomic data are not available in MAMEL.

3.2.1 Modernizing Medicine EHR

Modernizing Medicine’s suite of applications are at the center of innovation in mobile, cloud, and data for medicine. By being mobile-first, physicians can quickly record precise clinical information, and spend more time advising and interacting with patients. By hosting the entire application in the cloud, all customers can easily be on the latest version of the product, and the system architecture can scale with ease. By providing structured input methods for most data points, the data can be consumed by machines and algorithms to both improve the application and provide the ability for population-level health research.

Modernizing Medicine’s Electronic Medical Assistant (EMA) DermatologyTM product collects structured, real-world data from thousands of dermatology providers across the U.S. The cloud-based and HIPAA-compliant EMA database houses the data of millions of patients according to accepted data standards. For data elements that do not have widely-used industry standards, a team of practicing dermatologists, employed by Modernizing Medicine, maintains databases of clinical elements to allow for dermatology-specific standardized data capture across all users (see Table 3.1).

During clinic visits, dermatology providers obtain medical history, assess clinical presentation, perform physical examination, and make treatment recommendations

Table 3.1: High-level data elements

Concept	Standard(s)
Patient Demographics	HL7 ADT
Medical/Family/Social History	ICD, SNOMED
Medications	NDC, RxNorm
Allergies	RxNorm
Biopsy Results	Proprietary
Cancers	SNOMED, HL7 CCD
Patient Intake/History	Proprietary
Exams	Proprietary
Diagnoses	ICD, Proprietary
Procedures	CPT/HCPCS, SNOMED, Proprietary

based on their clinical judgment specific to individual patients at that visit. All data are entered directly into the EHR by providers and their staff at the point-of-care.

A typical patient encounter workflow includes intake by a medical assistant, who records medical history, chief complaint information, and assesses clinical presentation. Then, the physician records medical exams, findings, and treatments. Body locations are captured throughout the workflow for intake, findings, and treatments. They are recorded with an interactive anatomical atlas, a zoomable, 3D layered way to document thousands of detailed body locations.

3.2.2 Data Processing Architecture

The entire data processing architecture is hosted in the cloud, enabling rapid development and data access. Figure 3.1 shows a simplified view of the data architecture.

Sharded Application Servers

For load-balancing and performance, application servers are sharded by groups of customers (medical practices). Each shard is a cluster of application servers, but only services a specific number of customers. This allows the architecture to scale by adding new shards as a certain number of new customers are added to the system.

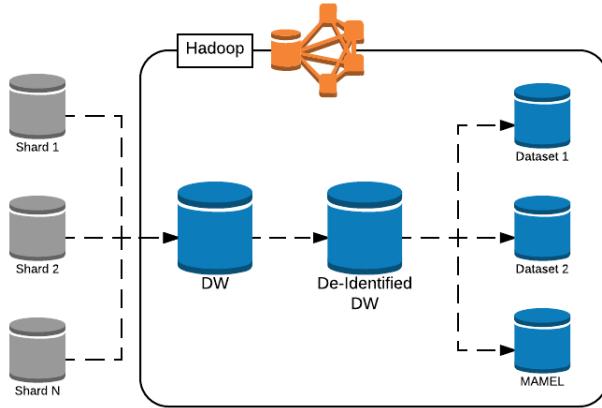


Figure 3.1: Big data processing architecture

Each shard has identical SQL database schemas, ensuring all customers have the same experience and are using the same version of the application. These application shards operate on a traditional relational database management system (RDBMS).

Big Data Processing

Each application shard houses terabytes of data containing billions of rows of structured clinical data. Due to the sheer size of the data and the fact that data is stored on disparate servers, aggregate analysis with traditional database technology is limited and time-consuming.

We utilized Apache Spark on top of Hadoop clusters to perform fast and fault-tolerant data processing. Spark has connectors to many different data sources and file formats, allowing for nightly batch loads from the RDBMS servers into Hadoop. Additional processing is performed to clean and format the data into a centralized Data Warehouse (DW).

De-identification

For population-level healthcare research to be performed, patient data is de-identified in accordance with the HIPAA Privacy Rule’s Expert Determination Method [5,162]. This includes removal or masking of individually identifiable information to minimize re-identification risk. The resultant De-Identified Data Warehouse is used for all analyses that must be performed using de-identified data.

Dynamic Dataset Generation

To facilitate research dataset creation and metadata management, a dynamic dataset generation tool was created with Apache Spark. This tool sources from the De-Identified Data Warehouse and creates domain-specific datasets, of which MAMEL is one. These domain-specific datasets have many shared data elements, such as patient demographics, clinical characteristics, and vital signs. More detailed data points can be shared between several domain-specific datasets, and furthermore, some data points may be specific to only one dataset. This presents a data management challenge, as multiple unique datasets need to be created that share data points and also have their own unique data points. These datasets must be documented well to facilitate accurate consumption by data scientists and researchers, and must be analyzed to ensure quality of the data.

3.3 EXPLORATORY DATA ANALYSIS

The data architecture described in Section 3.2.2 was used to generate the MAMEL dataset. This is a large, clinically relevant, and statistically powered dataset to enable decision support research for dermatology patients. There are two general types of data collected for each patient: patient demographics and clinical characteristics, and data recorded in a patient encounter. Demographics and clinical characteristics include information about the patient as a whole and aggregate lists of clinical

information (such as past diagnoses and pathology results). Patient encounters, or visits, include detailed clinical elements for specific diseases the patient is treated for. High-level data elements, along with the coding standards used for each, are given in Table 3.1. For elements that use a coding standard and a Modernizing Medicine proprietary standard, data can be cross-walked between the Modernizing Medicine and industry standards.

Melanoma Identification

In this section, we perform exploratory analysis for a subset of the MAMEL population with a confirmed diagnosis of melanoma (the described data elements are available for all patients, not only those with melanoma). For a patient to be included for the analysis, an ICD10 code for melanoma must be present in the patient's problem list (C43, D03, Z85.820). The time period for inclusion includes all historical data through the 2016 calendar year (November 2011 through December 2016). While ICD codes are used for standardized disease identification purposes, the EHR records more detailed melanoma subtypes for each patient. The frequency of these subtypes is presented in Table 3.2.

Demographics

Available demographic data are provided in Table 3.3. Of the 567,660 melanoma patients, approximately half (49.1%) are female, and the majority (74.1%) are white. Patient home locations are spread across the U.S. with most (43.5%) living in the South.

Medical, Family, and Social History

Medical history elements include conditions (diseases/comorbidities) the patient has, and past surgeries the patient has had. These conditions are marked by patients and

Table 3.2: Melanoma subtypes

History of malignant melanoma	327,536 (57.7%)
History of malignant Melanoma in situ	138,164 (24.3%)
Malignant melanoma in situ	47,059 (8.3%)
Malignant melanoma	39,706 (7.0%)
Melanoma in situ	29,404 (5.2%)
History of lentigo maligna	13,792 (2.4%)
Melanoma	19,100 (3.4%)
Lentigo maligna	16,283 (2.9%)
Metastatic melanoma	1,602 (0.3%)
Superficial spreading melanoma	2,160 (0.4%)
Recurrent malignant melanoma	1,175 (0.2%)
Amelanotic melanoma	832 (0.1%)
Nodular melanoma	391 (0.1%)
Lentigo maligna melanoma	361 (0.1%)
Melanoma metastatic to lymph node	276 (0.0%)
Acral melanoma	306 (0.1%)

Table 3.3: Demographics

Variable	Value	# of Patients (%)
Total Patients		567,660
Age (years), (mean \pm SD)		66.1 \pm 14.1
Sex		
	Male	288,883 (50.9%)
	Female	278,588 (49.1%)
	Unknown	189 (0.0%)
Race		
	African American	745 (0.1%)
	Asian	489 (0.1%)
	Hispanic	8,312 (1.5%)
	Other	27,239 (4.8%)
	Unknown	110,261 (19.4%)
	White	420,614 (74.1%)
Home Region		
	Midwest	93,654 (16.5%)
	Northeast	96,786 (17.0%)
	Other	3,495 (0.6%)
	South	247,148 (43.5%)
	West	126,577 (22.3%)

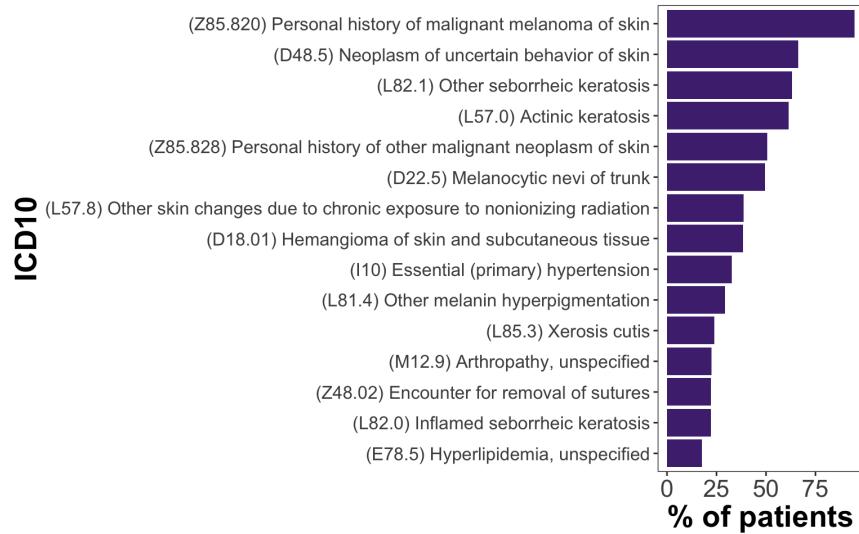


Figure 3.2: Medical conditions (ICD10)

medical staff from a standardized list, and can be mapped to ICD or SNOMED codes. Additionally, ICD and SNOMED codes can be added to a patient’s medical history separately from these standardized lists. Any diagnoses that occur in an encounter also generate ICD codes. Family history records conditions that the patient’s family members have (or have had). Social history includes various items related to a patient’s lifestyle, such as smoking status, sunscreen use, alcohol consumption, exercise status, and more. Family history and social history are also recorded by standardized lists, and can be mapped to SNOMED codes.

For uniformity in research applications, all medical conditions are mapped to ICD10 codes, while all other items (family history, procedures, social history) are mapped to SNOMED codes. Figure 3.2 presents the most frequent conditions. Figure 3.3 presents the most frequent SNOMED codes, broken down by category. Note that alcohol consumption is recorded by the number of alcoholic drinks consumed per day and is not mappable to SNOMED codes.

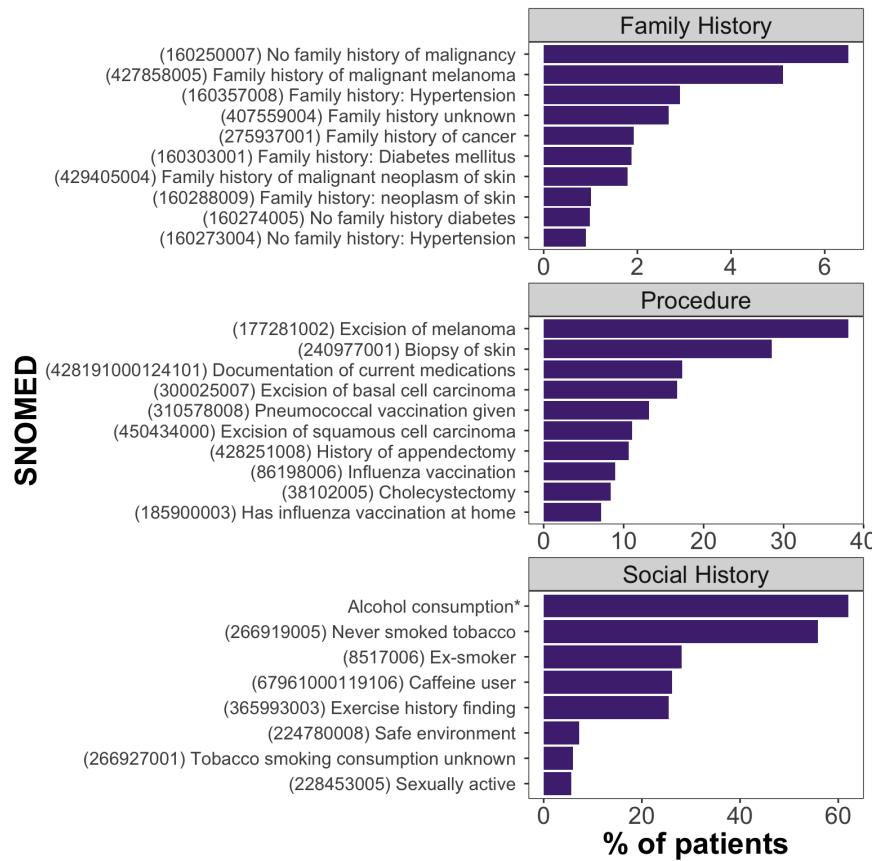


Figure 3.3: Family history, procedures, and social history (SNOMED)

* Alcohol consumption not mapped to SNOMED

Medications

Medications that are prescribed by a physician using the EHR, as well as any other medications a patient may be taking, are available. The medications adhere to both RxNorm and NDC standards, as provided by FirstDataBank¹. In addition, any drug allergies recorded for a patient are available.

Pathology

While clinical lab results are standardized according to LOINC, there does not exist a coding standard for pathology results, particularly for skin biopsies. Modernizing Medicine maintains a structured database of clinical biopsy results, allowing physicians to easily record the results of patient biopsies and allowing for clinical research with these results.

For pathology results that come back as cancerous or precancerous, a cancer log is used to track the cancer and submit data to state cancer registries. Variables include diagnostic information, various tumor characteristics and treatment history. At each visit, cancer interval history may be recorded to track any changes to the cancer site. Many of these variables, however, are optional to record. Figure 3.4 illustrates the pairwise completion rates of various variables for melanoma cancer entries.

Patient Intake and History

Chief complaints, history of present illness, and review of systems (ROS) encompass the various data points that are recorded as part of patient intake in an encounter. Generally, this is recorded by a medical assistant asking the patient a series of questions about why they are visiting the physician, the nature of their illness, and a review of their bodily systems. All responses are recorded by body location, select or checkbox selections, and numeric input. Any free-text input is not available in this

¹<http://www.fdbhealth.com/>

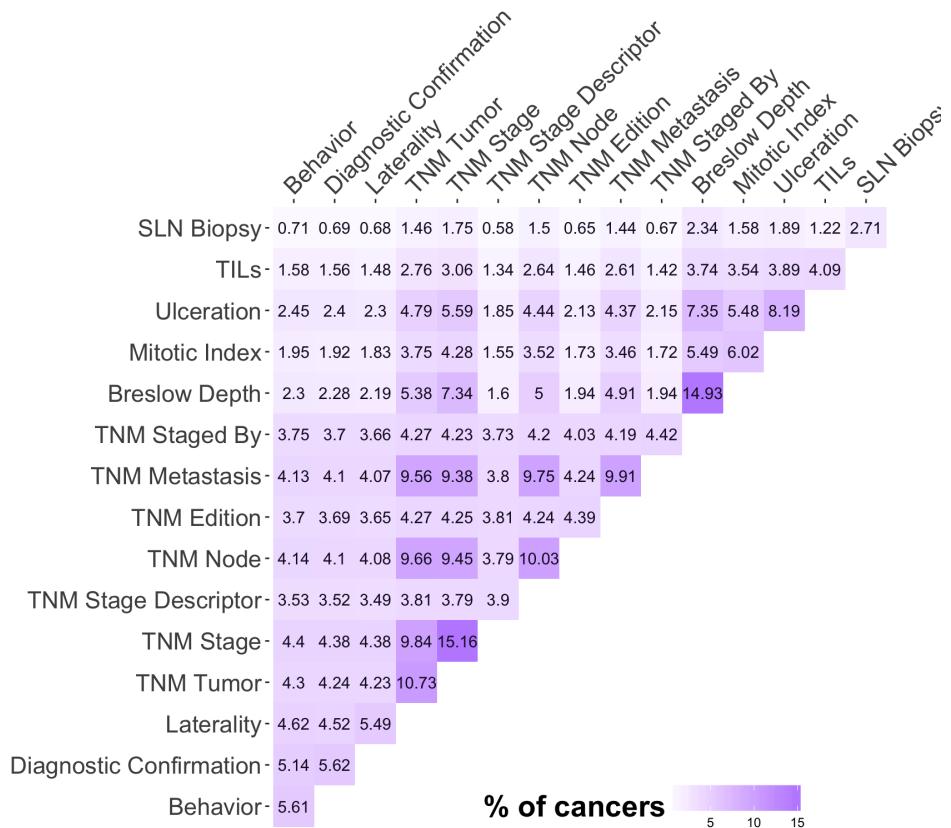


Figure 3.4: Pairwise completion rates of cancer variables. Bottom diagonal indicates single variable completion rate. See [20] for more information about melanoma staging variables.

Table 3.4: Top chief complaints

Chief Complaint	# of Patients (%)	# of Questions	Avg. # of Values*
Skin lesion	270,639 (47.7%)	13	14.9
Evaluation of skin lesion(s)	224,537 (39.6%)	14	13.4
History of malignant melanoma	163,221 (28.8%)	12	15.4
Skin lesions	162,666 (28.7%)	13	13.6
Full body skin examination	112,264 (19.8%)	10	8.6
Secondary complaint	97,848 (17.2%)	7	18.8
Rash	72,378 (12.8%)	12	8.8
Skin check	61,905 (10.9%)	6	7.0
History of non-melanoma skin cancer	45,594 (8.0%)	8	11.8
Procedure (skin surgery)	31,375 (5.5%)	10	12.3

* Number of select/checkbox values for each question (excludes numeric questions).

dataset. The most frequently used chief complaints are presented in Table 3.4

The most frequently recorded ROS questions are presented in Table 3.5. Individual practices may add their own custom questions; however, these are not available in MAMEL. Various vital signs, including blood pressure, height, weight, pulse, respiration, and temperature, may also be recorded in each patient encounter.

Exam, Diagnosis, and Procedure

Exams, diagnoses, and procedures (or plans) are all recorded by a physician as part of a patient encounter. An exam is performed by the physician prior to any diagnoses or treatments, which involves the physician noting the body elements that were examined (Table 3.6). A diagnosis is selected after the exam to note unusual findings, and to select an appropriate procedure or plan to perform. Only one exam is performed per visit, but multiple diagnoses can be recorded in a visit, and multiple procedures or plans can be performed for each diagnosis. Body locations, morphologies (Figure 3.5), and outcome measurements are available for each diagnosis.

Table 3.5: Top ROS questions

Question	Yes*	No*
Allergy to adhesive	76,604 (2.13%)	1,223,899 (34.0%)
Allergy to lidocaine	4,129 (0.11%)	1,278,455 (35.5%)
Blood thinners	323,756 (8.99%)	1,028,097 (28.6%)
Defibrillator	12,471 (0.35%)	1,255,446 (34.9%)
Joint aches	82,731 (2.30%)	407,896 (11.3%)
Pacemaker	46,627 (1.30%)	1,260,166 (35.0%)
Problems with bleeding	142,270 (3.95%)	1,469,346 (40.8%)
Problems with healing	84,078 (2.34%)	1,573,965 (43.7%)
Problems with scarring (hypertrophic or keloid)	55,706 (1.55%)	1,481,044 (41.1%)
Rash	99,816 (2.77%)	1,070,650 (29.7%)

* # of Visits (%) with yes/no response

Table 3.6: Top exam elements

Exam Element	# of Exams (%)
Head (including face)	2,573,010 (88.4%)
Mood and affect	2,299,077 (79.0%)
General appearance of the patient	2,298,697 (78.9%)
Orientation to time, place and person	2,266,715 (77.8%)
Neck	2,223,682 (76.4%)
Left upper extremity	2,121,534 (72.9%)
Right upper extremity	2,119,424 (72.8%)
Back	2,031,371 (69.8%)
Chest	2,021,923 (69.4%)
Scalp (including hair inspection)	1,943,372 (66.7%)

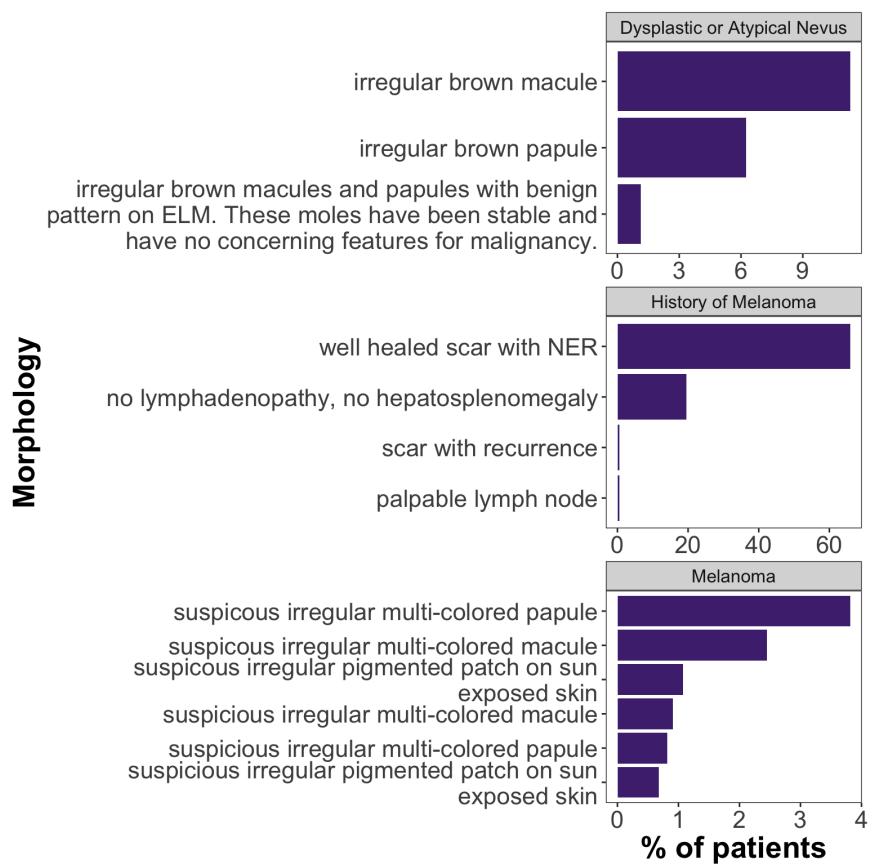


Figure 3.5: Top morphologies by diagnosis



Figure 3.6: Top procedures by diagnosis

Additional data for each procedure includes body locations and other procedure-specific variables (recorded by select, checkbox, and numeric inputs). Figure 3.6 shows the most frequent procedures for various diagnoses. For biopsy or excision procedures that require pathology testing, the pathology record is tracked separately from the visit, because results are obtained after the patient encounter has concluded.

Billing

Billing information is recorded as part of each visit, which includes CPT codes for each procedure, ICD diagnosis codes, and CPT modifiers. The EHR only records bills that are generated from each patient encounter, not what was actually paid by the patient or insurance.

3.4 DISCUSSION

3.4.1 Related Datasets

A PubMed search was conducted to identify review papers for melanoma retrospective studies within the past 5 years using the following search string:

```
(melanoma [MeSH Terms] OR melanoma [ All Fields ])  
AND (survey [Title] OR review [Title])  
AND (Review[ptyp])  
AND (2012/03/27[PDat] : 2017/03/25[PDat])
```

Of the 538 results, 9 reviews of retrospective and/or surveillance studies gave descriptive statistics of the dataset in each study [30,37,38,41,42,63,104,138,177]. In all studies reviewed, the largest sample size of melanoma patients was 162,078 [54]. To the best of our knowledge, the 567,660 melanoma patients in MAMEL constitutes the largest real-world observational database of melanoma patients in the world.

As seen from the above literature search, many institutions and academic centers that study melanoma have datasets similar to MAMEL. Particularly, these datasets contain clinical data about patients and diagnostic and treatment data about their cancers. Institution-specific datasets have several limitations, however. While larger, well-known centers can attract patients from different geographic regions, the care is still localized to the specific institution. There can be bias in diagnostic and treatment strategies due to the unique experiences of the physicians and researchers at the center. Generalizability of these datasets is also limited, as each institution employs different data collection strategies and captures disparate sets of variables. MAMEL addresses these limitations by collecting the same variables for all patients in dermatology offices spread throughout the U.S. This allows for research conducted with MAMEL to be generalized, as it is representative of thousands of diverse patients and medical practices.

Another source of cancer data is through cancer registries, traditionally maintained by governing bodies. Typically, this involves physicians submitting a specific set of data points about each new cancer they diagnose. In the U.S., each state maintains a cancer registry and requires diagnosing physicians to report cancers to the registry [12]. Generally, a minimal number of variables are collected in these registries, leading to selection bias and missing data [106]. In 2010, the American Academy of Dermatology performed a survey during its annual meeting and found that over half of dermatology providers were not aware of reporting requirements for melanoma [31]. Additionally, Raji et al. found that most cancer registry data is obtained from hospitals rather than private practices [122]. This results in under-reporting of less severe, early-stage melanomas. MAMEL addresses these issues by providing a large nationwide database of dermatology patients, and includes many variables beyond those collected in cancer registries.

3.4.2 Necessity of Structured and Available Clinical Data

The field of cancer risk modeling can benefit most by increasing the amount of data that is available to researchers and machine learning experts. This advancement is hindered by the lack of structured clinical data available in EHR systems, as many still record free-text clinical notes. Medical providers must also utilize all the functionalities available in an EHR system to capture the most complete and valuable data. Paré et al. studied family practice physicians in Canada, and found that the majority of them did not utilize all available features in their EHR systems, which included e-prescribing, electronic lab ordering, secure data transmissions, and more [112]. Additionally, data privacy concerns often result in institutions or cloud-based EHR systems keeping terabytes of data locked away in private servers, especially if the data is free-text, as it is especially difficult to de-identify clinical notes [150]. Research in anonymization techniques must continue to help alleviate these concerns [81], as

well as policy advancements to allow for more data sharing without breaching patient privacy.

While the adoption of EHRs has increased due to governmental requirements, the EHR industry is fragmented and data sharing is difficult. Standards need to be developed and enhanced to allow sharing of detailed clinical information. Through a study of mental health patients in Massachusetts, Madden et al. found that over half of the incidents of outpatient care were not captured in the patients' EHR system, as they occurred outside of the medical practice [94]. These data points were still covered, however, by insurance claims data. Ahmadian et al. specifically studied the data standards used in clinical decision support systems, and found that many users of these systems were limited by incomplete data sharing standards and capabilities [10].

Due to practical necessity and compliance requirements, EHR systems record patient information beyond demographics such as family history, smoking status, and alcohol use. These can provide valuable insights for modeling clinical data, as there may be hidden biomarkers that contribute to medical conditions. Additionally, EHR systems record real-world clinical data at the point-of-care, making models built from these datasets more generalizable to the public. Clinical trials and prospective observational studies may have small cohort sizes and can be biased towards the patients in the study.

Data must also be shared between clinical and non-clinical settings. For example, four studies from South Korea used data that were linked from a physical health examination, the national cancer registry, and the national death registry [50, 111, 145, 179]. This allowed for large-scale population health analysis, and they were able to build personalized predictive models for many different types of cancers. Razavi et al. were able to use linked data from the breast cancer registry, tumor registry, and death registry from Sweden [123].

Due to the overhead of prospective data collection, privacy and legal issues, and

modeling difficulty, studies often analyze data from many years in the past. This is not desirable, as clinical guidance is constantly changing based on medical breakthroughs and clinical trial results. A model built from data that is ten years old will be biased towards the treatments used and knowledge from that era, and may not be as accurate for current patients. Operational, policy, and data management efforts must be made to enhance the speed at which models can be built from current data. Additionally, online models can be built to utilize real-time data coming from EHR systems. While this requires major enhancements in infrastructure and data management, it will provide the most valuable models for predicting the risk and recurrence of different types of cancers. All experiments conducted as part of this dissertation utilized data from MAMEL that was updated no more than one year before the time of the experiment. Therefore, the risk models produced were able to capture up-to-date treatment patterns and patient information.

3.5 CHAPTER SUMMARY

We presented the Modernizing Analytics for MELanoma (MAMEL) dataset: a real-world, dermatology-specific research dataset specifically crafted to advance data mining and machine learning research in the field of melanoma diagnosis, analysis, and treatment. This dataset was collected and curated from Modernizing Medicine’s EMA Dermatology™ application, a cloud-based Electronic Health Record (EHR) platform. A big data processing architecture, built on Apache Hadoop and Apache Spark, was used to collect all patient data, identify patients for the MAMEL dataset, and create and document all data elements. This chapter outlined the application and data processing architectures and provided an exploratory analysis of data elements available in MAMEL. Subsequent chapters utilize datasets derived from MAMEL for experimentation.

CHAPTER 4

PREDICTING SENTINEL LYMPH NODE METASTASIS IN MELANOMA

4.1 BACKGROUND AND MOTIVATION

Metastasis of the sentinel lymph node (SLN), the closest lymph node to a cancer on the skin, is one of the most important prognostic indicators for melanoma survival [95]. The 5-year survival rate for patients with an early melanoma detection is 98%, while the rate drops to 62% if the cancer spreads to a lymph node. Therefore, early detection and treatment of melanoma is paramount. If there is positive sentinel lymph node metastasis, this means that cancer cells have spread beyond the location of the melanoma on the skin [84]. Metastasis is determined by taking a biopsy of the sentinel lymph node, which adds additional cost to treatment (beyond the initial surgical excision of the tumor) [98]. If the biopsy returns positive for metastasis, a patient can undergo an elective lymphadenectomy (lymph node dissection) to remove potential cancerous lymph nodes. For thin melanomas ($<1\text{mm}$), the risk of metastasis is low, so an SLN biopsy is only recommended when additional risk factors are present for the patient [174]. Traditional guidance is to not perform a sentinel lymph node biopsy for these tumors, but some studies have shown that thinner melanomas can metastasize [182].

The goal of a sentinel lymph node metastasis prediction model is to guide physicians on whether or not to suggest an SLN biopsy for a newly diagnosed melanoma. Current guidelines suggest that thick melanomas ($>1\text{mm}$) should be biopsied [182]; however, if a model can accurately predict low risk for these cancers, the procedure

can be foregone, saving healthcare costs. More importantly, high risk patients with thin melanomas ($<1\text{mm}$) can be identified for an SLN biopsy. Then if the biopsy returns positive, the patient can undergo more targeted treatments for the cancer before it spreads further. This latter case has higher consequences than the first, as an SLN biopsy is a generally safe procedure, so false positives from the model are not that dangerous. A false negative, however, means that a patient with a metastasized cancer may not know about it until it is too late for treatment. Therefore, a model with high sensitivity (TPR) is the ultimate goal, especially for thin melanomas. Most patients with thick melanomas ($>1\text{mm}$) are recommended for an SLN biopsy, so a model with high specificity for thick melanomas would save healthcare costs by avoiding an unnecessary procedure.

