

2021

## Interpretable Machine Learning Model for Clinical Decision Making

Ali El-Sharif

Nova Southeastern University, elsharifali@hotmail.com

Follow this and additional works at: [https://nsuworks.nova.edu/gscis\\_etd](https://nsuworks.nova.edu/gscis_etd)



Part of the [Computer Sciences Commons](#)

## Share Feedback About This Item

---

### NSUWorks Citation

Ali El-Sharif. 2021. *Interpretable Machine Learning Model for Clinical Decision Making*. Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, College of Computing and Engineering. (1159)

[https://nsuworks.nova.edu/gscis\\_etd/1159](https://nsuworks.nova.edu/gscis_etd/1159).

This Dissertation is brought to you by the College of Computing and Engineering at NSUWorks. It has been accepted for inclusion in CCE Theses and Dissertations by an authorized administrator of NSUWorks. For more information, please contact [nsuworks@nova.edu](mailto:nsuworks@nova.edu).

Interpretable Machine Learning Model for Clinical Decision Making

by

Ali El-Sharif

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

in

Information Systems

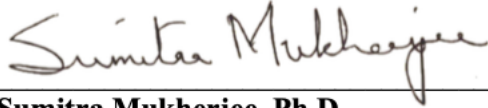
College of Computing and Engineering

Nova Southeastern University

Fort Lauderdale, FL

2021

**We hereby certify that this dissertation, submitted by Ali El Sharif conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.**



**Sumitra Mukherjee, Ph.D.**  
**Chairperson of Dissertation Committee**

11/29/21  
**Date**



**Michael J. Laszlo, Ph.D.**  
**Dissertation Committee Member**

11/29/21  
**Date**



**Frank J. Mitropoulos, Ph.D.**  
**Dissertation Committee Member**

11/29/21  
**Date**

**Approved:**



**Meline Kevorkian, Ed.D.**  
**Dean, College of Computing and Engineering**

11/29/21  
**Date**

**College of Computing and Engineering**  
**Nova Southeastern University**

**2021**

An Abstract of a Dissertation Submitted to Nova Southeastern University  
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

### Interpretable Machine Learning Model for Clinical Decision Making

Despite machine learning models being increasingly used in medical decision-making and meeting classification predictive accuracy standards, they remain untrusted black-boxes due to decision-makers' lack of insight into their complex logic. Therefore, it is necessary to develop interpretable machine learning models that will engender trust in the knowledge they generate and contribute to clinical decision-makers intention to adopt them in the field. The goal of this dissertation was to systematically investigate the applicability of interpretable model-agnostic methods to explain predictions of black-box machine learning models for medical decision-making. As proof of concept, this study addressed the problem of predicting the risk of emergency readmissions within 30 days of being discharged for heart failure patients. Using a benchmark data set, supervised classification models of differing complexity were trained to perform the prediction task. More specifically, Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM) models were constructed using the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database (NRD). The precision, recall, area under the ROC curve for each model were used to measure predictive accuracy. Local Interpretable Model-Agnostic Explanations (LIME) was used to generate explanations from the underlying trained models. LIME explanations were empirically evaluated using explanation stability and local fit ( $R^2$ ).

The results demonstrated that local explanations generated by LIME created better estimates for Decision Trees (DT) classifiers.

## Acknowledgements

First and foremost, I wish to thank my dissertation chair, Dr. Sumitra Mukherjee. I appreciate his advice, guidance, and support throughout the dissertation process.

I thank members of my dissertation committee members, Dr. Laszlo, and Dr. Mitropoulos for their valuable feedback and suggestions.

I am grateful to Dr. Marti Snyder. I appreciate her mentorship and advice throughout the program.

I thank my good friends and fellow NSU alumni, Mutharasu Narayanaperumal, John McConnell, and Richard McCrae, for their support and encouragement.

I wish to recognize the Aggregate Intellect team for offering me the opportunity to engage in a meaningful exchange of research ideas that were instrumental in shaping my dissertation. I am fortunate to have collaborated with Amir Feizpour, Suhas Pai, Xiang Chen, Serg Masis, and Muhammad Rehman Zafar.

I wish to thank my wife Dena and our three amazing children, Moeen, Omar, and Yasmeen, for their love and support.

Lastly, I wish to thank my late father and loving mother. They instilled in me the values of hard work and commitment to education.

## Table of Contents

### Chapters

#### 1. Introduction 1

- Background 1
- Black-Box Models 2
- Machine Learning Interpretability 2
- Intrinsic vs. Post-Hoc Interpretability 3
- Global vs. Local Interpretability 4
- Model-Specific vs. Model-Agnostic Interpretability 5
- Local Interpretable Model-Agnostic Explanations (LIME) 5
- Classification Predictive Accuracy Metrics 5
- Post-Hoc Explanation Quality Metrics 7
- Problem Statement 8
- Dissertation Goal 9
- Research Questions 9
- Relevance and Significance 10
- Related Studies 12
- Summary 13

#### 2. Review of the Literature 14

- Introduction 14
- Cost-Sensitive Learning for an Imbalanced Datasets 14
- Ensemble Methods 14
- Boosting 15
- Bagging 15
- Random Forests (RF) 16
- Decision Trees (DT) 16
- Logistic Regression (LR) 17
- Interpretability Characteristics 18
- Predictive Accuracy vs. Interpretability 22
- Intrinsic vs. Post-hoc Explanations 23
- Model-Specific vs. Model-Agnostic 24
- Model Approximation 28
- Local Interpretable Model-Agnostics Explanations (LIME) 29
- LIME Variants and Alternatives 33
- Summary 35

#### 3. Methodology 36

- Overview 36
- HCUP Dataset 36
- Pre-Process 38
- Predict 50
- Explain & Evaluate 51
- Resources 53

Summary 53

**4. Results 54**

Overview 54  
LIME Explanations 55  
LIME Explanation Instability 58  
Experimental Results 59  
Logistic Regression (LR) 59  
Random Forest (RF) 61  
Decision Tree (DT) 63  
LightGBM (GBM) 65  
Google Cloud Platform (GCP) – Auto-ML 67  
Classifiers Summary Results 68  
Summary 69

**5. Conclusions, Implications, Recommendations, and Summary 71**

Overview 71  
Conclusions 71  
Implications 72  
Recommendations 73  
Summary 73

**Appendices**

A. 2016 NRD Core File Schema 75  
B. 2016 NRD Severity Measures Schema 80  
C. 2016 NRD Hospital File Schema 81  
D. 2016 NRD File Specifications 83  
E. ICD-10 Code Mapping 84

**References 91**

## **List of Tables**

### **Tables**

1. Confusion Matrix 6
2. Average AUC Comparison for Related Studies 13
3. Pre-Processed Features 42
4. Selected Top DRG Categories 44
5. Selected Top APRDRG Categories 44
6. ICD-10 Mapped Medical Conditions 45
7. Top 20 Procedure Codes 46
8. Top 15 External Causes of Morbidity 48



## List of Figures

### Figures

1. Post-hoc Interpretability 4
2. LIME Decision Boundary 29
3. The LIME Process 30
4. Explaining a Prediction with LIME 31
5. LIME Tabular Classifier Explanation 32
6. Dataset Preprocessing 37
7. Study Outline 38
8. Pre-Processed Dataset (prior to feature engineering and encoding) 43
9. Example LIME Explanation for Readmission Prediction 55
10. Example LIME Explanation for Readmission Prediction - Low Confidence 58
11. LIME Instability Example 59

## Chapter 1

### Introduction

#### **Background**

Hospital readmissions refer to unplanned hospitalizations that occur within 30 days of discharge. Jencks et al. (2009) estimated an annual cost of avoidable Medicare readmissions of \$17.4 billion, and 20 percent of Medicare fee-for-service patients had readmissions within 30 days of discharge. These findings established managing preventable readmissions as a goal for policymakers to save costs and improve healthcare quality (Betancourt et al., 2015).

The Patient Protection and Affordable Care Act of 2010 instituted the Hospital Readmissions Reduction Program (HRRP), requiring the Centers for Medicare & Medicaid Services (CMS) to reduce payments to hospitals with excess readmissions starting the fiscal year 2013 for select clinical conditions. A 2016 report to the U.S. Congress noted that HRRP imposed \$420 million in penalties against 78 percent of hospitals (Medicare Payment Advisory Commission, 2016).

Heart failure is among the conditions targeted in the HRRP and is the leading cause of death in the United States. Cardiovascular disease accounted for over 17.3 million deaths in 2013 and is expected to account for over 23.6 million deaths by 2030 (Mozaffarian et al., 2016). Additionally, heart failure is the leading cause of hospital admissions and readmissions in the United States among patients over the age of 65 (Arundel et al., 2016; Joynt & Jha, 2011).

The increasing availability of electronic patient data presents opportunities to leverage machine learning (ML) methods to predict patients at high risk of readmission

and consequently aid clinical decision-making (Bayati et al., 2014). While not all readmissions are preventable, a high readmission rate has been established as an indicator of poor quality of care (Goldfield et al., 2008). Additionally, the active management of discharged patients has been established to have a significant bearing on outcomes (Verhaegh et al., 2014). Several studies have reported successful readmission reduction outcomes by allocating limited post-discharge resources such as scheduled outpatient visits and the application of telehealth and patient education (Bayati et al., 2014; Messina, 2016; Ross et al., 2009). Therefore, accurately and reliably predicting patients at high risk of readmissions is useful to healthcare practitioners to identify patients targeted for post-discharge intervention.

### **Black-Box Models**

Black-box models refer to automated decision systems that map user features into a decision class without exposing how and why they arrive at a particular decision (Montavon et al., 2017; Pedreschi et al., 2019). The internals of black-box models are either unknown or not clearly understood by humans (Carvalho et al., 2019; Guidotti et al., 2018). The terms black box, grey box, and white box refer to the level of exposure of the internal logic to the system user (Adadi & Berrada, 2018).

### **Machine Learning Interpretability**

ML interpretability is of paramount importance in high-stakes decision-making to maintain human oversight over black-box models. Although ML interpretability can be intentionally obstructed to protect secrets and maintain a competitive advantage (Burrell, 2016), black-box models' opacity can arise from the distinct difficulty of interpreting classification results leveraging large datasets and achieving accuracy through model

complexity. Interpretability approaches can be classified based on the following attributes: Intrinsic vs. Post-hoc; Global vs. Local; Model-Specific vs. Model-Agnostic (Carvalho et al., 2019).

### **Intrinsic vs. Post-Hoc Interpretability**

Intrinsic interpretability refers to transparent models in which the inner logic is represented by an interpretable model structure (Murdoch et al., 2019). Intrinsic interpretability is partly achieved by constraining model complexity, which can lower predictive accuracy (Du et al., 2019; Murdoch et al., 2019).

Post-hoc interpretability takes a trained model as input and extracts the underlying relationships that the model had learned by querying the model (Murdoch et al. 2019), observing the model's output on a large number of inputs, and constructing a white-box surrogate model (Burkart & Huber, 2020). Post-hoc explanations mimic model distillation (Tan et al., 2018) as they transfer the knowledge from a large, complex model (teacher) into a simpler model (student), representing an explanation of what the model is doing but not how the model is doing it. Although the approximate explanation is not an exact match of what the model is doing, it is close enough to be useful in understanding the model's logic. Post-hoc methods do not place constraints on the underlying model, hence explain the output of the black box model without negatively impacting predictive accuracy (Burkart & Huber, 2020; Du et al., 2019).

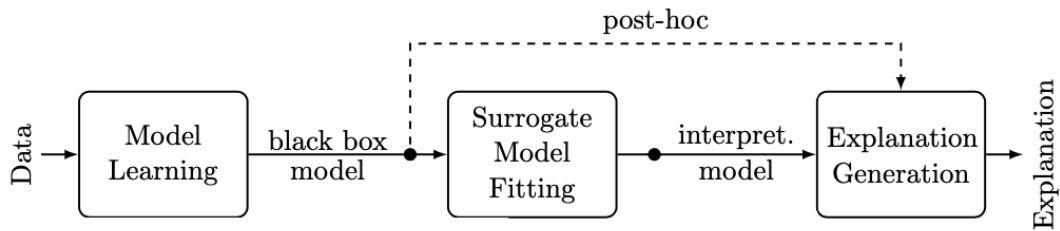


Figure 1. Post-hoc Interpretability

From: “Burkart, N., & Huber, M. F. (2020). A Survey on the Explainability of Supervised Machine Learning. *arXiv preprint arXiv:2011.07876*.”

### Global vs. Local Interpretability

Global interpretability explains the whole logic of a model and the reasoning behind all possible outcomes (Guidotti et al., 2018; Lakkaraju et al., 2019). Global model interpretability explains a model through the most important rules learned from the training data and represents the explanation through the model’s structure and parameters (Du et al., 2019). Examples of global interpretability rules are the coefficients in a linear regression model or rules encoded by a path from the root node to the leaf nodes in a decision tree model. Global model interpretability explains population-level decisions (Yang et al., 2018). However, they are not optimized for individual samples as they provide feature importance that is averaged across the entire input space (Yoon et al., 2018).

Local interpretability explains model characteristics and the impact of input features for a specific prediction (Adadi & Berrada, 2018; Du et al., 2019; Guidotti et al., 2018). Because small sections of the model are more likely to be linear, local models expressed as a linear function of input features can be more accurate than global models (Hall et al., 2017).

### **Model-Specific vs. Model-Agnostic Interpretability**

Model-specific interpretability refers to explanations that are exclusive to the classifier used and derive their explanation by using the internal model representation or learning process (Adadi & Berrada, 2018; Du et al., 2019; Robnik-Šikonja & Bohanec, 2018).

Model-agnostic explanatory methods approximate the behavior of underlying ML models to generate end-user explanations that are independent of the internal logic used to generate predictions (Ribeiro et al., 2016a). Model-agnostic explanations enable the use of black-box models for tasks requiring the high accuracy of black-box models without sacrificing the need for interpretability (Ribeiro et al., 2016a).

### **Local Interpretable Model-Agnostic Explanations (LIME)**

Local interpretable Model-Agnostic Explanations (LIME) is a post-hoc method that generates explanations for any underlying classifier prediction. The LIME explanations are extracted from the underlying model by learning a simpler linear model around the prediction. The LIME linear model is constructed by generating perturbed random samples around the instance and establishing local feature importance representing the primary drivers supporting the prediction. LIME allows the user to generate an explanation budget by pre-defining the number of features used in the explanation (Ribeiro et al., 2016a).

### **Classification Predictive Accuracy Metrics**

The predicted label of a binary classifier falls into one of four categories: true positive (TP), false positive (FP), false negative (FN), or true negative (TN) (Metz, 1978; Fawcett, 2006; Linden, 2006; Sokolova & Lapalme, 2009). A confusion matrix generally represents the frequencies of the classification label across the four key measures. Key

empirical metrics derived from these measures include accuracy, sensitivity (recall), specificity, precision, error rate, F-score, and Area Under the Receiver Operator Characteristic (ROC) Curve (AUC) derived from the Confusion Matrix parameter in Table 1 (Fawcett, 2006; Huang & Ling, 2005; Linden, 2006; Metz, 1978). The study reports precision, recall, area under the ROC curve as the accuracy metrics.

<b>Confusion Matrix</b>	<b>Predicted Condition Positive</b>	<b>Predicted Condition Negative</b>
<b>Actual Condition Positive</b>	TP	FN
<b>Actual Condition Negative</b>	FP	TN

*Table 1. Confusion Matrix*

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

$$Sensitivity/Recall/TP Rate = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Fscore = \frac{2}{\left(\frac{1}{Precision}\right) + \left(\frac{1}{Recall}\right)}$$

$$TN Rate = \frac{TN}{TN + FP}$$

$$FP Rate = \frac{FP}{TN + FP}$$

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

$$Error Rate = \frac{FP + FN}{TP + TN + FP + FN}$$

### Post-Hoc Explanation Quality Metrics

The following empirical metrics have been identified in the literature to empirically evaluate the quality of post-hoc local explanatory models (Islam et al., 2019; Robnik-Šikonja & Bohanec, 2018; Shankaranarayana & Runje, 2019; Yoon et al., 2019):

#### Explanation Stability

The random perturbation used by LIME introduces the risk that the local model may generate a different explanation for the same instance when the sampling process is repeated multiple times (Visani et al., 2021; Zafar & Khan, 2019). As explanations in LIME are expressed in terms of input features, a stable LIME explanation would consistently select the same input features for the same instance over multiple iterations as defined and experimentally demonstrated by (Zafar & Khan, 2019) using the average Jaccard similarity distance for a fixed number of iterations. The Jaccard coefficient is

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Where  $S_1, S_2$  are two explanation sets

$J(S_1, S_2)$  ranges from 0 to 1; 0 means the sets are identical; 1 means the sets are highly dissimilar. The value closer to 0 means the explanations are less unstable

The Jaccard similarity distance is:

$$J_{distance} = 1 - J(S_1, S_2)$$



### Local fit ( $R^2$ )

Also known as the coefficient of determination:

$$R^2 = 1 - \frac{SSE}{SST}$$

Where:

- $SSE$  (Sum of Squares)
- $SST$  (Sum of Squared Total)

$R^2$  measure ranges from 0 to 1; the closer to 1, the better the fit.

### **Problem Statement**

Clinical decision-making is evidence-based, probabilistic, fraught with uncertainty, and needs to balance conflicting decision criteria (Broekhuizen et al., 2015). While ML algorithms can improve decision-making and provide insight, their use introduces added uncertainty due to their inherent complexity and lack of interpretability (Choi et al., 2016). Managing the uncertainty introduced by ML models is necessary to assure healthcare practitioners that their adoption will yield better decisions and can be trusted (Ahmad et al., 2018). Numerous studies have demonstrated the improved accuracy metrics of ML methods in predicting the risk of unplanned hospital readmissions to support clinical decision-making. However, these studies were limited to a small non-publicly available dataset (Bayati et al., 2014) or offered global interpretable risk factors (Yang et al., 2016). Considering the increased penetration of ML models in medical decision making, it is necessary to develop interpretable ML models that will engender trust in the knowledge they generate and contribute to individual clinical decision-makers intention to adopt them in the field (Biran & Cotton, 2017; Burkart & Huber, 2020; Guidotti et al., 2018; Holzinger et al. 2017). There are no known

readmission studies that applied ML methods on large publicly available datasets using post-hoc model-agnostic local ML interpretability techniques.

### **Dissertation Goal**

The goal of this dissertation was to systematically investigate the applicability of local model-agnostic methods to explain the predictions of complex ML models used for medical decision-making. As a proof of concept, the dissertation addressed the binary classification problem of predicting the risk of emergency readmission within 30 days of discharge for heart failure patients based on the information available at the time of discharge.

Using a benchmark dataset, supervised classification models of differing complexity were trained to perform the prediction task. Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM) models were constructed using the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database (NRD). The precision, recall, F1-score, area under the ROC curve for each model were used to measure predictive accuracy. Local Interpretable Model-Agnostic Explanations (LIME) was used to interpret the predictive features of each trained model. Explanation stability and local fit ( $R^2$ ) were used to measure the quality of the explanation generated by LIME.

### **Research Questions**

The following research questions guided the study:

*Research Question 1 (RQ1): Can the ML predictions generate intelligible results to guide clinical decision-making?*

*Research Question 2 (RQ2): What are the most useful features in predicting hospital readmissions for heart failure patients?*

*Research Question 3 (RQ3): Will using the model-agnostic explanatory method (LIME) generate high-quality explanations as measured by explanation stability and local fit?*

### **Relevance and Significance**

ML techniques are increasingly being applied to support a broad range of health care decisions (Dey & Rautaray, 2014; Kaur & Singh, 2014; Valdes et al., 2016). Examples include rationalizing the allocation of limited healthcare resources (Bayati et al., 2014), diagnosing medical conditions (Foster et al., 2014; Nie et al., 2015), classifying stroke risk (Letham et al., 2015), and medical image analysis in digital pathology (Litjens et al., 2017; Madabhushi & Lee, 2016).

ML classification algorithms use training data to build models that map input feature values into a finite number of categorical outputs (Abu-Mostafa et al., 2012). ML methods improve at a given task through experience gained by learning from data. The learning is manifested in the form of tuning input parameters to generate the desired output. The parameter tuning is used to derive generalized ML models to produce predictions on new unseen data. ML models are designed to improve quantitative performance metrics, such as maximizing predictive accuracy and minimizing error categories (Jordan & Mitchell, 2015; Mitchell, 1997). The complexity of high-performing ML algorithms can make them inscrutable to humans resulting in perceiving them as untrusted black-boxes unsuitable for adoption in high stakes decision-making (Henelius et al., 2014; Lipton, 2018; Miotto et al., 2018).

An example of the risks posed by using black-box ML models is noted by Caruana (2017). In this example, the ML algorithm was tasked with predicting the

probability of death of pneumonia patients. The black-box ML model predicted that pneumonia patients with asthma had a lower probability of death than their counterparts with no asthma. Medical experts attributed the lower probability of death for the asthmatic patient group to a higher medical intervention level. The explanation provided by domain experts identified the presence of a moderating variable (increased medical intervention). By contrast, the unexplained prediction of the ML model could have led to an incorrect decision path of minimal medical intervention. Additional examples of ML models failing in a clinical context by learning unintended patterns in the training data due to the inability to distinguish causal from correlation effects are noted by (Badgeley et al., 2019; Zech et al., 2018).

