Machine Learning Approach to Predict Treatment Outcome Using Shockwave Lithotripsy in Management of Urinary Stone

Reihaneh Moghisi

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE

REQUIREMENTS

FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY

TORONTO, ONTARIO

Abstract

In Ontario, shock wave lithotripsy (SWL) is a regionalized resource and St. Michael's Hospital is one of only three centers in the province offering this service. As such, many of the patients travel a great distance to receive this noninvasive treatment. Our objective is to implement ensemble learning technique to predict treatment outcome based on the patients' demographic information and stone characteristics. In order to construct a rigorous machine learning model that can be confidently applied to assist in decision making process, we built our model based on the whole dataset of patients ages over 18 for the years from 1998 to 2016. Our objective is to build a classification model to predict treatment outcome using SWL prior to making any decision on treatment modality. The success or failure was based on having retreatment plan for the same patient within less than 90 days of initial treatment. We also compared six machine learning algorithms' performance on dataset in terms of their accuracy using t-test with 95% confidence interval.

In addition, we performed a retrospective comparison of three shock wave lithotripsies (SWL) that has been used in SMH during the past two decades in terms of their successfulness. Furthermore, we looked at changing trends over time in terms of stone size, location, and patient BMI, and site of origin, gender, age, etc.

Acknowledgements

I would like to thank Professor Jimmy Huang for his excellent guidance and assistance throughout the semesters. Moreover, his help in motivating me to pursue further academic work and showing me the path is well appreciated. I feel privileged to learn the skills necessary to conduct such a project with potential to improve care for many patients with kidney stones.

I would also like to express my appreciation to Dr. Kenneth Pace who is the chief of Division of Urology at St. Michael's Hospital for providing me with this dataset, introducing me to the field of urology and all his supportive help throughout this project.

Next, I would like to thank Dr. Gary Spaarkman who stood as my exam committee member and Dr. Chris El Morr who gave me my first chance and motivation to enter to the world of research and academics and was a major influence on my research skill development in my early days.

The next people I like to thank are my family who have supported and encouraged me in every step of my life. My parents whose support, love and guidance has been felt throughout my whole life and I cannot express in words how grateful I am for all they have done for me.

And last but foremost I find it difficult to express my appreciation to my beloved husband Dr. Mohammad Hajiha because it is so boundless. Without him, this whole journey wouldn’t be possible. He has made countless sacrifices to help me to get to this point. He is my most enthusiastic cheerleader; he is my best friend; and he is an amazing husband

and father. He has shared this entire amazing journey with me, so it only seems right that I dedicate this thesis to him.

This thesis concludes my Masters of Art degree in Information System and Technology at York University in Toronto. It was written at the school of technology during the Autumn of 2018 under the supervision of Professor Jimmy Huang.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

1.1 Urolithiasis

1.1.1 Background

Urinary stone disease (USD) also known as urolithiasis is a disease that happens when a solid particle is formed inside renal, kidney or urinary tract system. When the urine is concentrated urinary stones develop by crystalizing minerals. The particles are of different chemical compositions including calcium oxalate, calcium phosphate, uric acid, cysteine, and struvite. Small stones usually pass through the body by themselves. However larger stone that may even cause blockage in urinary tract system, need more invasive medical interventions and sometimes surgery.

1.1.2 Risk Factors

Dehydration caused by low fluid intake is the main source of stone formation. Obesity is another leading risk factor associated with stone formation. Dietary intake can also affect the chance of getting stone. Urolithiasis is also depending on underlying metabolic physical condition of patient or genetic disorders such as abnormal kidney form including horseshoe or medullary sponge kidney, which can lead to higher chances of stone disease due to restraining and prolonging the passage of crystals through kidney (Gambaro, Fabris, Puliatta, & Lupo, 2006).

1.1.3 Symptoms

Signs and symptoms of urinary stone disease can vary based on the nature of stone and physical characteristics of patients. Some large stones can have no symptoms but if the stone irritates the kidney or ureter wall it may cause some symptoms such as but not limited to: renal colic which is abdominal pain that can grow to inner thigh and groin, hematuria which is blood in urine, cloudy or abnormally dark urine, urinary urgency, nausea, vomiting, sweating.

1.1.4 Prevalence

A prevalence of urolithiasis has been of great interest to researchers over the past years. Considering the great diversity in ethnicity and race in northern America. A recent systematic review suggests the increasing prevalence of urinary stone disease in northern America over the past three decades. It was reported that 0.24% of all annual hospitalization in Canada in 1970 were due to urinary calculi. Stamatelou found that the prevalence of kidney stone in people aging between 20 to 74 has been increased significantly from 3.8% from 1976 to 1980 compared to 5.8% from 1988 to 1994 in United States. It was also reported that the prevalence is higher among males compared to females and is increased by age. Nephrolithiasis is a prevalent disease in Canada with a lifetime risk of 10% among both men and women, whereas there is 75% chance of recurrence in twenty years (Moe 2006) (Stamatelou et al. 2003).