In this chapter, we examine MAMEL data related to melanoma patients and lymph node metastasis [128]. We explore a heuristic model that reflects current clinical practice as well as several machine learning algorithms trained on the dataset. Validation of the models show that the machine learning algorithms achieve an AUC comparable to the heuristic model, but significantly higher sensitivity for thin melanomas and significantly higher specificity for thick melanomas. This shows that to calculate the probability of SLN positivity, the heuristic model is an accurate measure for the majority of melanomas. To discriminate between positive and negative metastasis, however, the heuristic model does not work well, especially for thin melanomas. In this case, the models we propose can provide valuable aid to physicians when deciding whether or not to recommend an SLN biopsy for patients with thin melanoma. This study is not the first effort to build a model to predict SLN status; we present an existing model in Section 4.2. Descriptive statistics about the SLN dataset, experimental design, and methods are provided in Section 4.3, and results and discussion in Section 4.4.

The screenshot shows a web-based form titled "Enter Your Information". At the top right are "Clear" and "Calculate" buttons. A note states "All fields are required unless noted optional".

How old are you?
56 years (20 to 95)

What was the thickness of your melanoma?
2.3 mm (0.1 to 10)
Note: If the tumor thickness is less than 0.1 mm, enter as 0.1 mm.

What was your Clark level?
Level III
Note: This prediction tool applies only to Clark levels II to V.
► [More on Clark levels](#):

Where was your melanoma located?
Trunk

Was there ulceration reported in your pathology report?
 Yes No
► [What is ulceration?](#)

Calculate **Clear**

Figure 4.1: Nomogram model online form¹

4.2 RELATED WORKS

Wong et al. developed a predictive model to determine sentinel lymph node status in melanoma patients [175], and several other studies applied the model to different melanoma populations [113, 115, 176]. The fitted model was developed into a nomogram [19], a visual tool to make predictions from several variables. Additionally, the model is publicly available for use by patients and physicians through an online form (Figure 4.1). Throughout our study, we refer to this model as the “nomogram model.”

The study was published in 2005, and compares the predictive accuracy of a logistic regression model to staging guidance from the American Joint Committee on Cancer. The five selected variables were age, thickness, Clark level, body location, and ulceration. This study did not have many patients with thin melanomas (186 out of 979 total patients) that would benefit from this model. It has been shown that thicker melanomas (>1mm) warrant an SLN biopsy, so it is difficult to generalize

¹<https://www.mskcc.org/nomograms/melanoma/sentinel-lymph-node-metastasis>

the nomogram model. Additionally, an author from Wong et al. suggested that the model is not used in clinical practice, but rather a rough calculation using the tumor depth is used to determine the probability of SLN metastasis (D. Coit, personal communication, June 27, 2017).

While a prediction model for SLN metastasis does exist, building a new model with the MAMEL dataset addresses several limitations identified in Wong et al. The nomogram model was built using a patient cohort from a single cancer center, and has been validated on four other localized cohorts. The MAMEL data includes patients from dermatology offices throughout the U.S., reducing bias in the data and allowing for more generalized testing and validation. The nomogram model was built from 979 patients, and validated on cohort sizes of 124 [176], 218 [115], 543 [113], 3,108 [175], and 3,286 [175] patients. This study includes over 5,000 patients from the MAMEL dataset with a recorded SLN status. The nomogram model collected 13 variables to build the model, and selected 5 in the final model. MAMEL collects thousands of structured clinical variables that can be used to build a model. While not all 13 variables that Wong et al. initially used are available in MAMEL, the final 5 variables are. Additionally, since the data is available as part of an EHR system, a small number of variables do not need to be selected to simplify a web input form or nomogram. If the prediction model is built into the EHR application, any number of variables can be used as input since they are readily available in the application's database. The nomogram model only considered patients with \geq 1mm or Clark level II-V, resulting in only 19% of patients having thin melanomas. We include all melanomas with an SLN biopsy result recorded, 54% of which are thinner than 1mm.

The dataset in this study, derived from the MAMEL dataset, includes real-world data from diverse practices throughout the country. This allows a model built from this data to be generalized to a wide range of patients. Additionally, we do not limit inclusion based on tumor thickness, providing guidance for thinner and less advanced

melanomas.

4.3 MATERIALS AND METHODS

Data

Patients were included in this study if they had a cancer log entry for melanoma and the following data points collected: (1) SLN biopsy result, and (2) tumor thickness (Breslow thickness). Since the data is collected from a real-world EHR system, records with inconsistent data were excluded. This could be due to a patient following up with a dermatologist over an extended period of time with variations in data that are recorded from multiple patient intakes or cancer log records.

For each patient, the following structured data was collected: age, sex, race, melanoma family history, degree of relation to the relative with melanoma, geographic location (U.S. state), height, weight, Fitzpatrick skin type, diagnoses (ICD10 codes), procedures (SNOMED codes), family history (SNOMED codes), drug allergies, and prescriptions. For each individual melanoma the following structured data was collected: subtype, date of biopsy, body location, Clark level, tumor depth, mitotic rate, ulceration, presence of tumor-infiltrating lymphocytes, and SLN biopsy result. Tables 4.1-4.2 outline several important demographic and clinicopathologic features for the patients in the study.

Since literature recommends that melanomas thinner than 1mm may not warrant an SLN biopsy, we split the data into three subsets to explore classifier performance on each group: (1) all records, (2) melanomas $\leq 1\text{mm}$, and (3) melanomas $> 1\text{mm}$. Table 4.3 shows the class distributions for each dataset.

Machine Learning Techniques

For each dataset, we built a logistic regression, decision tree, and random forest model to predict the SLN biopsy result. Due to the class imbalance present, particularly in

Table 4.1: Demographics

Variable	Value	n (%)
Total Patients*		5,105
Age (years)	(mean \pm SD)	63.2 \pm 14.8
Sex	Female	2,428 (47.56%)
	Male	2,677 (52.44%)
Race	Asian or African American	11 (0.3%)
	Hispanic	67 (1.3%)
	White	4,008 (78.5%)
	Other	281 (5.5%)
	Unknown	738 (14.5%)
Home Region	Midwest	1,144 (22.4%)
	Northeast	843 (16.5%)
	South	2,102 (41.2%)
	West	994 (19.5%)
	Unknown	22 (0.4%)
Home Division	East North Central	747 (14.6%)
	East South Central	310 (6.1%)
	Middle Atlantic	586 (11.5%)
	Mountain	559 (11.0%)
	New England	257 (5.0%)
	Pacific	435 (8.5%)
	South Atlantic	1,401 (27.4%)
	West North Central	397 (7.8%)
	West South Central	391 (7.7%)
	Unknown	22 (0.4%)
Skin Type	I	107 (2.1%)
	II	1,378 (27.0%)
	III	91 (1.8%)
	IV-VI	8 (0.2%)
	Unknown	3,521 (69.0%)
Height (m)	(mean \pm SD)	1.7 \pm 0.1 (88.9% unknown)
Weight (kg)	(mean \pm SD)	83.2 \pm 18.5 (87.8% unknown)

* 24 patients have more than one melanoma recorded, resulting in 5,126 total records.

Table 4.2: Tumor characteristics

Variable	Value	n (%)
Total Records		5,126
Subtype	Acral Melanoma	14 (0.3%)
	Lentigo Maligna Melanoma	90 (1.8%)
	Melanoma (Not Subtyped)	4,160 (81.2%)
	Nodular Melanoma	184 (3.6%)
	Superficial Spreading Melanoma	678 (13.2%)
Clark Level	I	15 (0.3%)
	II	234 (4.6%)
	III	290 (5.7%)
	IV	612 (11.9%)
	V	19 (0.4%)
	Unknown	3,956 (77.2%)
Thickness	T1	2,757 (53.8%)
	T2	1,526 (29.8%)
	T3	704 (13.7%)
	T4	139 (2.7%)
Mitotic Rate	< 1/mm ²	1,037 (20.2%)
	≥ 1/mm ²	498 (9.7%)
	Unknown	3,591 (70.1%)
Ulceration	Absent	3,270 (63.8%)
	Present	515 (10.0%)
	Unknown	1,341 (26.2%)
TILs	Absent	610 (11.9%)
	Brisk	364 (7.1%)
	Non-Brisk	736 (14.4%)
	Unknown	3,416 (66.6%)
SLN Biopsy Result	Negative	4,599 (89.7%)
	Positive	527 (10.3%)

Table 4.3: Class distributions

Thickness	Negative SLN Biopsy	Positive SLN Biopsy	Class Distribution
≤1mm	2,665	92	3.3%
>1mm	1,934	435	18.4%
All	4,599	527	10.3%

the <1mm dataset, we performed RUS to achieve three different positive class ratios: 50%, 35%, and 20%. We also performed a run of each model without sampling to determine its impact. All categorical feature values were converted to one-hot encoded features, resulting in 2,100 total features available for the experiment. For binary variables, the second extracted feature was removed since it contains the exact inverse of the first extracted variable. Missing values in categorical features were treated as a value (extracted as a new feature), and mean imputation was performed for missing values in numeric features. Features with zero variance (same value in all samples) in each training split were removed. Feature ranking was performed using the χ^2 statistic and the top K features were selected using K values of 5, 10, 20, 30. Additionally, one run was performed without the feature ranking technique (only remove features with zero variance).

Models were built using 5 iterations of 5-fold cross-validation and evaluated using the area under the ROC curve. Across the 3 datasets, 3 classifiers, 4 class balances, and 5 different feature selection strategies were applied, resulting in over 4,500 total models built. The best performing configuration (according to AUC) was selected for each dataset and model, and 30 runs of a 70/30 stratified train-test split was used to validate the models.

Tumor thickness is an important prognostic factor for melanoma, and many clinical decisions are made based on this thickness. Traditional guidance is for physicians to biopsy a tumor $>1\text{mm}$, but there are not many studies providing evidence-based guidance for the thinner melanomas. Therefore, a dermatologist may only suggest an SLN biopsy for a thin melanoma if there are other risk factors present in the pathology of the tumor or the patient’s background. A rough estimate of SLN positivity is to multiply the thickness by 10 to achieve a percentage (or divide by 10 to get the

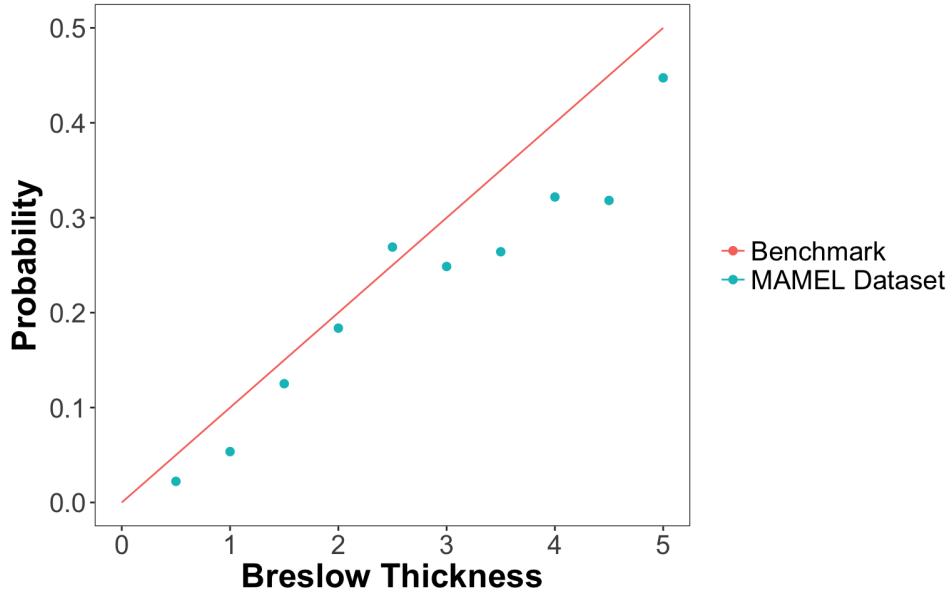


Figure 4.2: Probability of SLN metastasis vs. tumor thickness. The line indicates estimated probability from the heuristic model. The points indicate the actual probability within each 0.5mm group of thickness ($R^2 = 0.934$).

decimal value):

$$P(SLN = \text{positive}) = \frac{\text{thickness}}{10} \quad (4.1)$$

Upon investigation of this heuristic model, we found it to have good predictive performance, especially for a model using a single variable. Figure 4.2 shows the probability of metastasis from the MAMEL dataset versus the probability calculated by this heuristic model. This benchmark is referred to as BT (Breslow thickness) in subsequent figures and text.

In addition to model selection, the cross-validation results were used to select the decision threshold for each model. For the full and $\leq 1\text{mm}$ datasets, the best threshold was one in which the sensitivity is maximized without a large drop in specificity. For the $> 1\text{mm}$ dataset, the opposite was achieved. This is due to the nature of clinical

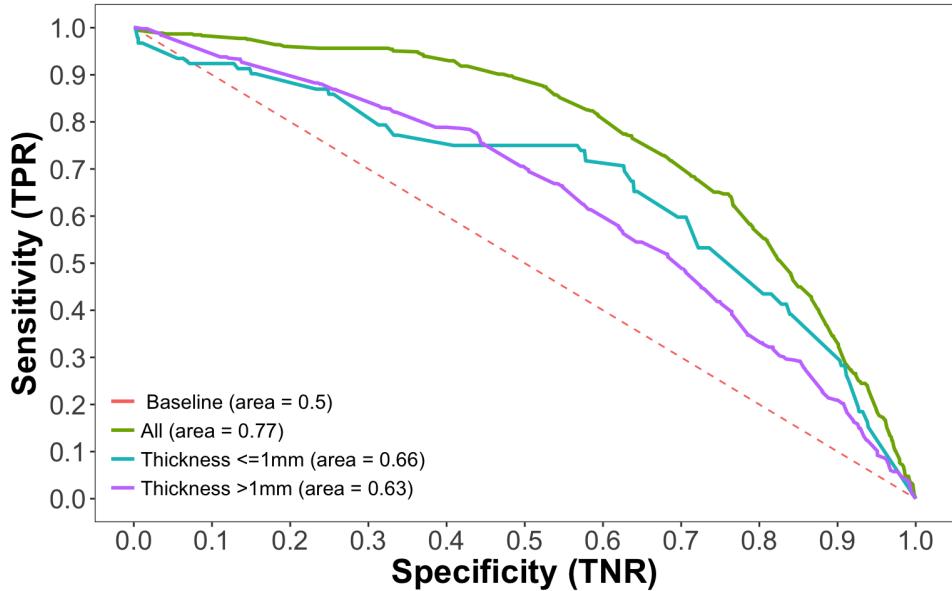


Figure 4.3: ROC curves for the benchmark model

response to differing thicknesses of melanoma; melanomas $\leq 1\text{mm}$ that are marked as negative could be metastatic melanomas that are missed. For those $>1\text{mm}$, most physicians recommend an SLN biopsy, so the metastasis would be detected. A higher specificity for thicker melanomas helps avoid having to perform the biopsy for very low-risk patients.

4.4 RESULTS AND DISCUSSION

Each model built in this study was compared to the benchmark heuristic model presented in Equation 4.1. ROC curves for the benchmark model on each dataset are presented in Figure 4.3. The model achieves good AUC on the full dataset (0.769) and poorer results when the dataset is stratified on tumor thickness (0.666 and 0.636). To simulate clinical decision-making, the discrimination threshold for the benchmark model was set to 0.1, which is equivalent to a 1mm thickness [182]. The effect of this discrimination threshold is illustrated in Figure 4.4; the model achieves 0% sensitivity for the $\leq 1\text{mm}$ and 0% specificity for the $>1\text{mm}$ dataset.

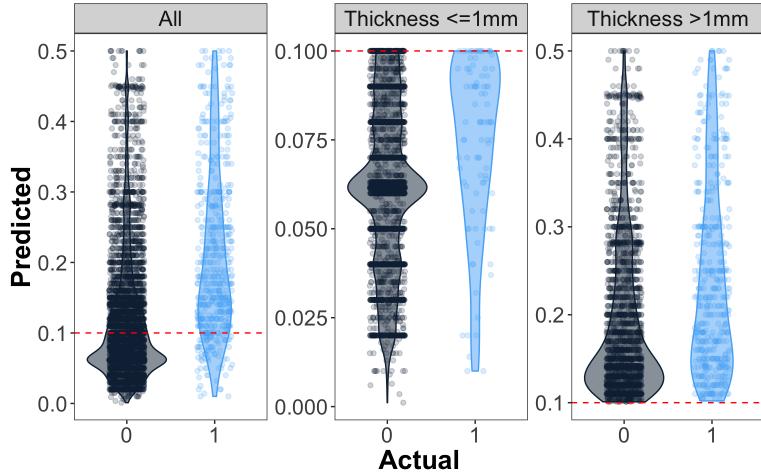


Figure 4.4: Predicted probabilities of each sample using the benchmark model. The dotted red lines indicate the discrimination threshold (0.1). Samples above the line are classified as positive, while samples below the line are classified as negative.

The best performing models (as measured by cross-validation AUC) were chosen for each classifier and dataset, along with a discrimination threshold to use in the validation step. The selected model configurations are displayed in Table 4.4. The AUC values of the models compared to the benchmark are shown in Figure 4.5, along with the sensitivities and specificities in Figure 4.6. Table 4.5 displays all validation metrics and the result of a two-sample t-test between the benchmark results and each model.

Most machine learning models for each dataset did not have a significantly higher or lower AUC than the benchmark. Logistic regression on the >1mm dataset was the only machine learning model that had a significantly higher AUC than the benchmark. Regarding sensitivity and specificity, however, the machine learning models significantly outperformed the benchmark in certain crucial scenarios. For thin melanomas, the benchmark had 100% specificity but 0% sensitivity. This is problematic as all thin melanomas that indeed have SLN metastasis would be missed. The random forest model was able to achieve a 78.9% sensitivity and a 49% specificity, catching these

Table 4.4: Model configurations

Dataset	Classifier	Undersampling Ratio	Top K Features	Threshold
All	LR	0.5	5	0.4
	Tree	None	10	0.1
	RF	0.5	All	0.4
$\leq 1\text{mm}$	LR	None	All	0.02
	Tree	None	All	0.04
	RF	0.35	All	0.3
$>1\text{mm}$	LR	0.2	5	0.2
	Tree	0.2	5	0.2
	RF	0.5	All	0.5

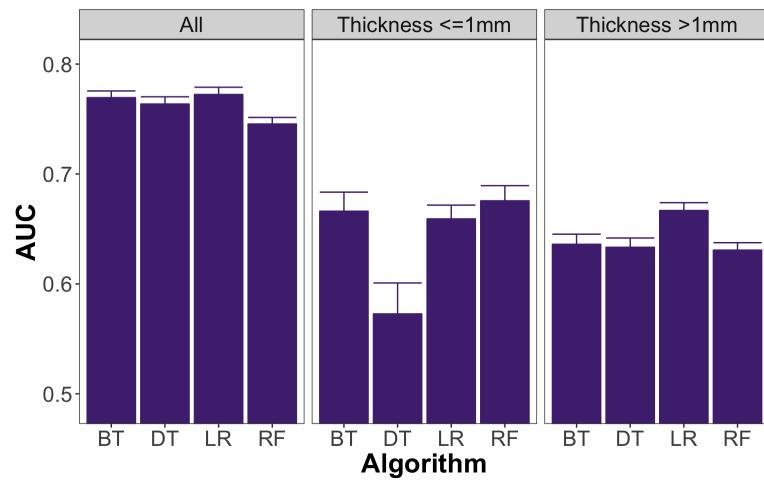


Figure 4.5: AUC values of each selected model compared to the benchmark. Lines above the bars indicate the end of the 95% confidence interval.

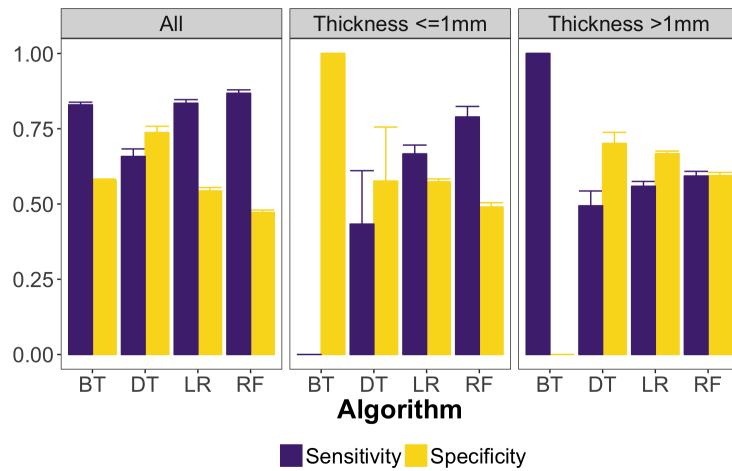


Figure 4.6: Sensitivity and specificity values of each selected model compared to the benchmark. Lines above the bars indicate the end of the 95% confidence interval.

Table 4.5: Validation results

Dataset	Model	Sensitivity	Specificity	Balanced Accuracy	AUC
All	BT	0.830	0.578	0.704	0.769
	LR	0.835	<i>0.544</i>	<i>0.689</i>	0.772
	DT	<i>0.658</i>	0.737	0.698	0.764
	RF	0.868	<i>0.472</i>	<i>0.670</i>	<i>0.745</i>
Thickness \leq 1mm	BT	0.000	1.000	0.500	0.666
	LR	0.667	<i>0.574</i>	0.620	0.659
	DT	0.433	<i>0.576</i>	<i>0.510</i>	<i>0.573</i>
	RF	0.789	<i>0.490</i>	0.640	0.676
Thickness >1mm	BT	1.000	0.000	0.502	0.636
	LR	<i>0.559</i>	0.667	0.613	0.667
	DT	<i>0.494</i>	0.701	0.598	0.633
	RF	<i>0.593</i>	0.594	0.594	0.631

Bold or italic text indicate values significantly better or worse than the benchmark, respectively ($p < 0.05$).

Table 4.6: LR model coefficients: Full dataset

Variable	Weight
Intercept	0.1174
Thickness	0.8513
Ulceration=Present	0.4203
Age	-0.0215
Mitotic rate $\leq 1\text{mm}^2$	-0.9622
ICD10=L82.0	-1.061

very important cases. For thick melanomas, the benchmark had 100% sensitivity and 0% specificity since all cases are marked as positive. Only 18% of thick melanomas were actually metastatic (Table 4.3), resulting in a very high false-positive rate. The logistic regression model was able to achieve more balanced hit rates: 55.9% sensitivity and 66.7% specificity.

No tree models achieved a significantly higher AUC than the benchmark, and a couple performed significantly worse. This shows that there is a strong linear relationship between the tumor thickness and probability of SLN metastasis, as seen in the model coefficients for the LR model on the full dataset (Table 4.6). While the selected LR configuration for the $\leq 1\text{mm}$ dataset included all non-zero variance features (1,615), Table 4.7 shows the coefficients for an LR model with the top 10 features selected ($\text{AUC}=0.651$, $p=0.153$). It is interesting to note here that the tumor thickness is not part of the model, indicating that there are indeed other factors contributing toward SLN metastasis for thin melanomas.

We were not able to compare models built in this study to the nomogram developed by Wong et al. [175], as we were not able to retrieve the model coefficients from the authors. Although MAMEL is a de-identified dataset, we opted to not run each sample through the online model due to patient privacy concerns and time constraints. Through personal communication with an author, however, we understood that the team instead estimates the probability of a positive sentinel node to

Table 4.7: LR model coefficients: $\leq 1\text{mm}$

Variable	Weight
Intercept	-0.2166
ICD10=L85.3	1.195
Clark level=IV	1.073
Home division=Pacific	0.5977
ICD10=D48.5	-0.3681
ICD10=L82.0	-0.7769
ICD10=Z87.2	-0.8275
ICD10=D23.5	-0.9812
Body zone=Face	-0.9996
SNOMED=450434000	-1.077
Mitotic rate $\leq 1\text{mm}^2$	-1.545

be 10 times the tumor thickness (D. Coit, personal communication, June 27, 2017).

We found that this heuristic actually performed on par with any machine learning model we built for the entire population. Considering clinical practice, we found it necessary to focus specifically on thin melanomas as these are the cases that would benefit from a prediction model. Most physicians will always recommend an SLN biopsy for thick melanomas [182]. The population in Wong et al. was collected from a cancer institution, which limited the number of low-risk melanomas that were seen. MAMEL, however, is collected from a private-practice dermatology EHR system that sees many more average and low-risk melanomas. In Wong et al., 19% of cases were $\leq 1\text{mm}$ thick compared to 54% of cases in the present study. Additionally, there was a higher percentage of SLN positivity in Wong et al. versus this study (16% to 10%).

There are several limitations in this study, mainly relating to the nature of the data collection. Since the data is collected from a real-world private-practice EHR system, there are many missing values in variables that are clinically relevant for melanoma (Table 4.2). As this was a retrospective real-world analysis, the data completeness does not have the same rigor of a prospective study, potentially reducing the quality of the data collected. Additionally, since traditional guidance is to not biopsy thin

melanomas, the sample size of thin melanomas with SLN positivity is very small.

Future work involves improving the algorithms to be able to deploy a predictive model in the clinical setting. We have shown that there is a strong linear relationship between tumor thickness and the probability of SLN metastasis; models that exploit this linear relationship should be studied further. We can also employ regularization methods to reduce the impacts of anomalous feature values. More advanced machine learning techniques can be used, such as model-based feature selection techniques and deep learning algorithms. Rather than predicting sentinel lymph node status alone, a model can be developed that also predicts hematogenous spread. It has been shown that about 50% of metastatic melanomas are distant, satellite, or in-transit metastases [86]. A model that predicts SLN metastases in addition to hematogenous metastases would help physicians understand the progression of melanoma and treat patients accordingly.

4.5 CHAPTER SUMMARY

Sentinel lymph node metastasis is one of the most important prognostic indicators for melanoma survival. We presented several machine learning models for predicting SLN metastasis using data from MAMEL. The class label is the result of a sentinel lymph node biopsy, an elective procedure that can be performed for newly-diagnosed melanoma patients to determine if there is metastasis in the nearest lymph node. We showed that a simple model, using solely Breslow thickness, can achieve predictive performance ($AUC=0.769$) comparable to a logistic regression model using 5 features ($AUC=0.772$, $p=0.518$). Current clinical recommendations are to perform a biopsy for patients with melanomas thicker than 1mm; however, when applying this 1mm threshold to the simple thickness model, it achieved 0% sensitivity for melanomas $<1\text{mm}$. Using a random forest model, we achieved 78.9% sensitivity ($p<0.001$) for melanomas $<1\text{mm}$. This chapter shows that the probability of sentinel lymph node

positivity is indeed linearly correlated to the tumor thickness ($R^2=0.934$), and that machine learning models can effectively detect thin melanomas that warrant an SLN biopsy.

CHAPTER 5

PREDICTING MELANOMA RISK

5.1 BACKGROUND AND MOTIVATION

The goal of cancer risk prediction is to determine if a given patient will develop cancer (or recur) at some point in the future [152]. The problem is distinct from patient identification (also called phenotyping [146]), as the goal is not to determine if a patient has a certain disease at the present moment, but to determine if the patient will develop it in the future. This task can be formulated as a supervised learning problem, where the input data are certain demographic and clinical elements (e.g. age, sex, and treatment history), and the output variable is the probability that the patient will develop the cancer at some point in the future. This probability can be tracked over time, assigning risk as time increases. The problem can also be formulated as a binary classification task, attempting to ascertain whether or not a patient will develop cancer at a specified point in time (i.e. developing the cancer within the next five years). A prediction model is built by supplying historical data from patients that did, or did not, develop the cancer in question. Statistical and machine learning techniques are used to fit a model to this historical data (i.e. training data). Then, to prove the model will be generalizable to different patient populations, a validation set (or multiple validation sets) is used to determine the performance of the model. When the performance of the model is adequate, based on several metrics, it can be deployed into clinical settings to help inform patients and providers. For more information about predictive modeling for medicine in general, see [25, 152].

Accurate models are clinically relevant, as they can provide personalized treatment

plans for patients at risk for a new cancer or recurrence of cancer in remission. There are various types of cancers, many of which have a very low incidence rate. It is not economically feasible to screen all patients visiting a doctor for a wide range of different diseases [50, 179]. Thus, a model that can predict future development of cancer based on regularly captured clinical biomarkers, demographic, and lifestyle information is of high value to a healthcare system. As the model is built and tested, it can be used to flag high-risk patients for enrollment in a surveillance program, catered towards each patients' individual risk and clinical profile [147]. Therefore, a model must be applicable to large populations of patients, given that cancer is still a relatively rare disease but one of high importance to humanity.

If a large quantity of consistent patient data can be collected for a predictive model, computational challenges arise when transforming the data and training a machine learning algorithm. First, data elements must be extracted from the data collection system and transformed into a tabular format to be passed to a machine learning model. The size of the dataset and complexity of the machine learning algorithm can subsequently introduce computational challenges. The cloud enables users to launch machines of varying size with prebuilt libraries for machine learning algorithms. This technology can be utilized to evaluate a wide range of algorithms to produce the most accurate model. When dealing with big data, or data that cannot be processed through traditional architectures, predictive accuracy is not the only consideration when choosing classifiers and machine learning techniques; computational complexity and cost must also be factored in the selection process.

In this section, we use MAMEL data to build classification models for melanoma risk. Data from real-world outpatient dermatology visits are used to identify variables that are most indicative of the patient developing melanoma in the future. We take a data-first approach by creating a large vector of data for each patient and letting the algorithms determine which features are important. This allows us to utilize the

full breadth of data collected by the EHR system to inform patients and providers about melanoma risk using the patient’s own data. Therefore, it is not required for the dermatology patient to present with any concerns for melanoma. The model examines data from routine office visits and provides a risk score for the patient, to inform future follow-up visits. We explore the features selected by the model, and provide examples of predictions made using real-world patient data. A comprehensive literature review of cancer risk models is presented in Section 5.2, followed by our own experiments in Sections 5.3-5.4, with a dedicated analysis of the interpretability of the various classifiers produced (Section 5.5).

5.2 LITERATURE REVIEW

In this review, a distinction is made between models that attempt to predict if a patient will develop a cancer in the future (risk prediction), and those that predict whether or not a patient will relapse after a potentially-curative treatment (recurrence prediction). These problems are distinct in that they often have different types of input data. For example, a risk prediction model will not have any variables about cancer in the patient, as the patient has not yet developed cancer (although family histories of cancer would be relevant). For recurrence models, as will be seen in the papers studied, information about the tumor and treatments for the cancer are often chosen for inclusion in the models [78]. While the problem scenarios are distinct, the approaches to solve them can be very similar.

5.2.1 Methodology

We conducted a comprehensive review of literature related to data mining for health-care applications, and filtered the list of works to those relevant for the experiments conducted in this work. Therefore, works focusing on other diseases besides cancer, and those using non-clinical data (such as genomic or proteomic data) or primarily

free-text clinical notes were excluded.

Papers were first identified by browsing through related journals, followed by a breadth-first search of articles using Pubmed¹ and Google Scholar². Keywords used included but were not limited to: “cancer risk”, “cancer recurrence”, “cancer prediction”, “machine learning”, “data mining”, and permutations of these keywords. Then, each paper identified was reviewed for relevance and a decision to keep or remove the paper was made. For each paper that was kept, related articles and articles citing the paper (utilizing search features available in both Pubmed and Google Scholar) were reviewed for relevance. This process was repeated until no new papers could be identified, and the results of the review are presented in Sections 5.2.2-5.2.3. There are many different types of cancers, with different risk factors and treatment options, resulting in researchers with specific and invaluable knowledge of a specific type of cancer. Therefore, each paper focuses on a particular type of cancer for modeling, with the exception of Bayati et al., who attempted to predict cancer in general [24]. No melanoma risk studies were identified through this search, but a follow-up investigation revealed several papers that are reviewed separately in Section 5.2.4.