Trust in automated systems plays a leading role in the willingness of humans to use them in a mission-critical domain such as healthcare (Biran & Cotton, 2017; Hoff & Bashir, 2014; Ustun & Rudin, 2014; Vorm, 2018). The lack of trust in ML black box models was addressed legislatively by the European Parliament's General Data Protection Regulation (GDPR). The legislation included the "right to an explanation" mandating human interpretation of ML decisions (Goodman & Flaxman, 2016; Selbst & Powles, 2017). The legislation mandated opening ML black-box models for inspection, highlighting the importance of human interpretation as a condition of adoption and granting subjects the right to opt-out of automated decision making. While the extent of the legal protections offered by GDPR to data subjects under the right to explanation is not fully established (Wachter et al., 2017), it is evident that the drive for regulatory safeguards requires human control over automated systems is a manifestation of distrust in ML algorithmic decisions.

Hence, the transparency afforded by ML interpretability is necessary to protect from discriminatory biases (Hutchinson & Mitchell, 2019; Kim et al., 2018; Lepri et al., 2018; Obermeyer et al., 2019; Wiens et al., 2019; Oliver et al., 2018), support model debugging (Du et al., 2019; Kulesza et al., 2015), provide feedback for improving models (Ahmad et al., 2018; Rudin, 2019), and support overall transparency and human autonomy in decision making (Datta et al., 2016; Pedreschi et al., 2019).

### Related Studies

Recent studies have attempted to predict hospital readmissions for heart failure patients. However, these studies were limited to a small dataset that is not publicly available (Bayati et al., 2014), reported interpretable risk factors that are global to the population (Yang et al., 2016), did not utilize interpretability techniques (Allam et al., 2019; Liu et al., 2020). There is no known readmission study that applies ML methods on a large publicly available dataset based on local model-agnostic explanations. A summary of the accuracy metrics for related studies is shown below:

Source	Classifier	Dataset	Instances	AUC
(Bayati et al., 2014)	Logistic Regression with LASSO	Hospital EHR	1,172	0.66
(Yang et al., 2016)	Logistic Regression with LASSO	NRD 2015	142,527	0.657
(Yang et al., 2016)	GBM	NRD 2015	142,527	0.663
(Yang et al., 2016)	DNN	NRD 2015	142,527	0.662
(Allam et al., 2019)	Logistic Regression with LASSO	NRD 2013	272,778	0.643
(Allam et al., 2019)	Recurrent Neural Networks combined	NRD 2013	272,778	0.642

Source	Classifier	Dataset	Instances	AUC
	with Conditional Random Fields			
(Liu et al., 2020)	Hierarchical Logistic Regression	NRD 2014	303,233	0.580
(Liu et al., 2020)	XGBoost	NRD 2014	303,233	0.602
(Liu et al., 2020)	Feed-Forward Neural Networks	NRD 2014	303,233	0.604
(Liu et al., 2020)	Medical Code Embedding Deep Set Architecture	NRD 2014	303,233	0.618

*Table 2. Average AUC Comparison for Related Studies*

## Summary

The goal of this dissertation was to systematically investigate the applicability of local model-agnostic methods to explain the predictions of black-box machine learning models used for medical decision-making. As proof of concept, this study addressed the problem of predicting the risk of emergency readmissions within 30 days of being discharged for heart failure patients. The precision, recall, area under the ROC curve for each model were used to measure predictive accuracy. Local Interpretable Model-Agnostic Explanations (LIME) was used to generate explanations from the underlying trained models. Explanation stability and local fit ( $R^2$ ) were used to measure LIME's explanation quality.

## Chapter 2

### Review of the Literature

#### **Introduction**

This chapter surveys the literature to provide a brief overview of ML concepts and algorithms used to generate the clinical decision models and their corresponding explanations.

#### **Cost-Sensitive Learning for an Imbalanced Datasets**

Imbalanced data refer to datasets with unequal distribution between classes where a class is outnumbered and underrepresented (Fernandez et al., 2018). Imbalanced data can result in low predictive accuracy for the minority class due to classifiers being optimized to minimize overall misclassification errors (Galar et al., 2012; He & Garcia, 2009; Lipitakis & Kotsiantis, 2014). The class imbalance problem can be addressed using sampling techniques such as minority-class over-sampling, majority-class under-sampling, and combining of minority-class over-sampling and majority-class under-sampling (Batista et al., 2004; Estabrooks et al., 2004). An alternative technique to address the class imbalance problem is cost-sensitive learning (He & Garcia, 2009; Ling & Sheng, 2008). Cost-sensitive learning can be used to assign a higher cost to misclassifying the minority class (Ling & Sheng, 2008), resulting in a classifier that is less biased towards the majority class.

#### **Ensemble Methods**

For classification problems, rule ensembles combine the prediction of multiple base learners to generate new classifiers (Bauer & Kohavi, 1999). Empirical studies have demonstrated that ensemble methods often produce more accurate predictions than base

learners (Bauer & Kohavi, 1999; Freund & Schapire, 1996). The ensemble approaches used in this dissertation are boosting and bagging (Breiman, 1996).

### **Boosting**

Boosting is a sequential learning algorithm that identifies errors in the early predictions as hard examples to focus on in subsequent iterations. The emphasis on hard examples is implemented by assigning higher weights to the incorrectly classified examples and lower weights to the correctly classified examples. The iterative process combines many individual weak learners into a complex predictor (Schapire & Freund, 2012). The committee of weak learners is individually tasked with developing rough rules of thumb that perform slightly better than random. (Schapire, 2003). Boosting has been shown to provide high accuracy results on benchmark datasets and competitive challenges such as the Netflix prize (Chen & Guestrin, 2016). This dissertation used the LightGBM implementation of Gradient Boosting Decision Trees (Ke et al., 2017).

### **Bagging**

Bagging is an acronym for bootstrap aggregation. Bagging is a sampling method that trains multiple base learners, each using different parts of the data set drawn at random with replacement. The sample size used by each learner is equal to the size of the original training set. The learners are combined through a majority vote to predict a class (Breiman, 1996). Bagging does not change the distribution of the training set based on the performance of previous classifiers (Bauer & Kohavi, 1999); each learning instance is chosen with equal probability. (Rokach, 2010). The independence of individual classifiers in Bagging characterizes it as a parallel ensemble method that exploits the independence of base learners to reduce the generalization error (Zhou, 2012). Bagging



has been demonstrated to be well suited for handling noisy and imbalanced data (Khoshgoftaar et al., 2011) and for use with tree-based methods that are characterized as high-variance, low-bias (Friedman et al., 2001). This study utilized the Random Forests (RF) (Breiman, 2001a; James et al., 2013) bagging implementation.

### **Random Forests (RF)**

Random Forest (RF) is an ensemble algorithm that extends the boosting of decision trees. RF is comprised of randomly constructed trees and makes the final prediction through a majority vote (Breiman, 2001a; James et al., 2013). RF creates many randomized decision trees and averages their predictions to fit the input data (Biau & Scornet, 2016). For many problems, RF achieves the same performance as boosting but is simpler to train and tune (Friedman et al., 2001).

### **Decision Trees (DT)**

Decision Tree (DT) algorithms use observed input attributes to classify objects. The tree is constructed top to bottom through a sequence of decision splits, starting with the root until a leaf is reached, representing a decision class. Candidate branching decision variables are chosen based on criteria such as maximizing information gain. The closer the feature is to the root, the more relevant it is for the prediction. The paths from the root to leaves can be linearized into a set of if-then classification rules (Frank & Witten, 1998; Quinlan, 1987).

DTs are induction algorithms where rules are derived from training examples (Mitchell, 1997). As multiple DT can be derived from an example set, simpler rules are preferred as they are expected to generalize better to unseen examples and avoid overfitting. This principle is referred to as Occam's razor (Blumer et al., 1987). In this

context, simple trees are usually small trees. Implementations of DT include Classification and Regression Trees (CART) (Breiman et al., 1984), Iterative Dichotomizer 3 (ID3) (Quinlan, 1986), and C4.5 (Quinlan, 1993). DTs are considered interpretable classification models because they can model non-linear relationships while maintaining a simple structure (Breiman et al., 1984); have a graphical structure that assists in visualizing the rules; select a subset of features that identify the most relevant attributes; have a hierarchical structure that indicates the relative importance of features. The disadvantages of DT include that prediction accuracy is sensitive to the presence of irrelevant features, and they are prone to overfitting if not pruned (Breiman et al., 1984; Freitas, 2014; James et al., 2013; Kohavi & John, 1997). Additionally, the structure of DT can be highly sensitive to small data perturbation (Breiman, 2001b).

### **Logistic Regression (LR)**

Logistic Regression (LR) is a probabilistic binary classification algorithm. The LR algorithm uses sigmoid transformation functions to assign the predicted output a probability of belonging to a class between 0 and 1. The sigmoid function produces an S-shaped curve combined with the decision threshold to determine binary class assignment (Friedman et al., 2001; James et al., 2013).

LR is a classic prediction method originating from the statistics field credited to (Cox, 1958) and has been well established and widely used in the medical literature (Bagley et al., 2001). LR is considered an interpretable model as the explanatory variables are assigned coefficients that measure their impact on the probability. Additionally, since LR is a probabilistic model, the level of confidence in the prediction can be gleaned from the probability assigned to the prediction. The primary disadvantage

of LR is that its performance has been reported as lower than more complex methods such as Artificial Neural Networks (Tu, 1996). As a result, using LR has traditionally been a choice of intelligibility at the expense of accuracy.

The Least Absolute Shrinkage and Selection Operator (LASSO) is a feature selection technique that reduces the number of prediction parameters and contributes to model interpretability (Tibshirani, 1996). While the initial presentation of LASSO in (Tibshirani, 1996) was for regression models, the technique has been used in the literature for classification problems (Ghosh & Chinnaiyan, 2005).

### **Interpretability Characteristics**

There is no consensus in the literature on a definition of interpretability (Bibal & Frénay, 2016; Doshi-Velez et al., 2017; Du et al., 2019; Gilpin et al., 2018; Lipton, 2016; Murdoch et al. 2019; Rudin, 2019). Rather, interpretability is context-dependent (Ahmad et al., 2018), varies depending on the problem domain (Guidotti et al., 2018) and end-user profile (Tomsett et al., 2018). Absent such definition, the following interpretability characteristics have been identified in the literature:

#### Time Sensitive

The explanation is available based on timing that is aligned with the task. Urgent decisions require simple, easy-to-understand explanations, while non-urgent decisions might warrant a more exhaustive and complex explanation (Guidotti et al., 2019b).

#### Understand Feature Contribution

The contribution of individual features to the final prediction is clearly understood (Caruana et al., 2015).

### Explainable to Humans

The ability to represent information in the way humans think and understand at their experience level intuition (Doshi-Velez et al., 2017; Kim, 2015). Example intuitive representations include natural text and images (Guidotti et al., 2019b).

### Aligns with User Expertise

The detail and level of explanation are aligned with the expertise of the user performing the task (Guidotti et al., 2019b).

### Support Case-based Reasoning

The ability to explain the decision based on similarities to previous cases and incorporate domain knowledge back into the system (Adhikari et al., 2019; Chen et al., 2019; Kim, 2015). Case-based reasoning explanations are represented through sample instances and not on feature importance (Plumb et al., 2018).

### Inspecting Individual Predictions

The ability to inspect through textual or visual artifacts that provide a qualitative explanation between the model inputs and resulting prediction for a single case (Ribeiro et al., 2016b).

### Comprised of Cognitive Chunks

The ability to form basic explanation units and define the interaction between them (Doshi-Velez et al., 2017).

### Expose Internal Logic

Combine the classification presentation with a user-facing explanation of the internal ML algorithm logic (Burrell, 2016).

### Human Simulatability

Human simulatable models provide a description of their calculations and can be fully understood and performed by a human in a reasonable timeframe (Lipton, 2018; Plumb et al., 2018).

Can be Edited by Experts

Domain experts have the option to identify an anomaly in the data and manually intervene to prevent incorrect or biased predictions (Caruana et al., 2015).

Generate Knowledge

Extract relevant knowledge about domain relationships contained in data (Murdoch et al., 2019).

Identify Causal Associations

The ability to distinguish between causal associations and non-causal associations (Lipton, 2018; Holzinger et al., 2019).

Human Precision

The percentage of predictions made by humans that correctly identify model output on unseen instances (Ribeiro et al., 2018).

Human Coverage

The percentage of instances predicted by the user after seeing the explanation (Ribeiro et al., 2018).

Mimics Human-based Reasoning

The level of agreement with independent expert judgment and intuition (Doshi-Velez et al., 2017; Kim, 2015).

Contrastive

Contrastive explanations provide the reason for a prediction was made instead of another prediction (Buhrmester et al., 2019); highlight the difference between a prediction and another instance prediction (Lipton, 1990; Miller, 2019); and align with questions in the form of “why this output instead of that output?” (Waa et al., 2018). Contrastive explanations are also labeled as counterfactual explanations (Wachter et al., 2017) and differential explanations (Du et al., 2019). In the context of medical decision-making, contrastive explanations identify of how a predicted outcome (heart disease) can be different if a feature has a different value (smoking).

### Fidelity

As the post-hoc models work differently than the underlying models they are explaining, differences are expected in their respective predictive outputs. The larger the difference, the less faithful the explanation is to the underlying model (Yang et al., 2019). Fidelity measures the level of alignment between the interpretable model and the black-box model (Adhikari et al., 2019; Guidotti et al., 2018; Lakkaraju et al., 2019). Fidelity is also defined as descriptive accuracy, “the degree to which an interpretation method objectively captures the relationships learned by machine learning models” (Murdoch et al., 2019). Explanation fidelity compares the explanatory model’s prediction accuracy vs. the underlying model to validate that the extracted explanation correctly represents the reasoning of the underlying black-box model (Adhikari et al., 2019; Yoon et al., 2019). Explanation fidelity is an established measure for generally evaluating post-hoc explanation methods (Adhikari et al., 2019) and for evaluating the quality of the explanation generated by LIME (Shi et al., 2020). Explanation fidelity is measured as the percentage of test-set instances in which the explanatory model classifications agree with the model it is explaining (Craven & Shavlik, 1996).

### Stability

The concept of stability is closely tied with model reliability or robustness: small changes to input should not result in large changes of a model selected or predicted class and has been widely reported in the literature as a measure for model quality (Breiman, 2001b; Carvalho et al., 2019; Doshi-Velez & Kim 2017; Van Assche & Blockeel, 2007; Yeh et al., 2019). The small changes to input could be outliers (points far from the majority of the points in the dataset), and a robust model can minimize the negative influence of outliers on the output (Björklund et al., 2019). In the context of post-hoc

explanatory methods, robustness is “the resilience of an ML system’s correctness in the presence of perturbations” (Zhang et al., 2020). Parts of a model that are not stable to perturbations are not considered interpretable (Guidotti & Ruggieri, 2019a; Murdoch et al., 2019). Post-hoc methods have been reported to be unstable (Alvarez-Melis & Jaakkola, 2019b) and vulnerable to adversarial attacks (Ghorbani et al., 2019). LIME specifically has been reported to exhibit instability issues, defined as the repeated application of the explainer under the same conditions yielding different outcomes (Visani et al., 2021). The *Jaccard coefficient* has been used to measure the stability of LIME explanations (Zafar & Khan, 2019).

#### Sparsity and Monotonicity Constraints

The interpretability of a model can be enhanced through sparsity and monotonicity constraints (Du et al., 2019). Sparsity constraints are established by selecting a subset of important features in the decision and presenting them as key drivers behind the prediction (Kim, 2015). Monotonicity constraints are established when a change in value in one or more input values monotonically increases or monotonically decreases the probability of the prediction label belonging to a class (Freitas, 2014).

#### **Predictive Accuracy vs. Interpretability**

The primary characteristics of successful ML predictions are accuracy and interpretability. Predictive accuracy establishes “what” is the correct label on unseen data, while interpretability answers “why” a prediction was made and what features influenced the prediction (Baehrens et al., 2010). The tradeoff between accuracy and interpretability has been established in the literature (Ahmad et al., 2018; Bratko, 1997). Interpretable models provide meaningful insight into the decision-making process but may not have the expressive power to capture the underlying relationship between the

input features and the output. Models that accommodate more complex functional relationships have more predictive power but are often difficult to interpret (Breiman, 2001b; Carvalho et al., 2019; Choi et al., 2016).

### **Intrinsic vs. Post-hoc Explanations**

Intrinsic explanations assume access to the model and generally explain transparent/white-box models such as decision trees, rule-based models, and linear models (Holzinger et al., 2017; Lipton 2018). White-box models are self-explanatory as the model represents the explanation (Du et al., 2019). White-box models are interpretable by design, where feature contribution and model logic can be determined by examining the model's parameters and structure. Examples of white-box models include:

- Bayesian List Machine (BLM) (Letham & Rudin, 2012)
- Supersparse Linear Integer Models (SLIM) (Ustun et al., 2013; Ustun & Rudin, 2016)
- Threshold-Rule Integer Linear Model (TILM) (Ustun & Rudin, 2014)
- Falling Rule Lists (FRL) (Wang & Rudin, 2015)
- Decision sets (Lakkaraju et al., 2016)
- Two-Level Boolean Rules (TLBR) (Su et al., 2016)
- Certifiably Optimal Rule Lists (CORELS) (Angelino et al., 2017)
- Scalable Bayesian Rule Lists (SBRL) (Yang et al., 2017)

Post-hoc model-agnostic explanation methods can be applied to any supervised machine learning model. These methods generate post-hoc explanations that are human interpretable and capture the causal relationship between inputs and outputs (Robnik-



Šikonja & Bohanec, 2018). Post-hoc explanatory methods treat the previously trained model as a black-box irrespective of how it is generated (white-box or black-box) (Carvalho et al., 2019). Post-hoc explanatory methods explain the model without changing it (Murdoch et al., 2019) and without insight on how the model predicts (Ahmad et al., 2018).

### **Model-Specific vs. Model-Agnostic**

Model-specific interpretability methods are exclusively tied to the specific class model (Adadi & Berrada, 2018; Du et al., 2019). Example model-specific interpretability methods include:

- TREPAN, explain neural networks with decision trees (Craven & Shavlik, 1996; Krishnan et al., 1999)
- Decision Tree extractor (DecText) extracts Decision Trees from trained feedforward Neural Networks (Boz, 2002).
- Conditional variable importance for random forests (Strobl et al., 2008)
- Feature contribution for random forest classification (Palczewska et al., 2014)
- Computer vision explanations for convolutional networks (Zeiler & Fergus, 2014)
- Genetic extraction of a single, interpretable model (GENESIM): use a genetic algorithm to transfer the learning from ensemble models into a single decision tree (Vandewiele et al., 2016).
- (Reverse Time Attention) model for recurrent neural networks (RNN) (RETAIN) (Choi et al., 2016)
- Additive Tree Models (ensembles of decision trees) interpreter (Hara & Hayashi, 2016)

- TreeView for Deep Neural Networks (Thiagarajan et al., 2016)
- Layer-wise Relevance Propagation (LRP) for interpreting deep neural networks (Binder et al., 2016)
- Extended the usage of Layer-wise Relevance Propagation (LRP) feed-forward neural network classification decisions (Arras et al., 2017)
- Deep Learning Important FeaTures (DeepLIFT) for interpreting neural networks (Shrikumar et al., 2017)
- Gradient-weighted Class Activation Mapping (Grad-CAM) for interpreting Convolutional Neural Networks (CNN) (Selvaraju et al., 2017)
- Integrated Gradients a method that attributes the prediction of deep neural networks for local explanations (Sundararajan et al., 2017)
- Scalable Bayesian Rule Lists (Yang et al., 2017)
- Prediction difference analysis for visualizing deep neural network decisions (Zintgraf et al., 2017)
- Explainable Neural Network Architecture (xNN) (Vaughan et al., 2018)
- Learning to Explain (L2X) (Chen et al., 2018) and INstance-wise VARIable SElection (INVASE) (Yoon et al., 2018) are neural networks that provide an interpretable explanation of its individual predictions.
- Contrastive Explanations with Local Foil Trees (Waa et al., 2018)
- Quantitative Testing with Concept Activation Vectors (TCAV) for interpreting neural networks (Kim et al., 2018)
- Randomized Input Sampling for Explanation (RISE) for explaining deep neural networks for image classifiers (Petsiuk et al., 2018)

- Interpretable trees (inTrees) for interpreting ensembles of decision trees (Deng, 2019)
- Extremal Perturbations (EP) for explaining deep neural network on computer vision classification tasks (Fong et al., 2019)
- GNN Explainer: post-hoc explanations of Graph Neural Networks (Ying et al., 2019)
- TreeSHAP for explaining tree-based models (Lundberg et al., 2020)

Model-agnostic explanatory methods approximate the behavior of underlying ML models to generate end-user explanations that are independent of the internal logic used to make predictions (Ribeiro et al., 2016a). Model-agnostic explanations enable the use of black-box models for tasks requiring explanations. We can take advantage of the accuracy offered by the black box model without sacrificing the need for interpretability (Ribeiro et al., 2016a). Examples of model-agnostic interpretability methods include:

- Local model-agnostic explanations for classification methods that output class probabilities (Robnik-Sikonja & Kononenko, 2008)
- Interactions-based Method for Explanation (IME) (Štrumbelj et al., 2009)
- Leveraging concepts from coalition game theory to explain individual predictions (Strumbelj & Kononenko, 2010)
- Local gradient explanation vector that describes the movement needed for a data point to change its predicted label (Baehrens et al., 2010)
- Sensitivity Analysis (Cortez & Embrechts, 2013)
- GoldenEye (Henelius et al., 2014)