1.1.5 Procedures

The most common procedures for managing urolithiasis are shockwave lithotripsy (SWL), Ureteroscopy (URS) and percutaneous nephrolithotomy (PCNL). Historically SWL has been a predominant and most commonly used procedure for treating upper tract urolithiasis due its noninvasive nature, lower cost, fewer side effects and faster recuperation (Antonelli et al. 2014). However, for some stones which are larger than 20mm in diameter, SWL is not recommended as a first line of treatment. Furthermore, depending on patient's characteristics and preferences, alternative treatments may be suggested. For example, SWL is not usually recommended for children and also for patients who had previous surgical treatments on their kidney and may have vulnerable tissue.

1.2 Extracorporeal Shockwave Lithotripsy

1.2.1 History

Before the introduction of Extracorporeal Shockwave Lithotripsy (ESWL) in 1980's, large number of the renal stones were treated using endoscopic or open surgery, resulting in severe complications, longer hospital stays and more long-term side effects. Since urinary stone disease is usually recurrent, patients had to underwent multiple high risk open surgeries in their life, increasing their chance of complications. Since the first introduction of Extracorporeal Shockwave Lithotripsy (ESWL) 30 years ago, it has been recognized as the cornerstone for treating urinary calculi.

The noninvasiveness, effectiveness, shorter recovery time and less complications of this treatment made it one of the most favorable therapeutic choices specially for stones

located in upper/middle calix and renal pelvis measuring 20 mm or less in diameter. The idea of SWL is to generate sound waves from outside the body (extracorporeal) to pulverize stones in vivo using lithotripters ("lithos" is Greek for "stone" and "tripsis" is Greek for "breaking") into smaller fragments so that they can easily pass through the body. Each machine also uses ultrasound positioning system to locate the stone inside the body.

The first lithotripter was invented in 1980's, since then three generations of lithotripters have been developed. The first generation or electro hydraulic lithotripters differ significantly on several aspects from the second and third generations, however the principles of shockwave therapy remain the same in all lithotripters. The first-generation lithotripter known as Dornier HM3, uses an ellipsoidal reflector lying underneath the water cushion. The ellipsoid-shaped chamber generates acoustic pulse at the stone residing in its focal point. The shockwave is generated underwater and patient needs to sit in a water bath, therefore the anesthesia posed a challenge due to immersion in the water. Second and third generation of lithotripters however do not need water bath. They use acoustic lens to focus the wave on the stone, the function similar to optical lens.

Dornier MFL 5000 is a multifunctional third generation lithotripter unit that has been employed in many locations for treating kidney stones. This lithotripter was the first-generation lithotripters that was installed and used in SMH at 1980. Although this lithotripter showed significant results in treating stones of less than 20mm, however one study conducted at university of Iowa in 2000, argues the effectiveness of this unit for stones burden more than 50 mm^2. The study demonstrated the successfulness of this unit for only 47% of patients, whereas success was defined as residual stone fragments of less

than 2 mm. More recent studies comparing the effectiveness of this lithotripter shows less success rate of Dornier MFL 5000 compared to Dornier HM3, the first generation of lithotripters, for nephrolithiasis in children. (Penn et al. 2009). St.Michael hospital which is one of the 13 lithotripsy centers in Canada was equipped with three different lithotripsy technologies within the last three decades (from 1980 to 2016). Dornier MFL 5000- from 1980's until, Philips LithoTron – March 12th 2001 until, Storz Modulith SLX-F2- from January 12 2010 until now are the three generations of lithotripters that has been implemented in this location.

1.2.2 SWL vs. Other Procedures

Among the available techniques both shockwave lithotripsy (SWL) and ureteroscopy (URS) that are the most common noninvasive approaches in management of urinary stones, have their own benefits and downsides and varying complications and success rate. In 1995 it was reported that although the success rate of PCNL in managing urinary stone is 26% higher compared to SWL with less auxiliary treatments required, however the overall cost of procedure is 1342$ higher than SWL (Jewett, Bombardier, and Menchions 1995).