5.2.2 Models in Practice

Several prognostic and predictive models are used, or available for use, in clinical practice. Some of these are not based on statistical or machine learning models, but rule-based methods or clinical guidelines.

Cancer Staging

The TNM Classification of Malignant Tumors is an international standard developed and maintained by the American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC) to describe the stage of a cancer tumor when

¹<http://www.ncbi.nlm.nih.gov/pubmed>

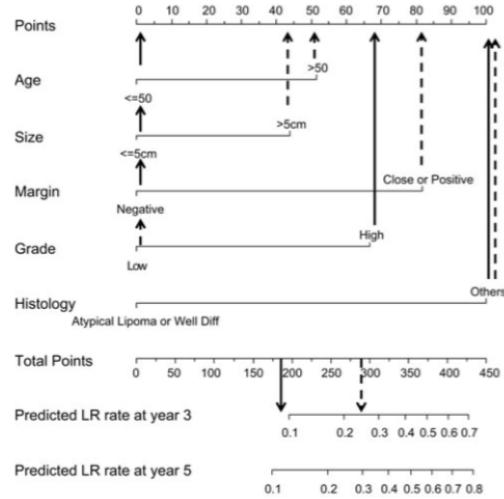
²<https://scholar.google.com/>

it is diagnosed. This standard measures the size of the primary tumor (T), spread to regional lymph nodes (N), and the presence of distant metastasis (M) [48]. The staging is used to bucket patients into mutually exclusive groups based on their tumor characteristics, providing a means to determine prognosis of the disease, including the risk of recurrence [28,29]. Several papers, namely Cahlon et al. [29], Weiser et al. [169], Bochner et al. [67], and Marelli et al. [97], built models to predict the risk of cancer recurrence, and found that their models were more accurate than using TNM staging alone.

Nomograms

A nomogram is a graphical calculating device that allows a mathematical equation to be answered by aligning a straight-edge across values of different inputs, with the end of the straight-edge pointing to the result of the equation (see Figure 5.1a). Nomograms for oncology, widely studied by researchers at the Memorial Sloan Kettering Cancer Center, can produce succinct formulas that determine a patient's risk for certain clinical events, including the development or recurrence of a cancer [19]. Rather than utilize the archaic means of aligning a ruler to a page, they publish these nomograms as online forms to be used by both physicians and patients³ (see Figure 5.1b). These nomograms were built using regression techniques, such as Cox proportional hazards or competing risk survival analysis, with the aim to use the minimum number of variables necessary to produce accurate results. Nomograms specific to cancer recurrence prediction were developed for: sarcoma (Cahlon et al. [29]), colon cancer (Weiser et al. [169]), breast cancer (Rudloff et al. [139]), and bladder cancer (Bochner et al. [67]).

³<https://www.mskcc.org/nomograms>



(a)

Sarcoma Nomogram: Local Recurrence Risk after Limb-Sparing Surgery without Radiation Nomogram

TEXT SIZE A A

This tool can be used to predict the chance of soft tissue sarcoma returning at the site of initial surgery after the tumor is removed through limb-sparing surgery if the patient does NOT receive radiation. The probability of local recurrence is calculated for both three years and five years after surgery.

Enter Your Information

Age Select the age group.	50 or younger
Tumor Size Select the tumor size.	More than 5 cm
Margin Select the tumor margin.	Positive or Close
Grade Select the tumor grade.	High grade
Histology Select the histology.	ALT/Well-diff lipo

Clear **Calculate ▶**

Your Results

Learn more about your results below.

Probability of Local Recurrence	3 Year 10%
	5 Year 12%

Print These Results

Make an Appointment

Call us to schedule an appointment or contact us online

Contact Us ▶

(b)

Figure 5.1: Example Nomogram [29]. (a) Manual nomogram. Lines are drawn from each feature to a particular score at the top line depending on the value of that feature. These points are then added up to reveal the predicted recurrence probability at either three or five years. The two styles of arrows indicate two different predictions made using the nomogram. (b) Online version of the nomogram.

Breast Cancer Recurrence Models

Kim et al. [78] built a model for predicting breast cancer recurrence, and compared it to several other established guidelines: St. Gallen, Nottingham Prognostic Index (NPI), and Adjuvant! Online. The St. Gallen International Expert Consensus, in 2009, published several factors that contribute to a low-risk of recurrence, thus informing the use of adjuvant therapies post-surgery [59]. Researchers at the Nottingham City Hospital, in 1982, conducted retrospective multivariate analysis of breast cancer patients to build a prognostic model for survival, resulting in the NPI [57]. Kim et al. used this score to group patients into risk groups for recurrence. Cirkovic et al. also used the NPI index as an input to their breast cancer relapse prediction model [36]. Adjuvant! Online is a web-based tool for determining survival and recurrence rates based on several factors⁴ [107]. Kim et al. found their SVM model to be superior to the three established models, indicating that there is more research to be done to build clinically effective recurrence predictors.

5.2.3 Cancer Risk Models

We found several studies that trained statistical and machine learning models to predict cancer risk [126, 131]. All articles in this review build predictive models to determine if a patient will develop a cancer, or recur, in the future. The techniques used, however, differ between studies. Generally, a study used either classical statistical methods, such as regression and survival analysis, or machine learning methods, such as ANN, SVM, or tree models. A few studies used hybrid approaches or compared statistical and machine learning methods. Studies produced by the same institution tended to use the same methods. For instance, four studies from Memorial Sloan Kettering all used survival analysis techniques, and four studies from the National Cancer Center in Korea also used survival analysis techniques.

⁴<https://www.adjuvantonline.com/>

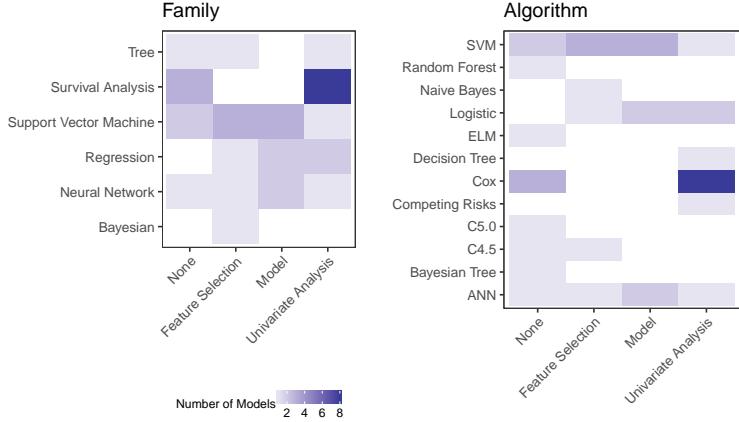


Figure 5.2: Feature selection and model algorithm methods. Studies with more than one method are counted multiple times. Feature Selection indicates use of a feature ranker or feature subset-selector. Left: Model Algorithms are grouped by their algorithm family. Right: Each model algorithm is outlined.

The goal of our analysis in this section is to provide a snapshot of the current techniques used in the literature and discuss gaps in research, but not to extensively describe the theory and implementation of the models. Section 2.1 can be referenced for algorithm theory.

Figure 5.2 illustrates the different statistical, machine learning, and feature selection methods used in the articles reviewed. The most widely used model combination is Cox regression, and most of those models utilized univariate analysis to select important features. SVM models tended to use feature rankers, subset selectors, or model-based feature selection.

Statistical and Machine Learning Models

Modeling of disease risk or recurrence is easily framed as a survival analysis problem, and many studies utilized survival analysis techniques to construct their predictive models. Cox proportional hazards [39] is typically the model of choice, as it allows for time censoring and multivariate analysis. To handle a large number of patient deaths

not due to the recurrence of sarcoma, Cahlon et al. used a competing risk survival analysis model [79], treating non-recurrent death as a competing risk [29]. This study is the only one profiled performing survival analysis with a model different than the Cox proportional hazards model.

El-Serag et al. used logistic regression models to predict the development of Hepatocellular Carcinoma (HCC), a form of liver cancer, within 6 months of an α -fetoprotein (AFP) test [49]. Among other models, Cirkovic et al. built a logistic regression model to predict recurrence after surgery for breast cancer [36]. Bayati et al. compared a traditional LR model to their own improved LR models based on multi-task learning, as their model attempts to predict risk of multiple different diseases (of which cancer is one) [24].

Tseng et al. found that their C5.0 decision tree model performed best when selecting two features to model the risk of recurrent cervical cancer [159]. Radespiel-Tröger et al. constructed decision trees to model the recurrence of colon cancer within five years of curative resection [121]. Cirkovic et al. and Ahmad et al. both utilized a C4.5 model (among others) to predict recurrent breast cancer [9, 36]. Singal et al. utilized a random forest to predict the development of HCC in patients with cirrhosis [147].

Kim et al. used an SVM model to predict recurrence for breast cancer patients [78]. Tseng et al. [159], Cirkovic et al. [36], Liang et al. [89], and Ahmad et al. [9] also used SVMs to predict cancer recurrence.

Jerez-Aragonés et al. constructed neural networks to predict the recurrence of breast cancer after surgery [68]. They constructed multiple models with different network topologies based on different time intervals, with the theory that recurrence risk is dependent on the amount of time after surgery, and not all features will have the same weight at different follow-up times [68]. Tseng et al. used a modification of an ANN, called extreme learning machine (ELM), that randomly assigns the input

weights while modeling the output weights of the network [159]. This makes the ELM model much faster to train than a typical ANN. Razavi et al. [123], Kim et al. [78], Cirkovic et al. [36], and Ahmad et al. [9] also used an ANN to model disease risk.

The papers employing machine learning models tended to use decision tree, neural network, or SVM models. Decision tree models, similar to regression models, are easy to interpret, but they can lack in predictive performance. SVMs and ANNs are difficult to interpret, but can achieve good classification results. Other models, such as naïve Bayes and random forest, were only used once in the papers studied. While algorithm choice can improve model performance, there can be a bottleneck related to the quality of the input data and how it is structured. The following sub-sections explore these ideas.

Feature Reduction

In most papers profiled, the authors have access to a dataset with a certain number of attributes, and these attributes are examined in the context of the research problem. In nearly every case, a domain expert, such as a physician or oncology researcher, informed the analysts about features he or she believes will be important to the model. The studies then only focus on these features, and perform a univariate analysis to find which covariates have a statistically significant correlation with the output variable. Then, only the significant features are used for training a model. Methods include the Pearson correlation coefficient, mutual information, or distance correlation. This generally results in less than 10 features input to the model, which is desirable to allow interpretation of regression models.

Cirkovic et al. combined three different feature rankers from the Weka ML toolkit [62] (mRMR, ReliefF, and Information Gain), to select the top 20 most relevant features for use in their ML models. Razavi et al. applied canonical correlation analysis (CCA) to reduce their feature set in the context of breast cancer recurrence

prediction. CCA is a subset selection technique that finds the subset of features that most correlate with an output set of features. In CCA, the output must be a set of features, rather than a single variable, so the authors broke down the recurrence variable into different types of recurrence (loco-regional recurrence or distant metastasis) for the feature selection step. Once the most informative features were selected, they included those features in a neural network to predict the binary outcome of recurrence. Liang et al. utilized two feature subset selection techniques, namely genetic algorithm (GA) [178] and simulated annealing (SA) [7] to reduce the feature space provided to their SVM model.

Several predictive models, such as decision trees, effectively perform feature selection as part of the model building process. The p-values from a statistical model can also be used as a form of feature selection, by only selecting those features that have significant p-values (often <0.05). Jerez-Aragonés et al. used a decision tree model to first select important features, then built ANNs to predict recurrence of breast cancer [68]. Tseng et al. [159], Radespiel-Tröger et al. [121], Cheng et al. [35], and Singal et al. [147] built models with trees or forests, limiting the features used to those in the resultant trees. Li et al. built a logistic regression model using features that were found to be statistically significant from a Cox survival model [88]. In addition to feature subset selection techniques alone, Liang et al. combined both the GA and SA algorithms with random forest to create a hybrid model and subset-based feature selection approach [89].

While manual feature reduction (i.e human experts) is useful for making models more interpretable, it can negatively impact the performance of the model. Nearly all studies reviewed used features that were deemed useful by a domain expert. A machine learning model, however, can often pick up on hidden patterns in data that humans cannot. Therefore, it can be advantageous to at least try building a model with all available features, or use a feature selection algorithm to automatically reduce

the feature set. Additionally, many studies only included covariates that were found to be statistically significant through univariate analysis. This can also decrease model performance, as a variable can still provide value to a model even if the p-value is not significant. Liang et al. found that univariate analysis resulted in only one significant variable, but their feature selection techniques selected four features to be included in their predictive models [89]. The models that utilized feature selection had better results than those that used the single significant variable.

Missing Data

Missing data is an important problem in all modeling efforts, especially in the health-care domain. If certain patient data is missing, such as tumor information or treatment history, the results can be significantly skewed. In addition to dropping patients, some studies will merely ignore variables that are available because there is not enough information filled in. El-Serag et al. chose to ignore lab results because they were too sparsely recorded in the input data [49]. It is important that all clinical variables are present, as to not bias the model, but it is also important to have a large sample size to make the model more generalizable for future instances. While most studies drop patients with missing clinical variables, there are several techniques that can help keep as many patients as possible in the model. The benefit of using a Cox proportional hazards model, as opposed to simpler survival models, is that Cox models allow for censoring of patients that drop out of the study without experiencing the event in question. This may be due to death not related to the cancer, or simply not following up at the clinic.

Prospective databases and clinical registries can help produce the most integrous data, as they can make certain fields mandatory for practitioners to populate as they see patients. There is a trade-off however, if too many fields are required, the participating investigators may simply not submit data as it takes away from their

time seeing patients. Additionally, there is an overhead of regulation and management in dealing with prospective studies, as compared to retrospective studies from EHR systems that are used regularly in practice [82]. Cahlon et al. [29], Tseng et al. [159], Singal et al. [147], and Bochner et al. [67] all used prospective databases or registries and do not mention the problem of missing data. This does not mean they did not encounter missing data, however, as they could have been filtered from their cohort counts beforehand. Weiser et al. were able to fill in some missing tumor information by having pathologists review the original slides [169].

Algorithmic techniques can be used to fill in missing values, such as mean imputation, or the expectation-maximization (EM) method. In mean imputation, a certain variable with missing data is filled in by taking the mean of the other instance's values [47]. Bayati et al. utilize mean imputation to substitute values for missing lab tests [24]. Naturally, this technique can only be used for continuous variables, and may not be desirable as it may bias those instances that do not have the value recorded by reducing the variance between values. EM is another method to impute missing values, and involves iteratively maximizing the log-likelihood of certain parameter values [102]. Radespiel-Tröger et al. dropped patients with more than one missing variable, but imputed values for one variable with EM as to not drop too many patients [121]. Ahmad et al. dropped patients with certain missing values, but used EM to impute other values [9].

5.2.4 Melanoma Risk Models

Several studies have built predictive models for melanoma risk. Most are case-control studies that compare differences between patients with and without melanoma to identify risk factors for the cancer [163], which is why they were excluded from the formal review in Section 5.2.3. Bakos et al. studied 117 patients from a hospital and outpatient clinic in Brazil [18]. Their model identified five risk factors for melanoma,

with hair color having the most predictive value. Fears et al. administered detailed questionnaires and skin exams to 1,663 patients in a case-control study of patients from two dermatology clinics in the U.S. [52]. The most informative risk factors were severe solar damage for men and number of moles for women.

Usher-Smith et al. conducted a review of the literature and found 25 studies that built risk prediction models for melanoma [163]. They found that all models built from these experiments fit along an ROC curve with an area of 0.755, meaning they all had similar discriminatory power with the sensitivity and specificity influenced by threshold selection. The various models assessed 144 different risk factors, and the largest study included 1,663 patients from two clinics in Philadelphia and San Francisco [52]. Most studies determined these risk factors through patient questionnaires or physical examinations. Additionally, only two studies used validation cohorts to determine the applicability of the models to different populations [56, 173].

5.2.5 Section Summary

A literature search retrieved several reports of predictive models for cancer risk. We identified several shortcomings:

- *Availability of structured clinical data:* Structured data points regarding patient history and encounters are limited. Many data-capture systems record free-text notes that are difficult to standardize across several patient charts. Data sharing among healthcare providers is lacking, limiting holistic views of patient history.
- *Old data:* Most studies were published five or more years after the end of the study period. This results in stale models that might not reflect the current state of diagnosis and treatment.
- *Advanced modeling methods:* Researchers often only use one or two familiar algorithms, possibly because of a lack of experience with various tools or limita-

tions in computing power. Some studies applied feature selection methods, but we were not able to find any studies that addressed the issue of class imbalance in their clinical datasets.

5.3 CLINICAL RISK MODEL

Here, we present a cloud-based approach to learning from big data and demonstrate its effectiveness on melanoma risk prediction from EHR system data [130, 133]. We evaluate methods for practical cost savings while maintaining model accuracy by using various types of computing infrastructures and data sampling techniques.

Among 4,061,172 patients who did not have melanoma through the 2016 calendar year, 10,129 were diagnosed with melanoma within one year. A gradient-boosted classifier achieved the best predictive performance with cross-validation ($AUC = 0.799$, sensitivity = 0.753, specificity = 0.688). Compared to a model built on the original data, a dataset two orders of magnitude smaller could achieve statistically similar or better performance with less than 1% of the training time and cost.

5.3.1 Materials and Methods

The models in this section were built to predict melanoma diagnosis within 12 months of a given patient encounter. We included patients in the experiments if they had no evidence of melanoma (defined as ICD9 V10.82, ICD9 172.*¹, ICD10 Z85.820, ICD10 C43.*¹, ICD10 D03.*¹, melanoma SNOMED, biopsy result, or cancer log entry) through 2016. We then tracked the patients through 2017 to determine if they received a diagnosis of melanoma. This served as the binary class label for the prediction problem: “melanoma” or “no melanoma.” The visit from which predictions were made (the “index visit”) was selected based on the following criteria for positive and negative cases. These constraints were inspired by Avati et al’s approach for a hospital mortality prediction problem [16].

- *Positive cases*: visit at least 6 months, and at most 12 months, before the earliest melanoma diagnosis in 2017. The earliest visit meeting this criterion was selected as the index visit.
- *Negative cases*: visit at least 12 months, and at most 24 months, before any visit in 2017. The latest visit meeting this criterion was selected as the index visit.

The goal of the constraints is to provide a consistent window of prediction time for positive and negative cases. The 6-month lower-bound for positive cases was selected to ensure the index visit was not a presentation for a melanoma biopsy or excision, and is a large enough time window to enact a change in screening patterns for early cancer detection. Furthermore, the lower bound of follow-up time for the negative cases must be greater than the upper bound of follow-up time for the positive cases. This is to ensure that the negative cases truly did not develop melanoma within a year. Otherwise, they may have developed it at some point after the observation time. We selected the 24-month upper-bound for negative cases to ensure that patients were consistently following up with their dermatologist. Patients that did not have an index visit matching these criteria were excluded from the study. We aggregated data from visits and prescriptions through each patient’s index visit to use as independent variables.

For each patient, we collected three types of data from the EHR system: Patient Data, Visit Data, and Historical Visit Data. Patient Data refers to static patient data that is not collected longitudinally, such as age, sex, race, melanoma family history, geographic location (i.e., U.S. state), family history conditions, and drug allergies. Visit Data represents the data recorded in the patient encounter of the index visit: chief complaints, review of systems, vitals, skin exams, diagnoses, procedures, body locations evaluated, prescriptions, biopsy results, and medical codes generated from the visit. Historical Visit Data contains the same elements as the Visit Data category,

but the features are aggregated across all visits prior to the index visit. As most data elements are categorical in nature, we aggregated each by counting the number of occurrences of each feature value across all visits. For the few numeric variables, we calculated summary statistics of each feature across the visits (minimum, maximum, mean, median, standard deviation). Table 5.1 describes each data element with a description, number of categories for each feature, and percentages of missing data. The missingness of each variable in a random sample of patients is given in Figure 5.3. Most Patient Data elements are required, and a visit generally contains complete data about the exam, diagnosis, procedure, and related ICD and CPT codes. There is more missing data for the Historical Visit Data, as not all patients had visits recorded before their index visit.

Table 5.1: Data elements

Group	Name	Description	% Missing	# Categories	Matrix Value
Patient Data	static_year_of_birth	Year the patient was born	<0.1%	-	(integer)
	static_sex	Birth sex (male, female, other)	0.00%	3	One-hot encoded categorical
	static_ethnic_group	Ethnic group (Hispanic or Latino, not Hispanic or Latino, other)	0.00%	3	One-hot encoded categorical
	race	Race (African-American, Asian, White, other)	21.20%	4	One-hot encoded categorical
	static_state_home	U.S. state of home address (including D.C.)	0.50%	51	One-hot encoded categorical
	static_melanoma_fh	If patient has a family history of melanoma	3.20%	- (0/1)	
	static_allergy	Drug allergen descriptions	67.00%	7,406	(integer) Number of times allergy was recorded across all visits
Visit Data	static_family_history_snomed	Family history SNOMED codes	70.60%	418	(integer) Number of times family history condition was recorded across all visits
	cpt	Standard codes describing medical procedures for billing purposes	7.80%	577	(integer) Number of units billed for each CPT code
	icd9	Diagnostic codes used for disease classification and billing (9th edition)	7.90%	896	(integer) Number of times each ICD9 code was referenced in the bill
	icd10	Diagnostic codes used for disease classification and billing (10th edition)	7.90%	2,819	(integer) Number of times each ICD10 code was referenced in the bill
	snomed	Standardized medical terminology covering terms beyond only procedures or diagnoses	38.90%	180	(0/1) If SNOMED code was associated with the visit
	loinc	Identifiers for laboratory orders	98.20%	1,953	(integer) Number of times each LOINC code was ordered in the visit
	cash_charge	Direct charges to the patient for non-medical procedures (categories represent a diagnosis/procedure combination that was charged for)	96.00%	5,477	(float) Dollar amount of cash charges in the bill
Clinical Data	vital	Height, weight, temperature, blood pressure (systolic/diastolic), pulse, respiration	90.80%	7	(float) Numeric value of each measurement
	ros	Series of questions to identify symptoms the patient is presenting with (ex. problems with healing, rash, hay fever, sore throat)	79.40%	131	(0/1) If ROS question response was yes
	chief_complaint	Reason why the patient is visiting the dermatologist (ex. skin lesion, skin lesion follow up, rash, acne)	29.10%	405	(0/1) If each chief complaint was documented in the visit
	follow_up_diagnosis	Previous diagnosis the patient is following up on	64.90%	1,700	(0/1) If each diagnosis was documented as the follow-up diagnosis for the visit
	exam	Body elements that the physician examined (ex. scalp, head, chest, neck, back)	9.10%	95	(0/1) If each body element was examined
	diagnosis	Findings noted in the exam to associate procedures with (ex. benign nevi, psoriasis, acne, melanoma, history of melanoma)	0.20%	2,048	(integer) Number of times each diagnosis was documented in the visit
	procedure	Procedures and plans performed during the visit (ex. liquid nitrogen, counseling, reassurance, biopsy)	0.50%	2,117	(integer) Number of times each procedure was documented in the visit
Administrative Data	bl_zone	Body locations associated with a finding and/or procedure (ex. head, face, trunk, scalp, leg)	6.40%	64	(integer) Number of times each body location was documented in the visit
	biopsy_result	Result of a biopsy/excision performed in the visit (ex. dysplastic nevus, basal cell carcinoma, melanoma)	87.40%	708	(integer) Number of times each result was received

Group	Name	Description	% Missing	# Categories	Matrix Value
Historical Visit Data	medication	Medication name of a prescription written during the visit	72.00%	1,107	(integer) Number of times each medication was prescribed during the visit
	ndc	NDC code of a prescription written during the visit	72.00%	5,086	(integer) Number of times each NDC was prescribed during the visit
	hist_num_visits	Number of visits documented before the current visit	0.00%	-	(integer)
	hist_earliest_visit_diff_days	Number of days between earliest historical documented visit and current visit	0.00%	-	(integer)
	hist_latest_visit_diff_days	Number of days between latest historical visit documented and current visit	0.00%	-	(integer)
	hist_visit_range_days	Number of days between earliest/latest historical visit	0.00%	-	(integer)
	hist_cpt	Standard codes describing medical procedures for billing purposes	23.20%	905	(integer) Number of times each CPT code was referenced across all historical visits
	hist_icd9	Diagnostic codes used for disease classification and billing (9th edition)	23.60%	1,445	(integer) Number of times each ICD9 code was referenced across all historical visits
	hist_icd10	Diagnostic codes used for disease classification and billing (10th edition)	26.80%	3,212	(integer) Number of times each ICD10 code was referenced across all historical visits
	hist_snomed	Standardized medical terminology covering terms beyond only procedures or diagnoses	44.10%	244	(integer) Number of times each SNOMED code was referenced across all historical visits
	hist_loinc	Identifiers for laboratory orders	94.70%	3,899	(integer) Number of times each LOINC code was ordered across all historical visits
	hist_cash_charge	Direct charges to the patient for non-medical procedures (categories represent a diagnosis/procedure combination that was charged for)	93.30%	6,602	(float) Dollar amount of cash charges in the bill across all historical visits
	hist_vital	Height, weight, temperature, blood pressure (systolic/diastolic), pulse, respiration	87.30%	49	(float) Min/max/mean/median/std of the numeric values of each measurement across all historical visits
	hist_ros	Series of questions to identify symptoms the patient is presenting with	72.60%	129	(integer) Number of times each ROS question response was yes across all historical visits
	hist_chief_complaint	Reason why the patient is visiting the dermatologist	22.30%	425	(integer) Number of times each chief complaint was documented across all historical visits
	hist_follow_up_diagnosis	Previous diagnosis the patient is following up on	56.10%	1,826	(integer) Number of times each diagnosis was documented as the follow-up diagnosis across all historical visits
	hist_exam	Body elements that the physician examined	22.90%	104	(integer) Number of times each body element was examined across all historical visits
	hist_diagnosis	Findings noted in the exam to associate procedures with	21.60%	2,143	(integer) Number of times each diagnosis was documented across all historical visits
	hist_procedure	Procedures and plans performed during the visit	21.60%	2,206	(integer) Number of times each procedure was documented across all historical visits
	hist_bl_zone	Body locations associated with a finding and/or procedure	23.30%	75	(integer) Number of times each body location was documented across all historical visits
	hist_biopsy_result	Result of a biopsy/excision performed in the visit	68.60%	797	(integer) Number of times each result was received across all historical visits
	hist_medication	Medication name of a prescription written during the visit	52.80%	1,833	(integer) Number of times each medication was prescribed across all history
	hist_ndc	NDC code of a prescription written during the visit	54.30%	9,224	(integer) Number of times each NDC was prescribed across all history

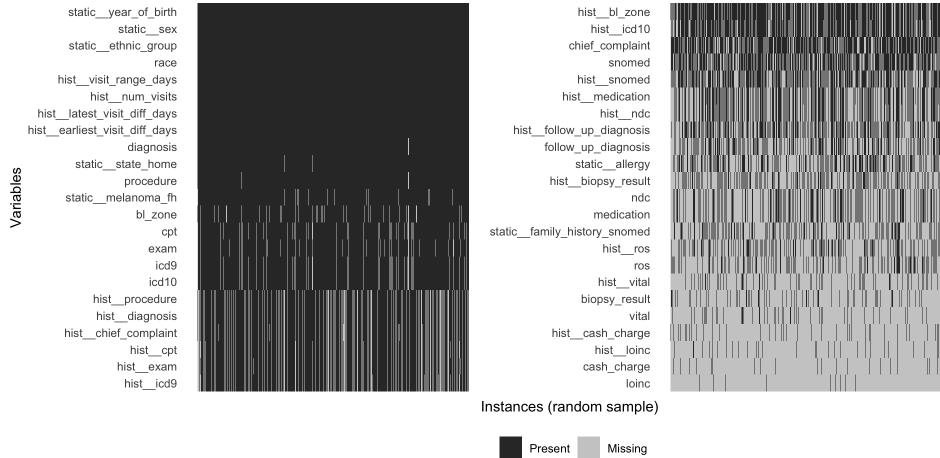


Figure 5.3: Variable missingness for a random sample of patient records ordered by most present variables.

The data extraction and matrix creation process is outlined in Figure 5.4. Given that the Patient Data variables are not longitudinal, these data were extracted separately from longitudinal data such as patient encounters (or visits) and prescriptions. The longitudinal data were aggregated for each patient by using count vectors of the data elements. For example, if a patient had four visits with a CPT code of 99201, the entry for “CPT 99201” in their vector would be “4.” Because there are more than 100,000 discrete data points, most entries in the aggregated patient data were zero, resulting in a collection of sparse vectors. The count vectors also account for missing data: if a patient did not have a record of a particular feature value, the count resulted in zero. We performed mean imputation for the small number of patients that did not have a recorded year of birth. The “# Categories” column in Table 5.1 shows the number of new features that were added by creating count vectors of the source features, and the “Matrix Value” column corresponds to the actual number that is stored in the matrix for the patient/feature entry.

⁵Icons made by *smalllikeart* on www.flaticon.com.

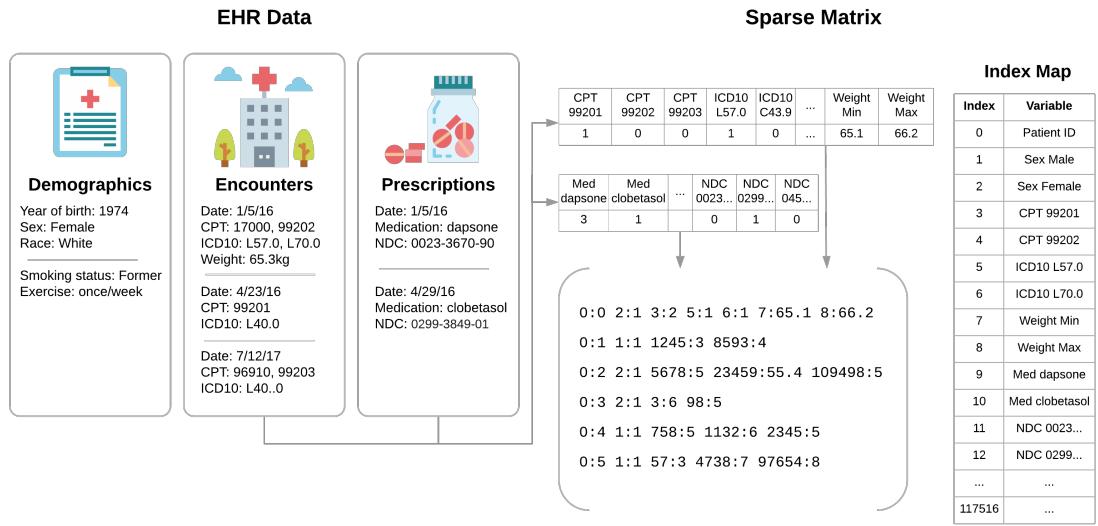


Figure 5.4: Sparse matrix creation process. First, longitudinal data are aggregated to create one vector for each patient. Then, vectors are collected into a sparse matrix using an index map to relate vector indexes to clinical variables. Note: Values in this figure are fictional and do not represent actual patients in the dataset.⁵

Model Training and Evaluation

We built logistic regression, random forest, and XGBoost models to evaluate performance across the original and sampled datasets. Model hyperparameters were selected based on a grid search; LR: L1 penalty (LASSO), regularization parameter $C = 0.5$, RF: 500 trees, no maximum depth, XGB: learning rate 0.1, 500 trees, maximum depth 3. Given that LR models can be affected by high dimensionality, we selected the top 1,000 features ranked according to the χ^2 statistic. We did not perform any feature selection for the RF or XGB models, because these models inherently select features. To evaluate model performance on smaller datasets and given the high class imbalance present in the training dataset, we used RUS to create multiple sampled datasets. We evaluated models on the original dataset along with sampled datasets according to the following target positive class ratios: 0.01, 0.1, 0.25, and 0.5.