- Layer-Wise Relevance Propagation (LRP) for interpreting image classification for multilayered feed-forward neural networks (Bach et al., 2015)
- Gradient feature auditing (GFA) (Adler et al., 2016)
- Model Explanation System (MES) (Truner, 2016)
- Single Tree Approximation (STA) (Zhou & Hooker, 2016)
- Quantitative Input Influence (QII) (Datta et al., 2016)
- Automatic STRucture IDentification (ASTRID) (Henelius et al., 2017)
- Black Box Explanations through Transparent Approximations (BETA) (Lakkaraju et al., 2017)
- Interpretability via extracting a decision tree to approximate the underlying model (Bastani et al., 2017)
- SHapley Additive eXplanations (SHAP) is based on coalition game theory and sets variable combinations as cooperating and competing coalitions to maximize the payoff of an accurate prediction. Kernel-SHAP is a model agnostic post-hoc interpretability method (Lundberg & Su-In, 2017b)
- Meaningful Perturbation (MP) for image data (Fong & Vedaldi, 2017)
- Real-time image saliency for black-box classifiers (Dabkowski & Gal, 2017)
- Influence Functions (Koh & Liang, 2017)
- Feature Importance (Adadi & Berrada, 2018)
- Local Rule-based Explanations (LORE) (Guidotti et al., 2018)
- Anchors (Ribeiro et al., 2018)
- Model Agnostic Supervised Local Explanations (MAPLE) (Plumb et al. 2018)

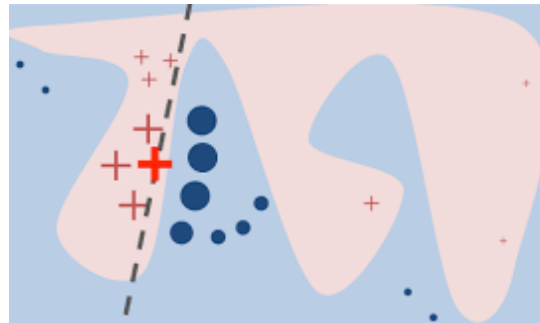
- Asymmetric Shapley values (ASV), which is based on cooperative game theory and incorporate causal knowledge into explanations (Frye et al., 2019)
- Causal explanation (CXPlain) (Schwab & Karlen, 2019)
- Contextual Local Explanation (CLE) (Zhang et al., 2019)
- Local Rule-based Model Interpretability with k-optimal Associations (LoRMiKA) (Rajapaksha et al., 2020)
- Local Example and Feature importance-based model AGnostic Explanations. (LEAFAGE) (Adhikari et al., 2019)
- Sparse Linear Subset Explanations (SLISE) (Björklund et al., 2019)
- Model Understanding through Subspace Explanations (MUSE) (Lakkaraju et al., 2019)

### **Model Approximation**

Explaining black box models through local approximation methods such as LIME (Ribeiro et al., 2016b) is categorized as a proxy method (Gilpin et al., 2018). The approach approximates large complex models (Ex: Ensemble or Neural Network) into smaller, simpler models (Ex: decision tree, rule-based model, or linear model) that are easier to interpret (Gilpin et al., 2018). The approach is also labeled in the literature as model compression (Bucila et al., 2006; Wang et al., 2019), knowledge distillation (Frosst & Hinton, 2017; Hinton et al., 2015; Liu et al., 2018), model extraction (Bastani et al., 2017), model distillation (Tan et al., 2020), and mimic learning (Che et al., 2015; Du et al., 2019).

### Local Interpretable Model-Agnostics Explanations (LIME)

The model-agnostic explanation in this study was generated using Local Interpretable Model-Agnostics Explanations (LIME) (Ribeiro et al., 2016b). LIME falls under the broader category of removal-based explanations that establish feature importance by systematically simulating removing features to quantify their influence (Covert et al., 2020). The main characteristics of LIME are model-agnostic and local. The LIME localized explanations zoom in to the input space region relevant to the individual prediction and identify an interpretable model locally faithful to the classifier without attempting to generalize or establish global rules for other instances in the input space (Ribeiro et al., 2016b). While the global decision boundary might be complex and squiggly, the localized explanation can be achieved through a linear approximation close to the decision point, as illustrated in Figure 2.

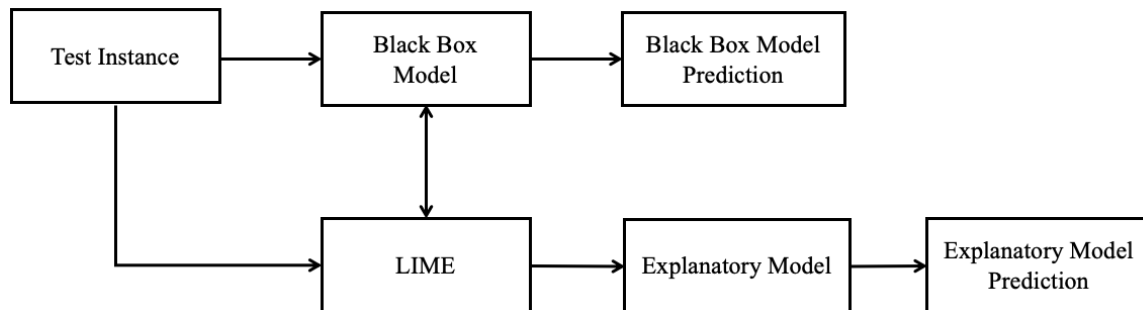


*Figure 2. LIME Decision Boundary*

From “Why Should I Trust You?” Explaining the Predictions of Any Classifier”, by M. T. Ribeiro, S. Singh, and C. Guestrin, 2016, ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Figure 2 illustrates how LIME is applied to a binary classifier tasked with separating instances into a red or blue class. The bold red cross represents the decision being explained, the blue and pink regions represent the global decision boundary, and

the dashed line represents the localized LIME explanation. The localized explanation identifies parts of the input necessary for the prediction, contributing to an intuitive and easy-to-understand explanation (Ribeiro et al., 2016b). In the context of medical decision-making, localized explanations can identify the specific attributes of the patient's condition that drive a decision. While the global decision model needs to account for complex edge cases such as rare medical conditions, a simple, localized explanation would suffice for most patients. Additionally, highlighting the input parameters that drove the decision, such as the presence of a medical condition or the number of recent emergency admissions, allows the decision-maker to recognize potential flaws in the model's logic. The LIME approximation process is depicted in Figure 3 and summarized below:



*Figure 3. The LIME Process*

Input Parameters: choose an instance to explain along with the number of input features used to provide an explanation.

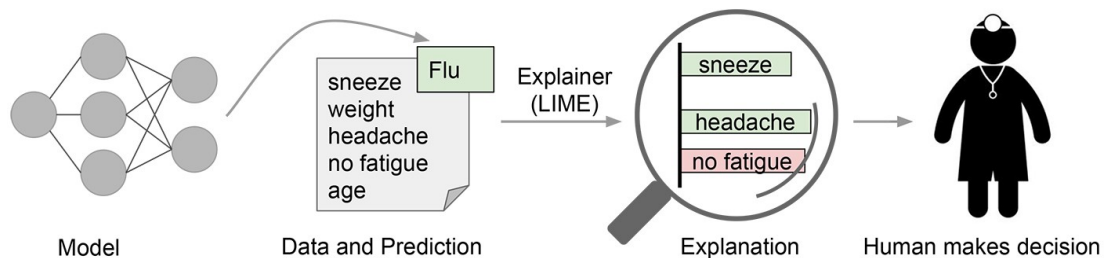
Perturbations: create a new synthetic dataset by randomly sampling points around the input instance; obtain the black box predictions of these new samples.

Weighting: use the proximity to the instance being explained as weights and a measure of similarity.

Fitting: fit a weighted, interpretable model on the perturbed dataset; fit a linear regression in the local area.

Explanation: construct a simple linear regression model against the perturbed dataset with the coefficients of the features used as the explanation. The explanation is expressed in terms of how input features influence the model in choosing a class. The coefficients can have positive or negative values indicating the direction of the relationship between the features and the predicted class. Coefficients values express the magnitude of feature contribution. The larger the coefficient value, the more significant the contribution to the underlying model's prediction.

Figure 4 illustrates the use of LIME to explain the flu/not-flu classifier. The patient is classified as having the flu, with the symptoms of sneeze and headache supporting the prediction. While the absence of fatigue symptoms contradicts the prediction of the flu, the influence of the supporting features is greater.



*Figure 4. Explaining a Prediction with LIME*

From "Why Should I Trust You?" Explaining the Predictions of Any Classifier", by M. T. Ribeiro, S. Singh, and C. Guestrin, 2016, ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Figure 5 illustrates the LIME's output explaining an instance from a tabular dataset for a classifier predicting if a mushroom is edible or poisonous. The leftmost graph provides prediction probabilities for each class, the middle graph provides feature importance visualization, and the rightmost graph provides the feature values for the explained instance.



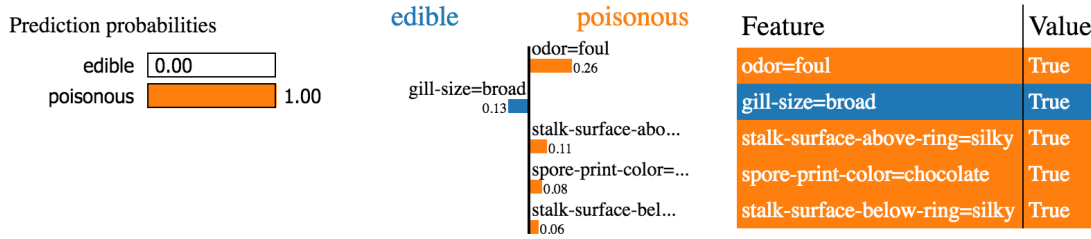


Figure 5. LIME Tabular Classifier Explanation

From: <https://github.com/marcotcr/lime>

The output of LIME includes  $R^2$  (fit statistic), which is a measure of the fit of the linear model in the local region. LIME also outputs the important features based on a pre-determined threshold.

The interpretation task performed by LIME can be summarized below (Asano et al., 2019; Ribeiro et al., 2016b; Visani et al., 2020):

To interpret the prediction for an instance  $x$  for an underlying model represented by function  $f_{model} : X \rightarrow Y_{model}; x \in X$ ;

- $x$ : explained instance
- $X$ : input feature space
- $Y_{model}$ : predicted target class for the underlying model

LIME locally approximates decisions made by  $f_{model}(x)$  with  $g_{LIME}(x)$ ;

The coefficient parameters  $p$  generated for  $g_{LIME}(x)$  represent the feature importance of the local model.

LIME is formalized as an optimization problem balancing local fidelity loss and interpretability with the objective function

$$\xi(x) = \operatorname{argmin} L(f_{model}, g_{LIME}, \Pi_x) + \Omega(g_{LIME})$$

- $\Omega(g_{LIME})$ : penalty function for the complexity of  $g_{LIME}$
- $L$ : loss function
- $\Pi_x$ : weight assigned based on proximity to instance  $x$

## Limitations of LIME

Several drawbacks of the LIME method have been reported in the literature. First, LIME is computationally expensive as it requires generating a local model for each instance with a large number of samples (Ahmad et al., 2018; Schwab & Karlen, 2019). Hence, generating individual explanations for the entire dataset can be impractical (Lundberg et al., 2020). Second, the presence of uncertainty in LIME explanations due to randomness in the sampling procedure (Zhang et al., 2019) results in LIME explanations lacking stability and producing different explanations for the same instance (Visani et al., 2021; Zafar & Khan, 2019). Third, LIME makes no claims on generating causal explanations. Additionally, LIME assumes local linearity, which means it may not faithfully approximate local non-linear decision boundaries. Finally, as demonstrated by (Slack et al., 2020), LIME is vulnerable to adversarial attacks allowing adversaries to hide underlying biases of a classifier by gaming a post-hoc perturbation-based technique such as LIME to generate an arbitrary explanation of their choice.

## LIME Variants and Alternatives

Explanatory methods based on LIME making extensions or revisions and reporting comparative results to LIME in the literature include:

*Modified Perturbed Sampling (MPS-LIME)* alters the perturbed sampling of LIME to consider the correlation between features and apply it to an image classifier. MPS LIME was reported to have higher local fidelity than LIME (Shi et al. 2020).

*Minimal Pattern (MP-LIME)* generates all non-redundant feature sets providing visibility to the combination of features that drove the decision. MP-LIME was reported to have higher precision than LIME (Asano et al., 2019).

Deterministic LIME (DLIME) substitutes the random perturbations with agglomerative Hierarchical Clustering to group training data. K-Nearest Neighbour (KNN) (Jerez et al., 2010) is used to select a relevant cluster of an explained instance. DLIME was reported to generate stable explanations compared to LIME's unstable explanations (Zafar & Khan, 2019; Zafar & Khan, 2021).

K-LIME partitions local using unsupervised clustering into K-clustered partitions and fit local generalized linear model (GLM) within each cluster. K-LIME is utilized in the commercial product Driverless AI by H2O (Hall et al., 2017).

Locally Interpretable Models and Effects based on Supervised Partitioning (LIME-SUP) variant of K-KLIME using supervised partitioning vs. unsupervised partitioning performed by K-LIME and reporting improved model fit metrics compared to K-LIME (Hu et al., 2018).

Autoencoder Based Approach for Local Interpretability (ALIME) reported improved stability and local fidelity using an autoencoder model as the weighting function (Shankaranarayana & Runje, 2019).

Optimized LIME Explanations for Diagnostic Computer Algorithm (OptiLIME), a framework designed to address the lack of stability of LIME explanations (Visani et al., 2020).

LIME-G uses generative models to explain image classifiers (Agarwal & Nguyen, 2021).

QLIME Quadratic Local Interpretable Model-Agnostic Explanation (Bramhall et al., 2020).

## Summary

The interpretability of machine learning algorithms is required to support critical healthcare decisions. Although there is no universally agreed-upon definition of interpretability, stability and local fit ( $R^2$ ) have been identified in the literature as empirical metrics measuring the quality of post-hoc local explanatory models.

The dissertation utilized LIME, a model agnostic, local, post-hoc explanatory method. LIME explains the underlying model by fitting a sparse linear model over synthetically created perturbed instances in the region of the predicted instance. The coefficients of the sparse linear model represent the relative feature importance for the prediction and can be used to understand the relationship between input features and the prediction outcome. LIME explanation complexity is enforced via regularization.

## Chapter 3

### Methodology

#### Overview

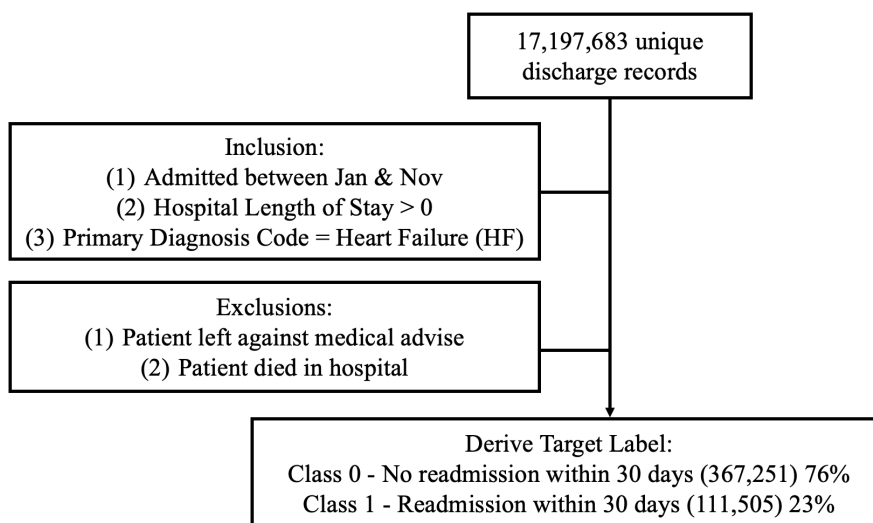
This study utilized ML techniques to predict hospital emergency readmissions for heart failure patients within 30 days of being discharged. The prediction task was formulated as the following binary classification problem: At the time of discharge, the heart failure patient instance  $I$  is represented by a feature vector  $\bar{x}$ . The predicted binary class label is represented by  $y \in \{0,1\}$ . Readmission within 30 days of discharge is represented by label  $y = 1$ ; No readmission within 30 days of discharge is represented by label  $y = 0$ . This chapter provides an overview of the HCUP datasets followed by a description of experimental steps (Pre-Process, Predict, Explain & Evaluate).

#### HCUP Dataset

The dataset for this study was derived from the 2016 National Readmissions Database (NRD) made available by the Healthcare Cost and Utilization Project (HCUP) and sponsored by the Agency for Healthcare Research and Quality (AHRQ), under the Department of Health and Human Services (DHHS). The use of the NRD is governed by the HCUP Data Usage Agreement (DUA). Patient records were deidentified in compliance with the Health Insurance Portability and Accountability Act (HIPAA). HCUP contains the most extensive collection of all-payer hospital care data in the US. The NRD's intended use is to support the analysis of repeat hospital visits within a year and includes both patient and hospital characteristics. While the general schema outline for the NRD has remained consistent for the yearly releases, one key difference across the years was the migration from the International Classification of Diseases, Ninth Edition ICD-9 to ICD-10 as of October 1, 2015. Note that patient IDs do not carry over

from one year to the next, so there is no ability to track a patient being discharged from a year and readmitted in the subsequent calendar year.

The 2016 NRD dataset is comprised of three files: Core, Hospital, and Severity. The dataset contains 17,197,683 unique discharge records that correspond to 12,602,866 unique patients and 2,355 unique hospitals. The Core file (17,197,683 rows, 103 features) contains admission/discharge, patient demographics, and clinical information on the encounter, including discharge diagnoses, recorded procedures based on ICD-10. The Severity file (17,197,683 rows, 5 features) contains attributes related to the severity of the patient's condition, such as Diagnosis Related Groups and the risk of mortality. The Hospital file (2,355 rows, 12 features) includes attributes such as ownership, number of beds, teaching hospital status, and regional characteristics. The schema of the 2016 NRD is detailed in appendix A. Preprocessing of the dataset was performed to identify patients with the primary discharge diagnosis of Heart Failure (HF) and retain features and discharge records that could be used for readmission prediction.



*Figure 6. Dataset Preprocessing*

An outline of the approach that was followed in the study is depicted below:

Pre-Process		Predict	
<b>Step 1</b> Exploratory Data Analysis		<b>Step 6</b> Construct Predictive Models	
<b>Step 2</b> Filter Records		<b>Explain &amp; Evaluate</b>	
<b>Step 3</b> Derive Target Variable		<b>Step 7</b> Construct LIME Explainers	
<b>Step 4</b> Feature Engineering and Encoding		<b>Step 8</b> Extract Model Specific Global Explanations	
<b>Step 5</b> Split Data		<b>Step 9</b> Accuracy Metrics	
		<b>Step 10</b> Interpretability Metrics	

Figure 7. Study Outline

## Pre-Process

### Step 1 – Exploratory Data Analysis

Outline the schema for the tabular NRD files (Core, Severity, Hospital), sample data, data distribution, check for missing or invalid codes. The analysis included identifying missing or invalid records as noted in the following HCUP coding practices (HCUP - Healthcare Cost and Utilization Project, 2020) defined the dataset coding practices:

- *Missing Data*: negative 9-filled value (-9, -99, -999, etc.) for numeric data elements; " " (blank) for character data elements.
- *Invalid Data*: negative 8-filled value (-8, -88, etc.) for numeric data elements; "A" for character data elements.
- *Data Unavailable from Source*: negative 7-filled value (-7, -77, etc.) for numeric data elements.
- *Inconsistent Data*: negative 6-filled (-6, -66, etc.) value for numeric data elements.

- *Not Applicable Data*: negative 5-filled value (-5, -55, etc.) for numeric data elements.

### Step 2 – Filter Records

The dataset was filtered to only include initial admissions between January 1, 2016, to November 30, 2016. Initial admissions in December were not included as admission records of January 2017 were not available. The following inclusions were applied: (1) Admitted between January and November; (2) Hospital length of stay > 0; (3) Non-elective admission; (3) Primary Diagnosis Code (I10\_DX1) corresponds to Heart Failure (HF) condition based on ICD 10 codes identified in appendix E. The following exclusions were applied: (1) Patient left against medical advice; (2) Patient died in hospital.

### Step 3 – Derive Target Variable (Readmissions within 30 Days)

The readmission logic was implemented based on (Yoon, Sheng, Jiang, Steiner, & Barrett, 2017). The “*nrd\_visitlink*” feature was used to identify a unique patient across multiple visits. To mask the identity of patients, the dataset included a randomly selected date of admission instead of the actual admission record. Therefore, requiring the creation of a “Pseudo Date” to be calculated based on the “days to the event” and “length of stay.” The “Pseudo Date” is required to calculate the readmission events. Pseudo Date is assigned: “Days To Event” + “Length of Stay.” Subsequent visits were used to calculate the difference between visits in days. The numeric difference between admission dates was used to establish the binary target variable for the prediction (1 for readmissions less than



30 days and 0 for no readmission within 30 days). Patient readmissions after 30 days are considered new admissions.

The data frame for the predictive model was established by merging elements from the Core, Severity and Hospital files using their common keys as shown below for the 2016 NRD dataset: Merge Severity File and Hospital file based on HOSP\_NRD field; Merge Core File with Severity/Hospital based on the KEY\_NRD field. The features were manually selected based on the ease of encoding them for an ML predictor.