At the time of its first introduction the success rate of 93.8% was reported, which resulted in replacing the conventional surgical procedure for ureteral stones in hospitals with lithotripters (Zehntner, Ackermann, and Zingg 1987). Hematuria, urinary tract infection, pain while passing fragments and blockage of urine flow -as a result of stone fragments stuck in the urinary tract- are known as some of the minor complications associated with ESWL. There has been no proven long-term side effect of ESWL, hence

the treatment can be performed repeatedly multiple times on a patient. In summary some of the advantages of ESWL listed below:

1. ESWL is one of the noninvasive stone therapies besides medical treatment
2. ESWL can be done with light sedation and local or general anesthetics and can be performed as outpatient procedure and even in a mobile setup
3. ESWL can be safely applied repeatedly without effort
4. ESWL is considered a safe procedure with low complications and short-term side effects
5. ESWL has no proven long-term side effects
6. ESWL is a procedure with a fast patient recovery and return to daily activity

Chapter Two: Literature Review

2.1 Motivation for the Study

The fundamental conception and the key principle behind medical field is the emphasizes on persistent progress in effectiveness, safety, efficiency and quality of care. This endeavor justifies the close relation of medical field to engineering and data science. In Ontario shock wave lithotripsy (SWL) is a regionalized resource at St. Michael's Hospital and is one of the only 3 centers in the province offering this service. As such, long wait times are expected and also many of the patients have to travel a great distance to access treatment.

Considering the intolerability of stone disease's pain and because of long wait times, some patients opt for more invasive therapies to gain access to faster treatment. Wait time to access lithotripsy in Canada ranges from one day to one year with a mean wait time of 8.4 weeks. In Toronto the mean wait time is 8 weeks however in Ottawa patients can be booked for lithotripsy within days of initial visit.

The fact that 20% of patients are willing to travel to access faster treatments corroborates the idea that the distribution of patients' referral pattern to lithotripsy centers across Canada and specially Ontario province is not optimized and may be enhanced according to the patients' geographical location, willingness to travel and resources available at each location.

Table below shows quantity and location of lithotripsy centers in Canada. Some of the data were collected by calling the hospitals directly and orally confirm the existence of this technology in their location.

After the introduction of non-invasive methods for treating urinary calculi, finding the optimal approach to manage stones grew into an important topic for researchers and decision makers. The ultimate goal is to achieve patients with stone free state with the least invasive method, using few shockwaves and as low power level as possible, whereas no other auxiliary treatments are needed. Besides patients' preferences, several factors including stone characteristics (size, location, stone density, skin to stone distance), patient's condition, surgeon's experience and the need for anesthesia should be taken into account when making a decision on treatment modality.

Aside from optimization in the management of patients' referral, some enhancements can be applied on current guidelines for managing nephrolithiasis using SWL to make better use of this technology for best candidates. In other words, considering the reported failure rate of SWL after first session in recent researches that ranges from around 30% to 60% (Altok et al. 2016; Javanmard et al. 2016; Yamashita et al. 2017), we can reduce this number significantly by identifying candidates who would benefit most from this treatment, and provide alternative therapy for those who won't. This goal is one of the main objectives of this thesis and we tried to contribute to enhancing and optimizing treatment result for SWL candidates.

Table 1. Quantity and distribution of lithotripsy centers in Canada

Quantity and locations of lithotripters in Canada	
Province	ESWL Locations
British Colombia	Vancouver: Vancouver General Hospital Victoria: Royal Jubilee Hospital Prince George: Prince George Regional Hospital
Alberta	Edmonton: Misericordia Community Hospital Calgary: Rockyview General Hospital
Saskatchewan	Saskatoon: St. Paul's Hospital
Manitoba	Winnipeg: Health science center
Ontario	St. Michael's Hospital Ottawa: Ottawa Hospital Riverside Campus London: St. Joseph's Healthcare
Quebec	Montreal: CHUM – St. Luc Hospital Montreal: MUHC – Royal Victoria Hospital Quebec City: St. François d'Assise
New Brunswick	St.John: St. Joseph's Hospital
PEI	Charlottetown: Queen Elizabeth Hospital (Traveling)
Nova Scotia	Halifax: Victoria General Hospital
Newfoundland & Labrador	St.John's: Health Sciences Centre
Yukon / NWT/ Nunavet	N/A

Current studies focused on statistical analysis of patients using bivariate and/or multivariate analysis, however this approach is heavily restricted by assumptions about data and its distribution and also is more suitable for small populations. The machine learning approach overcome these barriers and its ability to make predictions and provide rules based on attributes, makes it more practical and easy to implement for decision makers and practitioners. In general machine learning approach is more liberal in terms of approach and techniques, while traditional statistical approaches are more conservative. The other benefit of machine learning approach is that not only it does not promote data reduction prior to modeling, but it also promulgates the abundance ideology:" The more data, the better".

On the other hand, traditional statistical analysis encourages data reduction as much as possible before modeling: sampling, less input features and etc. Furthermore, redundancy and replacement of samples in data is allowed in machine learning but requires to be removed or fixed prior to modeling in statistical analysis.