We trained and evaluated all models using five-fold cross-validation repeated five times. An example model pipeline is provided in Figure 5.5. We performed data sampling and feature preprocessing separately in each training fold rather than for the whole dataset beforehand. This resulted in 25 runs that can be used for statistical tests. The dataset with the highest AUC for each classifier was selected for further examination with the following metrics: sensitivity (recall), specificity, precision, and AUPRC.

Because all experiments were conducted using Amazon EC2, we were able to directly calculate the cost of training each model configuration. Running time is not the best comparison across different classifiers, because the same model configuration can be run on more advanced hardware that would speed up running time. In addition, using an instance with more CPUs would only benefit models that support multithreading. Therefore, we estimated the cost of training a model as follows:

$$\text{Train cost} = \frac{\text{Train time}}{3600} * \text{Hourly price} \quad (5.1)$$

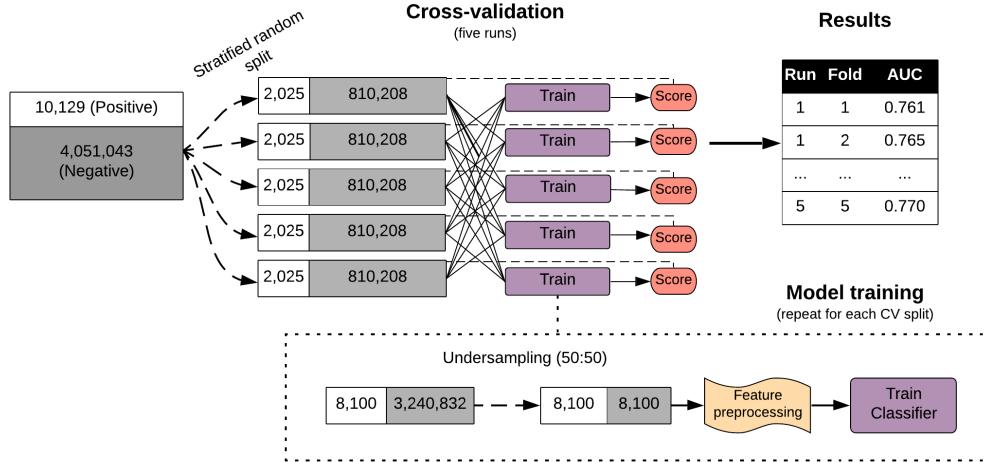


Figure 5.5: Example ML pipeline. Within each cross-validation run, the data are sampled and processed. Then, the results from each run are collected together.

Table 5.2: EC2 instance types

Instance Type	CPUs	Memory (GB)	Hourly Price (\$)	Models
r4.2xlarge	8	61	0.532	LR
c4.8xlarge	36	60	1.591	RF, XGB

Where train time is the average time (in seconds) spent training a single model (i.e., one training fold in cross validation) and hourly price is the Amazon EC2 hourly cost. The instance types used along with resource specifications are outlined in Table 5.2 (instance and cost data as of June 27, 2018). Amazon offers discounted rates by using Spot instances, but those prices are not constant over time, so we used the on-demand hourly rate for comparisons.

Table 5.3: Patient population

Variable	Value	No Melanoma (n, %)	Melanoma (n, %)	p
Total Patients		4,051,043 (99.75)	10,129 (0.25)	
Age (years)	(mean \pm SD)	57.00 \pm 19.88	68.31 \pm 12.59	<0.001
Sex	Female	2,403,446 (59.33)	4,025 (39.74)	<0.001
Race/Ethnicity	African-American	59,367 (1.47)	-	<0.001
	Asian	32,714 (0.81)	-	
	Hispanic	108,989 (2.69)	135 (1.33)	
	White	2,818,378 (69.57)	7,623 (75.26)	
	Other	1,031,595 (25.46)	2,362 (23.32)	
Geographic Region	Midwest	743,581 (18.36)	1,496 (14.77)	<0.001
	Northeast	778,382 (19.21)	1,621 (16)	
	South	1,700,964 (41.99)	4,686 (46.26)	
	West	802,258 (19.8)	2,269 (22.4)	
	Other	25,858 (0.64)	57 (0.56)	
Family History of Melanoma		484,882 (11.97)	1,387 (13.69)	<0.001
History (Number of Visits) ¹	Mean; Median (Q1-Q3)	3.95; 2 (1-5)	5.02; 3 (1-7)	<0.001
History (Days) ²	Mean; Median (Q1-Q3)	484.13; 371 (31-798)	557.72; 454 (40-932)	<0.001

¹ Number of visits recorded prior to index visit.

² Days between earliest visit recorded and index visit.

5.3.2 Results

Population

There were a total of 4,061,172 patients, 10,129 of whom were diagnosed with melanoma within one year (Table 5.3). Compared to the no melanoma class, the melanoma class had a lower proportion of females (59.33% vs. 39.74%), and higher proportions of white race (69.57% vs. 75.26%) and family history of melanoma (11.97% vs. 13.69%).

Table 5.4: Average results for each dataset and classifier

Dataset	Total (N)	Negative (N)	Positive (%)	Average AUC			Average Training Cost (\$)		
				LR	RF	XGB	LR	RF	XGB
4m	4,061,172	4,051,043	0.25	0.7617	0.6949	0.7988	1.1466	2.6703	0.9648
1m	1,012,900	1,002,771	1	0.7651	0.7415	0.7991	0.1373	0.5312	0.1347
100k	101,290	91,161	10	0.771	0.7682	0.7989	0.0152	0.0244	0.0168
40k	40,516	30,387	25	0.7713	0.7734	0.7971	0.006	0.0111	0.0099
20k	20,258	10,129	50	0.7674	0.7736	0.7921	0.0038	0.0072	0.0068

Table 5.5: Additional metrics

Classifier	Best Size	AUC	AUPRC	Sensitivity (Recall)	Specificity	Precision
LR	40k	0.7713	0.01	0.6961	0.7045	0.0059
RF	20k	0.7736	0.0095	0.7032	0.6952	0.0057
XGB	1m	0.7991	0.0136	0.7529	0.6877	0.006

Performance

Table 5.4 outlines the sizes of the original dataset and each sampled dataset as well as the average performance for the three classifiers; these results and error bars for the minimum/maximum values across all runs are plotted in Figure 5.6. The greatest AUC (0.7991) was achieved by the XGB model on the 1m dataset, but this was not significantly better than that on the original 4m dataset (0.7988, $p = 0.846$). Training the XGB model with the 40k dataset achieved statistically comparable results to the full dataset (0.7971 , $p = 0.1797$). The AUCs for the LR and XGB models were relatively unaffected by the reduction in dataset size, while the performance of the RF model actually improved when sampling was introduced. The best RF model had an AUC of 0.7736 on the 20k dataset compared to the baseline of 0.6949 ($p < 0.001$). Additional performance metrics for each best performing model are provided in Table 5.5. XGB had the highest AUPRC (0.0136), sensitivity (0.7529), and precision (0.0060), while LR had a slightly higher specificity (0.7045).

The RF model was the most expensive of the three classifiers, costing an average

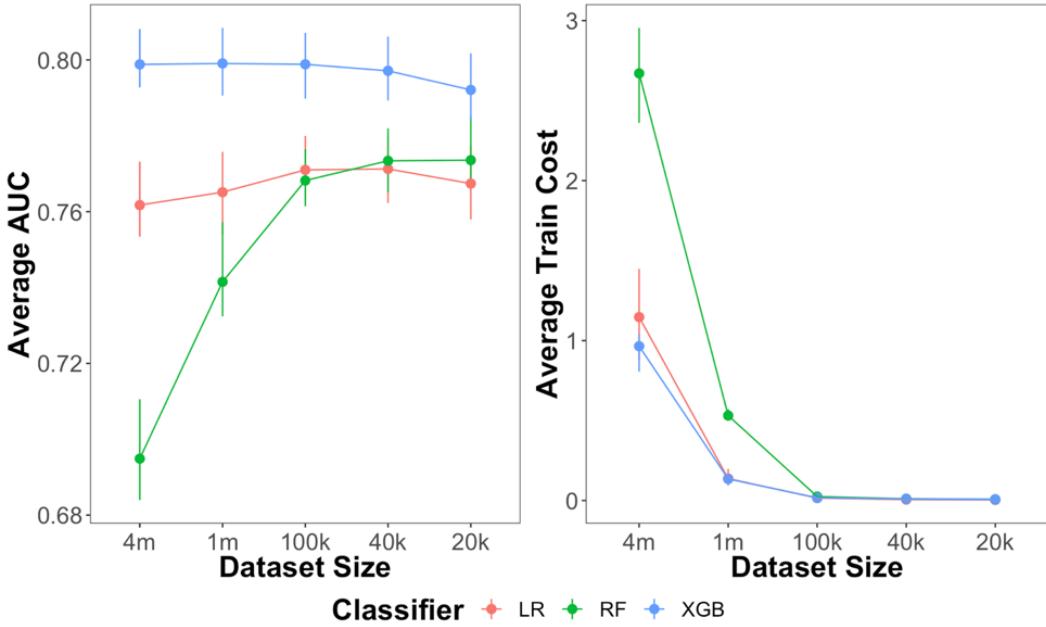


Figure 5.6: Average results for each classifier and dataset

of \$2.6703 for a single model fit using the full 4m dataset. For a run of 5-fold cross-validation with 5 repeats, this would cost over \$66. Meanwhile, the XGB model cost \$0.9648 per model fit, and the LR model cost \$1.1466 per model fit on the same dataset. When run on smaller datasets, all models were significantly cheaper than the respective baselines, with datasets smaller than 1m costing <\$0.1000 per model fit. The LR model had the cheapest cost with \$0.0038 for the 20k dataset, while the RF and XGB models were \$0.0072 and \$0.0068 for the smallest dataset, respectively.

5.3.3 Discussion

The results of this section offer several perspectives on the intersection of risk models, EHR systems, and big data. Datasets for specific biomedical and health applications can be small because of limited data sharing between institutions, strict inclusion criteria, and a lack of structured clinical data. Risk models are often built with data collected from individual healthcare or academic institutions. While large centers can attract patients from different geographical areas, the treatment and data col-

lection processes are still localized to the center and might not pair well with data from elsewhere. Furthermore, institutions generally do not share data, resulting in many different models being built from fragmented datasets. The dataset used in the present study is unique in that it provides over 100,000 structured data points from over 4 million dermatology patients throughout the U.S. Though the dataset is unique, the model can still have wide applicability due to the number of patients treated by physicians using the EHR system. Distributed computing is often required to deal with big data. While packages such as MLLib [100] provide distributed ways to train machine learning algorithms, there are many more libraries and algorithms available on non-distributed infrastructures (i.e., single machines). We found that the best approach for our scenario was to use Spark to perform data collection and transformation and then save the data into a sparse format to use with single machines. By doing so, we were able to use multiple machines to train and evaluate multiple machine learning models in parallel rather than using a cluster of machines to train one model at a time. For other datasets that are large, high-dimensional, and dense, cluster computing may still be required for model training.

Although the machine learning community often assumes that more data means better models [164], we hypothesize that this might not be true in cases with truly massive amounts of data. Here, we found that using datasets with tens of thousands of instances could achieve statistically similar (i.e., XGB models) or better (i.e., RF models) performance than when using the full dataset ($n = 4,061,172$). This is likely because of a high level of homogeneity among instances in the negative class, which means that less instances need to be used to produce a generalizable model. We will explore this hypothesis in Chapter 6. The RF performance increase using less data may be explained by tree overfitting in the forest. The datasets with less data had shallower trees, resulting in more generalizability when evaluating test data. Using fewer instances means that less sophisticated computing infrastructure can be used,

which allows researchers to continue to use known methods and tools rather than worrying about how to handle big data in their machine learning workloads.

Limitations of this dataset include selection bias and data quality. Because all patients in this study have already visited a dermatologist, we might be missing key patients who do not regularly visit a dermatologist. Increased interoperability and data sharing between institutions can help reduce this limitation, which is a key goal of the 21st Century Cures Act [6]. There is, however, clinical utility of this model on a dermatology population. We evaluated the chief complaint of the earliest visit for each patient in the dataset and found that 1,090,042 (26.91%) patients in the no melanoma class had a chief complaint for other conditions, such as acne, verruca vulgaris, or a rash. Even in the melanoma class, 1,167 (11.52%) of patients did not initially present for a skin check. These patients are ideal candidates for our model, as they may be high-risk for melanoma and not know it. While EHR systems provide a structured data input solution, variable input might differ substantially among providers, limiting the depth of data available. Accordingly, the size of the current dataset helps to alleviate concerns regarding consistency and missing data. While many factors selected by these models indicate patients that may be already presenting for skin checks, the model can still provide value as the predictions are personalized for each patient.

5.3.4 Section Summary

We described a case study of learning from big data to produce an effective melanoma risk prediction model based on data collected from a large representative dermatology EHR system covering millions of patients across the U.S. Our study provides a reference framework for machine learning studies using large, high-dimensional, and imbalanced EHR data. We used a distributed processing infrastructure for collecting and formatting the data as well as a non-distributed infrastructure for machine learn-

ing. Then, we achieved statistically similar or better performance using a sampled dataset versus the original data, saving hundreds of dollars in cloud computing costs for model experimentation.

5.4 ADVANCED MACHINE LEARNING TECHNIQUES

In this section, we build upon the modeling problem presented in Section 5.3 by experimentally evaluating various advanced machine learning techniques to address the problems of dataset size, sparseness, and imbalance [129]. We explore the use of logistic regression, decision tree, and random forest classifiers with various feature selection and random undersampling techniques.

5.4.1 Methods

We utilize the same dataset and matrix preprocessing methods described in Section 5.3.1 with a slightly modified class labeling and inclusion process. To maximize the number of instances available for machine learning, we relaxed the criteria for the index visit and follow up times as described below.

Patients were included in this study if they had: (1) at least one dermatologist visit in 2016, and (2) had no history of melanoma through 2016. Available visits from January 2011 through December 2016 were used to predict whether or not a patient developed melanoma in the 2017 calendar year. Data from each patient’s last visit in 2016 were extracted to use as instances for the predictive model (index visit). The patients were then tracked through 2017 to determine if they were diagnosed with a new melanoma. Patients that did not develop melanoma in 2017 were labeled “no melanoma”, while those that did were labeled “new melanoma”. Therefore, we predict the development of melanoma within 1-24 months of the index visit.

The goal of this study is to determine what modeling techniques are most beneficial for learning from sparse, imbalanced clinical data. Therefore, we built a large number

of models with different configurations and machine learning techniques to assess the performance of each technique. We used LR, DT, and RF for the classifiers.

The dataset was randomly split into a train and test set, with 30% of the instances from each class in the test set. Models and parameters were selected using stratified 5-fold cross-validation on the train set based on AUC, with the same pipeline as described in Section 5.3.1 (Figure 5.5). We also calculated balanced accuracy, TPR, and TNR for each model configuration. We applied RUS and then feature selection within each fold of cross-validation. We tested various sampling ratios for RUS: 1:99, 10:90, 35:65, and 50:50. We selected the top K features as ranked by χ^2 to feed to the classifier. Tested values of K were: 10, 100, 1,000, and 10,000. We were limited in feature ranker choices due to the sparse format of the data; *scikit-learn* only supports χ^2 and mutual information rankers⁶, and we found the mutual information ranker was too computationally intensive for our dataset.

In addition to model selection, the cross-validation results were used to select the decision threshold for each model. The default threshold is 0.5, but it is advantageous to explore the distribution of predicted probabilities of each model to select a threshold that will result in better TPR and TNR. We did this experimentally by plotting the predicted probabilities of instances versus their actual class membership, and then explored the TPR and TNR of various thresholds. We chose a threshold where the TPR is maximized without a large drop in TNR. The goal of the model is to flag patients that are at high risk for developing melanoma; however, too many false positives would make it intractable to screen all patients. Therefore, a balance must be achieved between the two metrics.

High-level dataset statistics for the full dataset as well as the train set are provided in Table 5.6.

⁶http://scikit-learn.org/stable/modules/feature_selection.html

Table 5.6: Dataset statistics

	Full	Train
Number of instances	9,531,408	6,671,985
Number of features	117,516	117,516
Number of data elements	1.12×10^{12}	7.84×10^{11}
Sparsity of matrix*	7.94×10^{-4}	7.94×10^{-4}
Number of positive cases (new melanoma)	17,246	12,059
Class distribution	0.181%	0.181%

* Number of non-zero elements divided by the total number of elements

5.4.2 Results

Population

Table 5.7 describes the patient population that met the inclusion criteria. Out of 9,531,408 patients that did not have a history of melanoma in 2016, 17,246 (0.18%) of them developed melanoma in 2017. The average number of historical visits was 3.4 and 1.9 for patients with a new melanoma and no melanoma, respectively. For many patients (36.34% of new melanoma patients, and 53.6% of no melanoma patients), the visit in 2016 was their first dermatology visit recorded in the EHR, which means no data from historical visits was available. Figure 5.7 outlines the number of visits recorded and time range (from first visit to last visit) for patients with historical visits. Additionally, 4,970,348 (52.1%) patients did not have any follow up visits in 2017 (i.e. lost to follow up). This is considered a form of censorship for this experiment and the patients are included in the “no melanoma” group.

Performance

Figure 5.8 shows the AUC values from the cross-validation results of each model configuration on the train set. These results show that RUS and feature selection sizes have an impact on model performance across all three classifiers. To select the best performing models, we performed various ANOVA and HSD tests. The ANOVA

Table 5.7: Demographics and clinical characteristics

Variable	Value	No Melanoma n (%)	New Melanoma n (%)
Total Patients		9,514,162 (99.82%)	17,246 (0.18%)
Age (years)	(mean \pm SD)	51.50 \pm 21.76	67.40 \pm 12.87
Sex	Female	5,684,788 (59.75%)	7,138 (41.39%)
	Male	3,822,755 (40.18%)	10,099 (58.56%)
	Other	6,619 (0.07%)	9 (0.05%)
Race	African American	230,877 (2.43%)	8 (0.05%)
	Asian	130,123 (1.37%)	9 (0.05%)
	Hispanic	367,633 (3.86%)	246 (1.43%)
	White	5,874,443 (61.74%)	12,781 (74.11%)
	Other	2,911,086 (30.60%)	4,202 (24.37%)
Home Region	Midwest	1,731,205 (18.20%)	2,570 (14.90%)
	Northeast	1,814,746 (19.07%)	2,919 (16.93%)
	South	3,958,926 (41.61%)	7,857 (45.56%)
	West	1,918,661 (20.17%)	3,806 (22.07%)
	Other	90,624 (0.95%)	94 (0.55%)
Family History of Melanoma		997,013 (10.48%)	2,418 (14.02%)
New Patient		5,109,613 (53.60%)	6,267 (36.34%)

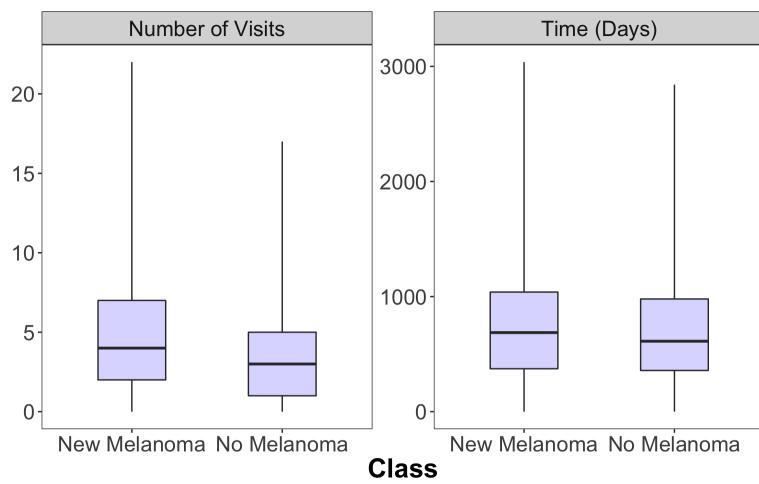


Figure 5.7: Distributions of the number of historical visits recorded and time range between the first and last visit, with respect to the class label. Note: this is only for patients with historical visits; those with only one visit are not represented.

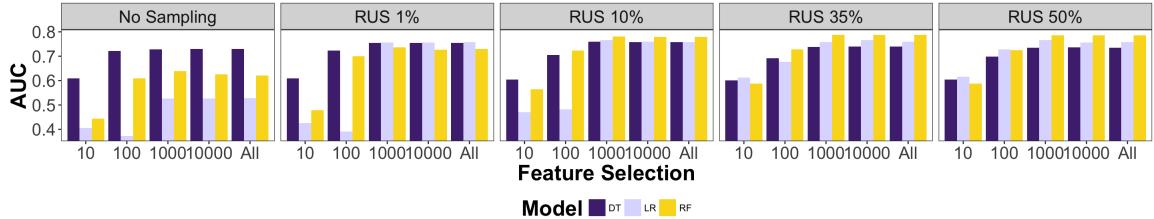


Figure 5.8: AUC of each model configuration

Table 5.8: ANOVA: All models

Source	DF	SS	MS	F value	Pr(>F)
Model	2	0.371	0.185	205.233	0
RUS	4	0.890	0.223	246.458	0
FS	4	2.034	0.508	563.135	0
Model:RUS	8	0.555	0.069	76.787	0
Model:FS	8	0.303	0.038	41.991	0
RUS:FS	16	0.142	0.009	9.856	0
Model:RUS:FS	32	0.237	0.007	8.205	0
Residuals	300	0.271	0.001		

tests shows that all factors and interactions are significant, most likely due to the sheer number of samples in the dataset. The ANOVA results for all models are in Table 5.8, and while we performed individual tests for each classifier, we only present the RF-specific results for brevity (Tables 5.9-5.10).

To select the best feature selection size and sampling ratio, we examined the results of HSD tests of the interaction between sampling ratio and feature selection size for each model. We first selected the feature selection size for each model, maximizing the number of features for the RF model, and minimizing the number of features for

Table 5.9: ANOVA: RF

Source	DF	SS	MS	F value	Pr(>F)
RUS	4	0.400	0.100	1348.701	0
FS	4	0.831	0.208	2804.633	0
RUS:FS	16	0.018	0.001	14.856	0
Residuals	100	0.007	0.000		

Table 5.10: HSD: RF

RUS:FS	Group	AUC	SD
RUS 35%:1000	A	0.788	0.004
RUS 35%:All	A	0.788	0.004
RUS 35%:10000	A	0.787	0.004
RUS 50%:10000	A	0.786	0.005
RUS 50%:All	A	0.786	0.004
RUS 50%:1000	A	0.786	0.004
RUS 10%:1000	A	0.782	0.004
RUS 10%:All	A	0.779	0.003
RUS 10%:10000	A	0.779	0.003
RUS 1%:1000	B	0.737	0.004
RUS 1%:All	B	0.730	0.005
RUS 35%:100	B	0.728	0.006
RUS 1%:10000	B	0.727	0.004
RUS 50%:100	B	0.724	0.004
RUS 10%:100	B	0.724	0.003
RUS 1%:100	C	0.701	0.004
No Sampling:1000	D	0.638	0.005
No Sampling:10000	DE	0.626	0.005
No Sampling:All	DE	0.621	0.004
No Sampling:100	E	0.609	0.005
RUS 35%:10	F	0.588	0.020
RUS 50%:10	F	0.587	0.018
RUS 10%:10	G	0.564	0.024
RUS 1%:10	H	0.479	0.013
No Sampling:10	I	0.443	0.004

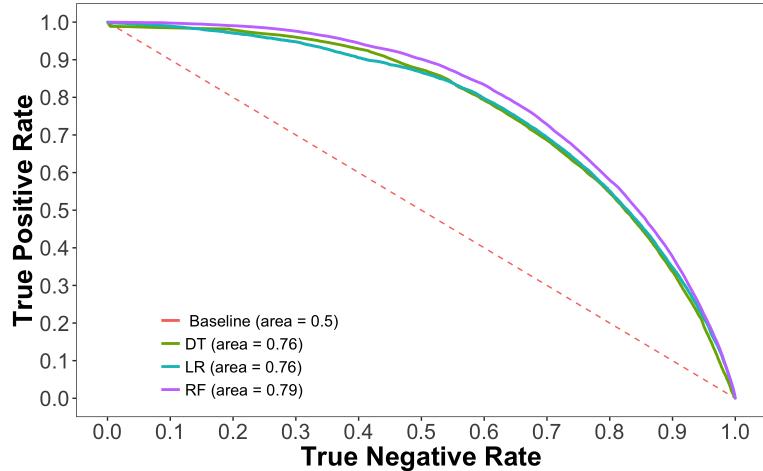


Figure 5.9: ROC curves for each classifier

the LR and DT models. The ensemble nature of the RF model benefits from large variability in instances and features, so it is advantageous to include as much data as possible. Conversely, the LR and DT models should have as few features as possible, so they do not overfit to features that are not generalizable to a larger population. Additionally, a small number of features is desired for global interpretability of these models. Then, we chose the sampling ratio that resulted in the highest AUC for each model. Figure 5.9 shows the ROC curve for these models.

Figure 5.10 plots the predicted probabilities of the instances versus their actual class membership for each of the selected models. Instances with a probability above the selected threshold (red dotted line) are classified as positive, while all probabilities below the line are classified as negative. Therefore, a good model will have a dense group of positive instances above the red line, and a dense group of negative instances below the red line. The RF model has the best spread of probabilities, indicating that the model has the most generalizability across various thresholds. If the default threshold of 0.5 was used for the DT and LR models, the models would have a very low TPR and be useless for clinical practice.

Table 5.11 shows the classification results for the selected models on the test set,

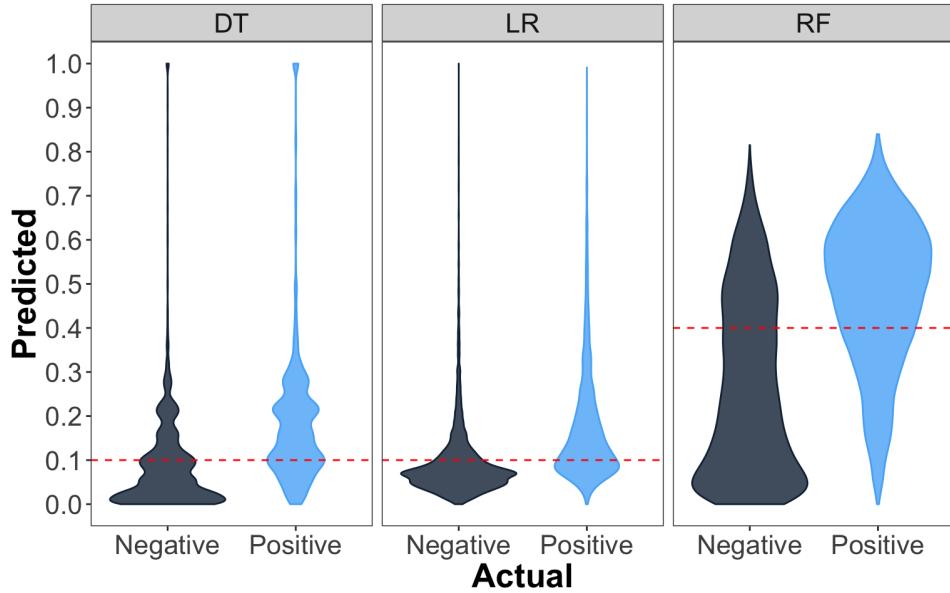


Figure 5.10: Density curves of the predicted probabilities versus actual classes for each selected model. The dotted lines represent the selected threshold for each model.

Table 5.11: Classification results

Model	TPR	TNR	Balanced Accuracy	AUC
RF	0.708	0.718	0.713	0.790
DT	0.869	0.493	0.681	0.749
LR	0.766	0.648	0.708	0.772

using the top 1,000 features for LR and DT, and RUS 35:65 for all classifiers. The RF model outperformed the other models across most metrics with an AUC of 0.790, TPR of 0.708, and TNR of 0.718. The DT and LR model have higher TPR but lower TNR than the RF model.

Figure 5.11 presents ROC curves for the RF model, broken down by patients that do and do not have historical visits. It is interesting to note that the AUC for new patients is higher than that for established patients, but the other metrics are slightly different as shown in Table 5.12. The model has a high TPR (0.799) and low TNR (0.570) for established patients but a high TPR (0.845) and low TNR (0.544) for new

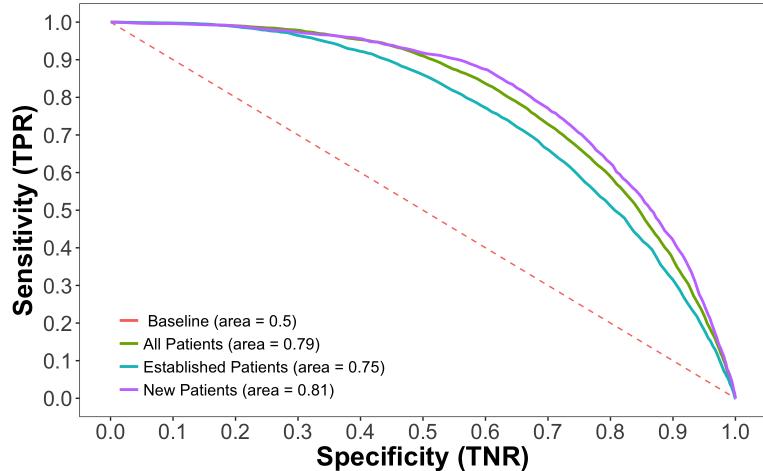


Figure 5.11: ROC curves for each selected model

Table 5.12: New vs. established patient results

	Sensitivity	Specificity	Balanced Accuracy	AUC
All	0.708	0.718	0.713	0.790
Established	0.799	0.570	0.684	0.753
New	0.544	0.845	0.695	0.807

patients. This shows that having data from multiple patient visits is beneficial for detecting patients with a high risk of melanoma.

5.4.3 Section Summary

In this section, we explored various machine learning techniques to build risk prediction models for melanoma from structured EHR data. We utilized sparse storage techniques and random undersampling to handle the large size and imbalance of the data. We found that the top 1,000 features with χ^2 significantly improved performance for LR and DT. Overall, RF achieved the best classification performance with a 35:65 class ratio and no prior feature selection.