<b>Feature Name</b>	<b>Description</b>	<b>Type</b>	<b>Number of Categories</b>	<b>Invalid/ Missing</b>
AGE	Age in years	Numeric	-	0%
TOTCHG	Total charges in dollars	Numeric	-	0%
LOS	Length of stay in days	Numeric	-	0%
I10_NECAUSE	Number of external causes of morbidity codes on the record	Numeric	-	0%
I10_NPR	Number of procedures coded	Numeric	-	0%
I10_PR1–I10_PR15	ICD-10-PCS Procedure Coding System, principal and secondary (15 features)	Categorical	*	*
I10_NDX	Number of ICD-10-CM diagnoses coded on the record	Numeric	-	0%
I10_DX1–I10_DX35	ICD-10-CM diagnoses, principal and secondary (35 features)	Categorical	*	*
AWEEKEND	Admission on weekend/weekday	Categorical	2	0%
DISPUNIFORM	Disposition of patient, uniform coding	Categorical	5	0%

<b>Feature Name</b>	<b>Description</b>	<b>Type</b>	<b>Number of Categories</b>	<b>Invalid/ Missing</b>
DMONTH	Discharge month	Categorical	12	0%
DQTR	Discharge quarter	Categorical	4	0%
FEMALE	Indicator of gender	Categorical	2	0%
HCUP_ED	HCUP indicator of emergency department record	Categorical	5	0%
PAY1	Expected primary payer	Categorical	8	0.107%
PL_NCHS	Patient Location: National Center for Health Statistics (NCHS)	Categorical	7	0.286%
REHABTRANSFER	Transfer to rehabilitation, evaluation, or other aftercare	Categorical	2	0%
RESIDENT	Patient is a resident of the State in which he or she received hospital care	Categorical	2	0%
SAMEDAYEVENT	Identifies transfer as same day event	Categorical	5	0%
ZIPINC_QRTL	Median household income for patient's ZIP Code	Categorical	6	1.461%
MDC	Major Diagnostic Category MDC in use on discharge date	Categorical	2	0%
MDC_NoPOA	Major Diagnostic Category (MDC) in use on discharge date, calculated without the use of the present on admission (POA) flags	Categorical	2	0%
DRG	Diagnosis Related Group (DRG)	Categorical	77	0%

<b>Feature Name</b>	<b>Description</b>	<b>Type</b>	<b>Number of Categories</b>	<b>Invalid/ Missing</b>
DRG_NoPOA	Diagnosis Related Group (DRG) without the use of the present on admission (POA) flags for the diagnoses	Categorical	77	0%
APRDRG	All Patient Refined Diagnosis Related Groups (APR-DRGs)	Categorical	25	0%
APRDRG_Risk_Mortality	All Patient Refined Diagnosis Related Groups: Risk of Mortality	Categorical	5	0%
APRDRG_Severity	All Patient Refined Diagnosis Related Groups Severity of Illness	Categorical	5	0%
HOSP_BEDSIZE	Hospital Bed Size	Categorical	3	0%
H_CONTRL	Hospital control/ownership	Categorical	3	0%
HOSP_URCAT4	Hospital urban-rural designation	Categorical	4	0%
HOSP_UR_TEACH	Teaching status of hospital	Categorical	4	0%

*Table 3 Pre-Processed Features*

*\* I10\_PR1–I10\_PR15 and I10\_DX1–I10\_DX35 are utilized to create engineered features as outlined in step 4.*

All instances with any missing/invalid features were removed. The preprocessing of the NRD dataset resulted in single merged dataset as outlined below.

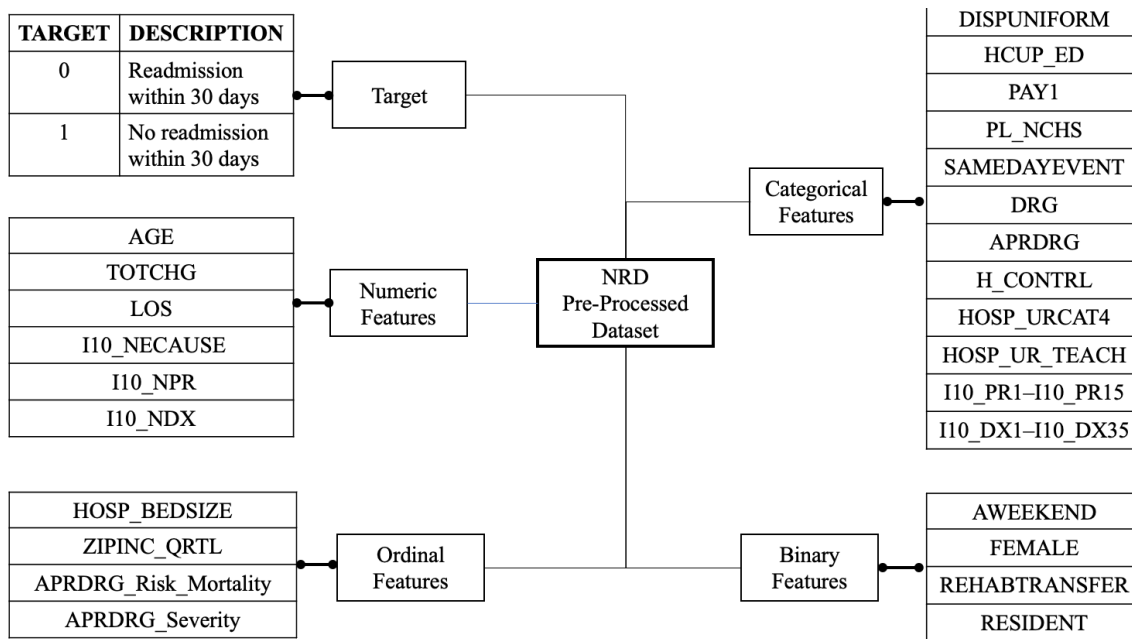


Figure 8 Pre-Processed Dataset (prior to feature engineering and encoding)

All invalid/missing instances removed resulting in (478,756 instances) with a target class distribution

Target Class	Number of Instances	% Distribution
Class 0 - Not Readmit <= 30	367,251	76%
Class 1 – Readmit <= 30	111,505	23%

Step 4 Feature Engineering and Encoding

AGE feature was binarized to “Not-Senior” for age distribution between 0 and 64 and “Senior” for age distribution of 65 to 99.

TOTCHG feature was binarized with bin distribution of 0 to 20K, 20k to 60K, and over 60K.

LOS feature was feature was binarized with bin distribution of 0 to 3, 3 to 6 and 6 to 344.

DRG categories were reduced to 11 categories by retaining the top 10 categories and assigning the remaining categories to OTHER as summarized below:

ICD 10 Code	Description
227	CARDIAC DEFIBRILLATOR IMPLANT WITHOUT CARDIAC CATHETERIZATION WITHOUT MAJOR COMPLICATION OR COMORBIDITY (MCC)
246	PERCUTANEOUS CARDIOVASCULAR PROCEDURES WITH DRUG-ELUTING STENT WITH MCC OR 4+ VESSELS OR STENTS
264	OTHER CIRCULATORY SYSTEM O.R. PROCEDURES
280	ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE WITH MCC
281	CUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE WITH CC
286	CIRCULATORY DISORDERS EXCEPT AMI, WITH CARDIAC CATETERIZATION WITH MCC
287	CIRCULATORY DISORDERS EXCEPT AMI, WITH CARDIAC CATETERIZATION WITHOUT MCC
291	HEART FAILURE AND SHOCK WITH MCC
292	HEART FAILURE AND SHOCK WITH CC
293	HEART FAILURE AND SHOCK WITHOUT CC/MCC
999	OTHER

*Table 4 Selected Top DRG Categories*

APRDRG was reduced to the top 6 categories by retaining the top 5 categories and assigning the remaining categories to OTHER.

ICD 10 Code	Description
161	CARDIAC DEFIBRILLATOR & HEART ASSIST IMPLANT
175	PERCUTANEOUS CARDIOVASCULAR PROCEDURES W/O ACUTE MYOCARDIAL INFARCTION (AMI)
180	OTHER CIRCULATORY SYSTEM PROCEDURES
192	CARDIAC CATHETERIZATION FOR ISCHEMIC HEART DISEASE
194	HEART FAILURE
999	OTHER

*Table 5 Selected Top APRDRG Categories*

All categorical features were processed with one hot encoding (Hackeling, 2017; Hancock & Khoshgoftaar, 2020). All non-primary diagnosis conditions present in

features I10\_DX2 through I10\_DX35 were encoded based on ICD-10 code mapping defined in appendix E to create the following binary features:

<b>Feature</b>	<b>Description</b>
IDXn_CP	Chest Pain
IDXn_HT	Hypertension
IDXn_SH	Shock
IDXn_DB	Diabetes
IDXn_PCI	Presence of coronary/cardiac implant and Percutaneous Coronary Intervention (PCI)
IDXn_STR	Stroke
IDXn_COPD	Chronic Obstructive Pulmonary Disease (COPD)
IDXn_ULC	Ulcer
IDXn_MI	Myocardial Infarction (MI)
IDXn_CVS	Cardiovascular System (CVS) disease
IDXn_PVS	Peripheral Vascular System (PVS) diseases
IDXn_LR	Liver or Renal Failure/Disease
IDXn_DM	Dementia
IDXn_CTD	Connective Tissue Disease (CTD)
IDXn_AIDS	Acquired immunodeficiency syndrome (AIDS) / Human Immunodeficiency Virus (HIV)

*Table 6 ICD-10 Mapped Medical Conditions*

The medical conditions were based on the readmission risk score calculator published by the Yale School of Medicine (Center for Outcomes Research & Evaluation (CORE), 2021).

The top 20 most frequently occurring procedure codes present in features I10\_PR1 through I10\_PR15 were identified and mapped into the following binary features:

<b>Feature</b>	<b>Description</b>
I10_PRn_B2111ZZ	Fluoroscopy of Multiple Coronary Arteries using Low Osmolar Contrast
I10_PRn_5A09357	Assistance with Respiratory Ventilation, Less than 24 Consecutive Hours, Continuous Positive Airway Pressure.
I10_PRn_5A1D60Z	Performance of Urinary Filtration, Multiple.
I10_PRn_4A023N7	Measurement of Cardiac Sampling and Pressure, Left Heart, Percutaneous Approach

<b>Feature</b>	<b>Description</b>
I10_PRn_02HV33Z	Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach
I10_PRn_30233N1	Transfusion of Nonautologous Red Blood Cells into Peripheral Vein, Percutaneous Approach
I10_PRn_B2151ZZ	Fluoroscopy of Left Heart using Low Osmolar Contrast
I10_PRn_4A023N8	Measurement of Cardiac Sampling and Pressure, Bilateral, Percutaneous Approach
I10_PRn_0W993ZZ	Drainage of Right Pleural Cavity, Percutaneous Approach
I10_PRn_5A09457	Assistance with Respiratory Ventilation, 24-96 Consecutive Hours, Continuous Positive Airway Pressure.
I10_PRn_4A023N6	Measurement of Cardiac Sampling and Pressure, Right Heart, Percutaneous Approach
I10_PRn_B246ZZZ	Ultrasonography of Right and Left Heart
I10_PRn_0W9B3ZZ	Drainage of Left Pleural Cavity, Percutaneous Approach
I10_PRn_5A1D00Z	Performance of Urinary Filtration, Single.
I10_PRn_B24BZZZ	Ultrasonography of Heart with Aorta.
I10_PRn_3E0234Z	Introduction of Serum/Tox/Vaccine into Muscle, Perc Approach
I10_PRn_3E0F7GC	Introduction of Other Therapeutic Substance into Respiratory Tract, Via Natural or Artificial Opening
I10_PRn_5A2204Z	Restoration of Cardiac Rhythm, Single
I10_PRn_0BH17EZ	Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening
I10_PRn_3E033GC	Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach

*Table 7 Top 20 Procedure Codes*

The top 20 most frequently occurring external causes of morbidity codes present in features I10\_ECAUSE1 through I10\_ECAUSE4 were identified and mapped into the following binary features:

<b>Feature</b>	<b>Encoding</b>	<b>Description</b>
I10_ECAUSEn_Y95	ECM_Y95	Nosocomial condition
I10_ECAUSEn_Y929	ECM_Y929	Unspecified place or not applicable
I10_ECAUSEn_Y92239	ECM_Y92239	Unspecified place in hospital as the place of occurrence of the external cause
I10_ECAUSEn_W19XXXA	ECM_W19XXXA	Unspecified fall, initial encounter
I10_ECAUSEn_Y92230	ECM_Y92230	Patient room in hospital as the place of occurrence of the external cause

<b>Feature</b>	<b>Encoding</b>	<b>Description</b>
I10_ECAUSEn_Y92009	ECM_Y92009	Unspecified place in unspecified non-institutional (private) residence as the place of occurrence of the external cause
I10_ECAUSEn_X58XXXA	ECM_X58XXXA	Exposure to other specified factors, initial encounter
I10_ECAUSEn_Y838	ECM_Y838	Other surgical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure
I10_ECAUSEn_W1830XA	ECM_W1830XA	Fall on same level, unspecified, initial encounter· External causes of morbidity. Slipping, tripping, stumbling and falls
I10_ECAUSEn_Y9289	ECM_Y9289	Other specified places as the place of occurrence of the external cause
I10_ECAUSEn_Y848	ECM_Y848	Other medical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure
I10_ECAUSEn_Y939	ECM_Y939	Activity, unspecified
I10_ECAUSEn_Y846	ECM_Y846	Urinary catheterization as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure
I10_ECAUSEn_Y831	ECM_Y831	Surgical operation with implant of artificial internal device as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure
I10_ECAUSEn_W010XXA	ECM_W010XXA	Fall on same level from slipping, tripping and stumbling without subsequent striking against object, initial encounter.
I10_ECAUSEn_Y832	ECM_Y832	Surgical operation with anastomosis, bypass, or graft as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure



<b>Feature</b>	<b>Encoding</b>	<b>Description</b>
I10_ECAUSEn_Y92019	ECM_Y92019	Unspecified place in single-family (private) house as the place of occurrence of the external cause
I10_ECAUSEn_Y830	ECM_Y830	Surgical operation with transplant of whole organ as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure - as a primary or secondary diagnosis code
I10_ECAUSEn_Y92238	ECM_Y92238	Other place in hospital as the place of occurrence of the external cause
I10_ECAUSEn_invl	ECM_invl	Invalid Code

*Table 8 Top 20 External Causes of Morbidity*

The pre-processing steps resulted in the following 77 categorical features:

<b>Feature Name</b>	<b>Number of Categories</b>
DISPUNIFORM	5
HCUP_ED	5
PAY1	6
PL_NCHS	6
SAMEDAYEVENT	5
H_CONTRL	3
HOSP_URCAT4	4
MDC	3
HOSP_UR_TEACH	3
DRG	11
APRDRG	6
AGE	2
TOTCHG	3
LOS	3
HOSP_BEDSIZE	3
ZIPINC_QRTL	4
Risk_Mortality	5
Severity	5
AWEEKEND	2

<b>Feature Name</b>	<b>Number of Categories</b>
FEMALE	2
REHABTRANSFER	2
RESIDENT	2
IDXn_CP	2
IDXn_HT	2
IDXn_SH	2
IDXn_DB	2
IDXn_PCI	2
IDXn_STR	2
IDXn_COPD	2
IDXn_ULC	2
IDXn_MI	2
IDXn_CVS	2
IDXn_PVS	2
IDXn_LR	2
IDXn_DM	2
IDXn_CTD	2
IDXn_AIDS	2
PRn_B2111ZZ	2
PRn_5A09357	2
PRn_5A1D60Z	2
PRn_4A023N7	2
PRn_02HV33Z	2
PRn_30233N1	2
PRn_B2151ZZ	2
PRn_4A023N8	2
PRn_0W993ZZ	2
PRn_5A09457	2
PRn_4A023N6	2
PRn_B246ZZZ	2
PRn_0W9B3ZZ	2
PRn_5A1D00Z	2
PRn_B24BZZZ	2
PRn_3E0234Z	2
PRn_3E0F7GC	2
PRn_5A2204Z	2
PRn_0BH17EZ	2
PRn_3E033GC	2

<b>Feature Name</b>	<b>Number of Categories</b>
ECM_Y95	2
ECM_Y929	2
ECM_Y92239	2
ECM_W19XXXXA	2
ECM_Y92230	2
ECM_Y92009	2
ECM_X58XXXXA	2
ECM_Y838	2
ECM_W1830XA	2
ECM_Y9289	2
ECM_Y848	2
ECM_Y939	2
ECM_Y846	2
ECM_Y831	2
ECM_W010XXA	2
ECM_Y832	2
ECM_Y92019	2
ECM_Y830	2
ECM_Y92238	2
ECM_invl	2

### Step 5 – Split Data

The dataset was split into a training set (80%) and a testing set (20%)

## **Predict**

### Step 6 – Construct Predictive Models

Generated predictive models against the dataset, with each model, fitted separately. The predictive models were used to eliminate features with low variance. The following models were trained: Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM). Hyperparameter tuning for each model was done through grid search with cross validation with K=5. The grid search was setup to maximize the AUC so that the search for optimal parameters is optimized to maximize the Area and the ROC curve score.

## **Explain & Evaluate**

### Step 7 – Construct LIME Models

Model interpretation was established by generating explanatory features for local predictions. Local Interpretable Model Agnostics Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016a) was used to extract local features for individual instances. The LIME hyperparameter of the maximum number of feature explainers was set to 10 features.

### Step 8 – Extract Global Model Specific Explanations

Global model-specific explanations were extracted and presented in terms of the coefficients for the top 10 features.

### Step 9– Accuracy Metrics

Predictive accuracy metrics were represented by a confusion matrix with key empirical metrics derived from these measures include accuracy, sensitivity (recall), specificity, precision, error rate F-Score, and Area Under the Curve (AUC).

### Step 10 – Interpretability Metrics

Interpretability metrics were represented by Explanation Fidelity and Stability. The metrics included the LIME hyperparameter of 5000 perturbed samples per explanation.

#### *Local Fit*

The LIME reported coefficient of determination  $R^2$  for these explanations was averaged for 500 randomly selected test instances.

#### Local fit ( $R^2$ )

$R^2$  is also known as the coefficient of determination is

$$R^2 = 1 - \frac{SSE}{SST}$$

Where:

- *SSE* (Sum of Squares)
- *SST* (Sum of Squared Total)

$R^2$  measure ranges from 0 to 1; the closer to 1, the better the fit.

### Stability

To measure stability, LIME explanations were generated for 100 randomly selected test instances. Each instance had 10 LIME explanations generated. The Jaccard similarity is based on the similarity of features generated for the same instance. The similarity measure is between 0 and 1. A value of 0 means highly similar. A value of 1 means highly dissimilar. The Jaccard coefficient is

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Where:

- $S_1, S_2$  are two explanation sets

$J(S_1, S_2)$  ranges from 0 to 1; 0 means the sets are dissimilar; 1 means the sets are identical.

The Jaccard similarity distance is:

$$J_{distance} = 1 - J(S_1, S_2)$$

**Resources**

This study was developed and tested on a MacBook Pro laptop equipped with a 2.7 GHz Intel Core i7 processor, 16 GB of memory, 500 GB of disk, and running macOS Sierra version 10.12.6. The development was implemented using Scikit-learn machine learning libraries (Pedregosa et al., 2011). The NRD dataset is publicly available for purchase through the AHRQ website.

**Summary**

The analysis of a large readmission database comprised of 17 million unique discharge records representing 12 million unique patients was performed. Almost 38,000 patient admissions with heart failure as the primary cause of admission were selected for analysis. Supervised classification models were trained to predict the risk of readmission: Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM).

## Chapter 4

### Results

#### Overview

The goal of this dissertation was to systematically investigate the applicability of local model-agnostic methods to explain the predictions of black-box machine learning models used for medical decision-making. As proof of concept, this study addressed the problem of predicting the risk of emergency readmissions within 30 days of being discharged for heart failure patients. As detailed in the methodology chapter, the study was based on the 2016 National Readmissions Database (NRD), containing a total of 17,197,683 unique discharge records that correspond to 12,602,866 unique patients and 2,355 unique hospitals. The pre-processing steps included feature engineering and manual feature selection resulting in 77 features, 478,756 instances with a target class distribution:

Target Class	Number of Instances	% Distribution
Class 0 – Not Readmit $\leq 30$	367,251	76%
Class 1 – Readmit $\leq 30$	111,505	23%

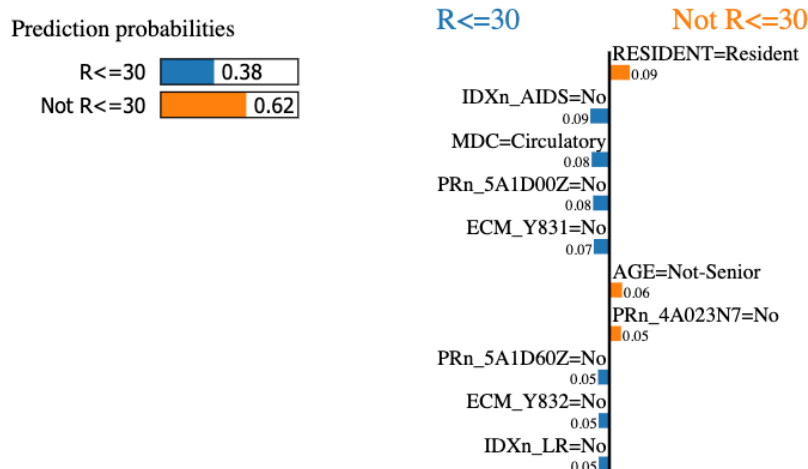
The precision, recall, area under the ROC curve for each model were used to measure predictive accuracy. Local Interpretable Model-Agnostic Explanations (LIME) was used to generate explanations from the underlying trained model. Explanation stability and local fit ( $R^2$ ) were used to measure LIME's explanation quality.