This factor makes a huge difference in applicability of data to ensure that as much data are included in the analysis as possible. Replacement of samples in modeling allows us to produce a less biased model and consequently increase more precise results by reducing the variance. This approach also reduces the risk of error in data and help to avoid over-fitting. Other related work include (Huang, Wen, Data, 2006, X. Huang, Zhong, TREC, 2005).

2.2 Definitions

Calculi is a solid particle in urinary system that causes pain and discomfort. Patients are diagnosed with urinalysis or radiologic imaging. The term calculi come from a Latin word "Pebble" which were first used for counting objects. The mathematical field of calculus also comes from this word. A urinary calculi is a pebble in a urinary system.

Ureter is a 10 to 12-inch-long tube that carries urine from the kidney to the urinary bladder. Human body has two ureters, one attached to each kidney on each side. The upper half of ureter is located in the abdomen the lower half is located in pelvic area.

SWL shockwave lithotripsy is a technique for treating urinary calculi. It uses shockwaves to break the kidney stone into smaller pieces so it can easily pass through the urine. The first shockwave was first invented in 1980's and three more generations of the machine has been developed since then. The main difference between the first generations and second and third generation of lithotripters is that the first generations required patients to sit in a water tub while newer generations do not.

Percutaneous nephrostomy is an interventional radiology/surgical procedure in which the skin in renal pelvis area is punctures in order to place small tube (catheter) to drain urine directly from kidney or passing stones. This procedure is done under local anesthesia and is considered to be one of the noninvasive approaches to manage kidney stone. The tubes are inserted through the body with the guidance of X-ray imaging screening.

Medullary sponge kidney (also known as Cacchi–Ricci disease) is a congenital disorder of the kidneys characterized by changes in a tubules or tiny tubes inside a fetus kidney. It

can happen in one or both kidneys. In a normal kidney, urine passes through these tubules as the kidney is formed during fetus growth. However in Medullary sponge kidney cysts, which are tiny fluid-filled sacs, are formed inside the kidney and prevent urine to flow freely through the tubules. Individuals with medullary sponge kidney are at increased risk for kidney stones and urinary tract infection (UTI).

Staghorn kidney stone is a term used to describe a large kidney stone that evolves into renal pelvis and that takes up more than two branches of renal pelvis kidney (calyces). Staghorn stones are best managed with complete surgical removal, however for patients with life threatening comorbidities, nonsurgical approaches may control the sequelae of untreated stones.

Horseshoe kidney also known as ren arcuatus (in Latin), renal fusion or super kidney, is a birth defect of kidney with the prevalence of one in 600 people, in which kidneys prevalence become attached together ("fuse") at the lower end or base and will result in a U shape or horseshoe shaped kidney. This condition is thought to be more prevalent in men than women.

Duplex kidney also called duplicated collecting system is a developmental condition in which one or both kidneys have two ureter tubes to drain urine, rather than a single tube. Most of the duplex kidneys do not require further medical intervention but some may be associated with conditions that need urology treatment. These include flow of a urine back to the kidney instead of bladder or obstruction in urinary tracts.

BMI is a body mass index or Quetelet index is a value derived from the ratio of body weight in kilogram to the square of body height in meter. BMI can be used as a screening tool but is not diagnostic of the body fatness or health of an individual. The National Institutes of Health (NIH) now defines normal weight, overweight, and obesity based on BMI rather than the traditional height/weight charts. Obesity is a BMI of 30 or more. Overweight is a BMI of 25 or more. Normal BMI is about 18.5 to 24.9 and underweight is a BMI of 18.5 or lower.

2.3 Contribution of Attributes to Classification Model

In this section we try to explain our understandings and findings of how these variables in the study connect with each other from other researcher's findings. A raw database that has been used for data mining consisted of 18 attributes related to patients' characteristics. However, there are more attributes in database that were not in the scope of this project hence were removed.

Gender, age, BMI, Side (Left/right), Location of stone, Frequency, Area of stone, Stent insertion, Stone treatment number, Lithotripter, Family history, Asymptomatic, Antibiotics, Number of stones, Urologist, Position, number of shocks are these attributes. Other attributes that were not used in algorithm production or used in descriptive statistics included patients' geographical location and some attributes were related to post treatment information that cannot be used as a predictor to predict treatment outcome.

We later reduced number of attributes and ended up with 12 attributes due to huge number of missing values in some attributes. Other researches also confirm the importance

of these attributes in determining the results of treatment. There are also some other factors such as stone density which are known to have an effect on outcome however are not collected in our database (Gücük & Uyetürk, 2014). The reason for this is that our database goes back to more than twenty years ago where some of the attributes were not collected or there were no means to obtain these features.