5.5 INTERPRETABILITY

For a model to be effectively incorporated into clinical practice, the physician and patient should have insight into how the model is making predictions (see Section 2.1.4). Global and local interpretability considerations for the models created in Sections 5.3-5.4 are provided below. In addition to ML and usability considerations, we examine the features selected to evaluate the clinical accuracy of the models produced.

5.5.1 Global Interpretability

The most important features in the LR model from Section 5.3 are provided in Tables 5.13-5.14. We found that models trained with patients that had historical visits (“history”) versus those that did not (“no history”) achieved similar predictive performance ($AUC = 0.7597$ vs. 0.7586) but utilized different features for each population. Therefore, we explore selected features separately for these populations. As we are not able to present all 1,000 features due to space constraints, we display the top ten features with positive weights and top ten features with negative weights. The highest predictors for melanoma risk are the presence of other cancerous and precancerous lesions such as basal cell carcinoma or actinic keratoses, neoplasms of uncertain behavior, and history of a malignant lesion, or treatments for these conditions (Mohs surgery, excision). Negative predictors of melanoma risk are related to lower-risk populations (African American, Hispanic, Female), and dermatology diagnoses that may be an indicator of young age (acne). For patients with history, we note that the model selects several variables covering historical visits (hist_*). Age, race, and presence of other lesions are fairly known general risk factors for the cancer, so we can confirm that the EHR dataset is capturing relevant clinical information for melanoma risk prediction.

Using 1,000 features in a linear model is still too many; it is difficult for a person to grasp contributions from more than 5 or 10 features at a time. We present the weights

Table 5.13: LR model weights: No history

Highest weights		Lowest weights	
Feature	β	Feature	β
icd10_D48.5	3.968976	chief_complaint_Pimples (Acne)	-1.168288
icd9_173.31	1.666087	exam_chest	-0.97172
procedure_ShaveBiopsy	1.659514	icd9_V65.49	-0.930093
icd10_L57.0	1.407742	diagnosis_Acne	-0.870827
diagnosis_Basal Cell Carcinoma	1.374019	procedure_NCounselingAcne	-0.745976
procedure_mipsQuality	0.792211	static_ethnic_group_HISPANIC_OR_LATINO	-0.568793
biopsy_result_Basal Cell Carcinoma	0.642071	static_sex_FEMALE	-0.487668
procedure_PunchBiopsy	0.482086	static_race_african_american	-0.45097
icd10_Z08	0.479657	procedure_Prescription	-0.273696
cpt_11100	0.410025	diagnosis_Milia	-0.244045

Intercept: -0.972721

Table 5.14: LR model weights: History

Highest weights		Lowest weights	
Feature	β	Feature	β
procedure_Mohs	2.637419	procedure_SutureRemoval	-1.542473
icd10_D48.5	2.58132	follow_up_diagnosis_Acne	-1.464221
hist_icd10_L57.0	2.307007	icd9_V65.49	-1.247825
procedure_ShaveBiopsy	1.589884	static_race_african_american	-1.156283
diagnosis_Basal Cell Carcinoma	1.52941	diagnosis_Acne	-1.020633
hist_icd10_D48.5	1.517638	hist_visit_range_days	-0.925938
procedure_Defer	1.452915	icd9_706.1	-0.654565
procedure_ExcisionMalignant	1.443212	icd10_Z71.89	-0.54742
hist_icd10_Z87.2	1.37611	procedure_TreatmentRegimen	-0.497152
cpt_17004	1.255758	static_sex_FEMALE	-0.476317

Intercept: -1.647285

Table 5.15: LR model weights: 10 features

Feature	Weight
Intercept	-0.849657489
Days between earliest visit and current visit	0.000339452
Days between latest previous visit and current visit	-0.00028521
Days between earliest visit and latest previous visit	0.000624662
History - sum of weight	-0.007420045
History - min systolic blood pressure	0.000363017
History - sum of systolic blood pressure	0.004628308
History - average diastolic blood pressure	-0.000295815
History - count of diastolic blood pressure	0.001046788
History - median height	-0.0000952
History - std of weight	-0.0000423

for an LR model from Section 5.4 using just 10 features in Table 5.15. Note that this model has much lower performance than the other models ($AUC = 0.616$), and we do not recommend using this model for clinical decisions. Based on the features included in the model, it is most likely overfit to numeric features in the training folds. Global interpretability of decision trees suffer from the same problem as logistic regression, in that it is difficult to grasp the full nature of the model when there are a large number of features. Figure 5.12 shows a small subset of the tree for the selected DT model.

Since a random forest model contains an ensemble of decision trees, it is not possible to directly explain the model from a global perspective (such as with a LR or DT). Feature importances, however, can be ascertained from the fitted model. The feature importance in a random forest model is a measure of how much impact the particular feature has on a prediction made from the model. All importances are positive as they do not explain the directional impact of the feature, just its importance relative to the other features. The sum of all importances add up to 1. Figure 5.13 shows the cumulative importance as the number of features is increased to 5,000, and Table 5.16 shows the top 15 features selected by the model produced in

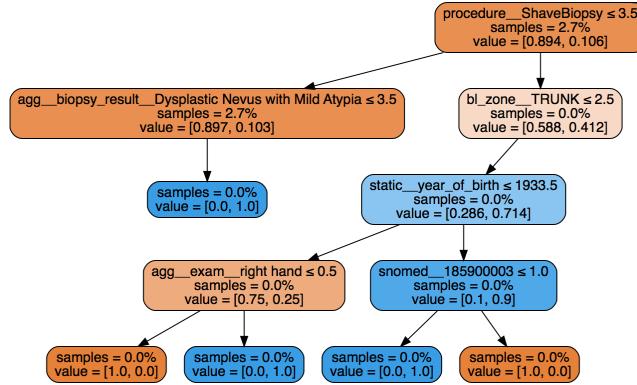


Figure 5.12: Sample nodes from the decision tree (some paths removed for display purposes). The left child node is traversed if the condition in the root node is met. The first number in the value array is the proportion of positive instances (new melanoma) in the training set.

Section 5.4.

5.5.2 Local Interpretability

Our goal is to build a risk model for personalized patient care; therefore, we examine more individualized feature importance using the RF and XGB models. Local interpretability is available from random forest models by exploring the path traversed

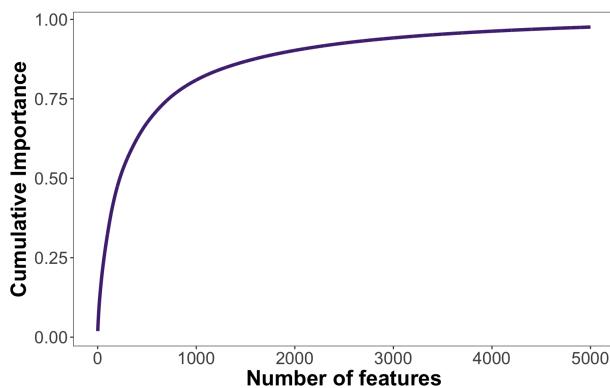


Figure 5.13: Cumulative feature importances for the RF model

Table 5.16: RF: Top 15 feature importances

Description	Importance
Patient year of birth	0.019057044
ICD9 702.0 (Actinic keratosis)	0.009366818
ICD10 L57.0 (Actinic keratosis)	0.008573076
Body location - trunk	0.006139061
Body location - face	0.005888611
Procedure - Liquid Nitrogen	0.005863717
CPT 17003 (Destruction premalignant lesion 2-14)	0.005844474
CPT 17000 (Destruction premalignant lesion 1st)	0.005828385
Historical ICD10 L57.0 (Actinic keratosis)	0.005595849
ICD10 D48.5 (Neoplasm of uncertain behavior of skin)	0.005068944
Procedure - Biopsy	0.005005304
Historical ICD9 (702.0) - Actinic keratosis	0.004913928
ICD9 238.2 (Neoplasm of uncertain behavior of skin)	0.004897940
Body location - arm	0.004733738
Sex - female	0.004539907

by each particular instance. This allows patients and providers to see why a specific prediction was made. The bias and contributions of each feature are added to make the prediction, as in a regression function. The bias for the RF model from Section 5.4 is 0.35, and example feature contributions for a positive and negative instance are provided in Tables 5.17 and 5.18. The “Prediction” column indicates the prediction given by only using the features up to the current row. Both examples displayed are represented by over 10 instances each in the test dataset. In both cases, over 2,000 features are required to determine the final prediction of the instance. This illustrates that the depth of the EHR data is crucial to making accurate predictions, and there are indeed a large number of factors that contribute to a patient’s risk of developing melanoma.

A summary plot of the top twenty features in the XGB model from Section 5.3 according to their mean SHAP values is provided in Figure 5.14. Each dot represents the impact of the particular feature for a given instance and is colored according

Table 5.17: Example prediction: Positive class

Rank	Feature		Value	Contribution	Prediction
1	Diagnosis - Squamous Cell Carcinoma		1	0.064	0.414
2	CPT 17262 - Destruction malignant lesion trunk/arm/leg 1.1-2.0cm	1	0.038	0.452	
3	ICD9 702.0 - Actinic keratosis		1	0.035	0.488
4	CPT 17000 - Destruction premalignant lesion 1st		1	0.034	0.522
5	Procedure - Liquid Nitrogen		1	0.033	0.555

Predicted value: 0.45. Number of features to achieve value: 2,562.

Table 5.18: Example prediction: Negative class

Rank	Feature		Value	Contribution	Prediction
1	Patient year of birth		1982	-0.030	0.320
2	Procedure - Treatment Regimen	1	-0.010	0.310	
3	ICD9 702.0 - Actinic keratosis	0	-0.010	0.301	
4	Chief complaint - Warts	1	-0.010	0.291	
5	CPT 99202 - Office outpatient new 20 minutes	1	-0.010	0.282	

Predicted value: 0.015. Number of features to achieve value: 3,211.

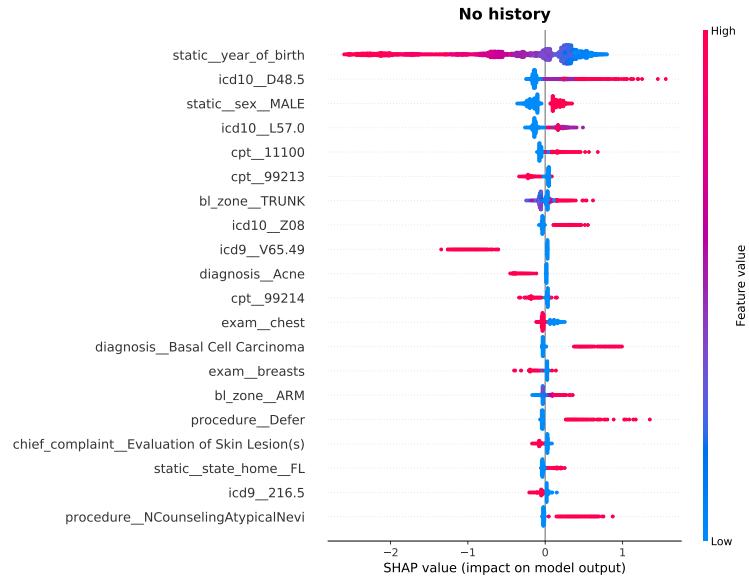
to what magnitude of value contributes to the model impact. For example, a high feature value of “static_sex_MALE” has a positive impact on model output, meaning male sex is a factor that increases melanoma risk for those instances. As with LR, we found that the features selected by a model on the “no history” ($AUC = 0.8173$) versus “history” ($AUC = 0.7934$) populations were different. The features selected in the XGB model have some substantial differences from the LR model. Particularly, year of birth (age) is the most important predictor, followed by actinic keratoses (L57.0), neoplasm of uncertain behavior (D48.5), and sex. Other risk factors not present in the LR model are melanoma family history, evaluations of various body locations (trunk, leg, chest, hands) and geographic location (home address in Florida). The plot provides an estimate of both individualized and global feature importance by plotting the SHAP values for a random sample of instances. Older age (lower year of birth) has the largest positive impact on risk for both populations, as well as different

magnitudes of negative risk decreasing as age decreases. The time between the index visit and the patient’s most recent visit (“hist_latest_visit_diff_days”) was the second most important factor for the history population, and seems to have an increasing impact on risk as the time increases. The top features for the history model are almost exclusively from historical visit data rather than the index visit. This shows that history is indeed important for building estimating risk, but other factors can also be used from the index visit if history is not available. These observations show that a simple global model does not necessarily provide the best estimate of melanoma risk.

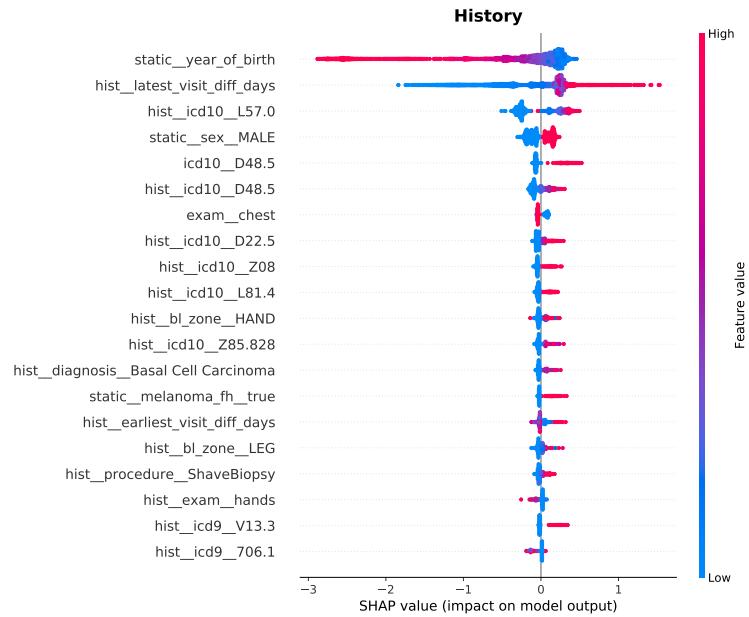
Age has an impact across the largest group of patients, but the other features appear to have effects only for localized groups, meaning that a large number of features must be included to produce the most effective model. To evaluate this hypothesis, we trained an XGB model on the 1m dataset with an increasing number of the most important features according to their SHAP values (Figure 5.15). We see that 300 features are required to achieve the best performing model, meaning that hundreds of different factors from the patient’s history can affect their risk of the disease. Deployment into the structured EHR system is ideal for this type of model, as a patient’s risk can be evaluated in real-time rather using an external risk evaluation tool to manually input data.

5.5.3 Section Summary

In this section, we presented various interpretability considerations of the classifiers produced in this chapter. Different models may be chosen based on the clinical scenario. For researchers looking to identify general risk factors for melanoma, globally interpretable models with a small number of features are desired. In a clinical risk prediction context, however, a locally interpretable model with the best performance is desired. As shown in our results, there are tradeoffs between these two scenarios.



(a) No history



(b) History

Figure 5.14: XGB SHAP values from a random sample of instances

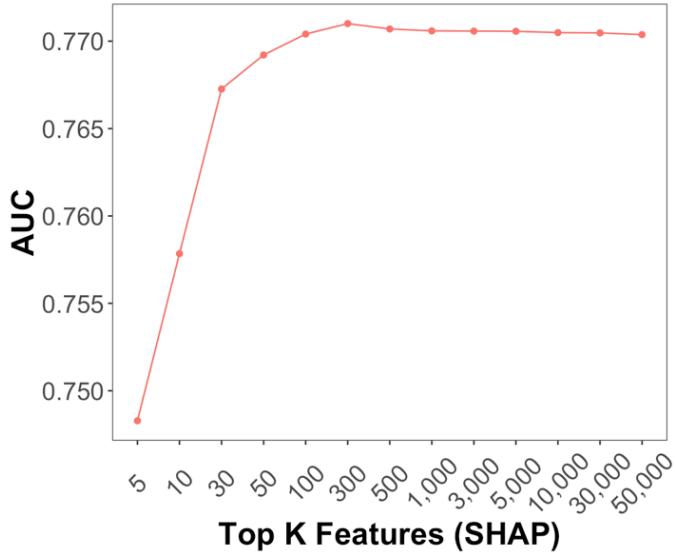


Figure 5.15: Model performance as more features are included in the XGB model using the 1m dataset. The top features are ranked according to the mean SHAP values from a trained model using all features.

A globally interpretable model, as shown with LR, tends to have worse performance than a model using more features, and the best performing ensemble models have limited global interpretability. SHAP values with these ensembles, however, are a promising new avenue for model interpretation and can be a great benefit when deploying prediction models into a clinical setting.

5.6 CHAPTER SUMMARY

This chapter opened with an introduction to cancer risk prediction models, followed by a review of previous literature building risk models for cancer, and then specifically for melanoma. We then described our two experiments that built risk models for melanoma using MAMEL data.

The structured-data EHR and cloud-based model training process described herein addressed the shortcomings identified in previous cancer risk modeling studies. *Availability of structured clinical data:* the structured, cloud-based EHR system provided

consistently collected data points across millions of patients at different practices. *Old data*: the data consistency allowed for rapid querying, de-identification and transformation of data to use for training machine learning models. The time between the end of the study period (December 2017) and experiment completions for both studies was less than one year. *Advanced modeling methods*: we conducted comprehensive studies of advanced classifiers such as XGBoost, as well as feature selection and data sampling methods. Risk factors were evaluated using LR weights and decision tree paths, and local interpretability was explored using tree traversal for RF and SHAP values for XGB. We showed that simple global models do not produce the results for clinical decision support, and a large number of features (>300) was required to produce the best model.

To the best of our knowledge, this work is the first to evaluate the impact of both data sampling techniques and feature selection for a melanoma risk prediction model. Additionally, we explored cost considerations when building models in the cloud, as well as interpretability of the various models. The average number of instances used in the reviewed works is 359,120, with the largest containing 2,975,369 [179]. Our largest dataset contains 9,531,408 instances (Section 5.4.1).

While models have been built from EHR data to predict risk for other cancers [60], none have been built for melanoma. To the best of our knowledge, this is the first study to automatically use raw features from an EHR system to build a melanoma risk prediction model, and certainly is the first to build a model using data from millions of real-world patient records. Only one study used more than 300 features [24] while most utilized less than 100. The raw features used in our experiments add up to over 100,000.

Our dataset contains a diverse patient population from dermatology offices located throughout the U.S. This makes our model more readily applicable to diverse patient populations than models developed using localized patient cohorts. Addition-

ally, using routinely collected EHR features allows more patients to be evaluated for melanoma as opposed to features collected through time-consuming questionnaires or examinations. This study does not negate the need for detailed case-control studies to investigate risk factors, but provides clinical decision support for evaluating an individual patient's risk for developing melanoma.

CHAPTER 6

LEARNING FROM LIMITED DATA

6.1 BACKGROUND AND MOTIVATION

The general consensus in the machine learning community is that more data means better models, and this assumption has not been made without experimental evidence [164]. The era of big data has enabled vast amounts of data to be processed and analyzed in a cost-efficient manner on a scale like never before. Limited data, however, is still a challenge even when dealing with big data. Just because there is a large amount of data, it is not necessarily the *right* data. In this chapter, we explore big data scenarios where specific types of data are limited.

Many datasets for machine learning can suffer from class imbalance, namely, when a particular class of interest is much less represented than other classes in a dataset [85, 166]. The estimated cancer incidence in the U.S. is 439.2 per 100,000 men and women [109]. Therefore, a predictive model for cancer risk would need to detect positive instances from a 0.44% class distribution (number of positive cases / number of total cases).

Supervised classification algorithms require that the data is *labeled*, meaning the class membership of each instance in the training data is known. For many applications, this requires expensive and time-consuming human annotation. Therefore, even if there is an infinite amount of computing power for model training, there is still a large cost that must be dedicated to labeling [73]. The question of “How much data is needed?” has been asked many times and explored through numerous studies, especially within the bioinformatics and biomedical community [55, 103]. More

recently, the problem of learning from limited labels has been formulated as an active area of research [141], even spawning a new research program funded by DARPA [3]. Generally, the problem of “limited labels” refers to when there is a large amount of unlabeled data available, but only a small amount of labeled data. Class imbalance is even more important when dealing with limited labels, as a theoretical cancer detection dataset with 10,000 labeled instances would only have 44 positive cases.

The representation learning nature of artificial neural networks leads many researchers to believe that enormous amounts of data are always needed to build effective models. With most problems, however, there is a point at which the law of diminishing returns takes effect, and the achieved classification performance hits a plateau with respect to dataset size, even for deep representational models [155]. This phenomenon can be visualized by creating a *learning curve*: training models on increasing sizes of data and plotting the data size versus classification performance on a graph.

In this chapter, we explore these two limited data problems: limited positive instances when there is class imbalance (Section 6.2), and limited labels when performing sample size determination (Section 6.3).

6.2 LIMITED POSITIVE SAMPLES

Datasets for machine learning are increasing in both availability and size. The field of big data has arisen in the last several years to be able to extract insight and build models from vast amounts of data. While big data has been historically defined by the 5 V’s (Volume, Velocity, Variety, Veracity, Value), it suffices to consider a dataset to be “big data” when traditional computing techniques and resources are unable to analyze or model the data. As computing technology continues to advance, certain datasets that used to be considered “big” can start be to handled by traditional methods.

The problem of class imbalance can be exacerbated when dealing with big data, as there can be millions of negative (majority) samples, but only hundreds or thousands of positive (minority) samples [76]. Certain techniques for handling class imbalance, such as random undersampling, will actually remove majority cases from a dataset. Therefore, a dataset can start out as “big data”, but if undersampling is performed, the data that the machine learning model is trained on may very well be “small” [135]. Machine learning research and model building is an experimental and iterative process; therefore, the cost and time to train a model can be significant for big data tasks. If a single model takes a long time to train, it can limit the amount of experimentation that can be done to achieve an exemplary model.

We present an in-depth study on learning from limited positive samples for the melanoma risk problem presented in Chapter 5. Using the dataset from Section 5.4, we applied several machine learning techniques for predicting individual risk of developing melanoma. We applied majority undersampling to these classifiers to determine how the various model configurations perform on the imbalanced big data task. K-means clustering of samples from each class shows that samples in the negative class have more homogeneity than those in the positive class. To the best of our knowledge, this is the first work to systematically study the impact of limited positive samples for a cancer risk model, as well as provide solutions for effectively learning from the data.

6.2.1 Related Works

The effect of class imbalance on machine learning models has been extensively studied in the literature [80, 142], but only a few studies in recent years have explored class imbalance for big data. The Evolutionary Computation for Big Data and Big Learning Workshop (ECBDL) 2014 hosted a competition to build the best predictive model on

a large imbalanced bioinformatics dataset¹. This dataset has 631 attributes and 32 million instances, 2% of which are in the positive class. Triguero et al., the winners of the competition, utilized random undersampling and evolutionary feature weighting to build a random forest model that achieved a 73% true positive rate and 73% true negative rate [158].

Fernandez et al. performed a review of studies that address class imbalance in big data and conducted their own experiment comparing machine learning model performance using the Hadoop MapReduce and Apache Spark frameworks using the ECBDL'14 data [53]. They applied random oversampling (ROS) and random undersampling (RUS) using Spark, and Synthetic Minority Over-sampling Technique (SMOTE) using Hadoop MapReduce. They found the Spark-based sampling methods performed better than SMOTE. Using both sampling methods, they only created a single balanced (50:50) class distribution. In our study, we evaluated various target class ratios in case 50:50 is not the optimal ratio.

6.2.2 Materials and Methods

In this section, we perform several extensive experiments to determine the impact of limited positive samples on a clinical risk problem, and solutions to mitigate this impact to develop accurate predictive models. We used six different classifiers and performed an initial grid search to determine appropriate parameters for each model in the subsequent experiments. Then, we evaluated the performance of the classifiers on both the original dataset and simulated imbalanced datasets. To improve classifier performance on these datasets, we performed undersampling of majority (negative) class samples.

¹<http://cruncher.ico2s.org/bdcomp/>

Simulating Imbalanced Datasets

To determine the impact of class imbalance on the melanoma risk problem, we created four simulated datasets with different levels of class imbalance from the original data. This is accomplished by sampling (without replacement) instances from the positive class. We created sample datasets with four different numbers of positive samples: 5000, 1000, 200, and 100. This is repeated five times for each dataset to reduce bias in the sampling process. Classifier results are then reported based on the average results across the five repeats. Along with the original dataset (containing 17,246 positive instances), this resulted in five different datasets that we trained all six classifiers on. Throughout this section, we refer to each dataset based on the number of positive samples in it.

Classifiers and Hyperparameters

We chose a wide range of classifiers from different families of algorithms to study the performance of each: logistic regression (LR), naïve Bayes (NB), support vector machine (SVM), decision tree (DT), random forest (RF), and regularized gradient boosted trees (XGB). Each model has some parameters (also known as hyperparameters) that must be chosen before training the model. We selected these parameters experimentally by performing a grid search with 5-fold cross-validation on the original dataset with RUS to a 35:65 class ratio, evaluated using AUC. The initial grid values were selected based on literature review, and then were evaluated graphically to find “peaks” in AUC values. If there was a large difference (>0.001) between peak and neighboring values, we explored additional values between them until the differences between values were small (<0.001). The parameters are outlined in the following paragraphs, and Table 6.1 outlines the grids and selected values for each. These selected values were used for all experiments in this section.

For LR, we tested various values of the L2 regularization parameter C during

Table 6.1: Hyperparameter grids

Classifier	Parameter	Grid Values	Selected
LR	C	0.001, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 10, 100	0.1
SVM	C	0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5	0.05
NB	α	0, 0.1, 0.5, 1, 2, 5, 10	5
DT	Max depth	None, 2, 3, 4, 5, 6, 7, 10, 20	6
RF	Max depth	None, 3, 5, 10	None
	Num trees	100, 200, 300, 500, 1000, 1500, 2000	500
XGB	Learning rate	0.01, 0.05, 0.1, 0.2	0.05
	Max depth	3, 5, 10	5
	Num trees	100, 200, 300, 500, 1000, 1500, 2000	500

the hyperparameter selection process. We used a multinomial Naïve Bayes model and selected from various values of the smoothing parameter α . Due to the size of the dataset we chose a linear kernel for SVM and tested various values of the L2 penalty parameter, C . Since SVM does not return class membership probabilities (as opposed to the other models used in this study), we calibrated probabilities with cross-validation using Platt’s method [116].

For DT, we chose to stop tree splitting based on the maximum depth of the tree, with the specific depth chosen by hyperparameter selection. We evaluated several values for the maximum depth of each tree in RF and the total number of trees for each forest. XGB utilizes tree models as the weak learners, and thus maximum depth and number of trees must be selected. In addition, we tested various values of the learning rate used in the boosting process.

Data Sampling

To alleviate the impact of class imbalance, we used RUS to preprocess the data with the following positive:negative sampling ratios to see which is most effective: 1:99,

Table 6.2: Negative samples by dataset and RUS Ratio

Dataset	No RUS	RUS	RUS	RUS	RUS	RUS
		0.01	0.1	0.25	0.35	0.5
100	9,514,162	9,900	900	300	186	100
200	9,514,162	19,800	1,800	600	371	200
1,000	9,514,162	99,000	9,000	3,000	1,857	1,000
5,000	9,514,162	495,000	45,000	15,000	9,286	5,000
17,246	9,514,162	1,707,354	155,214	51,738	32,028	17,246

10:90, 25:75, 35:65, and 50:50. Throughout the paper, we refer to these ratios based on the fraction of positive samples (i.e. 1:99 is represented as 0.01). Table 6.2 shows the number of negative samples for each dataset when the various RUS ratios are applied to them. We applied random undersampling to both the original dataset and simulated imbalanced datasets to understand the relationship between imbalance level and RUS target class ratio.

Feature Preprocessing

After any data sampling occurred, and before each classifier was trained, we performed preprocessing of features in each dataset. For all models, we removed features with zero variance (same value in all samples). Some classification models used in this study (NB, SVM, LR) depend on assumptions about distributions of independent variables, and can also perform poorly with high dimensions. For these models, we performed scaling and selected the top 1,000 features as ranked by the χ^2 statistic. Tree models are robust to varying feature distributions and perform internal feature selection in the model training process; therefore, for DT, RF, and XGB, we did not scale or select features before training the classifiers.

Evaluation

All experiments were performed and evaluated using stratified 5-fold cross-validation and AUC. Pipelines were created for each model configuration to ensure proper splitting of data in the cross-validation process. Figure 6.1 shows an example of a model pipeline including a simulated imbalanced dataset, undersampling, feature preprocessing, and finally classifier training and evaluation using cross-validation. Note that each preprocessing step (both sampling and feature processing) occurs *within* a single fold of cross-validation. Since each simulated imbalanced dataset is generated five times, this results in twenty-five runs of each model pipeline. For the original data, we repeated the 5-fold cross-validation five times to also achieve twenty-five total runs. Across all experiments in this study, 10,200 individual models were trained and scored, adding up to over 53 CPU core-days of computing time.

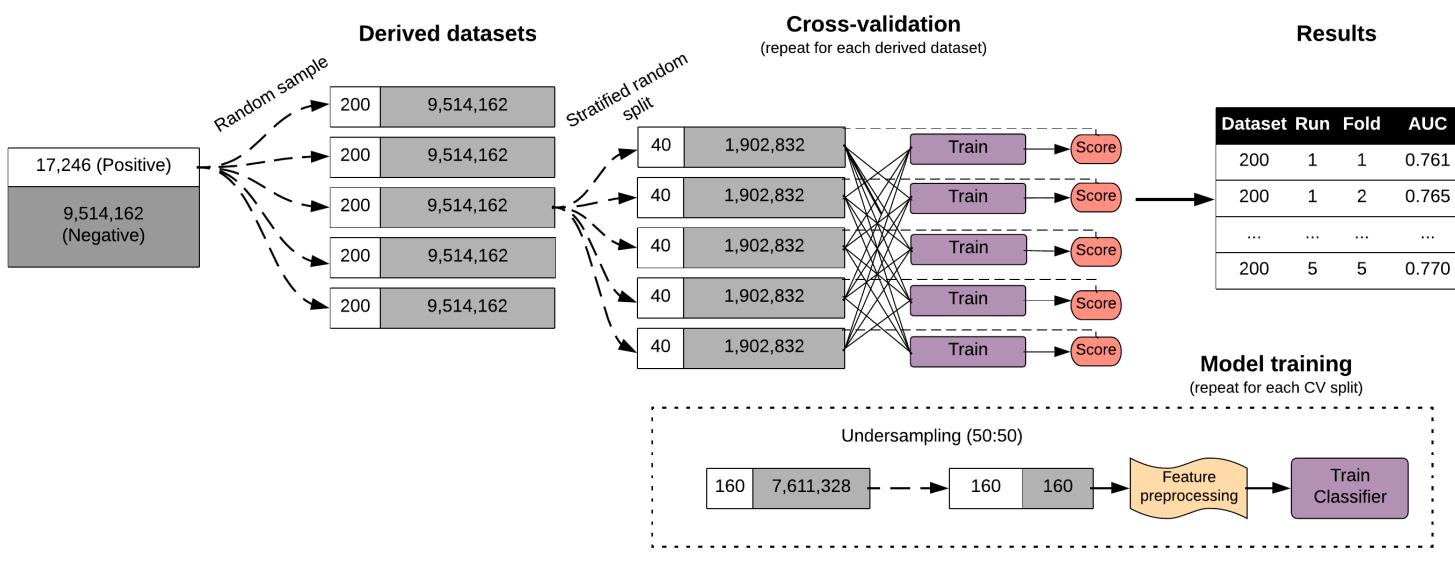


Figure 6.1: Machine learning pipeline for a model on the 200 dataset with undersampling to a 50:50 class ratio

To determine if observed differences in model performance were statistically significant, we performed a number of ANOVA tests. Factors and levels in the tests are:

- *Dataset*: 17,246, 5000, 1000, 200, 100 (based on the number of positive samples in each dataset)
- *Classifier*: LR, NB, SVM, DT, RF, XGB
- *RUS Ratio*: 0, 0.01, 0.1, 0.25, 0.35, 0.5 (target positive class ratio for RUS)

Some experiments do not utilize each machine learning technique, therefore, not all ANOVA tests presented will have all factors and levels described above. Additionally, we used Tukey's honestly significant difference (HSD) test to determine significant group differences for factors and interactions. Welch's t-test was used for head-to-head comparisons.