This chapter presents the experimental results of this study for the following classifiers: Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM). The remainder of the chapter is organized as follows. First, sample LIME explanations are presented along with a demonstration of the

useful insight they generate. Next, a visual illustration of the impact of LIME's instability is provided. Then accuracy and explanation metrics are reported for trained classifiers. Finally, the chapter concludes with a summary of results.

### LIME Explanations

The figure below demonstrates a visual explanation generated by LIME for a readmission prediction instance. The left graph provides prediction probabilities for each class label, the right graph provides feature importance visualization.



*Figure 9. Example LIME Explanation for Readmission Prediction*

In this example, the LIME local model predicted the class label probability of readmission within 30 days as 38%. The top 10 features influencing the prediction are identified, with each feature assigned a color code and a coefficient. The color code indicates if a feature supports or contradicts a class label. In this case, the feature-value pairs of AGE=Not-Senior and RESIDENT=Resident support the predictive outcome of no readmission within 30 days. The coefficient assigned to the RESIDENT=Resident feature-value pair is 0.09, and AGE=Not-Senior feature-value pair is 0.06 indicating their influence on the final predictive outcome by the LIME local model. Additionally, LIME supports extracting the coefficients showing the scale and direction of influence on the prediction as shown below:



```
exp_lr.as_list()
[('RESIDENT=Resident', 0.09440294885457647),
 ('IDXn_AIDS=No', -0.08789759972975307),
 ('MDC=Circulatory', -0.08196716155211808),
 ('PRn_5A1D00Z=No', -0.07524455345886892),
 ('ECM_Y831=No', -0.07198836949911053),
 ('AGE=Not-Senior', 0.05885684769469195),
 ('PRn_4A023N7=No', 0.053806809330566606),
 ('PRn_5A1D60Z=No', -0.05237189165982843),
 ('ECM_Y832=No', -0.049792981181576074),
 ('IDXn_LR=No', -0.049177806216800186)]
```

For this specific instance, the resulting explanation translates to:

Rank	Feature/Value Pair	Descriptive Explanation
1	RESIDENT=Resident	The patient was a resident of the state where the hospital admission occurred. This was the most influential feature supporting the prediction not likely to be readmitted within 30 days. The level of influence on the local prediction result is 9%.
2	IDXn_AIDS=No	The patient did not have a medical diagnosis of AIDS. This was the second most influential feature supporting a prediction of likely to be readmitted within 30 days with a level of influence on the local prediction result assigned to 8%.
3	MDC=Circulatory	The Major Diagnostic Category on the date of discharge (MDC) indicates the patient had an issue related to the Circulatory System. This is the third most influential feature supporting the prediction likely to be readmitted within 30 days. The level of influence on the local prediction result is 8%.
4	PRn_5A1D00Z=No	The patient did not have a medical procedure related to multiple urinary filtrations. This was the fourth most influential feature supporting the prediction likely to be readmitted within 30 days. The level of influence on the local prediction result is 7%.
5	ECM_Y831=No	The patient did not have an external cause of morbidity ( <i>surgical operation with an implant of an artificial internal device</i> ). This was the fifth most important feature supporting the prediction of likely to be readmitted within 30 days. The level of influence on the local prediction result is 7%.
6	AGE=Not-Senior	The patient was below the age of 65 at the time of admission. This was the sixth most important feature supporting the prediction outcome of not likely to be readmitted within 30 days. The level of influence on the local prediction result is 6%.
7	PRn_4A023N7=No	The patient did not have a procedure ( <i>Measurement of Cardiac Sampling and Pressure, Left Heart,</i>

Rank	Feature/Value Pair	Descriptive Explanation
		<i>Percutaneous Approach</i> ). This was the seventh most important feature supporting the prediction of likely to be readmitted within 30 days. The level of influence on the local prediction result is 5%.
8	PRn_5A1D60Z=No	The patient did not have a procedure ( <i>Performance of Urinary Filtration, Multiple</i> ). This was the eighth most important feature supporting the prediction of likely to be readmitted within 30 days. The level of influence on the local prediction result is 5%.
9	ECM_Y832=No	The patient did not have an external cause of morbidity ( <i>surgical operation with anastomosis, bypass, or graft</i> ). This was the ninth most important feature supporting the prediction of likely to be readmitted within 30 days. The level of influence on the local prediction result is 5%.
10	IDXn_LR=No	The patient did not have ( <i>Liver or Renal Failure/Disease</i> ). This was the tenth most important feature supporting the prediction of likely to be readmitted within 30 days. The level of influence on the local prediction result is 5%.

An interesting pattern noted in this example is that seven out of the ten most influential features to the local prediction were related to the absence of a medical condition or the absence of medical procedure, with six out of the seven features supporting a local prediction of likely to be readmitted within 30 days. This could be similar to the pattern reported by Caruana (2017), where the presence of certain medical conditions was associated with increased medical care and improved health outcomes. It is plausible for the presence of medical conditions and procedures to require extensive post-discharge follow up which would lead to the reduced likelihood of an emergency readmission. Although this association cannot be conclusively derived from the explanation, it is a useful insight that could be the basis of additional analysis and investigation.

Another example of a LIME explanation is shown below. In this example, the LIME explanation is not strongly weighted towards one class, as noted in the prediction probabilities for both classes being close in value (49%/51%). This could be labeled as a low confidence local prediction. Although, the confidence in the prediction is low, the explanation offers valuable insight as to which features support each prediction label. For example, AGE=Senior supports an increased risk of readmission within 30 days; LOS (Length of Stay) being longer than 6 days supports a predictive outcome of not likely to be readmitted within 30 days. The insight from the local explanation can be useful in validating the logic of the model.

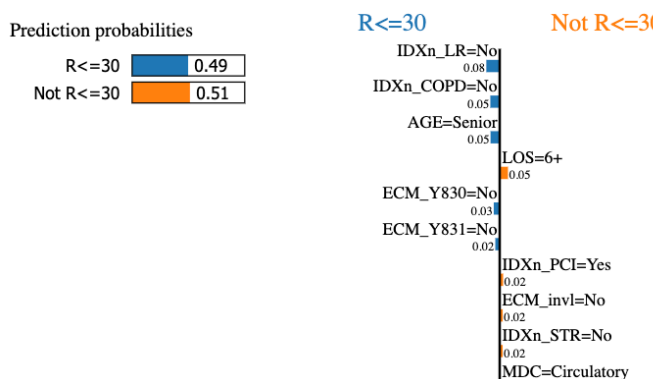


Figure 10. Example LIME Explanation for Readmission Prediction - Low Confidence

### LIME Explanation Instability

As detailed in the methodology chapter, LIME has been reported to generate explanations that are not stable where the repeated application of the explainer under the same conditions yields different outcomes (Visani et al., 2021). The figure below demonstrates LIME's instability in generating different features explanations for the same instance prediction by the same model.

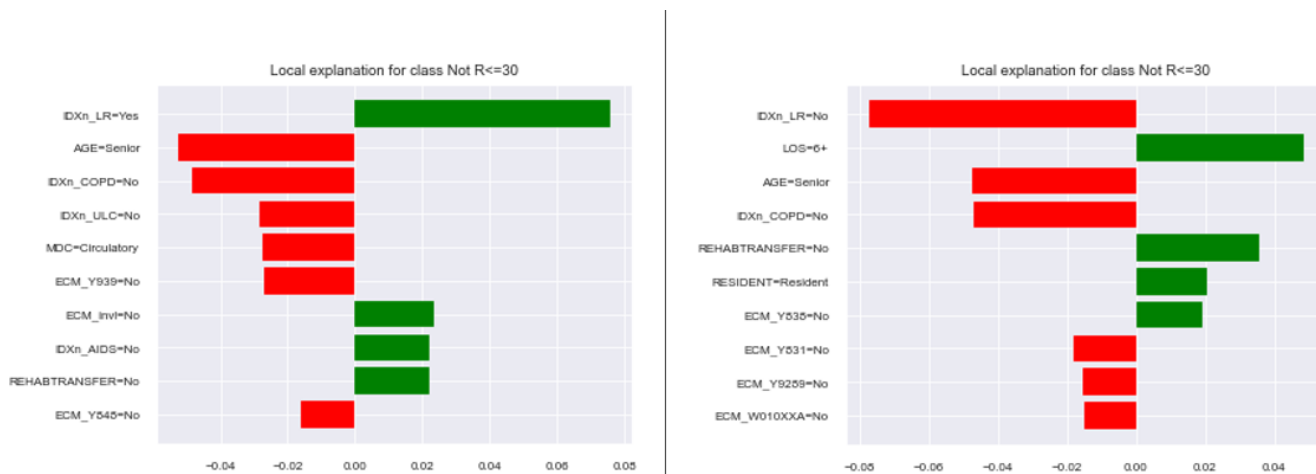


Figure 11. LIME Instability Example

## Experimental Results

This section summarizes the experimental results for the trained classifiers constructed using Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM) models. The dataset for all classifiers was split to train: test ratio of 80:20. As the binary class labels for the dataset were considered imbalanced. All four classifiers utilized the well-established k-fold cross validation method (Refaeilzadeh et al. 2016; Wong, 2015) with the number of folds = 5 to estimate classifier performance and to avoid over-fitting. The scikit-learn python library (Pedregosa et al., 2011) was used for LR, RF, and DT models. GBM was implemented through the open source LightGBM made available by Microsoft Research.

### Logistic Regression (LR)

The “*max\_iter*” (maximum iterations for solvers to converge) parameter is set to 500 to limit the a. The “*class\_weight*” parameter is set to “balanced” to automatically adjust weight and increase the penalty for misclassifying the minority class. The search for the best fit hyperparameters was done through scikit-learn’s GridSearchCV. The grid search parameter “*cv*” was set 5 to the K-fold cross validation, the “*scoring*” was set to

“roc\_auc” so that the search for optimal parameters is optimized to maximize the Area and the ROC curve score. The grid search hyperparameter search and resulting best fit hyperparameters were set to the following values:

<b>Classifier Hyperparameters Grid Search</b>	<b>Best Fit Model Hyperparameters</b>
<ul style="list-style-type: none"> <li>• C = [2, 10.0, 100.0, 1000.0]</li> <li>• solver = [liblinear, saga, newton-cg]</li> <li>• penalty = [11, 12]</li> </ul>	<ul style="list-style-type: none"> <li>• class_weight = balanced</li> <li>• C = 2</li> <li>• penalty = 11</li> <li>• solver = saga</li> </ul>

#### Accuracy Metrics

- Area under ROC curve: 0.5749
- Accuracy: 0.5732
- Weighted F1 score: 0.6054

<b>Class</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
Class 0 – Not Readmit <= 30	0.8145	0.5718	0.6719
Class 1 – Readmit <= 30	0.2941	0.5780	0.3898

#### Confusion Matrix

<b>True Label</b>	<b>Predicted Label</b>	
	0	1
0	40,251	30,144
1	9,170	12,2558

#### LIME Explanation Metrics

<b>Instance Selection</b>	<b>Stability</b>	<b>R<sup>2</sup></b>
500 Randomly Selected Test Instances	0.4328	0.2471
500 Randomly Selected Test Instances (Correctly Predicted)	0.4310	0.2400
500 Randomly Selected Test Instances (Incorrectly Predicted)	0.4292	0.2468

### Global Model Generated Feature Importance

<b>Rank</b>	<b>Feature</b>	<b>Description</b>	<b>Feature Importance</b>
1.	IDXn_CTD	Connective Tissue Disease	0.68
2.	PL_NCHS	Patient Location: National Center for Health Statistics	0.23
3.	PRn_3E033GC	Procedure: Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach	0.20
4.	APRDRG	All Patient Refined Diagnosis Related Groups	0.19
5.	ECM_X58XXXA	External cause of morbidity: Exposure to other specified factors, initial encounter	0.16
6.	PRn_02HV33Z	Procedure Code: Insertion of Infusion Device into Superior Vena Cava, Percutaneous Approach	0.11
7.	PRn_4A023N6	Procedure Code: Measurement of Cardiac Sampling and Pressure, Right Heart, Percutaneous Approach	0.09
8.	AGE	Age of the patient	0.09
9.	ECM_Y92238	External cause of morbidity: Other place in hospital as the place of occurrence of the external cause	0.09
10.	IDXn_PCI	Non-primary diagnosis Condition: Presence of coronary/cardiac implant and Percutaneous Coronary Intervention	0.07

### Experiment Observations

The LIME explanation metrics are stability average 0.43 and local fit (R2) averaging 0.24 were low. This was somewhat unexpected given that the underlying local approximation model generated by LIME is a logistic regression model.

### **Random Forest (RF)**

The “*bootstrap*” parameter is set to True, resulting in the model using bootstrap samples when building trees. The “*class\_weight*” parameter is set to “balanced” to automatically adjust weight and increase the penalty for misclassifying the minority class. The search for the best fit hyperparameters was done through scikit-learn’s

GridSearchCV. The grid search parameter “cv” was set 5 to the K-fold cross validation, the “scoring” was set to “roc\_auc” so that the search for optimal parameters is optimized to maximize the Area and the ROC curve score. The grid search hyperparameter search and resulting best fit hypermeters were set to the following values:

<b>Classifier Hyperparameters Grid Search</b>	<b>Best Fit Model Hyperparameters</b>
<ul style="list-style-type: none"> <li>• n_estimators [100, 300, 500, 800, 1000,1200, 2000, 2500, 3000]</li> <li>• min_samples_split = [8, 10, 12, 15, 20]</li> <li>• min_samples_leaf = [3, 4, 5, 15, 20]</li> <li>• max_features = ['auto', 'log2']</li> <li>• max_depth = [50, 70, 80, 90, 100, 110, None]</li> </ul>	<ul style="list-style-type: none"> <li>• n_estimators = 800</li> <li>• min_samples_split = 8</li> <li>• min_samples_leaf = 15</li> <li>• max_features = log2</li> <li>• max_depth = 50</li> </ul>

#### Accuracy Metrics

- Area under ROC curve: 0.5767
- Accuracy: 0.5740
- Weighted F1 score: 0.6068

<b>Class</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
Class 0 – Not Readmit <= 30	0.8101	0.5717	0.6731
Class 1 – Readmit <= 30	0.2920	0.5817	0.3888

#### Confusion Matrix

<b>True Label</b>	<b>Predicted Label</b>	
	0	1
0	43,395	30,265
1	8,979	12,484

#### LIME Explanation Metrics

<b>Instance Selection</b>	<b>Stability</b>	<b>R<sup>2</sup></b>
500 Randomly Selected Test Instances	0.4042	0.2453
500 Randomly Selected Test Instances (Correctly Predicted)	0.4081	0.2439
500 Randomly Selected Instances Test (Incorrectly Predicted)	0.4068	0.2407

### Global Model Generated Feature Importance

<b>Rank</b>	<b>Feature</b>	<b>Description</b>	<b>Feature Importance</b>
1.	ECM_Y92230	External cause of morbidity: Patient room in hospital as the place cause	0.02
2.	ECM_X58XXXA	External cause of morbidity: Exposure to other specified factors, initial encounter	0.02
3.	PAY1	Expected primary payer	0.02
4.	PRn_3E033GC	Procedure code: Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach	0.02
5.	ECM_Y95	External cause of morbidity: Nosocomial condition	0.02
6.	DRG	Diagnosis Related Group	0.02
7.	APRDRG	All Patient Refined Diagnosis Related Groups	0.01
8.	ECM_Y939	External cause of morbidity: Activity, unspecified	0.01
9.	ECM_Y92019	External cause of morbidity: Unspecified place in single-family (private) house as the place of occurrence of the external cause	0.01
10.	ECM_Y832	External cause of morbidity: Surgical operation with anastomosis, bypass, or graft as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure	0.01

### Experiment Observations

The LIME explanation metrics are stability average 0.4 and local fit (R2) averaging 0.24 were low.

### **Decision Tree (DT)**

The “*class\_weight*” parameter is set to “balanced” to automatically adjust weight and increase the penalty for misclassifying the minority class. The search for the best fit hyperparameters was done through scikit-learn’s GridSearchCV. The grid search parameter “*cv*” was set 5 to the K-fold cross validation, the “*scoring*” was set to



“*roc\_auc*” so that the search for optimal parameters is optimized to maximize the Area and the ROC curve score. The grid search hyperparameter search and resulting best fit hyperparameters were set to the following values:

<b>Classifier Hyperparameters Grid Search</b>	<b>Best Fit Model Hyperparameters</b>
<ul style="list-style-type: none"> <li>• criterion = ['gini','entropy']</li> <li>• splitter = ['best','random']</li> <li>• max_depth = [4,5,6,7,8,9,10,11,12,15,20,30,40,50,70,90,120,150]</li> </ul>	<ul style="list-style-type: none"> <li>• criterion=gini</li> <li>• splitter=best</li> <li>• max_depth=8</li> </ul>

#### Accuracy Metrics

- Area under ROC curve: 0.5625
- Accuracy: 0.5344
- Weighted F1 score: 0.5695

<b>Class</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
Class 0 – Readmit <= 30	0.8135	0.5098	0.6268
Class 1 – Not Readmit <= 30	0.2760	0.6125	0.3810

#### Confusion Matrix

<b>True Label</b>	<b>Predicted Label</b>	
	0	1
0	36,023	34,637
1	8,259	13,204

#### LIME Explanation Metrics

<b>Instance Selection</b>	<b>Stability</b>	<b>R<sup>2</sup></b>
500 Randomly Selected Test Instances	0.6927	0.4977
500 Randomly Selected Test Instances (Correctly Predicted)	0.6932	0.4969
500 Randomly Selected Test Instances (Incorrectly Predicted)	0.6935	0.4980

### Global Model Generated Feature Importance

Rank	Feature	Description	Feature Importance
1.	ECM_Y95	External cause of morbidity: Nosocomial condition	0.07
2.	ECM_X58XXXA	External cause of morbidity: Exposure to other specified factors, initial encounter	0.05
3.	DRG	Diagnosis Related Group	0.04
4.	ECM_Y92230	External cause of morbidity: Patient room in hospital as the place of occurrence of the external cause	0.04
5.	APRDRG	All Patient Refined Diagnosis Related Group	0.04
6.	PRn_5A2204Z	Procedure Code: Restoration of Cardiac Rhythm, Single	0.03
7.	AGE	Age of patient	0.02
8.	PAY1	Expected primary payer	0.02
9.	ECM_Y92019	External cause of morbidity: Unspecified place in single-family (private) house as the place of occurrence of the external cause	0.02
10.	ECM_W19XXXA	External cause of morbidity: Unspecified fall, initial encounter	0.02

### Experiment Observations

The LIME explanation metrics are stability average 0.69 and local fit (R2) averaging 0.49 were better for a DT generated model than for other models.

### **LightGBM (GBM)**

The “*is\_unbalance*” parameter is set to “true” to indicate to the classifier that the training dataset is not balanced. The “*objective*” parameter is set to “binary” to indicate the setup of a binary classifier. The search for the best fit hyperparameters was done through scikit-learn’s GridSearchCV. The grid search parameter “*cv*” was set 5 to the K-fold cross validation, the “*scoring*” was set to “roc\_auc” so that the search for optimal parameters is optimized to maximize the Area and the ROC curve score. The grid search hyperparameter search and resulting best fit hyperparameters were set to the following values:

Classifier Hyperparameters Grid Search	Best Fit Model Hyperparameters
<ul style="list-style-type: none"> <li>• colsample_bytree = [0.69,1,1.25]</li> <li>• learning_rate = [0.5, 0.10, 0.15]</li> <li>• n_estimators = [50,100,115,116,120]</li> <li>• num_leaves = [9,10,11,15,20]</li> <li>• subsample = [0.74,1,1.25]</li> </ul>	<ul style="list-style-type: none"> <li>• colsample_bytree=0.69</li> <li>• learning_rate=0.15</li> <li>• max_depth=3 (default value)</li> <li>• n_estimators=116</li> <li>• num_leaves=9</li> <li>• subsample=0.74</li> </ul>

#### Accuracy Metrics

- Area under ROC curve: 0.5737
- Accuracy: 0.5664
- Weighted F1 score: 0.5994

Class	Precision	Recall	F1-Score
Class 0 – Not Readmit <= 30	0.8155	0.5599	0.6639
Class 1 – Readmit <= 30	0.2909	0.5876	.3891

#### Confusion Matrix

True Label	Predicted Label	
	0	1
0	61,830	8,879
1	16,839	4,575

#### LIME Explanation Metrics

Instance Selection	Stability	R <sup>2</sup>
500 Randomly Selected Test Instances	0.416	0.511
500 Randomly Selected Test Instances (Correctly Predicted)	0.418	0.506
500 Randomly Selected Test Instances (Incorrectly Predicted)	0.418	0.509

#### Global Model Generated Feature Importance

Rank	Feature	Description	Feature Importance
1.	PRn_3E033GC	Procedure code: Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach	32
2.	PAY1	Expected primary payer	32

Rank	Feature	Description	Feature Importance
3.	ECM_Y92230	External cause of morbidity: Patient room in hospital as the place cause	22
4.	DRG	Diagnosis Related Group	19
5.	HOSP_BEDSIZE	Hospital Bed Size	13
6.	ECM_Y92019	External cause of morbidity:	12
7.	LOS	Length of Stay	11
8.	AGE	Age of patient	11
9.	ECM_Y939	External cause of morbidity:	10
10.	APRDRG	All Patient Refined Diagnosis Related Groups	10

### Google Cloud Platform (GCP) – Auto-ML

The consolidated and filtered input table was provided to the model as input with no feature engineering performed. The optimization objective was set to AUC ROC.