Different researches have determined various predictive factors for success of SWL. As such, Nakasato's analysis on 260 patients with solitary renal or kidney stone revealed that Hounsfield unit and stone location have been significant factors of determining success or failure of SWL. Stone volume, location, skin-to-stone distance, stone HU values, and stone composition were evaluated and assessed in this study with a multiple regression analysis to find a significant factors (Nakasato, Morita, & Ogawa, 2015).

Another study performed by researchers in Japan demonstrates that only stone size has significant impact on the success rate of SWL. Other prognostic factors that were evaluated in this study were BMI, stone position and hydronephrosis (Takahara et al., 2012). To further acknowledge the findings, Ahmad El Assmy et al. determined stone size (diameter less than 12mm) and stone attenuation as predictive factors of success rate of SWL for children (El-Assmy, El-Nahas, Abou-El-Ghar, Awad, & Sheir, 2013).

Other than patients' characteristics, some external factors also have an effect on success rate of SWL. Research shows that physician and specialist training, for example, significantly improved the success of SWL. Physician's skill in detecting and targeting the

stone as well as patient's therapeutic position are crucial components of successful stone removal procedure (Okada et al., 2013).

Furthermore, this matter also suggests that it is crucial to consider peripheral and outlying factors as well as patients' circumstances and characteristics when choosing the treatment modality in order to achieve the best results with no auxiliary treatments needed. However, some factors are not easy to manage and there might be some limitations that are inevitable. Training physicians and staff is a good example of that. Although research suggests that training may lead to more successful treatment but the objectives and depth of the training and also each person's perception from the lessons may be different and hard to evaluate and later control.

Aside from outlying factors, it is believed that shockwave frequency along with other factors is one of the key elements that affect stone fragmentation in patients with renal or ureteral stones. There are four frequencies that is commonly used in SWL therapy 30, 60 ,90 or 120 shocks per minute. Although studies show that the success rate of SWL can be enhanced by adjusting the shockwave frequency to 90 which is proved to be the optimal frequency (Yilmaz et al., 2005), however due to resource limitations in Toronto, they almost only apply 120 shocks per minute. In Toronto since the volume of patients is high and resources including shockwave lithotripsy machine, staff, physicians are limited, they use high frequency in order to reduce treatment duration so that they can treat more patients in less amount of time. Higher shocks per minute is also associated with more complications for patient and more pain perception.

Researchers could not find any difference between 30 and 60 shocks per minute however the pain perception is significantly higher in 30 shocks per minute. Also higher frequencies is only recommended for lower ureteral calculi specially with dimension of over 8 mm (Altok et al., 2016), however in Toronto higher frequency is applied to majority of patients regardless of their health situation and stone characteristics. Higher frequencies are more favorable by physicians and healthcare staff as it substantially expedites the treatment session. However, researches argue the benefit of this approach in regard to patients' health and following side effects.

Therefore, this factor can represent an important point of interest that can be further evaluated to achieve an optimum treatment outcome for patients while diminishing healthcare burden. In summary, based on other researches and our findings we believe that the elements that are included in this study are related to the outcome of SWL treatment.

Chapter Three: Supervised Learning Method To Predict Treatment Outcome

3.1 Ensemble Learning Technique

The method that has been used in this study to predict the treatment outcome for SWL candidates is a branch of ensemble learning technique called AdaBoost algorithm. A number of studies have shown the superiority of this technique over other classification models. Ensemble learning method had won multiple prizes in machine learning competitions over the last few years. The principal component of ensemble learning is to combine multiple weak classifiers in order to generate a more accurate and stronger classifier. This process is done by applying a weak classifier that only slightly does better job than a random guess of 50% iteratively. After each iteration a specific weight is assigned to each weak classifier and the final classification model is then generated as a weighted average of these weak classifiers. The whole process of ensemble learning technique and more specifically AdaBoost is discussed in this chapter. However, in order to get the comprehend look into the ensemble learning technique and AdaBoost algorithm we need to first talk about some fundamentals that are required to get a full understanding of ensemble learning technique. All of these prerequisites and related work about these fundamentals are explained in this chapter (Feng, Zhang, Hu, & Huang, 2014; Huang, Peng, Schuurmans, Cercone, & Robertson, 2003).

3.1.1 Introduction

Ensemble learning is a term that is been used a lot in the context of machine learning. Machine learning uses various algorithms such as linear regression, logistic regression, K-means, decision trees and etc. Ensemble learning is nothing but a combination of groups of algorithms in this context. In ensemble learning a group of various algorithms and models combine together to bring forth a model that is more accurate.