6.2.3 Results

Before applying any data sampling, we evaluated the performance of each classifier on the original and simulated imbalanced datasets. Results of this experiment are presented in Figure 6.2. ANOVA and HSD tests were performed individually for each classifier to test the differences in AUC across the five datasets. The color of each item in the figure is labeled according to its group in the HSD test. We can observe that the AUC generally increases across all classifiers as the number of positive samples in the dataset increases. DT and RF are the most affected by class imbalance, as the 17,246 dataset performs significantly better than all other datasets. This is likely due to overfitting in the tree models when there are not many positive instances available to the model. While RF can overcome this slightly due to the bagging approach, it is still affected and does not perform well when the number of available positive

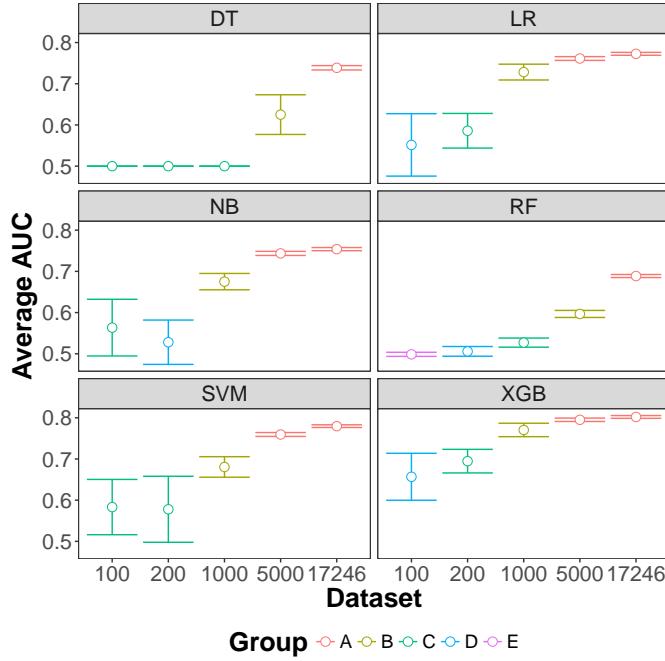


Figure 6.2: Average AUC for each classifier by dataset. Classifier/dataset combinations are labeled based on their group from Tukey's HSD test.

instances is low. LR, NB, SVM, and XGB are less affected by class imbalance, as they all have statistically equivalent performance on the 17,246 and 5000 datasets.

These results show that for big data, the class ratio is not an appropriate metric to use for the impact of class imbalance on a dataset. Even with a low class ratio (0.18%), an XGB model is able to achieve very good classification performance (AUC=0.80) because there are still an adequate number of positive samples available to the model (17,246).

Removing Majority Samples

We evaluated the performance of applying random undersampling before classification for both the original and derived datasets. Through ANOVA we found that the 200 and 100 datasets have statistically similar performance across all classifiers, so due to space constraints we only present results for the 17,246, 5000, 1000, and 200 datasets

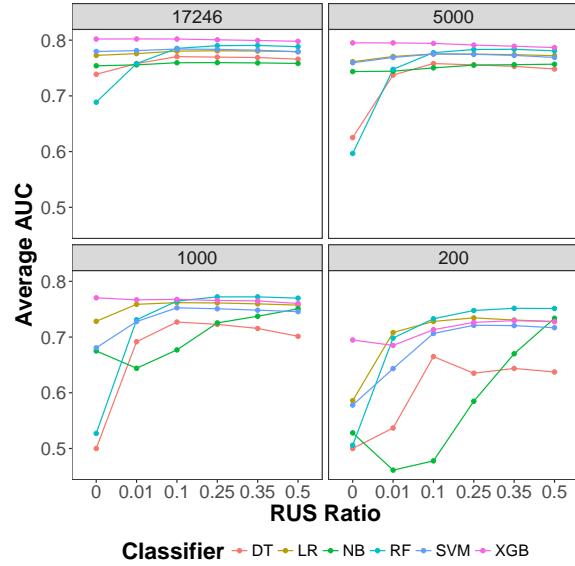
Table 6.3: ANOVA: Dataset/Classifier/RUS Ratio

Factor	DF	SS	MS	F	p
Dataset	3	6.735	2.245	3751.162	0.000
Classifier	5	2.733	0.547	913.249	0.000
RUS_Ratio	5	3.086	0.617	1031.277	0.000
Dataset:Classifier	15	1.173	0.078	130.614	0.000
Dataset:RUS_Ratio	15	1.127	0.075	125.596	0.000
Classifier:RUS_Ratio	25	2.465	0.099	164.747	0.000
Dataset:Classifier:RUS_Ratio	75	1.142	0.015	25.436	0.000
Residuals	3456	2.068	0.001		

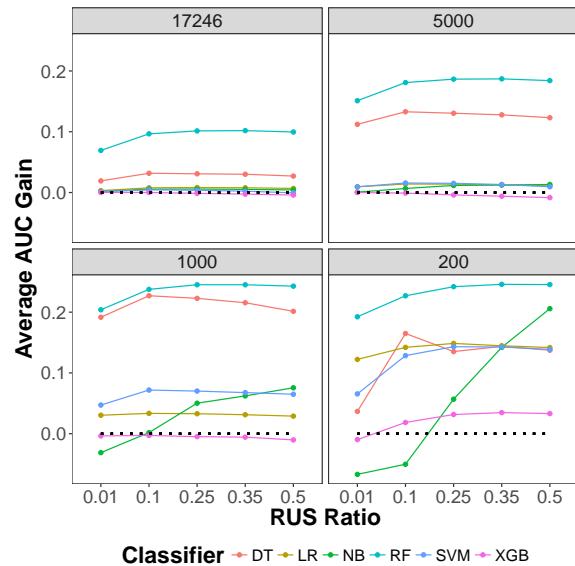
(Figure 6.3). Note that an RUS ratio of zero indicates that no RUS was performed. We consider this case to be the baseline for each dataset/classifier configuration, and Figure 6.3b shows the average gain in AUC for each configuration with respect to its baseline.

By observing these figures, we can see that the datasets with larger numbers of positive samples (less imbalance) tend to be less affected by RUS across most classifiers. RF and DT generally have large improvements over the baseline across all datasets. These models had relatively poor performance on the datasets without RUS, so RUS helps to improve their performance drastically. RF is even shown to have the highest AUC for the 1000 and 200 datasets with a 0.5 RUS ratio. The ANOVA test is presented in Table 6.3, and HSD tests are provided in Tables 6.4-6.5. The ANOVA shows that the Dataset, Classifier, and RUS Ratio factors and all interactions are significant. Across all classifiers and datasets, we can see that every RUS ratio achieves significantly higher AUC than no RUS (Table 6.4, Group D). While the 0.01 and 0.1 ratios perform better than no RUS, the other ratios (0.25, 0.35, and 0.5) perform significantly better than them and are not statistically different from each other (Group A). We can also see that XGB is the highest performing classifier across all datasets and RUS ratios (Table 6.5).

To simplify investigation of the Dataset and Classifier factors, we perform an-



(a) Average AUC



(b) Average AUC gain over baseline

Figure 6.3: RUS results by classifier and RUS ratio. An RUS ratio of 0 indicates no sampling (baseline).

Table 6.4: HSD: RUS Ratio

RUS_Ratio	Group	AUC	SD
0.5	A	0.753	0.040
0.35	A	0.752	0.043
0.25	A	0.749	0.054
0.1	B	0.741	0.069
0.01	C	0.719	0.083
0	D	0.670	0.106

Table 6.5: HSD: Classifier

Classifier	Group	AUC	SD
XGB	A	0.768	0.039
LR	B	0.752	0.046
SVM	C	0.742	0.056
RF	D	0.733	0.080
NB	E	0.697	0.095
DT	E	0.693	0.087

other ANOVA test in Table 6.6 by grouping all the top RUS ratios together into a new factor: “RUS”. This factor has two levels, FALSE for no RUS and TRUE for any of the selected RUS ratios (0.25, 0.35, or 0.5). All factors and interactions are also significant in this test. Table 6.7 presents results of the HSD test of Classifier/RUS interaction across all datasets. RF has the largest performance increase, going from the lowest AUC group without RUS to the highest AUC group with RUS. All classifiers perform significantly better when using RUS *except* for XGB. XGB without RUS even performs better or equivalent to all other classifiers with or without RUS (Group AB). As seen in Figure 6.3a, XGB is least affected by the number of positive samples and RUS ratios, and achieves top performance across most scenarios.

Across all classifiers, RUS significantly improves AUC results for each dataset (Table 6.8). Additionally, each dataset with RUS performs better than a dataset with more positive instances without RUS. The 5000 and 1000 datasets with RUS even perform better or equivalent to the original 17,246 dataset without RUS. While

Table 6.6: ANOVA: Dataset/Classifier/RUS

Factor	DF	SS	MS	F	p
Dataset	3	3.748	1.249	1805.922	0.000
Classifier	5	1.653	0.331	477.993	0.000
RUS	1	2.928	2.928	4233.318	0.000
Dataset:Classifier	15	0.445	0.030	42.886	0.000
Dataset:RUS	3	0.868	0.289	418.347	0.000
Classifier:RUS	5	1.735	0.347	501.658	0.000
Dataset:Classifier:RUS	15	0.399	0.027	38.501	0.000
Residuals	2352	1.627	0.001		

Table 6.7: HSD: Classifier/RUS interaction

Classifier:RUS	Group	AUC	SD
RF:TRUE	A	0.773	0.024
XGB:TRUE	A	0.770	0.033
XGB:FALSE	AB	0.766	0.046
LR:TRUE	BC	0.761	0.028
SVM:TRUE	C	0.755	0.033
NB:TRUE	D	0.729	0.055
DT:TRUE	E	0.718	0.059
LR:FALSE	E	0.712	0.078
SVM:FALSE	F	0.699	0.090
NB:FALSE	G	0.675	0.095
DT:FALSE	H	0.591	0.103
RF:FALSE	H	0.580	0.072

Table 6.8: HSD: Dataset/RUS interaction

Dataset:RUS	Group	AUC	SD
17246:TRUE	A	0.780	0.014
5000:TRUE	B	0.771	0.014
17246:FALSE	C	0.756	0.036
1000:TRUE	C	0.749	0.025
5000:FALSE	D	0.714	0.078
200:TRUE	E	0.705	0.064
1000:FALSE	F	0.647	0.102
200:FALSE	G	0.565	0.080

RUS significantly improves performance for all classifiers, Figure 6.3b shows that the magnitude of AUC gain gets higher as the number of positive samples in a dataset decreases. Besides RF and DT, the classifiers on the 17,246 and 5000 datasets are barely affected by RUS. The smaller datasets, however, need RUS to achieve good performance, and the performance on these are actually not much lower than the datasets with more positive samples. An RF model with an RUS ratio of 0.5 on the 200 dataset can achieve better performance than an RF model on the original data with no sampling (0.751 vs. 0.689, $p<0.001$), and close performance on the original data with 0.5 sampling (0.751 vs. 0.788, $p<0.001$). This is quite surprising because the 200 dataset with 0.5 sampling only has 400 instances (200 positive, 200 negative), while the 17,246 dataset without RUS has 9,531,408 instances. This shows that a large number of negative samples is not required for building an accurate model; rather, the number of positive samples controls how effective a model can be in discriminating between the two classes. Therefore, we can hypothesize that the heterogeneity of the positive samples is higher than that of the negative samples.

To test this hypothesis, we performed K-means clustering on random samples of positive and negative instances separately. We randomly sampled 100, 500, 1,000, 5,000, and 10,000 instances from each class twenty-five times and ran K-means clustering with $K = 5$ and the Euclidean distance function. Grid search was not performed

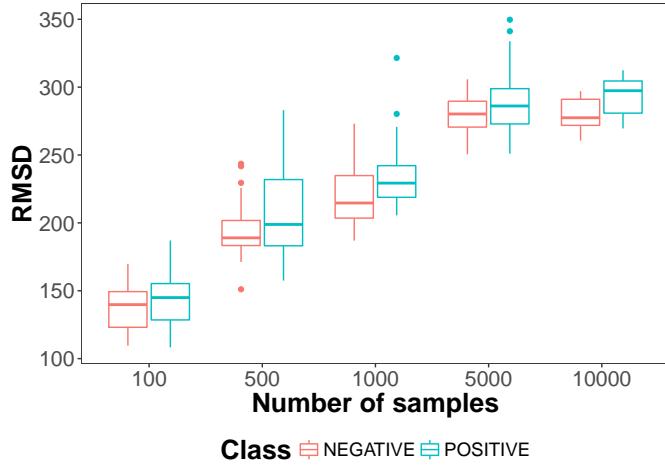


Figure 6.4: Distribution of root-mean-squared distance of each sample to its cluster center for five different samples by class.

on these parameters because the goal of the clustering was to make a relative comparison of the similarity among the positive and negative sample groups. Therefore, we selected general parameters and ensured consistency when clustering the positive and negative samples separately. We then measured the distance of each sample to its cluster center and aggregated all the distances to produce a metric that measures the heterogeneity of the instances. We take the root-mean-square of each of the distances for comparison purposes:

$$RMSD = \sqrt{\frac{\sum_{i=1}^n d(x_i, CC_i)^2}{n}} \quad (6.1)$$

Where n is the number of samples being clustered, CC_i is the cluster center for a particular instance, and d is the distance function. Figure 6.4 shows that across all random samples, a set of positive samples generally has greater RMSD than an equally-sized set of negative samples. This confirms our hypothesis that negative samples are more similar to each other than positive samples. Therefore, less negative (majority) samples are needed for classification purposes.

Table 6.9: EC2 instance types

Instance Type	# CPUs	Memory (GB)	Hourly Price (\$)
r4.2xlarge	8	61	0.532
r4.4xlarge	16	122	1.064
c4.8xlarge	36	60	1.591
m4.10xlarge	40	160	2.000

Model Training Costs

When dealing with big data, predictive accuracy is not the only consideration when choosing classifiers and machine learning techniques. With large amounts of data, computational complexity and cost must be a factor in the selection process. Since all experiments were conducted using Amazon EC2, we are able to directly calculate the cost of training each model configuration. Table 6.9 shows the different instance types used in the experiment along with the on-demand hourly cost of each as of June 27, 2018. Note that Amazon offers discounted rates by using Spot instances², but those prices are not consistent over time so we use the on-demand hourly rate for comparisons.

Running time is not the best comparison across different classifiers because the same model configuration can be run on more advanced hardware that would speed up running time. Additionally, using an instance with more CPUs would only benefit models that support multithreading. Therefore, we estimate the cost of training a model by multiplying the running time by the EC2 cost for that instance. EC2 has a fairly standard pricing model that increases along with hardware complexity. For example, the r4.4xlarge instance type has double the number of CPUs and memory as r4.2xlarge and is double the price. This allows us to use EC2 cost as a proxy for running time that is not affected by hardware configuration. We attempted to use the smallest possible instance type for each model, and increased the instance

²<https://aws.amazon.com/ec2/spot/>

type as we ran into memory errors or significantly long running time. The c4.8xlarge and m4.10xlarge instance types were only used for XGB and RF as they support multithreading and benefit from machines with a large number of CPUs.

Figure 6.5 shows the cost of a single model fit for each configuration. Generally, the cost increases as the model complexity and dataset size increases. Simpler models such as LR and NB do not incur large costs even for the largest datasets. Some classifiers can cost over \$2, and RF without sampling on the full dataset costs over \$10 for a single model fit. This is due to sheer size of the dataset without sampling (Table 6.2). The results shows that performing undersampling improves performance across all datasets and classifiers. This is advantageous because the models with undersampling will require significantly less time and money to train. Additionally, after the RUS is performed, each dataset becomes much smaller and does not require big data methods to learn from. The largest dataset with 0.5 sampling only has 34,492 instances; most modern laptops will be able to train a model on a dataset of that size.

6.2.4 Section Summary

This section examined the impact of limited positive samples on a model that predicts individual patient risk of developing melanoma. We created several datasets with limited numbers of positive cases to determine how this affects model performance across a number of classifiers. Additionally, we discussed the effect of data sampling to handle the imbalance of each dataset. From the full dataset, we derived four additional datasets simulating different levels of class imbalance (100, 200, 1,000, and 5,000 positive samples). We note increasing model performance (as measured by AUC) as the number of positive samples in each dataset increases. After applying RUS to each dataset, with various target class ratios, we find that RUS significantly improves model performance. This shows that while positive cases are indeed distinct

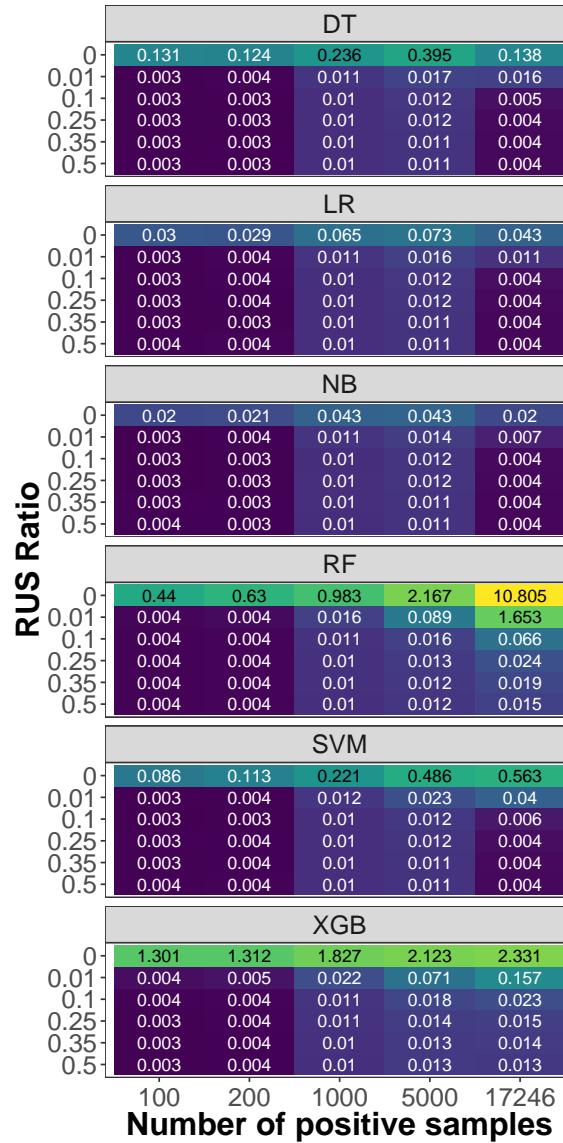


Figure 6.5: Average training cost (in dollars) for a single model fit by number of positive samples and RUS ratios for each classifier. Fill colors are based on log scale to highlight differences.

from negative cases, many of the negative cases are similar to each other and are not required for building an accurate model. There also appears to be more variation between instances in the positive class than between instances in the negative class, shown by the fact that the model performance increases when there are more positive cases available, but is not very affected when negative cases are limited. We performed K-means clustering individually on samples from each class, and show that there is a greater distance between samples and their cluster center for the positive samples versus the negative samples. This indicates greater heterogeneity between instances of the positive class than those of the negative class. The cost of building a model is significantly reduced by using RUS. Therefore, we can conclude that it is not necessary to use the full 9.5 million records in this study, and we can save time and cost by performing random undersampling of the majority class before building a classifier.

6.3 LEARNING CURVE APPROXIMATION WITH LIMITED LABELS

Labeling data for supervised learning can be an expensive task, especially when large amounts of data are required to build an adequate classifier. For most problems, there exists a point of diminishing returns on a learning curve where adding more data only marginally increases model performance (i.e. convergence). It would be beneficial to approximate this point for scenarios where there is a large amount of data available but only a small amount of labeled data. Then, time and resources can be spent wisely to label the sample that is required for acceptable model performance. In this section, we explore learning curve approximation methods on the MAMEL melanoma data as well as three other real-world biomedical datasets, spanning genomics, proteomics, and insurance claims data, all with millions of instances each and <2% class ratio. We evaluate a curve fitting method developed on small data using an inverse power law model, and propose a new semi-supervised method to take advantage of the large amount of unlabeled data [136]. We find that the traditional curve fitting method is

useful for most datasets, while the semi-supervised method provides a stable estimator of convergence as dataset sizes increase.

Approximating learning curves is a useful exercise for scenarios of limited labeled data where more labels can be gathered at a known cost [132, 134]. The shape of the curve along with the labeling cost can be used to estimate the point of diminishing returns: where it would not be worth it to collect more labeled data. We explore this scenario here for the four large, imbalanced biomedical datasets. We simulate the problem of limited labeled data by only making a small number of labels available for building learning curves, and evaluate the accuracy of these curves against one built on the actual labeled data. We apply two techniques for learning curve building: (1) fitting an inverse power law curve using nonlinear least squares optimization, and (2) building a semi-supervised learning curve by pseudo-labeling the unlabeled data from a classifier trained on the labeled data. These methods are compared to see how well they fit to the actual learning curve, and more importantly, how well they can identify convergence. To the best of our knowledge, this is the first study to apply inverse power law learning curve fitting to imbalanced big datasets, as well as the first to propose a semi-supervised method for learning curve approximation.

Section 6.3.1 outlines prior studies related to learning curves and sample size determination. The data and modeling methods are presented in Sections 6.3.2 and 6.3.3, respectively, followed by a discussion of the results in Section 6.3.4.

6.3.1 Related Works

Learning curves are a common component of machine learning research and are associated with several areas of research such as sample size determination, active learning, and progressive sampling.

Sample size determination is a core part of many statistical studies, particularly those in the healthcare and biomedical fields [93]. In short, sample size determination

identifies the number of samples that are needed to prove or disprove a particular hypothesis or test [61]. For machine learning problems, that can equate to the number of instances that are needed to build an adequate classifier. Figuero et al. explored the use of an inverse power law model to fit a learning curve using nonlinear weighted least squares optimization [55] and compared that to a non-weighted method devised by Mukherjee et al. [103]. They evaluated their method on two medical text datasets and one signal processing dataset with 7,016, 8,449, and 5,000 instances respectively. The positive class ratios for these datasets ranged from 0.240% to 0.400%. They found that between 80 to 560 labeled samples were required to fit an accurate learning curve compared to the actual data. We used Figuero’s inverse power law method as one technique for learning curve approximation.

Progressive sampling and active learning are related fields of research that use increasing sizes of labeled data to train models. Progressive sampling attempts to find a point of convergence where adding more data does not improve model performance, based on a pre-determined or adaptive sampling schedule [117]. The goal is to achieve the best performance with the minimum amount of computation. Active learning has a slightly different goal of selecting the most informative instances for training a model [144]. Active learning methods iteratively add more data to achieve better performance, and can be used to build a learning curve.

The current study attempts to provide a method for sample size determination, rather than selecting the best instances or minimizing computation. While we utilize some methods from the progressive sampling and active learning communities (such as curve comparison methods and convergence detection), this study does not attempt to contribute towards those fields.

Table 6.10: Datasets

	ECB DL	Medicare	Melanoma	Splice
Domain	Proteomics	Insurance claims	Clinical records	Genomics
Instances	7,998,231	3,692,555	9,531,408	4,627,840
Positive Instances	171,933	1,409	17,246	14,549
Class Ratio	2.15%	0.0381%	0.181%	0.314%
Features	985	123	117,513	100,000

6.3.2 Data

We used four datasets from several domains within biomedical and health informatics, three of which are derived from publicly-available data. All datasets are inherently imbalanced (i.e. class balance was not artificially altered for the experiments). Table 6.10 outlines each training dataset, and the following sections describe them in more detail, review previous work, and discuss data processing that was performed before experimentation.

ECB DL

The Evolutionary Computation for Big Data and Big Learning (ECB DL) workshop published a large protein contact map prediction dataset for a competition as part of the 2014 Genetic and Evolutionary Computation Conference (GECCO)³. The dataset was originally generated to train a predictor for a different competition: 9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction [17]. The competition provided separate train and test sets; in our study we derived our dataset from the test set (7,998,231 instances and 985 feature after one-hot encoding of categorical features). Each instance represents a pair of amino acids, and the class label is if the pair is in contact. Understanding protein structure is important for bioengineering research such as drug development, because the function of a protein depends on its 3-dimensional structure based on a sequence of amino

³<http://cruncher.ico2s.org/bdcomp/>

acids. There are many sub-problems in protein structure prediction, of which contact map prediction is one. Specifically, contact map prediction involves estimating if two amino acids are in contact in a 3D structure based on their sequence properties alone.

There are several studies in the literature that have used the ECBDL data for different experiments. Triguero et al. won the ECBDL competition [157] using a random forest model, random oversampling, and differential evolutionary feature weighting. Rio et al. [43] describe the same experiment as Triguero et al., and achieve the same best results (compare Rio Table VI to Triguero Table V). Rio does provide more results than Triguero, namely RF results without ROS, and random undersampling (RUS) results. Both studies were performed using Hadoop MapReduce and the Mahout library for its RF implementation.

Medicare

The Centers for Medicare & Medicaid Services (CMS) runs the Medicare insurance program, which covers all U.S. citizens age 65 and older (along with select other groups). CMS releases aggregated public use files every year covering various components of the insurance program such as provider demographics, payments, drug utilization, and more.

Bauder et al. used the Medicare Provider Utilization and Payment Data set covering 2012 to 2015⁴ to detect possible fraudulent medical providers [22,23,64]. The dataset contains physician information and aggregate payment information for each combination of Medicare physician and medical procedure, such as: provider specialty, provider gender, number of procedures performed, number of unique beneficiaries receiving the service, number of unique beneficiaries per day, average submitted charge amount (dollars), and average payment made to the provider per claim (dollars). This results in 3,692,555 instances and 123 features after one-hot encoding of categorical

⁴<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html>

data.

The class label was gathered from another publicly-available U.S. government dataset: List of Excluded Individuals/Entities (LEIE) from the Department of Health and Human Services⁵. The individuals and entities in this list are prohibited from participating in federally-funded healthcare programs such as Medicare due to various convictions related to medical practice, or license revocation or suspension. The LEIE database was linked to the Medicare dataset via National Provider Identifier (NPI), which is a unique identifier for each physician in the U.S. The original Medicare dataset contains one row for each physician/procedure performed, so Bauder et al. reduced the data to the physician-level by computing aggregated stats for the values in the data (i.e. min, max, mean, median, sum, standard deviation).

Melanoma

This section uses the same MAMEL melanoma risk dataset as Sections 5.4 and 6.2. The dataset contains information from routine dermatology office visits, occurring from 2011 to 2016, for 9,531,408 unique patients that did not have a diagnosis of melanoma through 2016. The class label for each instance is whether or not the patient developed melanoma in the subsequent year (2017).

Splice

Splice is a dataset for detecting human acceptor splice sites in DNA sequences, gathered from the LIBSVM [32] dataset repository⁶. Splice site detection is an important part of understanding gene structure [21]. The prediction problem involves detecting the border between introns and exons (splice site) in a DNA sequence. These splice sites occur after dinucleotides AG (acceptor site) or before dinucleotides GT (donor site), but these dinucleotides do not always mark a splice site. This machine learning

⁵https://oig.hhs.gov/exclusions/exclusions_list.asp

⁶<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#splice-site>

problem involves using neighboring sequences to discriminate between real and fake splice sites.

The dataset was originally created by Sonnenburg et al. to demonstrate the effectiveness of their COFFIN linear SVM training strategy [149]. They provided downloadable sequence data along with scripts to extract features for the model, but Agarwal et al. [8] materialized this full dataset of 50 million instances and 11,725,480 features and provided it for the LIBSVM repository. Due to the sheer size and dimensionality of the data, we derived our train/test sets from Sonnenburg/Agarwal’s test set for our experiments (4,627,840 instances, 0.3% positive) and selected the top 100,000 features as ranked by the χ^2 statistic.

The research using this dataset, led by Sonnenburg, Agarwal, and others, focused primarily on methods for large-scale computing, and did not attempt to approximate learning curves or sample the data. Agarwal even stated that the Splice dataset was “the largest public data set for which subsampling is not an effective learning strategy.” [8] To confirm or deny that statement is outside the scope of this study, although for the sampled dataset we used, we observe a performance plateau in the learning curve.

6.3.3 Methods

We conducted experiments to perform sample size determination on four large imbalanced datasets by approximating learning curves. To do so, we used a dataset where all labels are known, then created samples of data without labels to simulate limited label scenarios.

Learning curves

A learning curve is created by plotting the size of the dataset versus the classification performance as the size increases. Due to the large and imbalanced nature of the

dataset, we present all curves with the number of positive cases for brevity. In the text, we also refer to the sizes based on the number of positive instances and include the total number of instances in parentheses. Generally, a sampling schedule is set to determine which points along the curve models are built and evaluated for, in the same manner as progressive sampling [117]. We used an arithmetic schedule with a constant step size that resulted in 100 steps for each dataset. Since the datasets have a varying number of instances, the step size is different for each.

Each point in the learning curve is averaged over 10 repeats of 5-fold cross-validation, evaluated using AUC. The machine learning pipeline was consistent across all models for a given dataset. For feature pre-processing we removed features with zero variance, performed feature scaling, and selected the top features according to χ^2 . We selected 100 features for ECBDL and 1,000 features for Melanoma and Splice (no feature selection was performed for Medicare). The classifier was logistic regression with the L2 penalty. Note that the pre-processing steps were all performed independently within each repeat and fold of cross-validation.

Approximation methods

To determine sampling schedules for the approximation methods, we first examined the learning curves for each dataset and found the point in the curves where the AUC was within 1% of the AUC using the full dataset. Then, to reduce computational requirements, we built curves only up to that point, also using 100 steps for the reduced learning curve.

For the inverse power law approximation method, we trained and evaluated models with a schedule from 10 to 200 with a step size of 10. Then, these points were used to fit a curve using nonlinear least squares optimization according to Figueroa et al.’s method [55]. The optimization routine needed at least 3 training points to converge; therefore, we were able to start making predictions from 30 positive samples.

For the new proposed semi-supervised approximation method, we trained a classifier using LR on the small labeled data and used that to create pseudo-labels for the unlabeled data. We trained models with 30, 50, and 100 positive samples and then created pseudo labels for the rest of the samples. Then, the same learning curve creation process on the actual data was used for the pseudo-labeled data. A new model from the labeled data was trained for each of the 10 repeats. This involves using all the available data, but still only the small amount of labeled data. This helps to capture the variance and properties of the full dataset, even though labels are not available. The class distribution of the pseudo labels were drastically different than the original distribution; for each point in the learning curve we sampled instances from the positive and negative pseudo labels separately to match the target class distribution. If there were not enough instances to satisfy the desired sample size, we performed random oversampling.