#### Accuracy Metrics

- Area under ROC curve: 0.616
- Accuracy: 0.766
- Weighted F1 score: 0.6664

Class	Precision	Recall	F1-Score
Class 0 – Not Readmit $\leq 30$	0.766	0.999	0.867
Class 1 – Readmit $> 30$	0.564	0.003	0.006

#### Confusion Matrix

True Label	Predicted Label	
	0	1
0	35233	24
1	10771	21

#### Global Model Generated Feature Importance

Rank	Feature	Description	Feature Importance
1.	AGE	Age of the patient	0.1072
2.	IDXn_LR	Non-primary diagnosis condition: Liver or Renal Failure/Disease	0.1037
3.	LOS	Length of stay	0.1014

Rank	Feature	Description	Feature Importance
4.	IDXn_COPD	Non-primary diagnosis condition: Chronic Obstructive Pulmonary Disease	0.0664
5.	I10_NDX	Number of ICD-10-CM diagnoses coded on the record	0.0560
6.	DISPUNIFORM	Disposition of patient, uniform coding	0.0527
7.	PAY1	Expected primary payer	0.0420
8.	HCUP_ED	HCUP indicator of emergency department record	0.0385
9.	ZIPINC_QRTL	Median household income for patient's ZIP Code	0.0382
10.	IDXn_PCI	Non-primary diagnosis condition: Presence of coronary/cardiac implant and Percutaneous Coronary Intervention	0.0338

### Experiment Observations

The brute force approach of the GCP-Auto-ML solution required 46 node hours of model training and resulted in a highly biased classifier. Although the classifier had a higher AUC score than the other trained models, this was accomplished by almost predicting all labels belonging to the majority class. The majority class (Class 0) had a recall score of 0.999, and the minority class had a recall score of 0.003.

### **Classifiers Summary Results**

#### Accuracy and Local Explanation Metrics Summary

Model	AUC	Accuracy	Weighted F1 Score	F1 Class 0	F1 Class 1	Stability	R <sup>2</sup>
LR	0.5749	0.5732	0.6054	0.6719	0.3898	0.4328	0.2471
DT	0.5625	0.5344	0.5695	0.6268	0.3820	0.6927	0.4977
RF	0.5767	0.5740	0.6068	0.6731	0.3888	0.4042	0.2453
GBM	0.5737	0.5664	0.5994	0.6639	0.3891	0.4161	0.511
GCP Auto-ML	0.616	0.766	0.6664	0.867	0.006	*	*

\* *Explanatory metrics were not calculated*

The accuracy metrics for the classifiers had AUC results in the range of (0.57 to 0.61) and accuracy results ranging from (0.57 to 0.76). Accuracy results are consistent with previously reported readmission studies as summarized below:

Source	Classifier	Dataset	Instances	AUC
(Bayati et al., 2014)	Logistic Regression with LASSO	Hospital EHR	1,172	0.66
(Yang et al., 2016)	Logistic Regression with LASSO	NRD 2015	142,527	0.657
(Yang et al., 2016)	GBM	NRD 2015	142,527	0.663
(Yang et al., 2016)	DNN	NRD 2015	142,527	0.662
(Allam et al., 2019)	Logistic Regression with LASSO	NRD 2013	272,778	0.643
(Allam et al., 2019)	Recurrent Neural Networks combined with Conditional Random Fields	NRD 2013	272,778	0.642
(Liu et al., 2020)	Hierarchical Logistic Regression	NRD 2014	303,233	0.580
(Liu et al., 2020)	XGBoost	NRD 2014	303,233	0.602
(Liu et al., 2020)	Feed-Forward Neural Networks	NRD 2014	303,233	0.604
(Liu et al., 2020)	Medical Code Embedding Deep Set Architecture	NRD 2014	303,233	0.618

LIME explanation stability range [0.40 -0.69], and local fit ( $R^2$ ) [0.24 - 0.51].

### Global Model Generated Feature Importance Summary

LR	RF	DT	GBM	GCP-AutoML
IDXn_CTD (1)	ECM_Y92230 (3)	ECM_Y95(2)	PRn_3E033GC (3)	AGE (4)
PL_NCHS (1)	ECM_X58XXXXA (3)	ECM_X58XXXXA (3)	PAY1(4)	IDXn_LR (1)
PRn_3E033GC (3)	PAY1(4)	DRG (3)	ECM_Y92230 (3)	LOS (2)
APRDRG (4)	PRn_3E033GC (3)	ECM_Y92230 (3)	DRG (3)	IDXn_COPD (1)
ECM_X58XXXXA (3)	ECM_Y95 (2)	APRDRG (4)	HOSP_BEDSIZE (1)	I10_NDX (1)
PRn_02HV33Z (1)	DRG (3)	PRn_5A2204Z (1)	ECM_Y92019 (3)	DISPUNIFORM (1)
PRn_4A023N6 (1)	APRDRG (4)	AGE (4)	LOS (2)	PAY1 (4)
AGE (4)	ECM_Y939 (2)	PAY1(4)	AGE (4)	HCUP_ED (1)
ECM_Y92238(1)	ECM_Y92019 (3)	ECM_Y92019 (3)	ECM_Y939 (2)	ZIPINC_QRTL (1)
IDXn_PCI (2)	ECM_Y832 (1)	ECM_W19XXXXA (1)	APRDRG (4)	IDXn_PCI (2)
5 of the top 10 features are unique to the model	1 of the top 10 features is unique to the model	1of the top 10 features is unique to the model	1of the top 10 features is unique to the model	6 of the top 10 features are unique to the model

\* The numeric value next to the feature label (X) range from 1 to 5 and indicates the number of times a feature has been identified as a top 10 global feature in the five models used in this experiment.

### Summary

This chapter presented experimental results of supervised machine learning developed to predict the risk of emergency readmissions within 30 days of being

discharged for heart failure patients. Tuned hyperparameters and the accuracy metrics of F1 score, precision, recall, the area under the ROC curve for each model were reported. Local Interpretable Model-Agnostic Explanations (LIME) sample explanations were provided demonstrating the value and the limitations of local explanations for the prediction task. Explanation metrics were reported.

## Chapter 5

### Conclusions, Implications, Recommendations, and Summary

#### Overview

Previous chapters outlined the goals, method, and results of this dissertation. This chapter draws the conclusions of this dissertation relative to the research questions and in the context of reviewed literature. Implications of the findings, recommendations and future research direction are summarized. The chapter concludes with a summary of the dissertation.

#### Conclusions

The evaluation of experimental results was be guided by answers to the following research questions:

*Research Question 1 (RQ1): Can the ML predictions generate intelligible results to guide clinical decision-making?*

The combination of global feature importance generated by individual models and model agnostic feature importance generated by LIME provided insight explaining the logic of the model.

*Research Question 2 (RQ2): What are the most useful features in predicting hospital readmissions for heart failure patients?*

This study demonstrated that training five different classifiers capable of reporting global feature importance resulted in agreement between the classifiers on a subset of the features. Agreement in this context, refers to the same feature being reported by at least 3 out of the 5 models as a top globally important 10 feature for the model. The following features meet this criterion:



<b>Feature</b>	<b>Description</b>
AGE	Age of the patient
APRDRG	All Patient Refined Diagnosis Related Groups
DRG	Diagnosis Related Group
ECM_X58XXXA	External cause of morbidity: Exposure to other specified factors, initial encounter
ECM_Y92019	External cause of morbidity: Unspecified place in single-family (private) house as the place of occurrence of the external cause
ECM_Y92230	External cause of morbidity: Patient room in hospital as the place of occurrence of the external cause
PAY1	Expected primary payer
PRn_3E033GC	Procedure code: Introduction of Other Therapeutic Substance into Peripheral Vein, Percutaneous Approach

*Research Question 3 (RQ3): Will using the model-agnostic explanatory method (LIME) generate high-quality explanations as measured by explanation stability and local fit?*

LIME explanation stability ranged from (0.40 to 0.69), and local fit (R<sup>2</sup>) ranged from 0.24 to 0.51. The results demonstrated that local explanations generated by LIME created better estimates for Decision Trees (DT) classifiers as shown below:

<b>Model</b>	<b>Stability</b>	<b>R<sup>2</sup></b>
LR	0.4328	0.2471
DT	0.6927	0.4977
RF	0.4042	0.2453
GBM	0.4161	0.511

### **Implications**

The use of Cost and Utilization Project (HCUP) Nationwide Readmissions Database (NRD) to predict hospital readmissions for heart failure patients resulted in binary classifiers of moderate accuracy (AUC [0.57-0.61]). While the dataset contains some clinical features, they are limited into diagnosis and procedures reported as a result

of hospital admission records. It is plausible that better predictors can be constructed with access to clinical notes and medical history.

LIME explanation stability [0.40 - 0.69], and local fit ( $R^2$ ) [0.24 - 0.51] results were poor to moderate using LIME's default hyper-parameters. An interesting future research direction would be to attempt to optimize LIME's stability and local fit metrics through a systematic search of LIME's hyper-parameter space (kernel width and number of samples).

### **Recommendations**

The results demonstrated that local explanations generated by LIME created better estimates for Decision Trees (DT) classifiers with an accuracy metrics that are nearly identical more complex model such Random Forests (RF) and Gradient Boosting Machines (GBM). Accordingly, the use of Decision Trees (DT) classifiers is recommended due to ability to higher quality local explanations by LIME.

### **Summary**

This dissertation investigated the applicability of interpretable model-agnostic methods to explain predictions of black-box machine learning models for medical decision-making.

Supervised classification models of differing complexity were trained to perform the prediction task. Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Gradient Boosting Machines (GBM) models were constructed using the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database (NRD). The precision, recall, area under the ROC curve for each model were used to measure predictive accuracy. Local Interpretable Model-Agnostic Explanations (LIME) was used

to generate explanations from the underlying trained models. LIME explanations were empirically evaluated using explanation stability and local fit ( $R^2$ ). The results demonstrated that local explanations generated by LIME created better estimates for Decision Trees (DT) classifiers.

## Appendix A

## 2016 NRD Core File Schema

Category	Data Element Name	Description
Admission/ Discharge	<a href="#">AWEEKEND</a>	Admission on weekend: (0) admission on Monday–Friday, (1) admission on Saturday– Sunday
	<a href="#">DIED</a>	Indicates in-hospital death: 0) did not die during hospitalization, (1) died during hospitalization
	<a href="#">DISPUNIFORM</a>	Disposition of patient, uniform coding: (1) routine, (2) transfer to short term hospital, (5) other transfers, including skilled nursing facility, intermediate care, and another type of facility, (6) home health care, (7) against medical advice, (20) died in hospital, (99) discharged alive, destination unknown
	<a href="#">DMONTH</a>	Coded: (1) Jan; (2) Feb; (3) Mar; (4) Apr; (5) May; (6) Jun; (7) Jul; (8) Aug; (9) Sep; (10) Oct; (11) Nov; (12) Dec;
	<a href="#">DQTR</a>	Coded: (1) Jan–Mar, (2) Apr–Jun, (3) Jul–Sep, (4) Oct–Dec
	<a href="#">ELECTIVE</a>	Indicates elective admission: (1) elective, (0) non-elective admission
	<a href="#">HCUP_ED</a>	Indicator that discharge record includes evidence of emergency department (ED) services: (0) record does not meet any HCUP ED criteria, (1) ED revenue code was on SID record, (2) ED charge reported on SID record, (3) ED CPT procedure code on SID record, (4) other indication of ED services
	<a href="#">DISCWT</a>	Weight to discharges in the universe
	<a href="#">YEAR</a>	Discharge year
Clinical Information	<a href="#">DRG</a>	The Diagnosis Related Group (DRG) in use on discharge date

Category	Data Element Name	Description
	<a href="#">DRG_NoPOA</a>	DRG in use on discharge date, calculated without POA (present on admission)
	<a href="#">DRGVER</a>	Groupver version in use on discharge date
	<a href="#">DXVER</a>	Diagnosis version (indicating ICD-10-CM)
	<a href="#">I10_DX1-</a> <a href="#">I10_DX35</a>	ICD-10-CM diagnoses, principal and secondary
	<a href="#">I10_ECAUSE1-I10</a> <a href="#">ECAUSE4</a>	ICD-10-CM external cause of morbidity codes
	<a href="#">I10_NDX</a>	Number of ICD-10-CM diagnoses coded on the record
	<a href="#">I10_NECAUSE</a>	Number of external causes of morbidity codes on the record
	<a href="#">I10_NPR</a>	Number of procedures coded
	<a href="#">I10_PR1-I10_PR15</a>	ICD-10-PCS (Procedure Coding System) procedures, principal and secondary
	<a href="#">MDC</a>	MDC (Major Diagnostic Category) in use on discharge date
	<a href="#">MDC_NoPOA</a>	MDC assignment made without the use of the present on admission flags for the diagnoses
	<a href="#">PRDAY1-</a> <a href="#">PRDAY15</a>	The day on which the procedure is performed. A value of 0 indicates the day of admission.
	<a href="#">PRVER</a>	Procedure version (indicating ICD-10-PCS)
NRD Identifiers	<a href="#">HOSP_NRD</a>	NRD hospital identifier specific to the NRD and is not linkable to any other HCUP or external databases. HOSP_NRD can be used to add data elements from the Hospital file to records on the discharge-level files. The values of HOSP_NRD differ from year to year. An individual hospital cannot be tracked across data years.
	<a href="#">KEY_NRD</a>	Unique record identifier for the discharge in the NRD and not linkable to any other HCUP or external databases.

Category	Data Element Name	Description
		<p>KEY_NRD can be used to add data elements from the Severity and Diagnosis/Procedure Groups files to the records on the Core file within the same data year. The values of KEY_NRD are different in each data year 2010–2012 and 2015–2016 but are nonunique between 2013 and 2014.</p> <p>Please note that KEY_NRD is a record identifier and not a patient linkage number. NRD_VISITLINK is the patient linkage number specific to the NRD.</p>
Patient Demographics	<a href="#">AGE</a>	Age in years coded 0-90 years; any age greater than 90 was set to 90. Missing age was imputed using other records with the same patient linkage number. In the 2016 NRD, about 2,000 discharges (0.011 percent) had the age imputed.
	<a href="#">FEMALE</a>	Indicates sex: (0) male, (1) female. Missing sex was imputed using other records with the same patient linkage number. In the 2016 NRD, about 1,000 discharges (0.006 percent) had the sex imputed.
	<a href="#">PAY1</a>	Expected primary payer, uniform: (1) Medicare, (2) Medicaid, (3) private insurance, (4) self-pay, (5) no charge, (6) other
	<a href="#">PL_NCHS</a>	Patient location: National Center for Health Statistics (NCHS) urban-rural classification scheme for U.S. counties: (1) "Central" counties of metro areas of >=1 million population, (2) "Fringe" counties of metro areas of >=1 million population, (3) Counties in metro areas of 250,000–999,999 population, (4) Counties in metro areas of 50,000–249,999 population, (5) Micropolitan counties, (6) Not metropolitan or micropolitan counties

Category	Data Element Name	Description
	<a href="#">ZIPINC_QRTL</a>	<p>Median household income quartiles for patient's ZIP Code: (1) quartile 1 [lowest income], (2) quartile 2, (3) quartile 3, (4) quartile 4 [highest income].</p> <p>For 2016, the median income quartiles are defined as: (1) \$1–\$42,999; (2) \$43,000– \$53,999; (3) \$54,000–\$70,999; and (4) \$71,000 or more.</p>
Readmission Specific	<a href="#">DMONTH</a>	Discharge month coded from (1) January to (12) December
	<a href="#">NRD_DaysToEvent</a>	Count of days from randomly selected "start date" to admission date coded differently for each value of NRD_VisitLink
	<a href="#">LOS</a>	Length of stay (LOS) is calculated by subtracting the admission date (ADATE) from the discharge date (DDATE).
	<a href="#">SAMEDAYEVENT</a>	<p>One of two data elements that identify transfers, same-day stays, and combined transfer records in the NRD.</p> <p>Readmission analyses do not usually allow the hospitalization at the receiving hospital to be counted as a readmission. To eliminate this possibility, pairs of records representing a transfer are collapsed into a single "combined" record in the NRD.</p>
	<a href="#">NRD_VisitLink</a>	Patient linkage number specific to the NRD and not linkable to any other HCUP or external databases. The values of NRD_VISITLINK differ from year to year. An individual person cannot be tracked across data years.
	<a href="#">REHABTRANSFER</a>	A combined record involving transfer to rehabilitation, evaluation, or other aftercare: (1) yes, (0) no
	<a href="#">RESIDENT</a>	Identifies patient as a resident of the State in which he or she received hospital care: (1) resident, (0) non-resident

<b>Category</b>	<b>Data Element Name</b>	<b>Description</b>
	<a href="#">TOTCHG</a>	Total charges. Values are rounded to the nearest dollar



## Appendix B

## 2016 NRD Severity Measures Schema

Category	Data Element Name	Description
3M APR-DRG	<a href="#">APRDRG</a>	3M All Patient Refined DRG (Diagnosis Related Groups)
	<a href="#">APRDRG_Risk_Mortality</a>	3M All Patient Refined DRG: Risk of Mortality Subclass: (0) No class specified, (1) Minor likelihood of dying, (2) Moderate likelihood of dying, (3) Major likelihood of dying, (4) Extreme likelihood of dying
	<a href="#">APRDRG_Severity</a>	3M All Patient Refined DRG: Severity of Illness Subclass: (0) No class specified, (1) Minor loss of function (includes cases with no comorbidity or complications), (2) Moderate loss of function, (3) Major loss of function, (4) Extreme loss of function
	<a href="#">HOSP_NRD</a>	NRD hospital identifier specific to the NRD and is not linkable to any other HCUP or external databases. HOSP_NRD can be used to add data elements from the Hospital file to records on the discharge-level files. The values of HOSP_NRD differ from year to year. An individual hospital cannot be tracked across data years.
	<a href="#">KEY_NRD</a>	Unique record identifier for the discharge in the NRD and not linkable to any other HCUP or external databases. KEY_NRD can be used to add data elements from the Severity and Diagnosis/Procedure Groups files to the records on the Core file within the same data year.

## Appendix C

## 2016 NRD Hospital File Schema

Category	Data Element Name	Description
Admission/ Discharge	<a href="#">YEAR</a>	Discharge year
Hospital Information	<a href="#">H_CONTRL</a>	Control/ownership of hospital: (1) government, nonfederal [public], (2) private, not-for-profit [voluntary], (3) private, investor-owned [proprietary]
	<a href="#">HOSP_BEDSIZE</a>	Size of hospital based on the number of beds: (1) small, (2) medium, (3) large. The categories are defined using region of the U.S., the urban-rural designation of the hospital, in addition to the teaching status.
	<a href="#">HOSP_UR_TEACH</a>	Teaching status of hospital: (0) metropolitan non-teaching, (1) metropolitan teaching, (2) non-metropolitan
	<a href="#">HOSP_URCAT4</a>	Hospital urban-rural location: (1) large metropolitan areas with at least 1 million residents, (2) small metropolitan areas with less than 1 million residents, (3) micropolitan areas, (4) not metropolitan or micropolitan, (8) metropolitan, collapsed category of large and small metropolitan, (9) non-metropolitan, collapsed category of micropolitan and rural
	<a href="#">NRD_STRATUM</a>	NRD stratum for post-stratification based on geographic region, urban/rural location, teaching status, bed size, and control. Region is not identified. The values of NRD_STRATUM differ from year to year. An individual stratum cannot be tracked across data years.