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. These methods usually boost the accuracy in models and provide more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, whereas the winning solutions used ensemble methods. Ensemble methods can be divided into two groups:

- *sequential* ensemble methods where the base learners are generated consecutively (e.g. AdaBoost). The basic stimulant of sequential methods is to **make use of the dependence between the base learners.** The overall performance can be boosted by increasing the weight of previously mislabeled examples.
- *parallel* ensemble method is when the base learners are generated in parallel (e.g. Bagging, Random Forest). The basic stimulant of parallel methods is to **make use of independence between the base learners.** The error is then reduced significantly by averaging the error between base learners.

Three methods of most known ensemble learning methods are bagging, stacking and boosting.

Bagging which stands for bootstrap aggregating is a method to decrease the variance of prediction by generating additional data for training. These additional data are generated from original dataset using combinations with repetitions (replacement) to produce multisets of the same cardinality as the original data. Although this approach does not necessarily improve the predictive force of model, however it can move the prediction toward the expected outcome by decreasing the variance of model.

Stacking or stacked generalization same as boosting apply multiple models on training set. Stacking is a way of combining multiple models, that introduces the concept of a meta learner. It is less widely used than bagging and boosting. Unlike bagging and boosting, stacking may be (and normally is) used to combine models of different types. Stacking first split a data into two disjoint sets and train several base learners on first set and test them against the second set. It then applies the higher-level learner by using the predictions from previous level as an input and correct responses as an output. In stacking the output of one classifier is used as training data for other classifiers to approximate the same target function.

Boosting is a sequential ensemble method which tries to add new models to the initial model that perform better in areas where the previous model was deficient at. This is a general technique that combines rules of thumbs or several weak classifiers to improve the precision of any given classifier. The concept of boosting starts with applying a simple and

weak classifier such as one node decision tree -also known as decision stump- on the dataset. After identifying misclassified items, it increases the weight of those items. Therefore, in the next run of the upper level model, we concentrate more on mislabeled examples and less on correctly classified ones. It then repeats this process sequentially until an acceptable model is generated. The theory of boosting is developed for binary classification but can be extended to multiclass cases.

Data mining techniques and more specifically boosting methods has been vastly used in the recent studies specially for solving face detection problems, text retrieval tasks, and even for large datasets for image processing. And in many of these areas it has been shown that data mining techniques not only are flexible in terms of adapting and adjustments to data but also can surpass other conventional statistical methods in classification tasks (Huang, Wen, 2009.).

3.1.2 Boosting

Boosting is a kind of ensemble learning method which calls a weak hypothesis iteratively in order to achieve better fit model with higher accuracy. In boosting we first generate a really weak classifier that can perform slightly better than a random guess of 50%. Then in further steps this classifier is boosted by adding more emphasize and weight on mislabeled examples iteratively. At the final number of iterations the aggregation of these weak classifiers then represent a very strong classifier and the error is represented as the average of these classifier. Boosting algorithm can be represented as:

Equation 1

$$f(x) = \theta_0 + \sum_{m=1}^{M} \theta_m \phi_m(x)$$

Where ϕ is a basis weak classifier, which is picked by the user. Essentially boosting model is a linear combination of these M weak classifiers. Each of these weak classifiers perform only slightly better than coinflip in a sense that they do not need to classify with high accuracy. The most common and famous choice of base learners are tree models and decision stumps (Nielsen, 2016). We are going to discuss the AdaBoost algorithm in more details in the upcoming sections. Prior proceeding toward AdaBoost and its functions, we first are going to discuss some other concepts that are crucial fundamental knowledge for understanding AdaBoost. These are the Loss Function, convex optimization and Gradient Descent.

3.1.3 The Loss Function

All algorithms in machine learning work on a context of minimizing or maximizing a particular function which we call "objective function". A group of functions which need to be minimized in order to achieve a right classification algorithm are called "loss function". A loss function tells how good the prediction model does in terms of correctly classifying items to expected outcome. "Gradient descent" refers to a method to find the least minimum point in the loss function. We will talk about gradient descent in more details in next section. Understanding loss function is important in machine learning and specifically in studying Adaboost algorithm, since Adaboost has been seen as a method to minimize exponential loss function by researchers (Friedman, Hastie, & Tibshirani, 2000).

In the upcoming sections we are going to talk about Adaboost algorithm and its relation to loss function and gradient descent in more details.

In this section we are going to understand the fundamentals of loss function and its usage and applications in machine learning, boosting and Adaboost algorithm.

There are two types of loss functions:

- Regression loss: Mean square error/ Quadratic Loss, Mean absolute error, Huber Loss/ Smooth mean absolute error, log cosh loss, Quantile loss
- Classification loss: Log Loss, Focal Loss, Kl divergence/ Relative Entropy, Exponential Loss, Hinge Loss.