Evaluation

We used several methods to compare the actual learning curve to the two approximation techniques. First, we can compare the curves visually to note the trends in curve shape. To quantitatively compare the curves, we compute the mean absolute error (MAE) between points on the actual and the approximated curve. The cumulative MAE as the prediction size increases can be plotted to observe how accurate the methods are at approximating performance at larger dataset sizes.

While MAE measures the average distance between each actual and predicted point, it does not quantify well the differences in curve shape. To identify potential points of diminishing returns, we calculated point-wise slopes using linear regression with local sampling (LRLS) [117]. This method selects neighbors around a point in the curve and trains a linear model to calculate the slope of the curve at that point. We used 4 neighbors (2 before and 2 after) for LRLS. A confidence level of 0.95 was

used for all statistical tests.

6.3.4 Results

The experiment was conducted in a step-wise approach, first by analyzing the full learning curves for each dataset, then performing another round of experimentation to build and evaluate the learning curve approximation methods. For the approximation methods, we used at most 200 positive samples for training, then created a test curve for each dataset based on the behavior of the full learning curve. Sampling schedules for the full and test learning curves are provided in Table 6.11. In total, over 100,000 logistic regression models were trained and evaluated, adding up to approximately 717 CPU-days of compute time.

Learning curves

The full learning curves for each dataset are presented in Figure 6.6. For all curves, there is a point about halfway or before where the AUC is within 1% of the full data (indicated by the red and black dotted lines). It is interesting to note that the point of rapid performance increase for all datasets besides Medicare occurs before 2,500 positive instances. This shows that even for these big datasets, not all the data is required to achieve adequate classifier performance. Medicare has a very small range on the y-axis, indicating that adding more samples does not drastically increase performance. We hypothesize that this is due to the severely imbalanced nature of the dataset- there are only 1,409 positive samples available and the classifiers are not able to discriminate well between the two classes. For ECBDL, the curve begins at an AUC very close to the full AUC. This is because the period of rapid performance increase for ECBDL is before the start of the sampling schedule, as will be seen in subsequent analysis.

The approximated learning curves for both methods are presented in Figure 6.7.

Table 6.11: Sampling schedules

(a) Full learning curve

ECBTL	Medicare	Melanoma	Splice
1,719 (79,966)	14 (36,689)	172 (95,059)	145 (46,122)
3,438 (159,932)	28 (73,378)	344 (190,118)	290 (92,244)
5,157 (239,898)	42 (110,067)	516 (285,177)	435 (138,366)
6,876 (319,864)	56 (146,756)	688 (380,236)	580 (184,488)
8,595 (399,830)	70 (183,445)	860 (475,295)	725 (230,610)
...
165,024 (7,676,736)	1,344 (3,522,144)	16,512 (9,125,664)	13,920 (4,427,712)
166,743 (7,756,702)	1,358 (3,558,833)	16,684 (9,220,723)	14,065 (4,473,834)
168,462 (7,836,668)	1,372 (3,595,522)	16,856 (9,315,782)	14,210 (4,519,956)
170,181 (7,916,634)	1,386 (3,632,211)	17,028 (9,410,841)	14,355 (4,566,078)
171,900 (7,996,600)	1,400 (3,668,900)	17,200 (9,505,900)	14,500 (4,612,200)

(b) Test learning curve

ECBTL	Medicare	Melanoma	Splice
34 (1,581)	6 (15,724)	79 (43,661)	81 (25,765)
68 (3,162)	12 (31,448)	158 (87,322)	162 (51,530)
102 (4,743)	18 (47,172)	237 (130,983)	243 (77,295)
136 (6,324)	24 (62,896)	316 (174,644)	324 (103,060)
170 (7,905)	30 (78,620)	395 (218,305)	405 (128,825)
...
3,264 (151,776)	576 (1,509,504)	7,584 (4,191,456)	7,776 (2,473,440)
3,298 (153,357)	582 (1,525,228)	7,663 (4,235,117)	7,857 (2,499,205)
3,332 (154,938)	588 (1,540,952)	7,742 (4,278,778)	7,938 (2,524,970)
3,366 (156,519)	594 (1,556,676)	7,821 (4,322,439)	8,019 (2,550,735)
3,400 (158,100)	600 (1,572,400)	7,900 (4,366,100)	8,100 (2,576,500)

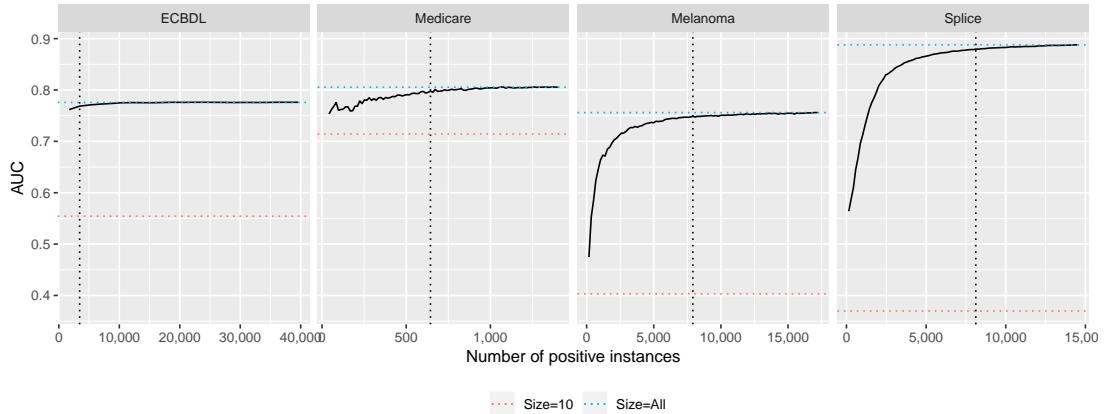


Figure 6.6: Full learning curves for each dataset. The horizontal dotted lines indicate the range of AUC from a small size (10) to the full dataset (last point). The dotted black line indicates the point where the AUC is within 1% of the full dataset. For ECBDL, we only show part of the curve as the AUC converges very early.

While the inverse power law method has a large number of potential fit configurations, we show fit sizes of 30, 50, and 100 to compare to the semi-supervised method. For all datasets except Melanoma, the inverse power law method tends to fit to the test curves quite well, while the semi-supervised curves tend to have the correct shape but larger absolute values.

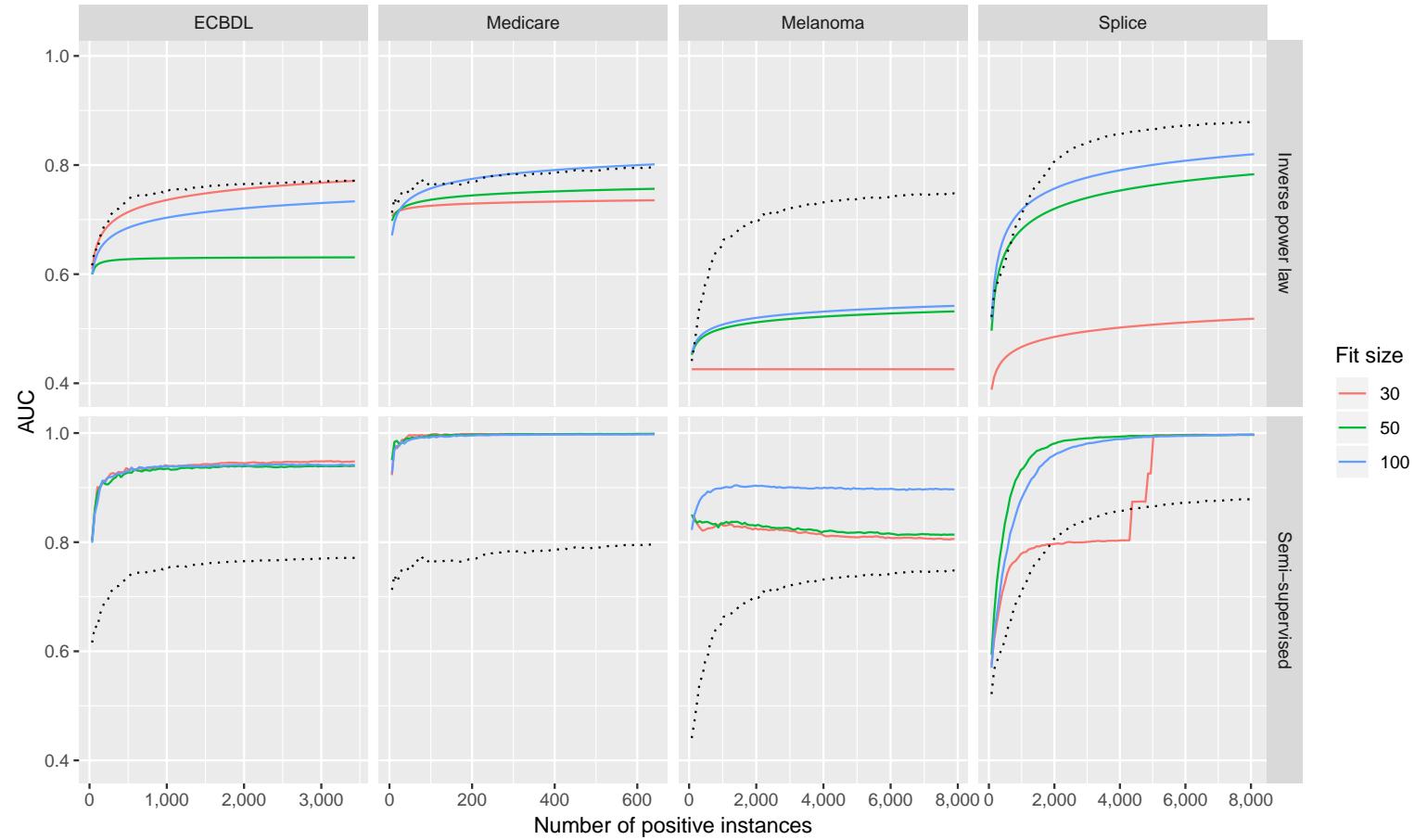


Figure 6.7: Approximated learning curves for each dataset using the inverse power law and semi-supervised methods. The dotted line indicates the test learning curve for each dataset.

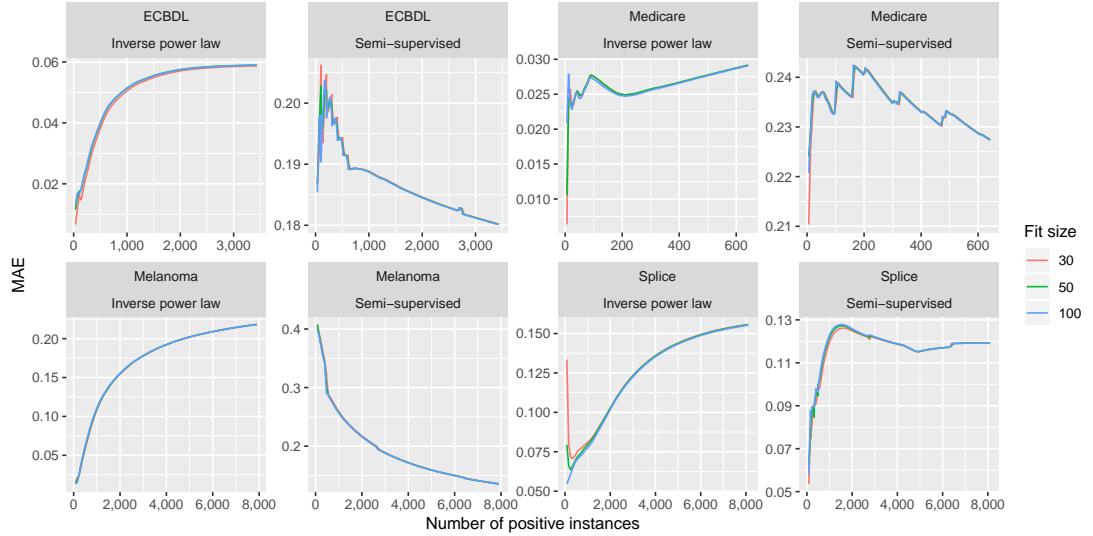


Figure 6.8: MAE as prediction size increases

The MAE as the number of positive samples increases is provided in Figure 6.8. Across all datasets, the power law method shows an increasing MAE as the size of the number of positive instances increases, while the semi-supervised method has a decreasing MAE. Note that the error is indeed larger for the semi-supervised method, so it is less accurate if the actual predicted AUC value is important. We believe that the more important scenario is the point of convergence, since that is what will be used to make a determination about how much data to label for an experiment. Therefore, the semi-supervised curves are still be valid for identifying convergence.

Inverse power law method

As seen in Figure 6.7, the inverse power law method fits well to the test curve for all datasets except Melanoma. Those curves all use a sampling schedule starting from 10 to 30, 50, or 100 for fitting the inverse power law curve. We explored varying this fit schedule by starting and ending at all possible points in the train data where the number of points for training is >2 . This results in a large number of possible fit schedules: 10-30, ..., 10-200, .., 100-130, 100-140, .., 100-200, 180-200, etc. The

Table 6.12: Best fitting inverse power law curves

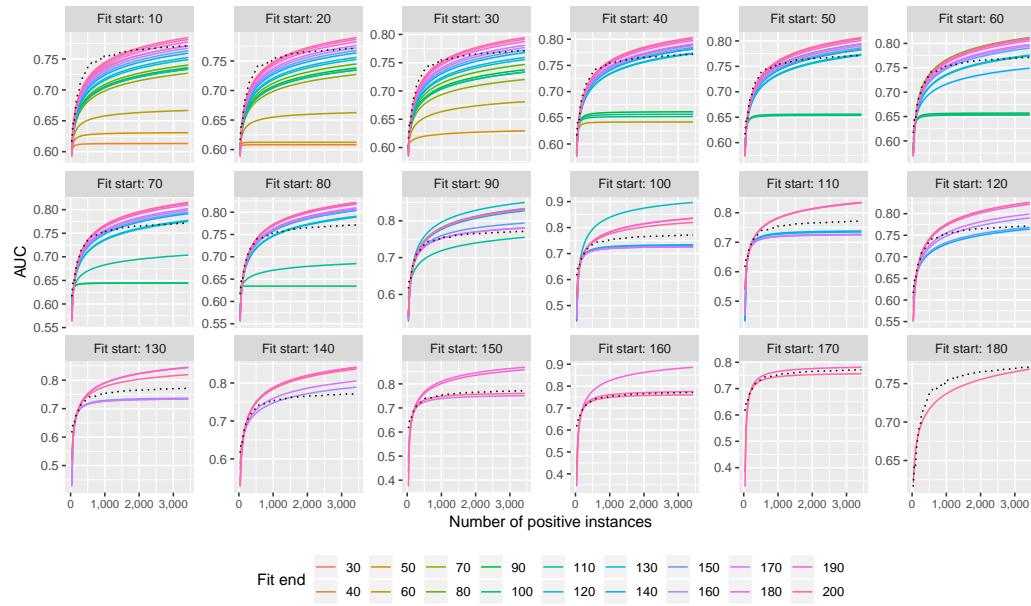
Dataset	Fit start	Fit end	MAE
ECBDL	90	170	0.006
Medicare	80	180	0.004
Melanoma	150	170	0.011
Splice	30	50	0.017

results are presented in Figure 6.9. Specifically for Melanoma, we see that the curves tend to fit better when the fitting starts with >120 positive instances. Table 6.12 shows the fit schedules for the best fitting curves for each dataset (according to MAE). Note that we are only able to know which curves fit best because we know the actual test curve. In a real scenario of sample size determination with limited labels, we would not know the actual learning curve. Therefore, we are unable to know which fit schedule produces the correct approximation for a learning curve. Future work is needed to develop methods that can evaluate the accuracy of these curves even when the actual learning curve is not known.

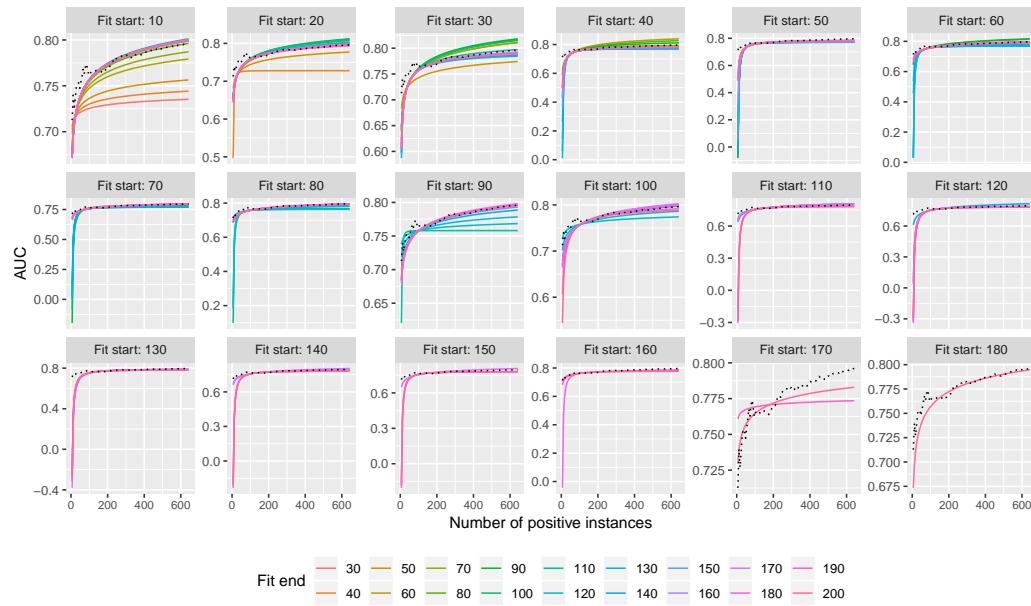
Semi-supervised method

Even if the absolute values of AUC are overly optimistic with the semi-supervised method, we believe that the most important point in a learning curve for sample size determination is the point of convergence, as that is what will be used to decide how much data to label for an experiment. The point-wise slopes as calculated by LRLS for the Splice data are shown in Figure 6.10, with a sample convergence point at a slope <0.0001 . This means that each addition of 100 positive (31,808 total) instances only increases the AUC of the model by 0.0001. The actual data hits the convergence point at 1,300 instances, while the semi-supervised curve hits at 1,000, and the inverse power law at 700. Therefore, if the point of convergence is the key consideration, the semi-supervised method is more accurate.

Most learning curves exhibit a trend where the initial portion of the curve shows

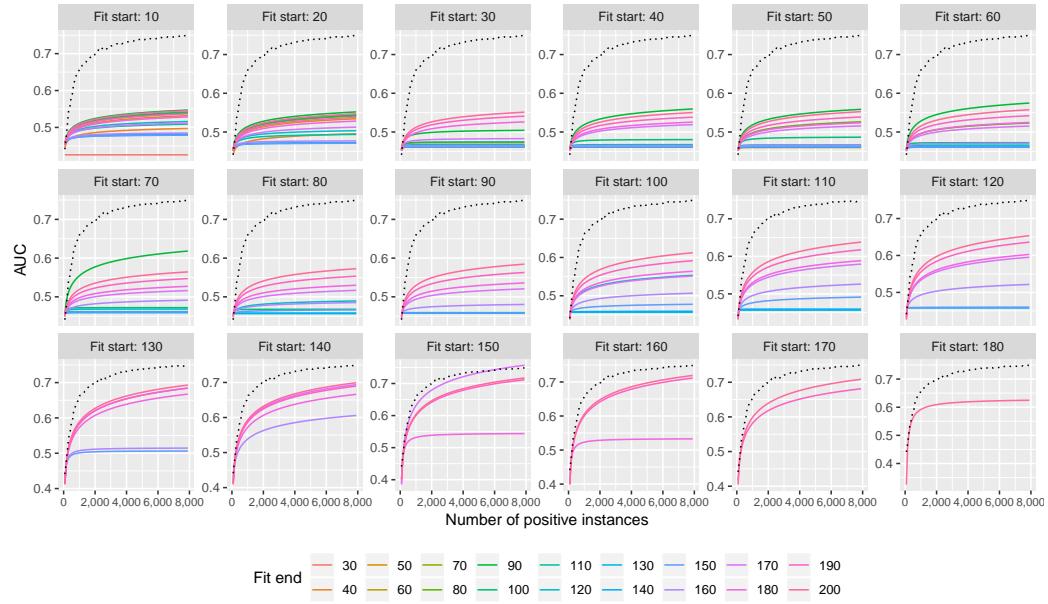


(a) ECBDL

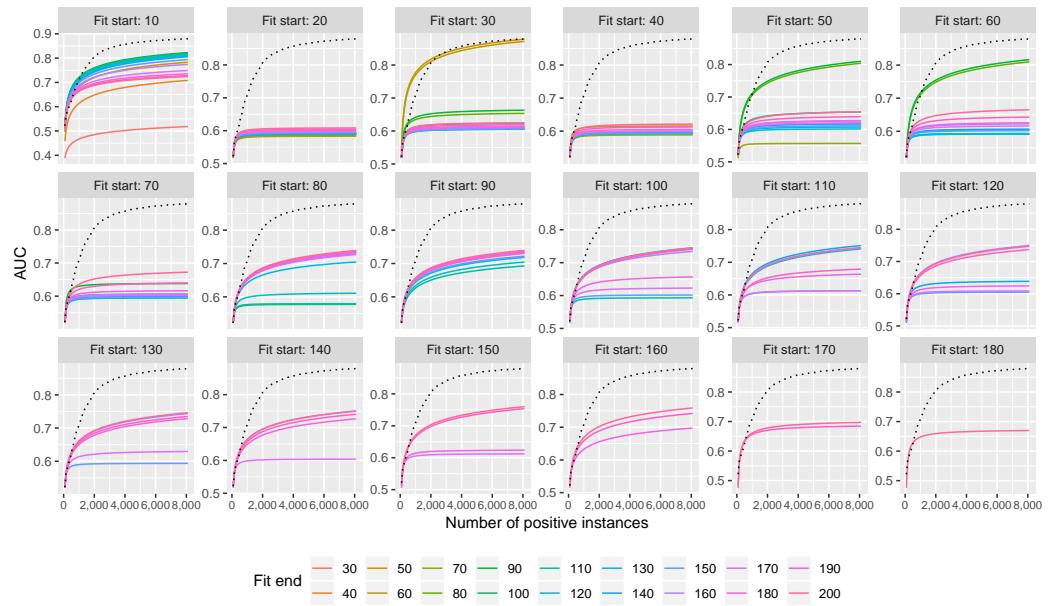


(b) Medicare

Figure 6.9: Inverse power law method with varying fit schedules



(c) Melanoma



(d) Splice

Figure 6.9: Inverse power law method with varying fit schedules (contin.)

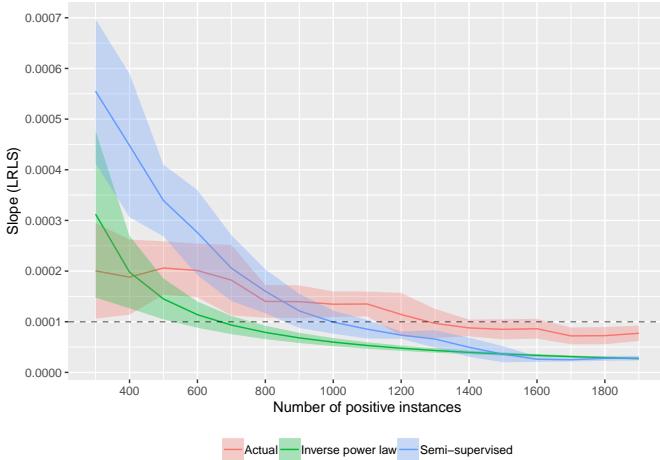


Figure 6.10: Point-wise slopes determined by LRLS (Splice dataset). The dotted line represents a point of convergence.

Table 6.13: Curve slopes by region

Region	Actual	Inverse power law	Semi-supervised
Exp	2.514 (1.554-3.474)	1.398 (1.16-1.635)	1.566 (1.059-2.074)
Increase	0.109 (0.098-0.121)	0.056 (0.046-0.067)	0.097 (0.07-0.123)
Plateau	0.012 (0.01-0.013)	0.011 (0.01-0.011)	0.002 (0.002-0.002)

exponentially increasing performance, followed by a period of gradual increase, then a plateau. We visually identified these portions in the actual learning curve then compared the slopes for each portion (Table 6.13). The Exp region is from 10 to 100 positive instances, the Increase region is from 100 to 2,500, and the Plateau is >2,500. For an approximation method, the most important part of the curve is the Increase region, as the end of that region indicates the start of the Plateau, or point of diminishing return. While the inverse power law method achieves nearly the exact slope as the actual during the Plateau region, the slope of the actual curve during the Increase region is within the confidence interval of the semi-supervised method. This shows that the semi-supervised method has a more accurate slope for the Increase region than the inverse power law method.

Using a semi-supervised method raises the question: is it possible to just train a

semi-supervised classifier rather than labeling more data? We explored this question by evaluating the performance of a classifier on a large set of data versus the same set of data with pseudo-labels created from a model on a small set. We repeated both experiments 25 times using the Splice dataset with 2,500 positive (795,216 total) instances and using 50 positive (15,904 total) for pseudo-labeling. The classifier from the actual data achieved an AUC of 0.843, while the semi-supervised classifier achieved an AUC of 0.544 ($p < 0.001$). This shows that the pseudo-labels are indeed noisy and not an accurate representation of the true labels. As shown by our experiment, however, the behavior of a learning curve on the pseudo-labeled data can be effectively used for sample size planning.

6.3.5 Section Summary

In this section, we evaluated two learning curve approximation methods for large imbalanced biomedical datasets in the context of sample size planning. These methods provide guidance for future machine learning problems that require expensive human-labeling of instances. A small number of labeled instances can be used to build a learning curve, and this can be used in conjunction with labeling costs to determine the number of samples that need to be labeled. We applied a traditional inverse power law model as well as a new proposed semi-supervised method through pseudo-labeling. We found that the inverse power law method was accurate for smaller sizes of data, and while the semi-supervised method had a larger absolute error, it was better at detecting convergence than the power law method. There is much opportunity for future work in this line of research, including evaluating these methods on more large, imbalanced datasets with additional classifiers. Additionally, more mathematical and semi-supervised methods for learning curve approximation can be built upon the foundation of those presented in this study.

6.4 CHAPTER SUMMARY

While we are in the era of big data, we still have challenges related to limited data. In this chapter we explored two such scenarios: (1) limited positive samples for binary classification and (2) limited labels for sample size determination. Both sections show that not all the data is necessary for specific tasks. When there is class imbalance (i.e. limited positive samples), many of the negative samples are not needed to build effective classifiers. Additionally, we found for various datasets in Section 6.3 that a small percentage of data was required to achieve classification performance within 1% of the full dataset. We evaluated two methods for learning curve approximation and presented scenarios where they would be useful for sample size determination with limited labels.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we presented data engineering and machine learning approaches to build clinical risk models for melanoma from structured electronic health records. We explored the use of various machine learning algorithms along with advanced methods to handle high-dimensionality and class imbalance. Additionally, we explored limited data scenarios with learning curve approximation. In the following sections, we summarize the conclusions drawn from this research and avenues for future work.

7.1 STRUCTURED CLINICAL DATA

We created the Modernizing Analytics for Melanoma (MAMEL) dataset: a clinically relevant dataset derived from real-world de-identified EHR data. This dataset was used for several studies in this work, and provided thousands of structured data points for millions of patients. It is important to continue to build datasets such as this, as data provided to clinical models must be structured, frequently captured, and relevant as to apply to large populations of patients.

7.2 SENTINEL LYMPH NODE METASTASIS

We studied the performance of several models for predicting sentinel lymph node metastasis in melanoma patients. We used the MAMEL data to build logistic regression, decision tree, and random forest models. These models were compared to a simple benchmark model using tumor thickness as a single predictor of SLN posi-

tivity. The machine learning models were not able to outperform the benchmark as measured by AUC, but were able to achieve higher sensitivity for thin melanomas and higher specificity for thick melanomas. Contributions of this work involve demonstrating that the tumor thickness is the single greatest predictor of SLN status, and that machine learning models can provide guidance for recommending an SLN biopsy for thin melanomas (<1mm). Future work includes using additional machine learning techniques, predicting other types of melanoma metastasis, and deploying the model in a clinical setting.

7.3 MELANOMA RISK

We reviewed existing literature for cancer risk modeling, and described several experiments to build melanoma risk prediction models from the MAMEL data. Our work provides a reference framework for machine learning studies using large, high-dimensional, and imbalanced EHR data. We used a distributed processing infrastructure for collecting and formatting the data as well as a non-distributed infrastructure for machine learning. Then, we achieved statistically similar or better performance using a sampled dataset versus the original data, saving hundreds of dollars in cloud computing costs for model experimentation. The structured-data EHR and cloud-based model training process described herein addressed the shortcomings identified in previous cancer risk modeling studies. *Availability of structured clinical data:* the structured, cloud-based EHR system provided consistently collected data points across millions of patients at different practices. *Old data:* the data consistency allowed for rapid querying, de-identification and transformation of data to use for training machine learning models. The time between the end of the study period (December 2017) and study completion was about one year. *Advanced modeling methods:* we used a familiar and feature-rich machine learning framework (scikit-learn) with advanced machine learning techniques such as random forest, XGBoost,

data sampling, and feature selection.

Future studies should aim to validate the data infrastructure choices on other clinical datasets and improve the accuracy of the melanoma risk models. More advanced algorithms such as artificial neural networks can be explored to take advantage of the longitudinal data available in EHR systems. While this current study does not utilize image data, future research should consider combining both EHR and image data to provide the best risk models for dermatology patients.

7.4 LEARNING FROM LIMITED DATA

We explored the problem of limited data from two perspectives: (1) limited positive samples for melanoma risk prediction (2) limited labels for sample size determination with four biomedical datasets. When there are limited positive samples, many of the negative samples are not needed to build effective classifiers. We also found that only a small percentage of data across all four datasets was required to achieve classification performance within 1% of the full dataset. We applied a traditional inverse power law model as well as a new proposed semi-supervised method for approximating learning curves with limited labeled data. We found that the inverse power law method was accurate for smaller sizes of data, and while the semi-supervised method had a larger absolute error, it was better at detecting convergence than the power law method. There is much opportunity for future work in this line of research, including evaluating these methods on more large, imbalanced datasets with additional classifiers. Additionally, more mathematical and semi-supervised methods for learning curve approximation can be built upon the foundation of those presented in this work.