Category	Data Element Name	Description
NRD Identifiers	<a href="#">HOSP_NRD</a>	NRD hospital identifier specific to the NRD and is not linkable to any other HCUP or external databases. The values of HOSP_NRD differ from year to year. An individual hospital cannot be tracked across data years.
Weighting	<a href="#">N_DISC_U</a>	Number of discharges in the target universe in the stratum
	<a href="#">N_HOSP_U</a>	Number of hospitals in the target universe in the stratum
	<a href="#">N_DISC_U</a>	Number of NRD discharges in the stratum
	<a href="#">N_HOSP_U</a>	Number of NRD hospitals in the stratum
	<a href="#">TOTAL_DISC</a>	Total number of discharges for this hospital in the NRD
	<a href="#">S_DISC_U</a>	Total number of inpatient discharges for the stratum
	<a href="#">S_HOSP_U</a>	Total number of hospitals in the stratum

## Appendix D

### 2016 NRD File Specifications

#### Core File

- Data Set Name: NRD\_2016\_CORE
- Number of Records: 17,197,683
- Number of Data Elements: 103
- [Record layout NRD 2016 Core](#)

#### Hospital File

- Data Set Name: NRD\_2016\_HOSPITAL
- Number of Records: 2,355
- Number of Data Elements: 12
- [Record layout NRD 2016 Hospital](#)

#### Severity Measure File

- Data Set Name: NRD\_2016\_SEVERITY
- Number of Records: 17,197,683
- Number of Data Elements: 5
- [Record layout NRD 2016 Severity](#)

## Appendix E

## ICD-10 Code Mapping

**Heart Failure (HF)**

<b>ICD-10 Codes</b>	<b>Description</b>
I50; I50.1-I50.9	Heart failure
I11.0; I11-I11.9	Heart failure due to hypertension
I13.0	Heart failure due to hypertension with chronic kidney disease
I13.2	Heart failure due to hypertension with chronic kidney disease
I97.130- I97.131	Heart failure following surgery
I09.81	Rheumatic heart failure
P29.0	Neonatal cardiac failure
i46.2- i46.9	Cardiac arrest

**Chest Pain (CP)**

<b>ICD-10 Codes</b>	<b>Description</b>
R07.1 -R07.9	Chest Pain

**Hypertension (HT)**

<b>ICD-10 Codes</b>	<b>Description</b>
I10	Hypertension
I12-I12.9	Hypertension with chronic kidney disease
I15-I15.9	Secondary hypertension
I16.0-I16.9	Hypertensive crisis
H35-H35.09	Essential (primary) hypertension involving vessels of eye
O10-O11.9	Hypertensive disease complicating pregnancy
O13-O13.69	Hypertensive disease complicating pregnancy
I13.0-i13.2	Heart failure due to hypertension with chronic kidney disease
I60-I69.998	Essential (primary) hypertension involving vessels of brain

**Shock (SH)**

<b>ICD-10 Codes</b>	<b>Description</b>
R57-R57.9	Shock
T78.2-T78.2XXS	Anaphylactic shock, unspecified
T78.0-T78.09XS	Anaphylactic reaction or shock due to adverse food reaction
T80.5-T80.59XS	Anaphylactic shock due to serum
T88.6-T88.6XXS	Anaphylactic shock due to adverse effect of correct drug or medicament properly administered
T75.4-T75.4XXS	Electric shock
O75.1	Obstetric shock
T81.1-T81.19XS	Postprocedural shock
F43.0	Psychic shock
O00-O07.4	Shock complicating or following ectopic or molar pregnancy
O08.3	Shock due to lightning
T75.0-T75.09XS	Traumatic shock
T79.4-T79.4XXS	Traumatic shock
A48.3	Toxic shock syndrome

**Diabetes (DB)**

<b>ICD-10 Codes</b>	<b>Description</b>
E08-E08.9	Diabetes mellitus due to underlying condition
E09-E09.9	Drug or chemical induced diabetes mellitus
E10-E10.9	Type 1 diabetes mellitus
E11-E11.9	Type 2 diabetes mellitus
E11.22	Type 2 diabetes mellitus with diabetic chronic kidney disease
E13-E13.9	Other specified diabetes mellitus

**Prior Percutaneous Coronary Intervention (PCI)**

<b>ICD-10 Codes</b>	<b>Description</b>
Z95.1-Z95.5	Presence of cardiac and vascular implants and grafts
Z95.818	Presence of coronary angioplasty implant and graft
Z95.82-Z95.9	Presence of other cardiac and vascular implants and grafts

**Stroke Ischemia (STR)**

<b>ICD-10 Codes</b>	<b>Description</b>
I63-I63.9	Cerebral infarction
P91.82-P91.829	Neonatal cerebral infarction

**Chronic Obstructive Pulmonary Disease (COPD)**

<b>ICD-10 Codes</b>	<b>Description</b>
J40	Bronchitis, not specified as acute or chronic
J41	Simple and mucopurulent chronic bronchitis
J42	Unspecified chronic bronchitis
J43	Emphysema
J44	Other chronic obstructive pulmonary disease
J45	Asthma
J47	Bronchiectasis

**Peptic Ulcer (ULC)**

<b>ICD-10 Codes</b>	<b>Description</b>
K27-K27.9	Peptic ulcer, site unspecified, unspecified as acute or chronic, without hemorrhage or perforation
P78.82	Peptic ulcer of newborn

**Dementia (DM)**

<b>ICD-10 Codes</b>	<b>Description</b>
F03-F03.91	Dementia

**Myocardial Infarction (MI)**

<b>ICD-10 Codes</b>	<b>Description</b>
i21-i21.9	Acute myocardial infarction
I21.A-I21.A9	Other type of myocardial infarction
I22-I22.9	Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction
I23-I23.8	Certain current complications following ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction (within the 28 day period)
I25.2	Prior MI-Old myocardial infarction

**Cardiovascular System Disease (CVS)**

<b>ICD-10 Codes</b>	<b>Description</b>
I05	Rheumatic mitral valve diseases
I06	Rheumatic aortic valve diseases
I07	Rheumatic tricuspid valve diseases
I08	Multiple valve diseases
I09	Other rheumatic heart diseases
I11	Hypertensive heart disease
I13	Hypertensive heart and chronic kidney disease
I20	Angina pectoris
I21	Acute myocardial infarction
I22	Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction
I23	Certain current complications following ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction (within the 28 day period)
I24	Other acute ischemic heart diseases
I25	Chronic ischemic heart disease
I26	Pulmonary embolism



<b>ICD-10 Codes</b>	<b>Description</b>
I27	Other pulmonary heart diseases
I28	Other diseases of pulmonary vessels
I30	Acute pericarditis
I31	Other diseases of pericardium
I32	Pericarditis in diseases classified elsewhere
I33	Acute and subacute endocarditis
I34	Nonrheumatic mitral valve disorders
I35	Nonrheumatic aortic valve disorders
I36	Nonrheumatic tricuspid valve disorders
I37	Nonrheumatic pulmonary valve disorders
I38	Endocarditis, valve unspecified
I39	Endocarditis and heart valve disorders in diseases classified elsewhere
I40	Acute myocarditis
I41	Myocarditis in diseases classified elsewhere
I42	Cardiomyopathy
I43	Cardiomyopathy in diseases classified elsewhere
I44	Atrioventricular and left bundle-branch block
I45	Other conduction disorders
I46	Cardiac arrest
I47	Paroxysmal tachycardia
I48	Atrial fibrillation and flutter
I49	Other cardiac arrhythmias
I51	Complications and ill-defined descriptions of heart disease
I52	Other heart disorders in diseases classified elsewhere

**Pulmonary Valve Stenosis (PVS)**

<b>ICD-10 Codes</b>	<b>Description</b>
I70	Rheumatic mitral valve diseases
I71	Aortic aneurysm and dissection
I72	Other aneurysm
I73	Other peripheral vascular diseases
I74	Arterial embolism and thrombosis
I75	Atheroembolism
I76	Septic arterial embolism
I77	Other disorders of arteries and arterioles
I78	Diseases of capillaries
I79	Disorders of arteries, arterioles and capillaries in diseases classified elsewhere

**Connective Tissue Disease (CTD)**

<b>ICD-10 Codes</b>	<b>Description</b>
L94-L94.9	Other localized connective tissue disorders
M30-M30.8	Polyarteritis nodosa and related conditions
M31-M31.9	Other necrotizing vasculopathies
M32-M32.9	Systemic lupus erythematosus (SLE)
M33-M33.9	Dermatopolymyositis
M34- M34.9	Systemic sclerosis [scleroderma]
M35-M35.9	Systemic disorders of connective tissue
M36- M36.8	Other systemic involvement of connective tissue

**Acquired immunodeficiency syndrome (AIDS)**

<b>ICD-10 Codes</b>	<b>Description</b>
B20	Human Immunodeficiency Virus (HIV)

**Liver or Renal Failure/Disease (LR)**

<b>ICD-10 Codes</b>	<b>Description</b>
K70-K70.9	Alcoholic liver disease
K71-K71.9	Toxic liver disease
K72-K72.9	Hepatic failure
K73-K73.9	Chronic hepatitis, not elsewhere classified
K74-K74.9	Fibrosis and cirrhosis of liver
K75-K75.9	Other inflammatory liver diseases
K76-K76.9	Other diseases of liver
K77-K77.9	Liver disorders in diseases classified elsewhere
N18.1-N18.6	Renal Failure/Chronic Kidney disease
E08.22	Diabetes mellitus due to underlying condition with diabetic chronic kidney disease
E13.2-E13.29	Other specified diabetes mellitus with kidney complications
Z94.0	Kidney transplant status
I12-I12.9	Hypertensive chronic kidney disease
I13.0-I13.2	Hypertensive heart and chronic kidney disease

## References

- Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). *Learning from data: A short course*. (Vol. 4). AMLBook.
- Adadi, A., & Berrada, M. (2018). Peeking inside the Black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160.  
<https://doi.org/10.1109/access.2018.2870052>
- Adhikari, A., Tax, D. M., Satta, R., & Faeth, M. (2019). LEAFAGE: Example-based and feature importance-based explanations for black-box ML models. *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.  
<https://doi.org/10.1109/fuzz-ieee.2019.8858846>
- Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016). Auditing black-box models for indirect influence. *2016 IEEE 16th International Conference on Data Mining (ICDM)*.  
<https://doi.org/10.1109/icdm.2016.0011>
- Agarwal C., Nguyen A. (2021). Explaining image classifiers by removing input features using generative models. *Proceedings of the Asian Conference on Computer Vision*. [https://doi.org/10.1007/978-3-030-69544-6\\_7](https://doi.org/10.1007/978-3-030-69544-6_7)
- Agency for Healthcare Research and Quality. (2018). *HCUP Databases. Healthcare Cost and Utilization Project (HCUP), Nationwide Readmissions Database (NRD)* [Data set]. Healthcare Cost and Utilization Project (HCUP). [www.hcup-us.ahrq.gov/databases.jsp](http://www.hcup-us.ahrq.gov/databases.jsp)

- Agency for Healthcare Research and Quality. (2017). *Clinical Classifications Software (CCS) for ICD-9-CM* [Computer software]. Healthcare Cost and Utilization Project (HCUP): <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>
- Agency for Healthcare Research and Quality. (2018). *HCUP summary statistics report: NRD 2016 core file means of continuous data elements* [Data set]. Healthcare Cost and Utilization Project (HCUP). <https://www.hcup-us.ahrq.gov/db/nation/nrd/nrdsummstats.jsp#2016>
- Ahmad, M. A., Eckert, C., Teredesai, A., & McKelvey, G. (2018, August 15). *Interpretable machine learning in healthcare* [Conference session]. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 559-560.  
<https://doi.org/10.1145/3233547.3233667>
- Allam, A., Nagy, M., Thoma, G., & Krauthammer, M. (2019). Neural networks versus Logistic regression for 30 days all-cause readmission prediction. *Scientific reports*, 9(1), 1-11. <https://doi.org/10.1038/s41598-019-45685-z>
- Alvarez-Melis, D., & Jaakkola, T. (2018b). Towards robust interpretability with self-explaining neural networks. *32nd Conference on Neural Information Processing Systems* [Conference proceeding]. 7775-7784. Montréal, Canada: NeurIPS.  
<https://arxiv.org/pdf/1806.07538.pdf>
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1), 8753-8830. <https://doi.org/10.1145/3097983.3098047>

- Arlot, S. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, 40-79. <https://doi.org/10.1214/09-SS054>
- Arras, L., Montavon, G., Muller, K. R., & Samek, W. (September, 2017). *Explaining recurrent neural network predictions in sentiment analysis* [Conference proceeding]. In Association for Computational Linguistics, Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 159–168. <https://doi.org/10.18653/v1/W17-5221>
- Arundel, C., Lam, P. H., Khosla, R., Blackman, M. R., Fonarow, G. C., Morgan, C., Zeng, Q., Fletcher, R. D., Butler, J., Wu, W., Deedwania, P., Love, T. E., White, M., Aronow, W. S., Anker, S. D., Allman, R. M., & Ahmed, A. (2016). Association of 30-Day all-cause readmission with long-term outcomes in hospitalized older Medicare beneficiaries with heart failure. *The American Journal of Medicine*, 129(11), 1178-1184. <https://doi.org/10.1016/j.amjmed.2016.06.018>
- Asano, K., Chun, J., Koike, A., & Tokuyama, T. (2019). *Model-agnostic explanations for decisions using minimal patterns* [Conference paper]. In Springer, Cham, International Conference on Artificial Neural Networks. [https://doi.org/10.1007/978-3-030-30487-4\\_19](https://doi.org/10.1007/978-3-030-30487-4_19).
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>

- Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., McConnell, M. V., Percha, B., Snyder, T. M., & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2(1). <https://doi.org/10.1038/s41746-019-0105-1>
- Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Muller, K. R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun), 1803-1831. <https://www.jmlr.org/papers/volume11/baehrens10a/baehrens10a.pdf>
- Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of clinical epidemiology*, 54(10), 979-985. [https://doi.org/10.1016/s0895-4356\(01\)00372-9](https://doi.org/10.1016/s0895-4356(01)00372-9)
- Bastani, O., Kim, C., & Bastani, H. (2017). Interpretability via model extraction <https://arxiv.org/pdf/1706.09773.pdf>
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29. <https://doi.org/10.1145/1007730.1007735>
- Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36(1), 105-139. <https://doi.org/10.1023/A:1007515423169>
- Bayati, M., Braverman, M., Gillam, M., Mack, K. M., Ruiz, G., Smith, M. S., & Horvitz, E. (2014). Data-driven decisions for reducing readmissions for heart

failure: General methodology and case study. *PLoS ONE*, 9(10), e109264.

<https://doi.org/10.1371/journal.pone.0109264>

Betancourt, J. R., Tan-McGrory, A., & Kenst, K. S. (2015). *Guide to preventing readmissions among racially and ethnically diverse Medicare beneficiaries.*

Massachusetts General Hospital, Mongan Institute for Health Policy. Centers for Medicare & Medicaid Services Office of Minority Health.

[https://eldercarebroker.com/pub/OMH\\_Readmissions\\_Guide.pdf](https://eldercarebroker.com/pub/OMH_Readmissions_Guide.pdf)

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197-227.

<https://doi.org/10.1007/s11749-016-0481-7>

Bibal, A., & Fréney, B. (2016). Interpretability of machine learning models and

representations: An introduction. In M. Verleysen (Ed.), *ESANN 2016*

*proceedings, 24<sup>th</sup> European Symposium on Artificial Neural Networks,*

*Computational Intelligence and Machine Learning* (pp. 27-29). Bruges

(Belgium). [https://researchportal.unamur.be/en/publications/interpretability-of-](https://researchportal.unamur.be/en/publications/interpretability-of-machine-learning-models-and-representations-a)

[machine-learning-models-and-representations-a](https://researchportal.unamur.be/en/publications/interpretability-of-machine-learning-models-and-representations-a)

Binder, A., Montavon, G., Bach, S., Muller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers.

<https://arxiv.org/pdf/1604.00825.pdf>

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A

survey. [http://www.cs.columbia.edu/~orb/papers/xai\\_survey\\_paper\\_2017.pdf](http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf)

Björklund, A., Henelius, A., Oikarinen, E., Kallone, K., & Puolamäki, K. (2019). *Sparse robust regression for explaining classifiers* [Conference presentation].



- International Conference on Discovery Science (351-366).  
[https://doi.org/10.1007/978-3-030-33778-0\\_27](https://doi.org/10.1007/978-3-030-33778-0_27)
- Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1), 245-271.  
[https://doi.org/10.1016/S0004-3702\(97\)00063-5](https://doi.org/10.1016/S0004-3702(97)00063-5)
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters*, 24(6), 377-380. [https://doi.org/10.1016/0020-0190\(87\)90114-1](https://doi.org/10.1016/0020-0190(87)90114-1)
- Boz, O. (2002). *Extracting decision trees from trained neural networks* [Conference presentation]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, (456-461).  
<https://doi.org/10.1145/775047.775113>
- Bramhall, S., Horn, H., Tieu, M., & Lohia, N. (2020). Qlime-a quadratic local interpretable model-agnostic explanation approach. *SMU Data Science Review*, 3(1), 4. <https://scholar.smu.edu/datasciencereview/vol3/iss1/4>
- Bratko, I. (1997). Machine learning: Between accuracy and interpretability. *Learning, Networks and Statistics*, 163-177. [https://doi.org/10.1007/978-3-7091-2668-4\\_10](https://doi.org/10.1007/978-3-7091-2668-4_10)
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.  
<https://doi.org/10.1007/bf00058655>
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1), 5-32.  
<http://dx.doi.org/10.1023/A:1010933404324>

- Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199-231.  
<https://doi.org/10.1214/ss/1009213726>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth and Brooks. <https://doi.org/10.1002/cyto.990080516>
- Broekhuizen, H., Groothuis-Oudshoorn, C. G., Van Til, J. A., Hummel, J. M., & IJzerman, M. J. (2015). A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics*, 33(5), 445-455. <https://doi.org/10.1007/s40273-014-0251-x>
- Bucila, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (535-541). New York, NY, USA: ACM.  
<https://doi.org/10.1145/1150402.1150464>
- Buhrmester, V., Münch, D., & Arens, M. (2019). Analysis of explainers of black box deep neural networks for computer vision: A survey.  
<https://arxiv.org/pdf/1911.12116.pdf>
- Burkart, N., & Huber, M. F. (2020). A survey on the explainability of supervised machine learning. <https://arxiv.org/pdf/2011.07876.pdf>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Carbonera, J. L., & Abel, M. (2015). *A density-based approach for instance selection* [Conference presentation]. 27th International Conference on Tools with Artificial Intelligence (ICTAI). <https://ieeexplore.ieee.org/document/7372210>

- Caruana, R. (2017). Intelligible machine learning for critical applications such as health care. *In 2017 AAAS Annual Meeting. (February 16-20, 2017)*. aaas.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission [Conference presentation]. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1721-1730)*.  
<https://doi.org/10.1145/2783258.2788613>
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832.  
<https://doi.org/10.3390/electronics8080832>
- Center for Outcomes Research & Evaluation (CORE). (2021). *Readmission risk score for heart attack*. [https://www.readmissionscore.org/heart\\_attack.php](https://www.readmissionscore.org/heart_attack.php)
- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28.  
<https://doi.org/10.1016/j.compeleceng.2013.11.024>
- Che, Z., Purushotham, S., Khemani, R., & Liu, Y. (2015). Distilling knowledge from deep networks with applications to healthcare domain.  
<https://arxiv.org/pdf/1512.03542.pdf>
- Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., & Rudin, C. (2019). This looks like that: Deep learning for interpretable image recognition.  
<https://arxiv.org/pdf/1806.10574.pdf>

- Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation.  
<https://arxiv.org/pdf/1802.07814.pdf>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, 3504-3512.  
<https://arxiv.org/pdf/1608.05745.pdf>
- Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225, 1-17. <https://doi.org/10.1016/j.ins.2012.10.039>
- Covert, I. C., Lundberg, S., & Lee, S. I. (2020). Explaining by removing: A unified framework for model explanation. <https://arxiv.org/pdf/2011.14878.pdf>
- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215-232.  
<https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>
- Craven, M. W., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8, 24-30.  
<https://papers.nips.cc/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf>

- Dabkowski, P., & Gal, Y. (2017). Real time image saliency for black box classifiers. *Advances in Neural Information Processing Systems*, 6967-6976.
- Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(1-4), 131-156.
- Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE symposium on security and privacy (SP)* (598-617). IEEE. <https://doi.org/10.1109/SP.2016.42>
- David Alvarez-Melis, & Jaakkola, T. S. (2018a). On the robustness of interpretability methods. *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. Stockholm, Sweden.
- Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7(4), 277-287. <https://arxiv.org/pdf/1408.5456.pdf>
- Dey, M., & Rautaray, S. S. (2014). Study and analysis of data mining algorithms for healthcare decision support system. *Planning*, 5(6). <http://www.ijcsit.com/docs/Volume%205/vol5issue01/ijcsit20140501100.pdf>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. <https://arxiv.org/pdf/1702.08608.pdf>
- Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68-77. <https://doi.org/10.1145/3359786>
- Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple Resampling method for learning from Imbalanced data sets. *Computational Intelligence*, 20(1), 18-36. <https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x>

- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905. <https://doi.org/10.1613/jair.1.11192>
- Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, (3429-3437). <https://doi.org/10.1109/iccv.2017.371>
- Fong, R., Patrick, M., & Vedaldi, A. (2019). Understanding deep networks via extremal perturbations and smooth masks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. <https://doi.org/10.1109/iccv.2019.00304>
- Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering OnLine*, 13(1), 94. <https://doi.org/10.1186/1475-925x-13-94>
- Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. <https://hdl.handle.net/10289/1047>
- Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1), 1-10. <https://doi.org/10.1145/2594473.2594475>
- Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *European conference on computational*

- learning theory* (pp. 23-37). Berlin, Heidelberg: Springer.
- [https://doi.org/10.1007/3-540-59119-2\\_166](https://doi.org/10.1007/3-540-59119-2_166)
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. 148-156.
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 916-954. <https://doi.org/10.1214/07-aos148>
- Friedman, J., Hastie, T., & Tibshiran, R. (2001). *The elements of statistical learning* (Vol. 1). Berlin: Springer.
- Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. <https://arxiv.org/pdf/1711.09784.pdf>
- Frye, C., Rowat, C., & Feige, I. (2019). Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. <https://arxiv.org/pdf/1910.06358.pdf>
- Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging, boosting, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 463-484. <https://doi.org/10.1109/tsmcc.2011.2161285>
- García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98, 1-29. <https://doi.org/10.1016/j.knosys.2015.12.006>

- Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 3681-3688. <https://doi.org/10.1609/aaai.v33i01.33013681>
- Ghosh, D., & Chinnaiyan, A. M. (2005). Classification and selection of biomarkers in Genomic data using LASSO. *Journal of Biomedicine and Biotechnology*, 2005(2), 147-154. <https://doi.org/10.1155/jbb.2005.147>
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of Interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. <https://doi.org/10.1109/dsaa.2018.00018>
- Goldfield, N. I., McCullough, E. C., Hughes, J. S., Tang, A. M., Eastman, B., Rawlins, L. K., & Averill, R. F. (2008). Identifying potentially preventable readmissions. *Health Care Financing Review*, 30(1), 75.
- Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a "right to explanation". <https://arxiv.org/pdf/1606.08813.pdf>
- Gosiewska, A., & Biecek, P. (2019). iBreakDown: Uncertainty of model explanations for non-additive Predictive Models. *arXiv preprint arXiv:1903.11420*.
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018b). Local rule-based explanations of black box decision systems. <https://arxiv.org/pdf/1805.10820.pdf>



- Guidotti, R., & Ruggieri, S. (2019a). On the stability of interpretable models. *2019 International Joint Conference on Neural Networks (IJCNN)*.  
<https://doi.org/10.1109/ijcnn.2019.8852158>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1-42. <https://doi.org/10.1145/3236009>
- Hackeling, G. (2017). *Mastering machine learning with scikit-learn*. Packt Publishing Ltd.
- Hall, P., Gill, N., Kurka, M., & Phan, W. (2017). *Machine Learning Interpretability with H2O Driverless AI*. Retrieved from <http://docs.h2o.ai>
- Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1). <https://doi.org/10.1186/s40537-020-00305-w>
- Hara, S., & Hayashi, K. (2016). Making tree ensembles interpretable.  
<https://arxiv.org/pdf/1606.05390.pdf>
- HCUP - Healthcare Cost and Utilization Project. (2020). *HCUP Coding Practices*. Retrieved from <https://www.hcup-us.ahrq.gov/db/coding.jsp>
- He, H., & Garcia, E. (2009). Learning from Imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284.  
<https://doi.org/10.1109/tkde.2008.239>
- Henelius, A., Puolamäki, K., Boström, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: Exploring classifiers by randomization. *Data Mining and Knowledge Discovery*, *28*(5-6), 1503-1529. <https://doi.org/10.1007/s10618-014-0368-8>

- Henelius, A., Puolamaki, K., & Ukkonen, A. (2017). Interpreting classifiers through attribute interactions in datasets. <https://arxiv.org/pdf/1707.07576.pdf>
- Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105-120.  
<https://doi.org/10.1016/j.techfore.2015.12.014>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. <https://arxiv.org/pdf/1503.02531.pdf>
- Hoff, K. A., & Bashir, M. (2014). Trust in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407-434.  
<https://doi.org/10.1177/0018720814547570>
- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?  
<https://arxiv.org/pdf/1712.09923.pdf>
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1312.  
<https://doi.org/10.1002/widm.1312>
- Hu, L., Chen, J., Nair, V. N., & Sudjianto, A. (2018). Locally interpretable models and effects based on supervised partitioning (LIME-SUP).  
<https://arxiv.org/pdf/1806.00663.pdf>

- Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), 299-310. <https://doi.org/10.1109/tkde.2005.50>
- Hutchinson, B., & Mitchell, M. (2019). 50 years of test (Un)fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, (49-58).  
<https://doi.org/10.1145/3287560.3287600>
- Islam, S. R., Eberle, W., & Ghafoor, S. K. (2019). Towards Quantification of Explainability in Explainable Artificial Intelligence Methods.  
<https://arxiv.org/pdf/1911.10104.pdf>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York, NY: Springer.
- Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *New England Journal of Medicine*, 360(14), 1418-1428. <https://doi.org/10.1056/NEJMsa0803563>
- Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2), 105-115. <https://doi.org/10.1016/j.artmed.2010.05.002>
- John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Machine Learning Proceedings 1994*, 121-129.  
<https://doi.org/10.1016/b978-1-55860-335-6.50023-4>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260. <https://doi.org/10.1126/science.aaa8415>

- Joynt, K. E., & Jha, A. K. (2011). Who has higher readmission rates for heart failure, and why? *Circulation: Cardiovascular Quality and Outcomes*, 4(1), 53-59.  
<https://doi.org/10.1161/circoutcomes.110.950964>
- Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. *International journal on recent and innovation trends in computing and communication*, 2(10), 3003-3008.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 3146-3154.  
<https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
- Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(1), 51. <https://doi.org/10.1186/1472-6947-11-51>
- Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2011). Comparing boosting and bagging techniques with noisy and Imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(3), 552-568.  
<https://doi.org/10.1109/tsmca.2010.2084081>
- Kim, B. (2015). Interactive and interpretable machine learning models for human machine collaboration (Doctoral dissertation, Massachusetts Institute of Technology).
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept

- activation vectors (tcav). *International conference on machine learning* (pp. 2668-2677). PMLR.  
<http://proceedings.mlr.press/v80/kim18d/kim18d.pdf>
- Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions.  
<https://arxiv.org/pdf/1703.04730.pdf>
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2), 273-324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 3, 249-269. <https://doi.org/10.1007/s10462-007-9052-3>
- Krishnan, R., Sivakumar, G., & Bhattacharya, P. (1999). Extracting decision trees from trained neural networks. *Pattern Recognition*, 32(12), 1999-2009.  
[https://doi.org/10.1016/s0031-3203\(98\)00181-2](https://doi.org/10.1016/s0031-3203(98)00181-2)
- Kulesza, T., Burnett, M., Wong, W., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, (126-137).  
<https://doi.org/10.1145/2678025.2701399>
- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (1675-1684).  
<https://doi.org/10.1145/2939672.2939874>
- Lakkaraju, H., Caruana, R., Kamar, E., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. <https://arxiv.org/pdf/1707.01154.pdf>

- Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (131-138).  
<https://doi.org/10.1145/3306618.3314229>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4), 611-627. <https://doi.org/10.1007/s13347-017-0279-x>
- Letham, B., & Rudin, C. (2012). Building interpretable classifiers with rules using Bayesian analysis. *Department of Statistics Technical Report, University of Washington*, 9(3). <https://arxiv.org/pdf/1511.01644.pdf>
- Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3). <https://doi.org/10.1214/15-aos848>
- Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: An introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12(2), 132-139. <https://doi.org/10.1111/j.1365-2753.2005.00598.x>
- Ling, C. X., & Sheng, V. (2008). Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning, 2011*, 231-235.  
[https://www.csd.uwo.ca/~xling/papers/cost\\_sensitive.pdf](https://www.csd.uwo.ca/~xling/papers/cost_sensitive.pdf)
- Lipitakis, A. D., & Kotsiantis, S. (2014). A hybrid machine learning methodology for imbalanced datasets. *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on Information Intelligence, Systems and*

- Applications* (252-257). Chania: IEEE.  
<https://doi.org/10.1109/IISA.2014.6878762>
- Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27, 247-266. <https://doi.org/10.1017/S1358246100005130>
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36-43. <https://doi.org/10.1145/3233231>
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., . . . Sanchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88. <https://doi.org/10.1016/j.media.2017.07.005>
- Liu, H., & Motoda, H. (2002). On issues of instance selection. *Data Mining and Knowledge Discovery*, 6(2), 115-130. <https://doi.org/10.1023/A:1014056429969>
- Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4), 491-502. <https://doi.org/10.1109/TKDE.2005.66>
- Liu, W., Stansbury, C., Singh, K., Ryan, A. M., Sukul, D., Mahmoudi, E., . . . & Nallamotheu, B. K. (2020). Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PloS one*, 15(4), e0221606. <https://doi.org/10.1371/journal.pone.0221606>
- Liu, X., Wang, X., & Matwin, S. (2018). Improving the Interpretability of deep neural networks with knowledge distillation. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. <https://doi.org/10.1109/icdmw.2018.00132>
- Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD international conference on*

*Knowledge discovery and data mining - KDD '12.*

<https://doi.org/10.1145/2339530.2339556>

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.*

<https://doi.org/10.1145/2487575.2487579>

Lundberg, S. M., & Su-In, L. (2017b). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765-4774.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., . . . & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 2522-5839.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*(250), 113-141.

<https://doi.org/10.1016/j.ins.2013.07.007>

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33, 170-175.

Medicare Payment Advisory Commission. (2016, March). *Medicare Payment Policy, Report to the Congress*. Retrieved from Medpac, Advising the Congress on Medicare issues: <http://www.medpac.gov/-documents-reports>

Messina, W. (2016). Decreasing congestive heart failure readmission rates within 30 days at the Tampa VA. *Nursing Administration Quarterly*, 40(2), 146-152.

<https://doi.org/10.1097/naq.0000000000000154>



- Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4), 283-298. [https://doi.org/10.1016/s0001-2998\(78\)80014-2](https://doi.org/10.1016/s0001-2998(78)80014-2)
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236-1246. <https://doi.org/10.1093/bib/bbx044>
- Mitchell, T. M. (1997). *Machine Learning*.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65, 211-222. <https://doi.org/10.1016/j.patcog.2016.11.008>
- Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., . . . & Turner, M. B. (2016). Executive summary: heart disease and stroke statistics-2016 update: a report from the American Heart Association. *Circulation*, 133(4), 447-454.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071-22080. <https://doi.org/10.1073/pnas.1900654116>
- Nie, L., Wang, M., Zhang, L., Yan, S., Zhang, B., & Chua, T. (2015). Disease inference from health-related questions via sparse deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(8), 2107-2119. <https://doi.org/10.1109/tkde.2015.2399298>

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. <https://doi.org/10.1126/science.aax2342>
- Palczewska, A., Palczewski, J., Marchese Robinson, R., & Neagu, D. (2014). Interpreting random forest classification models using a feature contribution method. *Integration of Reusable Systems*, 193-218. [https://doi.org/10.1007/978-3-319-04717-1\\_9](https://doi.org/10.1007/978-3-319-04717-1_9)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. 33, 9780-9784. AAAI Conference on Artificial Intelligence.
- Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3-14. <https://doi.org/10.1080/00220670209598786>
- Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. <https://arxiv.org/pdf/1806.07421.pdf>
- Plumb, G., Molitor, D., & Talwalkar, A. (2018). Model agnostic supervised local explanations. *Advances in Neural Information Processing Systems*, 2515-2524. <https://arxiv.org/pdf/1807.02910.pdf>
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106. <https://doi.org/10.1007/BF00116251>

- Quinlan, J. R. (1987). Generating production rules from decision trees. *Proceedings of the 10th international joint conference on Artificial intelligence*, 304-307.  
<https://dl.acm.org/doi/10.5555/1625015.1625078>
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publications. <https://doi.org/10.1007/BF00993309>
- Refaeilzadeh P., Tang L., Liu H. (2016) Cross-Validation. *Encyclopedia of Database Systems*. [https://doi.org/10.1007/978-1-4899-7993-3\\_565-2](https://doi.org/10.1007/978-1-4899-7993-3_565-2)
- Rajapaksha, D., Bergmeir, C., & Buntine, W. (2020). LoRMikA: Local rule-based model interpretability with K-optimal associations. *Information Sciences*, 540, 221-241.  
<https://doi.org/10.1016/j.ins.2020.05.126>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. <https://arxiv.org/pdf/1606.05386.pdf>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why should I trust you?". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1135-1144).  
<https://doi.org/10.1145/2939672.2939778>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1)  
<https://ojs.aaai.org/index.php/AAAI/article/view/11491>
- Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5), 589-600. <https://doi.org/10.1109/tkde.2007.190734>

- Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-based explanations of prediction models. *Human and Machine Learning*, 159-175.  
[https://doi.org/10.1007/978-3-319-90403-0\\_9](https://doi.org/10.1007/978-3-319-90403-0_9)
- Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2), 1-39. <https://doi.org/10.1007/s10462-009-9124-7>
- Ross, J. S., Chen, J., Lin, Z., Bueno, H., Curtis, J. P., Keenan, P. S., Normand, S. T., Schreiner, G., Spertus, J. A., Vidán, M. T., Wang, Y., Wang, Y., & Krumholz, H. M. (2010). Recent national trends in readmission rates after heart failure hospitalization. *Circulation: Heart Failure*, 3(1), 97-103.  
<https://doi.org/10.1161/circheartfailure.109.885210>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- Saeyns, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.  
<https://doi.org/10.1093/bioinformatics/btm344>
- Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, 149-171. [https://doi.org/10.1007/978-0-387-21579-2\\_9](https://doi.org/10.1007/978-0-387-21579-2_9)
- Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms*. Cambridge, MA: MIT Press.

- Schwab, P., & Karlen, W. (2019). CXPlain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*.  
<https://arxiv.org/pdf/1910.12336.pdf>
- Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4), 233-242.  
<https://doi.org/10.1093/idpl/ix022>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*.  
<https://doi.org/10.1109/iccv.2017.74>
- Shankaranarayana, S. M., & Runje, D. (2019). ALIME: Autoencoder based approach for local interpretability. *Intelligent Data Engineering and Automated Learning – IDEAL 2019*, 454-463. [https://doi.org/10.1007/978-3-030-33607-3\\_49](https://doi.org/10.1007/978-3-030-33607-3_49)
- Shi, S., Zhang, X., & Fan, W. (2020). A Modified perturbed sampling method for local interpretable model-agnostic explanation. <https://arxiv.org/pdf/2002.07434.pdf>
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3145-3153). JMLR. org.
- Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning. *Cutter Business Technology Journal*, 31(2), 47-53.
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.  
<https://doi.org/10.1145/3375627.3375830>

- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.  
<https://doi.org/10.1016/j.ipm.2009.03.002>
- Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5).  
<https://doi.org/10.1002/widm.1379>
- Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1), 307.  
<https://doi.org/10.1186/1471-2105-9-307>
- Štrumbelj, E., Kononenko, I., & Robnik Šikonja, M. (2009). Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering*, 68(10), 886-904. <https://doi.org/10.1016/j.datak.2009.01.004>
- Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, 1-18.
- Su, G., Wei, D., Varshney, K. R., & Malioutov, D. M. (2016). Learning sparse two-level boolean rules. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *arXiv* <https://arxiv.org/pdf/1703.01365.pdf>

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-compare. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

<https://doi.org/10.1145/3278721.3278725>

The Affordable Care Act, 42 U.S.C. § 18001 (2010).

<https://www.govinfo.gov/content/pkg/PLAW-111publ148/pdf/PLAW-111publ148.pdf>

Thiagarajan, J. J., Kailkhura, B., Sattigeri, P., & Ramamurthy, K. N. (2016). Peeking into deep neural networks via feature-space partitioning. *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*.

<https://arxiv.org/abs/1611.07429>

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 267-288.

<https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018).

Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, (pp. 8-14). Stockholm, Sweden.

Turner, R. (2016). A model explanation system. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*.

<https://doi.org/10.1109/mlsp.2016.7738872>

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11), 1225-1231. [https://doi.org/10.1016/s0895-4356\(96\)00002-](https://doi.org/10.1016/s0895-4356(96)00002-9)

Ustun, B., & Rudin, C. (2014). Methods and models for interpretable linear classification.

<https://arxiv.org/pdf/1405.4047.pdf>

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3), 349-391.

<https://doi.org/10.1007/s10994-015-5528-6>

Ustun, B., Traca, S., & Rudin, C. (2013). Supersparse linear integer models for interpretable classification. <https://arxiv.org/pdf/1306.6677.pdf>

Valdes, G., Luna, J. M., Eaton, E., Simone, C. B., Ungar, L. H., & Solberg, T. D. (2016). MediBoost: A patient stratification tool for interpretable decision making in the era of precision medicine. *Scientific Reports*, 6(1).

<https://doi.org/10.1038/srep37854>

Van Assche, A., & Blockeel, H. (2007). Seeing the forest through the trees: Learning a comprehensible model from an ensemble. *Machine Learning: European Conference on Machine Learning (ECML) 2007*, 418-429.

[https://doi.org/10.1007/978-3-540-74958-5\\_39](https://doi.org/10.1007/978-3-540-74958-5_39)

Vandewiele, G., Janssens, O., Ongenaes, F., De Turck, F., & Van Hoecke, S. (2016).

GENESIM: Genetic extraction of a single, interpretable model. *30th Conference on Neural Information Processing Systems (NIPS 2016)*. 30th Conference on Neural Information Processing Systems (NIPS 2016).

Vaughan, J., Sudjianto, A., Brahimi, E., Chen, J., & Nair, V. N. (2018). Explainable neural networks based on additive index models.

<https://arxiv.org/pdf/1806.01933.pdf>



- Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1), 175-186.  
<https://doi.org/10.1007/s00521-013-1368-0>
- Verhaegh, K. J., MacNeil-Vroomen, J. L., Eslami, S., Geerlings, S. E., De Rooij, S. E., & Buurman, B. M. (2014). Transitional care interventions prevent hospital readmissions for adults with chronic illnesses. *Health Affairs*, 33(9), 1531-1539.  
<https://doi.org/10.1377/hlthaff.2014.0160>
- Visani, G., Bagli, E., & Chesani, F. (2020). OptiLIME: Optimized LIME explanations for diagnostic computer algorithms. <https://arxiv.org/pdf/2006.05714.pdf>
- Visani, G., Bagli, E., & Chesani, F., Poluzzi, A., & Capuzzo, D. (2021). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 1-11.  
<https://doi.org/10.1080/01605682.2020.1865846>
- Vorm, E. S. (2018). Assessing demand for transparency in intelligent systems using machine learning. *2018 Innovations in Intelligent Systems and Applications (INISTA)*. <https://doi.org/10.1109/inista.2018.8466328>
- Waa, J. V., Robeer, M., Diggelen, J. V., Brinkhuis, M., & Neerinx, M. (2018). Contrastive explanations with local foil trees.  
<https://arxiv.org/pdf/1806.07470.pdf>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76-99.  
<https://doi.org/10.1093/idpl/ix005>

- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3063289>
- Wang, F., Rudin, C. (2015). Falling rule lists. *Artificial Intelligence and Statistics*, 1013-1022. <http://proceedings.mlr.press/v38/wang15a.pdf>
- Wang, J., Bao, W., Sun, L., Zhu, X., Cao, B., & Yu, P. S. (2019). Private model compression via knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 1190-1197.  
<https://doi.org/10.1609/aaai.v33i01.33011190>
- Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Author correction: Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(10), 1627-1627. <https://doi.org/10.1038/s41591-019-0609-x>
- Wong, T. (2015). Performance evaluation of classification algorithms by K-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), 2839-2846.  
<https://doi.org/10.1016/j.patcog.2015.03.009>
- Xie, N., Ras, G., van Gerven, M., & Doran, D. (2020). Explainable deep learning: A field guide for the uninitiated. <https://arxiv.org/pdf/2004.14545.pdf>
- Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2016). Predicting 30-day all-cause readmissions from hospital inpatient discharge data. *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. <https://doi.org/10.1109/healthcom.2016.7749452>

- Yang, C., Rangarajan, A., & Ranka, S. (2018). Global model interpretation via recursive partitioning. *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. <https://doi.org/10.1109/hpcc/smartcity/dss.2018.00256>
- Yang, F., Du, M., & Hu, X. (2019). Evaluating explanation without ground truth in interpretable machine learning. <https://arxiv.org/pdf/1907.06831.pdf>
- Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable Bayesian rule lists. *Proceedings of the 34th International Conference on Machine Learning* (pp. 3921-3930). JMLR.org. <https://dl.acm.org/doi/10.5555/3305890.3306086>
- Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. <https://arxiv.org/pdf/1903.03894.pdf>
- Yoon, F., Sheng, M., Jiang, H. J., Steiner, C. A., & Barrett, M. L. (2017). Calculating nationwide readmissions database (NRD) variances. *US Agency for Healthcare Research and Quality*, <https://www.hcup-us.ahrq.gov/reports/methods/2017-01.pdf>.
- Yoon, J., Arik, S. O., & Pfister, T. (2019). RL-LIM: Reinforcement learning-based Locally interpretable modeling. <https://arxiv.org/pdf/1909.12367.pdf>
- Yoon, J., Jordon, J., & van der Schaar, M. (2018). INVASE: Instance-wise variable selection using neural networks. *International Conference on Learning Representations*. [https://openreview.net/pdf?id=BJg\\_roAcK7](https://openreview.net/pdf?id=BJg_roAcK7)

Zafar, M. R., & Khan, N. M. (2019). DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems.

<https://arxiv.org/pdf/1906.10263.pdf>

Zafar, M. R., & Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable Explainability. *Machine Learning and Knowledge*

*Extraction*, 3(3), 525-541. <https://doi.org/10.3390/make3030027>

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K.

(2018). Confounding variables can degrade generalization performance of radiological deep learning models. <https://arxiv.org/pdf/1807.00431.pdf>

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding Convolutional

networks. *Computer Vision – ECCV 2014*, 818-833. [https://doi.org/10.1007/978-](https://doi.org/10.1007/978-3-319-10590-1_53)

[3-319-10590-1\\_53](https://doi.org/10.1007/978-3-319-10590-1_53)

Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 1-1.

<https://doi.org/10.1109/tse.2019.2962027>

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why should you trust my explanation?" understanding uncertainty in lime explanations.

<https://arxiv.org/pdf/1904.12991.pdf>

Zhang, Z., Yang, F., Wang, H., & Hu, X. (2019). Contextual local explanation for black box classifiers. <https://arxiv.org/pdf/1910.00768.pdf>

Zhou, Y., & Hooker, G. (2016). Interpreting models via single tree approximation.

<https://arxiv.org/pdf/1610.09036.pdf>

Zhou, Z. (2012). Ensemble methods: Foundations and algorithms.

<https://doi.org/10.1201/b12207>

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis.

<https://arxiv.org/pdf/1702.04595.pdf>