G.A.Young and Smith in their book argue that loss functions play essential role in statistical decision making theories (Young, young, 2005, n.d.). Machine learning in its own nature adds a wide, unifying perspective to the field of statistics. We can view the concept of statistical decision theory as a game against the whole space or nature (Robert, 2014). In this match we have to choose a particular action of $\hat{y}$ amongst a full set of possible actions, for example action space $\mathcal{A}$. This action is then evaluated against true outcome of nature $y \, \epsilon \, \mathcal{Y}$. The loss function is then denoted as:

$$L : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$$

Where it gives a quantity value of loss that happened when choosing the action $\hat{y}$ while the true outcome by nature was supposed to be y. Therefore, the lower this value, the better prediction. The loss function can also be applied in the concept of quantitative

predictions of specific parameter whereas the outcome has specific real value. In this case we calculate the loss function as the difference or the extent of discrepancy of our predicted value from the real outcome value. This is the definition and application of regression loss. However, in the concept of prediction models we are concerned of the quality of our prediction in the context of true outcome y which defines the classification loss.

In regression the loss function input, action and outcome spaces are all in $\mathbb{R}$. Regression loss functions depend on the value of residual $r = y - \hat{y}$ which is the difference between what you wanted to predict and what you actually predicted. It is a numeric value that needs to be added to your prediction in order to get the right prediction answer. Loss function is zero when residual is zero. Another characteristic of loss function is that they are translation-invariant. Meaning that if you shift the data to some certain point evenly or add some amount to data, the loss function will be the same.

Equation 2

$$l\,(\hat{y} + a, y + a) = l\,(\hat{y}, y)$$

Example. As mentioned above, one of the commonly used loss functions for the regression is Mean Squared Error Loss Function (MSE). Mean squared error is the sum of squared distances between target value and our prediction value. which is calculated by:

Equation 3

$$MSE = \; L(y, \hat{y}) = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

Mean Absolute Loss is another common loss function for regression which measures the sum of discrepancies between the target value and our prediction without considering their direction using:

Equation 4

$$MAE = L(y, \hat{y}) = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Also, for classification purpose one common loss function is misclassification or 0-1 loss function which is determined by:

Equation 5

$$L(y, \hat{y}) = I\ (y \neq \hat{y})$$

Where I is 0 when class is correctly assigned and is 1 when predicted class is incorrectly chosen. More generally one can assign higher loss function value when we have misclassified an item.

Example. Suppose we have a function Standard Squared Error (SSE) $J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^i) - y^i)^2$ which is a half of total squared error of all examples. Where:

$h_\theta(x)$: the predicted value for the i-th example

y: the actual value of i-th example

This is the cost function in which we are aiming to minimize. As we have discussed before the significance of Adaboost algorithm is not only about what it minimize, but rather it is more about how it minimize it. Therefore, in order to explain the loss function concepts in simpler way and to show how AdaBoost algorithm minimizes the loss function, we are

going to work on this function from now on and all definitions, terminologies and calculations are going to be performed on this function. However we should note that the loss function that is implemented in AdaBoost is an Exponential Loss Function (Schapire & Freund, 2012). The Exponential Loss function that AdaBoost is greedily trying to minimize is noted as

$$\frac{1}{m}\sum_{i=1}^{m} e^{-y_i F(x_i)}$$

Where $F(x) = \sum_{t=1}^{T} \alpha_t h_t\,(x)$ is the linear combination weak classifiers that were generated by AdaBoost. This loss function is basically defining an upper bound for the error of AdaBoost function and then tries to minimize this upper bound which will essentially lead to minimizing the error itself. Many other classification techniques such as support vector machine, logistic regression, neural network etc. can also be viewed as the process of minimizing some sort of loss function. Taking this approach to first define and then try to minimize a specific function has many advantages including it helps us to understand and state the goal of learning method explicitly leaving no room for confusion. By doing so we can easily understand what the learning algorithm is trying to do.

Although AdaBoost algorithm can be viewed as a function of greedily minimizing the exponential loss function, however AdaBoost was not developed setting this aim in mind. In fact, there are other learning methods that are trying to minimize the same loss function too, although they perform less effective than AdaBoost. Therefore, we can see that the objective of AdaBoost is not only summarized as what it is minimizing but rather on how it is doing so. On that count, we are going to minimize the Standard Squared Error-

which is a different loss function than what AdaBoost has- with the same method that AdaBoost minimize its exponential loss function.

Loss functions plays central role in assessing machine learning performance and specially boosting algorithms. Class prediction performance have a tight relation to loss functions, which is extensively studied in machine learning and boosting researches. There are various loss functions developed for classification problems in which the Hinge loss for support vector machine and exponential loss for AdaBoost (Friedman et al., 2000) are one of two common ones. Friedman et al. also proved the idea that Adaboost can also be interpreted as a function of minimizing the loss function.