BIBLIOGRAPHY

- [1] Cancers That Develop in Young Adults.
- [2] CDC - National Program of Cancer Registries (NPCR).
- [3] Learning with Less Labels (LwLL) - HR001118s0044 (Archived) - Federal Business Opportunities: Opportunities.
- [4] Melanoma - SkinCancer.org.
- [5] Health insurance portability and accountability act of 1996, 1996.
- [6] 21st century cures act, 2016.
- [7] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, Inc., New York, NY, USA, 1989.
- [8] A. Agarwal, O. Chapelle, M. Dudk, and J. Langford. A Reliable Effective Terascale Linear Learning System. *Journal of Machine Learning Research*, 15:1111–1133, 2014.
- [9] L. Ahmad, A. Eshlaghy, M. Ebrahimi, and A. Razavi. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, 04(02), 2013.
- [10] L. Ahmadian, M. van Engen-Verheul, F. Bakhshi-Raiez, N. Peek, R. Cornet, and N. F. de Keizer. The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey. *International Journal of Medical Informatics*, 80(2):81–93, Feb. 2011.
- [11] J. AK. Meaningful use of electronic health records: The road ahead. *JAMA*, 304(15):1709–1710, 2010.
- [12] American Cancer Society. Cancer surveillance programs and registries in the united states.
- [13] American Cancer Society. Cancer Facts & Figures 2019, 2019.
- [14] A. M. Association. *Current procedural terminology: CPT*. American Medical Association, 2007.

- [15] A.-M. Audet, D. Squires, and M. M. Doty. Where Are We on the Diffusion Curve? Trends and Drivers of Primary Care Physicians' Use of Health Information Technology. *Health Services Research*, 49(1pt2):347–360, Feb. 2014.
- [16] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah. Improving palliative care with deep learning. *arXiv preprint arXiv:1711.06402*, 2017.
- [17] J. Bacardit, P. Widera, A. Marquez-Chamorro, F. Divina, J. S. Aguilar-Ruiz, and N. Krasnogor. Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. *Bioinformatics*, 28(19):2441–2448, Oct. 2012.
- [18] L. Bakos, S. Mastroeni, R. R. Bonamigo, F. Melchi, P. Pasquini, C. Fortes, L. Bakos, S. Mastroeni, R. R. Bonamigo, F. Melchi, P. Pasquini, and C. Fortes. A melanoma risk score in a Brazilian population. *Anais Brasileiros de Dermatologia*, 88(2):226–232, Apr. 2013.
- [19] V. P. Balachandran, M. Gonen, J. J. Smith, and R. P. DeMatteo. Nomograms in oncology: more than meets the eye. *The Lancet Oncology*, 16(4):e173–e180, 2015.
- [20] C. M. Balch, J. E. Gershenwald, S.-j. Soong, J. F. Thompson, M. B. Atkins, D. R. Byrd, A. C. Buzaid, A. J. Cochran, D. G. Coit, S. Ding, et al. Final version of 2009 ajcc melanoma staging and classification. *Journal of clinical oncology*, 27(36):6199–6206, 2009.
- [21] A. Baten, B. Chang, S. Halgamuge, and J. Li. Splice site identification using probabilistic parameters and SVM classification. *BMC Bioinformatics*, 7(S5), 2006.
- [22] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin. An Empirical Study on Class Rarity in Big Data. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 785–790, Orlando, FL, Dec. 2018. IEEE.
- [23] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland. Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 784–790, San Jose, CA, USA, Nov. 2016. IEEE.
- [24] M. Bayati, S. Bhaskar, and A. Montanari. A Low-Cost Method for Multiple Disease Prediction. In *AMIA Annual Symposium Proceedings*, volume 2015, page 329. American Medical Informatics Association, 2015.
- [25] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, Feb. 2008.

- [26] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [27] L. Breiman. *Classification and regression trees*. Routledge, 2017.
- [28] H. B. Burke. Outcome Prediction and the Future of the TNM Staging System. *JNCI Journal of the National Cancer Institute*, 96(19):1408–1409, Oct. 2004.
- [29] O. Cahlon, M. F. Brennan, X. Jia, L.-X. Qin, S. Singer, and K. M. Alektiar. A Postoperative Nomogram for Local Recurrence Risk in Extremity Soft Tissue Sarcomas After Limb-Sparing Surgery Without Adjuvant Radiation. *Annals of Surgery*, 255(2):343–347, Feb. 2012.
- [30] S. Caini, M. Boniol, E. Botteri, G. Tosti, B. Bazolli, W. Russell-Edu, F. Giusti, A. Testori, and S. Gandini. The risk of developing a second primary cancer in melanoma patients: A comprehensive review of the literature and meta-analysis. *Journal of Dermatological Science*, 75(1):3–9, 2014.
- [31] T. V. Cartee, S. P. Kini, and S. C. Chen. Melanoma reporting to central cancer registries by US dermatologists: An analysis of the persistent knowledge and practice gap. *Journal of the American Academy of Dermatology*, 65(5):S124.e1–S124.e9, Nov. 2011.
- [32] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [34] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.
- [35] S. H. Cheng, C.-F. Horng, J. L. Clarke, M.-H. Tsou, S. Y. Tsai, C.-M. Chen, J. J. Jian, M.-C. Liu, M. West, A. T. Huang, and L. R. Prosnitz. Prognostic index score and clinical prediction model of local regional recurrence after mastectomy in breast cancer patients. *International Journal of Radiation Oncology*Biology*Physics*, 64(5):1401–1409, Apr. 2006.
- [36] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and N. D. Filipovic. Prediction models for estimation of survival rate and relapse for breast cancer patients. In *Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference on*, pages 1–6. IEEE, 2015.
- [37] S. Colantonio, M. B. Bracken, and J. Beecker. The association of indoor tanning and melanoma in adults: Systematic review and meta-analysis. *Journal of the American Academy of Dermatology*, 70(5):847–857.e18, 2014.

- [38] F. Costa Svedman, D. Pillas, M. Kaur, R. Linder, J. Hansson, and T. Aiki. Stage-specific survival and recurrence in patients with cutaneous malignant melanoma in Europe – a systematic review of the literature. *Clinical Epidemiology*, page 109, 2016.
- [39] D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.
- [40] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1), Dec. 2015.
- [41] K. D. Cromwell, M. I. Ross, Y. Xing, J. E. Gershenwald, R. E. Royal, A. Lucci, J. E. Lee, and J. N. Cormier. Variability in melanoma post-treatment surveillance practices by country and physician specialty: a systematic review. *Melanoma Research*, 22(5):376–385, 2012.
- [42] C. de Waure, G. Quaranta, M. Gualano, C. Cadeddu, A. Jovic-Vranes, B. Djikanovic, G. La Torre, and W. Ricciardi. Systematic review of studies investigating the association between dietary habits and cutaneous malignant melanoma. *Public Health*, 129(8):1099–1113, 2015.
- [43] S. del Ro, J. M. Bentez, and F. Herrera. Analysis of data preprocessing increasing the oversampling ratio for extremely imbalanced big data classification. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, volume 2, pages 180–185. IEEE, 2015.
- [44] D. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano. Random forest: A reliable tool for patient response prediction. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pages 289–296. IEEE, 2011.
- [45] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and J. Van Hulse. Feature selection algorithms for mining high dimensional dna microarray data. In *Handbook of Data Intensive Computing*, pages 685–710. Springer New York, 2011.
- [46] S. Doan, M. Conway, T. M. Phuong, and L. Ohno-Machado. Natural language processing in biomedicine: a unified system architecture overview. *Clinical Bioinformatics*, pages 275–294, 2014.
- [47] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, Oct. 2006.
- [48] S. Edge, D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, and A. Trotti. *AJCC Cancer Staging Manual*. Springer-Verlag New York, 7 edition, 2010.
- [49] H. B. El-Serag, F. Kanwal, J. A. Davila, J. Kramer, and P. Richardson. A New Laboratory-Based Algorithm to Predict Development of Hepatocellular

- Carcinoma in Patients With Hepatitis C and Cirrhosis. *Gastroenterology*, 146(5):1249–1255.e1, May 2014.
- [50] B. W. Eom, J. Joo, S. Kim, A. Shin, H.-R. Yang, J. Park, I. J. Choi, Y.-W. Kim, J. Kim, and B.-H. Nam. Prediction Model for Gastric Cancer Incidence in Korean Population. *PLOS ONE*, 10(7):e0132613, July 2015.
 - [51] E. Erdei and S. M. Torres. A new understanding in the epidemiology of melanoma. *Expert Review of Anticancer Therapy*, 10(11):1811–1823, 2010.
 - [52] T. R. Fears, D. Guerry, R. M. Pfeiffer, R. W. Sagebiel, D. E. Elder, A. Halpern, E. A. Holly, P. Hartge, and M. A. Tucker. Identifying Individuals at High Risk of Melanoma: A Practical Predictor of Absolute Risk. *Journal of Clinical Oncology*, 24(22):3590–3596, Aug. 2006.
 - [53] A. Fernndez, S. del Ro, N. V. Chawla, and F. Herrera. An insight into imbalanced Big Data classification: outcomes and challenges. *Complex & Intelligent Systems*, 3(2):105–120, June 2017.
 - [54] D. Feskanich, W. C. Willett, D. J. Hunter, and G. A. Colditz. Dietary intakes of vitamins A, C, and E and risk of melanoma in two cohorts of women. *British Journal of Cancer*, 88(9):1381–1387, May 2003.
 - [55] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo. Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1), Dec. 2012.
 - [56] C. Fortes, S. Mastroeni, L. Bakos, G. Antonelli, L. Alessandroni, M. A. Pilla, M. Alotto, A. Zappal, T. Manoorannparampill, R. Bonamigo, P. Pasquini, and F. Melchi. Identifying individuals at high risk of melanoma: a simple tool. *European Journal of Cancer Prevention*, 19(5):393–400, Sept. 2010.
 - [57] M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis. The Nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, 22(3):207–219, 1992.
 - [58] A. Gelman. Analysis of variance: Why it is more important than ever. *The Annals of Statistics*, 33(1):1–31, 2005.
 - [59] A. Goldhirsch, J. N. Ingle, R. D. Gelber, A. S. Coates, B. Thurlimann, H.-J. Senn, and Panel members. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2009. *Annals of Oncology*, 20(8):1319–1329, Aug. 2009.
 - [60] B. A. Goldstein, A. M. Navar, and M. J. Pencina. Risk Prediction With Electronic Health Records. *JAMA cardiology*, 1(9):976–977, Dec. 2016.
 - [61] K. Hajian-Tilaki. Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of Biomedical Informatics*, 48:193–204, Apr. 2014.

- [62] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
- [63] J.-F. Hamel, M. Pe, C. Coens, F. Martinelli, A. M. Eggermont, Y. Brandberg, and A. Bottomley. A systematic review examining factors influencing health related quality of life among melanoma cancer survivors. *European Journal of Cancer*, 69:189–198, Dec. 2016.
- [64] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder. Big Data fraud detection using multiple medicare data sources. *Journal of Big Data*, 5(1), Dec. 2018.
- [65] M. Herland, T. M. Khoshgoftaar, and R. Wald. A review of data mining using big data in health informatics. *Journal of Big Data*, 1(1):1–35, 2014.
- [66] A. N. Houghton and D. Polsky. Focus on melanoma. *Cancer cell*, 2(4):275–278, 2002.
- [67] International Bladder Cancer Nomogram Consortium. Postoperative Nomogram Predicting Risk of Recurrence After Radical Cystectomy for Bladder Cancer. *Journal of Clinical Oncology*, 24(24):3967–3972, Aug. 2006.
- [68] J. M. Jerez-Aragons, J. A. Gmez-Ruiz, G. Ramos-Jimnez, J. Muoz-Prez, and E. Alba-Conejo. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial intelligence in medicine*, 27(1):45–63, 2003.
- [69] C. Jhappan, F. P. Noonan, and G. Merlino. Ultraviolet radiation and cutaneous malignant melanoma. *Oncogene*, 22(20):3099–3112, May 2003.
- [70] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), Dec. 2019.
- [71] E. Jones, T. Oliphant, and P. Peterson. SciPy: open source scientific tools for Python, 2014.
- [72] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia. *Learning Spark: Lightning-Fast Big Data Analysis.* ” O'Reilly Media, Inc.”, 2015.
- [73] A. Karpathy. Software 2.0, Nov. 2017.
- [74] T. M. Khoshgoftaar and E. B. Allen. Logistic regression modeling of software quality. *International Journal of Reliability, Quality and Safety Engineering*, 6(04):303–317, 1999.
- [75] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse. An Empirical Study of Learning from Imbalanced Data Using Random Forest. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, pages 310–317, Patras, Greece, Oct. 2007. IEEE.

- [76] T. M. Khoshgoftaar, C. Seiffert, J. V. Hulse, A. Napolitano, and A. Folleco. Learning with limited minority class data. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 348–353, Cincinnati, OH, USA, Dec. 2007. IEEE.
- [77] T. M. Khoshgoftaar and N. Seliya. Fault prediction modeling for software quality estimation: Comparing commonly used techniques. *Empirical Software Engineering*, 8(3):255–283, 2003.
- [78] B. J. Kim, S.-w. Chung, and others. Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of breast cancer*, 2012.
- [79] D. G. Kleinbaum and M. Klein. Competing risks survival analysis. *Survival Analysis: A self-learning text*, pages 391–461, 2005.
- [80] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.
- [81] C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50:S82–S101, 2012.
- [82] W. W. LaMorte. Prospective versus Retrospective Studies, May 2016.
- [83] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1), Dec. 2015.
- [84] C. Lee, F. Collichio, D. Ollila, and S. Moschos. Historical review of melanoma treatment and outcomes. *Clinics in Dermatology*, 31(2):141–147, 2013.
- [85] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1), Dec. 2018.
- [86] U. Leiter, F. Meier, B. Schittekk, and C. Garbe. The natural course of cutaneous melanoma. *Journal of Surgical Oncology*, 86(4):172–178, July 2004.
- [87] G. Lematre, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.
- [88] S. Li, K.-D. Yu, L. Fan, Y.-F. Hou, and Z.-M. Shao. Predicting Breast Cancer Recurrence Following Breast-Conserving Therapy: A Single-Institution Analysis Consisting of 764 Chinese Breast Cancer Cases. *Annals of Surgical Oncology*, 18(9):2492–2499, Sept. 2011.

- [89] J.-D. Liang, X.-O. Ping, Y.-J. Tseng, G.-T. Huang, F. Lai, and P.-M. Yang. Recurrence predictive models for patients with hepatocellular carcinoma after radiofrequency ablation using support vector machines with feature selection methods. *Computer Methods and Programs in Biomedicine*, 117(3):425–434, Dec. 2014.
- [90] LOINC. LOINC: The freely available standard for identifying health measurements, observations, and documents.
- [91] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]*, Feb. 2018. arXiv: 1802.03888.
- [92] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [93] S. K. Lwanga, S. Lemeshow, W. H. Organization, et al. *Sample size determination in health studies: a practical manual*. Geneva: World Health Organization, 1991.
- [94] J. M. Madden, M. D. Lakoma, D. Rusinak, C. Y. Lu, and S. B. Soumerai. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *Journal of the American Medical Informatics Association*, page ocw021, Apr. 2016.
- [95] M. Madu, M. Wouters, and A. van Akkooi. Sentinel node biopsy in melanoma: Current controversies addressed. *European Journal of Surgical Oncology (EJSO)*, 43(3):517–533, Mar. 2017.
- [96] I. Maglogiannis and C. Doukas. Overview of Advanced Computer Vision Systems for Skin Lesions Characterization. *IEEE Transactions on Information Technology in Biomedicine*, 13(5):721–733, Sept. 2009.
- [97] D. Marrelli, A. De Stefano, G. de Manzoni, P. Morgagni, A. Di Leo, and F. Roviello. Prediction of Recurrence After Radical Surgery for Gastric Cancer: A Scoring System Obtained From a Prospective Multicenter Study. *Annals of Surgery*, 241(2):247–255, Feb. 2005.
- [98] T. Martnez-Menchn, P. Snchez-Pedre, J. Martnez-Escribano, R. Corbaln-Vlez, and E. Martnez-Barba. Evaluacin del coste econmico de la tcnica de la biopsia selectiva del ganglio centinela en melanoma. *Actas Dermo-Sifiliogrificas*, 106(3):201–207, 2015.
- [99] E. Meldolesi, J. van Soest, A. Damiani, A. Dekker, A. R. Alitto, M. Campitelli, N. Dinapoli, R. Gatta, M. A. Gambacorta, V. Lanzotti, P. Lambin, and

- V. Valentini. Standardized data collection to build prediction models in oncology: a prototype for rectal cancer. *Future Oncology*, 12(1):119–136, Jan. 2016.
- [100] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. MLlib: Machine Learning in Apache Spark. *J. Mach. Learn. Res.*, 17(1):1235–1241, Jan. 2016.
- [101] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1):213, 2009.
- [102] T. K. Moon. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60, 1996.
- [103] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov. Estimating Dataset Size Requirements for Classifying DNA Microarray Data. *Journal of Computational Biology*, 10(2):119–142, Apr. 2003.
- [104] V. K. Nahar, M. Allison Ford, R. T. Brodell, J. F. Boyas, S. K. Jacks, R. Biviji-Sharma, M. A. Haskins, and M. A. Bass. Skin cancer prevention practices among malignant melanoma survivors: a systematic review. *Journal of Cancer Research and Clinical Oncology*, 142(6):1273–1283, 2016.
- [105] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), Dec. 2015.
- [106] H. Nathan and T. M. Pawlik. Limitations of claims and registry data in surgical oncology research. *Annals of Surgical Oncology*, 15(2):415–423, 2008.
- [107] National Collaborating Centre for Cancer (Great Britain). *Early and locally advanced breast cancer diagnosis and treatment: full guideline*. National Collaborating Centre for Cancer, Cardiff, 2009.
- [108] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448, July 2011.
- [109] NIH. Cancer statistics, 2018.
- [110] W. H. Organization and others. International classification of diseases (ICD), 2012.

- [111] S. Park, B.-H. Nam, H.-R. Yang, J. A. Lee, H. Lim, J. T. Han, I. S. Park, H.-R. Shin, and J. S. Lee. Individualized Risk Prediction Model for Lung Cancer in Korean Men. *PLoS ONE*, 8(2):e54823, Feb. 2013.
- [112] G. Par, L. Raymond, A. O. d. Guinea, P. Poba-Nzaou, M.-C. Trudel, J. Marsan, and T. Micheneau. Electronic health record usage behaviors in primary care medical practices: A survey of family physicians in Canada. *International Journal of Medical Informatics*, 84(10):857–867, Oct. 2015.
- [113] S. Pasquali, S. Mocellin, L. Campana, A. Vecchiato, E. Bonandini, M. Montesco, S. Santarcangelo, G. Zavagno, D. Nitti, and C. Rossi. Maximizing the clinical usefulness of a nomogram to select patients candidate to sentinel node biopsy for cutaneous melanoma. *European Journal of Surgical Oncology (EJSO)*, 37(8):675–680, 2011.
- [114] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-sos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [115] A. Piero, M. Canteras, E. Ortiz, E. Martnez-Barba, and P. Parrilla. Validation of a Nomogram to Predict the Presence of Sentinel Lymph Node Metastases in Melanoma. *Annals of Surgical Oncology*, 15(10):2874–2877, 2008.
- [116] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.
- [117] F. Provost, D. Jensen, and T. Oates. Efficient progressive sampling. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, pages 23–32, San Diego, California, United States, 1999. ACM Press.
- [118] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman. Impact of feature selection techniques for tweet sentiment classification. In *Proceedings of the 28th International FLAIRS conference*, pages 299–304, May 2015.
- [119] J. R. Quinlan. Improved use of continuous attributes in C4. 5. *Journal of artificial intelligence research*, pages 77–90, 1996.
- [120] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
- [121] M. Radespiel-Trger, W. Hohenberger, and B. Reingruber. Improved prediction of recurrence after curative resection of colon carcinoma using tree-based risk stratification: Recurrence Prediction in Colon Ca. *Cancer*, 100(5):958–967, Mar. 2004.

- [122] K. O. Raji, L. Payne, and S. C. Chen. Reporting Melanoma: A Nationwide Surveillance of State Cancer Registries. *Journal of Skin Cancer*, 2015:1–5, 2015.
- [123] A. R. Razavi, H. Gill, H. Ahlfeldt, and N. Shahsavar. Canonical correlation analysis for data reduction in data mining applied to predictive models for breast cancer recurrence. *Studies in health technology and informatics*, 116:175, 2005.
- [124] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.
- [125] A. N. Richter, M. Crawford, B. Heredia, and T. M. Khoshgoftaar. Efficient Modeling of User-Entity Preference in Big Social Networks. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 982–988, Vietri sul Mare, Italy, Nov. 2015. IEEE.
- [126] A. N. Richter and T. M. Khoshgoftaar. Predicting Cancer Relapse with Clinical Data: A Survey of Current Techniques. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 369–376, Pittsburgh, PA, USA, July 2016. IEEE.
- [127] A. N. Richter and T. M. Khoshgoftaar. Modernizing Analytics for Melanoma with a Large-Scale Research Dataset. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 551–558, San Diego, CA, Aug. 2017. IEEE.
- [128] A. N. Richter and T. M. Khoshgoftaar. Predicting sentinel node status in melanoma from a real-world EHR dataset. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1872–1878, Kansas City, MO, Nov. 2017. IEEE.
- [129] A. N. Richter and T. M. Khoshgoftaar. Building and Interpreting Risk Models from Imbalanced Clinical Data. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 143–150, Volos, Nov. 2018. IEEE.
- [130] A. N. Richter and T. M. Khoshgoftaar. Melanoma Risk Prediction with Structured Electronic Health Records. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*, pages 194–199, Washington, DC, USA, 2018. ACM Press.
- [131] A. N. Richter and T. M. Khoshgoftaar. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine*, 90:1–14, Aug. 2018.
- [132] A. N. Richter and T. M. Khoshgoftaar. Approximating learning curves for imbalanced big data with limited labels. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2019. Under review.

- [133] A. N. Richter and T. M. Khoshgoftaar. Efficient learning from big data for cancer risk modeling: A case study with melanoma. *Computers in Biology and Medicine*, 110:29–39, July 2019.
- [134] A. N. Richter and T. M. Khoshgoftaar. Learning curve estimation with large datasets. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2019. Under review.
- [135] A. N. Richter and T. M. Khoshgoftaar. Melanoma risk modeling from limited positive samples. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 8(1), Dec. 2019.
- [136] A. N. Richter and T. M. Khoshgoftaar. Sample size determination for biomedical big data with limited labels. 2019. Under submission.
- [137] A. N. Richter, T. M. Khoshgoftaar, S. Landset, and T. Hasanin. A Multi-dimensional Comparison of Toolkits for Machine Learning with Big Data. In *2015 IEEE International Conference on Information Reuse and Integration*, pages 1–8, San Francisco, CA, USA, Aug. 2015. IEEE.
- [138] M. Rota, E. Pasquali, R. Bellocchio, V. Bagnardi, L. Scotti, F. Islami, E. Negri, P. Boffetta, C. Pelucchi, G. Corrao, and C. La Vecchia. Alcohol drinking and cutaneous melanoma risk: a systematic review and dose-risk meta-analysis. *British Journal of Dermatology*, 170(5):1021–1028, 2014.
- [139] U. Rudloff, L. M. Jacks, J. I. Goldberg, C. A. Wynveen, E. Brogi, S. Patil, and K. J. Van Zee. Nomogram for Predicting the Risk of Local Recurrence After Breast-Conserving Surgery for Ductal Carcinoma In Situ. *Journal of Clinical Oncology*, 28(23):3762–3769, Aug. 2010.
- [140] A. Saabas. Treeinterpreter. <https://github.com/andosa/treeinterpreter>, 2015.
- [141] S. Sam. Learning with Limited Labeled Data, Mar. 2019.
- [142] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano. Mining data with rare events: A case study. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*, ICTAI ’07, pages 132–139, Washington, DC, USA, 2007. IEEE Computer Society.
- [143] N. Seliya, T. M. Khoshgoftaar, and J. V. Hulse. A Study on the Relationships of Classifier Performance Metrics. In *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, pages 59–66, Newark, New Jersey, USA, Nov. 2009. IEEE.
- [144] B. Settles. Active Learning Literature Survey. *University of Wisconsin-Madison Department of Computer Sciences*, page 47, 2009.

- [145] A. Shin, J. Joo, H.-R. Yang, J. Bak, Y. Park, J. Kim, J. H. Oh, and B.-H. Nam. Risk Prediction Model for Colorectal Cancer: National Health Insurance Corporation Study, Korea. *PLoS ONE*, 9(2):e88079, Feb. 2014.
- [146] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, Mar. 2014.
- [147] A. G. Singal, A. Mukherjee, B. J. Elmunzer, P. D. Higgins, A. S. Lok, J. Zhu, J. A. Marrero, and A. K. Waljee. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *The American journal of gastroenterology*, 108(11):1723–1730, 2013.
- [148] SNOMED International. SNOMED CT: The global language of healthcare.
- [149] S. Sonnenburg and V. Franc. COFFIN: A computational framework for linear SVMs. In *ICML*, pages 999–1006, 2010.
- [150] I. Spasi, J. Livsey, J. A. Keane, and G. Nenadi. Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics*, 83(9):605–623, Sept. 2014.
- [151] S. D. Stellman, T. Takezaki, L. Wang, Y. Chen, M. L. Citron, M. V. Djordjevic, S. Harlap, J. E. Muscat, A. I. Neugut, E. L. Wynder, and others. Smoking and lung cancer risk in American and Japanese men: an international case-control study. *Cancer Epidemiology Biomarkers & Prevention*, 10(11):1193–1199, 2001.
- [152] E. W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health. Springer-Verlag New York, 1 edition, 2009.
- [153] E. W. Steyerberg, T. van der Ploeg, and B. Van Calster. Risk prediction with machine learning and regression methods: Risk prediction with machine learning and regression methods. *Biometrical Journal*, 56(4):601–606, July 2014.
- [154] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009.
- [155] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, Venice, Oct. 2017. IEEE.
- [156] W. T, T. E, and M. T. Screening for skin cancer: An update of the evidence for the u.s. preventive services task force. *Annals of Internal Medicine*, 150(3):194–198, 2009.

- [157] I. Triguero, S. del Ro, V. Lpez, J. Bacardit, J. M. Bentez, and F. Herrera. ROSEFW-RF: The winner algorithm for the ECBDL14 big data competition: An extremely imbalanced big data bioinformatics problem. *Knowledge-Based Systems*, 87:69–79, Oct. 2015.
- [158] I. Triguero, S. del Ro, V. Lpez, J. Bacardit, J. M. Bentez, and F. Herrera. ROSEFW-RF: The winner algorithm for the ECBDL’14 big data competition: An extremely imbalanced big data bioinformatics problem. *Knowledge-Based Systems*, 87:69 – 79, 2015. Computational Intelligence Applications for Data Science.
- [159] C.-J. Tseng, C.-J. Lu, C.-C. Chang, and G.-D. Chen. Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Computing and Applications*, 24(6):1311–1316, May 2014.
- [160] J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.
- [161] F. Turati, C. Galeone, M. Rota, C. Pelucchi, E. Negri, V. Bagnardi, G. Corrao, P. Boffetta, and C. La Vecchia. Alcohol and liver cancer: a systematic review and meta-analysis of prospective studies. *Annals of Oncology*, 25(8):1526–1535, Aug. 2014.
- [162] U.S. Department of Health and Human Services. Methods for de-identification of PHI.
- [163] J. A. Usher-Smith, J. Emery, A. P. Kassianos, and F. M. Walter. Risk Prediction Models for Melanoma: A Systematic Review. *Cancer Epidemiology Biomarkers & Prevention*, 23(8):1450–1463, Aug. 2014.
- [164] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, 14(1):137, 2014.
- [165] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science and Engg.*, 13(2):22–30, Mar. 2011.
- [166] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942. ACM, 2007.
- [167] J. Van Hulse, A. Napolitano, R. Wald, and T. M. Khoshgoftaar. Feature selection with high-dimensional imbalanced data. In *2009 IEEE International Conference on Data Mining Workshops(ICDMW)*, volume 00, pages 507–514, 12 2009.

- [168] C. Watts, M. Dieng, R. Morton, G. Mann, S. Menzies, and A. Cust. Clinical practice guidelines for identification, screening and follow-up of individuals at high risk of primary cutaneous melanoma: a systematic review. *British Journal of Dermatology*, 172(1):33–47, 2015.
- [169] M. R. Weiser, R. G. Landmann, M. W. Kattan, M. Gonen, J. Shia, J. Chou, P. B. Paty, J. G. Guillem, L. K. Temple, D. Schrag, L. B. Saltz, and W. D. Wong. Individualized Prediction of Colon Cancer Recurrence Using a Nomogram. *Journal of Clinical Oncology*, 26(3):380–385, Jan. 2008.
- [170] B. L. Welch. The generalization of Student’s problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.
- [171] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
- [172] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2019. R package version 0.8.0.1.
- [173] L. H. Williams, A. R. Shors, W. E. Barlow, C. Solomon, and E. White. Identifying Persons at Highest Risk of Melanoma Using Self-Assessed Risk Factors. *Journal of clinical & experimental dermatology research*, 2(6), 2011.
- [174] S. L. Wong, C. M. Balch, P. Hurley, S. S. Agarwala, T. J. Akhurst, A. Cochran, J. N. Cormier, M. Gorman, T. Y. Kim, K. M. McMasters, R. D. Noyes, L. M. Schuchter, M. E. Valsecchi, D. L. Weaver, and G. H. Lyman. Sentinel Lymph Node Biopsy for Melanoma: American Society of Clinical Oncology and Society of Surgical Oncology Joint Clinical Practice Guideline. *Journal of Clinical Oncology*, 30(23):2912–2918, Aug. 2012.
- [175] S. L. Wong, M. W. Kattan, K. M. McMasters, and D. G. Coit. A Nomogram That Predicts the Presence of Sentinel Node Metastasis in Melanoma With Better Discrimination Than the American Joint Committee on Cancer Staging System. *Annals of Surgical Oncology*, 12(4):282–288, 2005.
- [176] J. F. C. Woods, J. A. De Marchi, A. J. Lowery, and A. D. K. Hill. Validation of a nomogram predicting sentinel lymph node status in melanoma in an Irish population. *Irish Journal of Medical Science (1971 -)*, 184(4):769–773, 2015.
- [177] Y. P. Wu, L. G. Aspinwall, B. M. Conn, T. Stump, B. Grahmann, and S. A. Leachman. A systematic review of interventions to improve adherence to melanoma preventive behaviors for individuals at elevated risk. *Preventive Medicine*, 88:153–167, 2016.
- [178] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection*, pages 117–136. Springer, 1998.

- [179] A. Yu, S. M. Woo, J. Joo, H.-R. Yang, W. J. Lee, S.-J. Park, and B.-H. Nam. Development and Validation of a Prediction Model to Estimate Individual Risk of Pancreatic Cancer. *PLOS ONE*, 11(1):e0146473, Jan. 2016.
- [180] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association, 2012.
- [181] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud’10, pages 10–10, Berkeley, CA, USA, 2010. USENIX Association.
- [182] J. L. Zapas, H. C. Coley, S. L. Beam, S. D. Brown, K. A. Jablonski, and E. G. Elias. The risk of regional lymph node metastases in patients with melanoma less than 1.0 mm thick: recommendations for sentinel lymph node biopsy. *Journal of the American College of Surgeons*, 197(3):403–407, Sept. 2003.
- [183] H. Zhang. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, volume 2, 01 2004.
- [184] S. Zhong, T. M. Khoshgoftaar, and N. Seliya. Clustering-based network intrusion detection. *International Journal of reliability, Quality and safety Engineering*, 14(02):169–187, 2007.
- [185] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.