Loss function has a tight relationship with the error of the model. It means that by minimizing the cost function, we are basically minimizing the error which essentially leads to get higher accuracy in the model, which ultimately means better predictions. We are going to discuss Adaboost algorithm concepts and in a context of minimizing loss function using gradient descent method in a next section.

3.1.4 Convex optimization

Before proceeding to gradient descent algorithm, we first need to understand convex functions. In mathematics, a real-valued function defined on an interval is called convex if graph of function lies under or on the line segment between any two points on the graph. This definition is to be met in a Euclidean space or broadly a vector space with at least two dimensions. In mathematic arithmetic we say that:

Definition 1. A function $f : I \to \mathbb{R}$ is named convex if for all points $x, y \in I$ and all $\lambda \in [0,1]$

Equation 6

$$f\big((1-\lambda)\,x + \lambda\,y\big) \leq (1-\lambda)\,f(x) + \lambda\,f(y)$$

Geometrically, the convexity of a function is defined as if we draw a line between any two points in a graph of function, the graph of function lies under or on this line. For example, the two graphs below show a convex and non-convex functions.

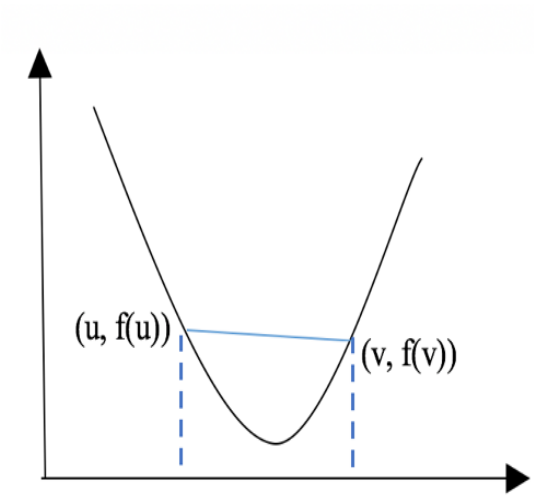

Figure 1. Convex function

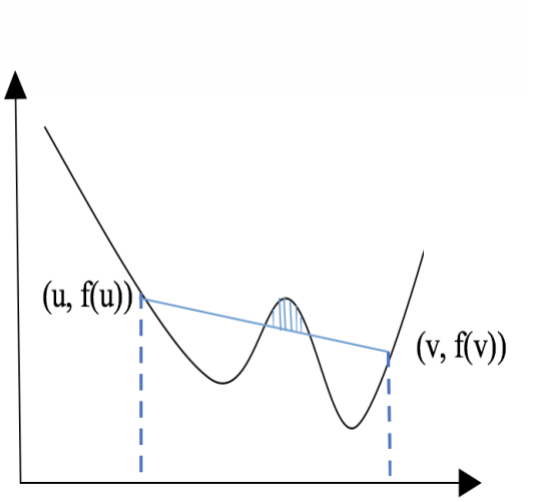

Figure 2. Non-convex Function

Example. Coming back to our cost function where $J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^i) - y^i)^2$ is a convex function.

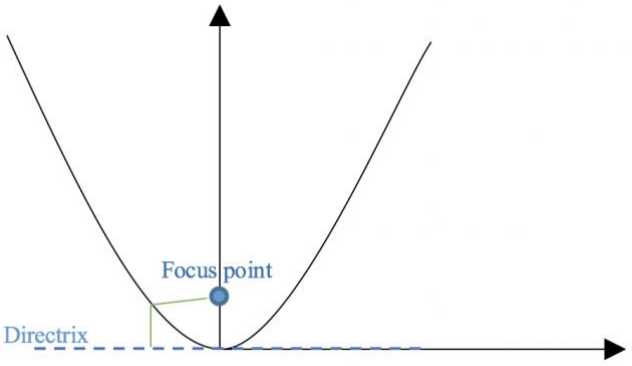

Figure 3. J(θ) a convex curve

Considering our predefined Standard Squared Error (SSE) function that we denoted as $J(\theta)$ in previous section, we can show that this function is proved to have a *convex* curve. Our overall aim is to minimize this convex cost function using gradient descent method that is used in Adaboost algorithm.

3.1.5 Gradient Descent

Majority of data science algorithms are focused on optimization tasks and one of the famous methods to do this is gradient descent algorithm. Gradient descent is the most used learning method in machine learning that almost all of the machine learning models, with small variations, depends on the concepts of gradient descent. Gradient descent is a method of minimizing a function for example loss function. It can be summarized in few easy to understand steps that we are going to discuss in this section. As we have discussed before, most of machine learning tasks are focused on minimizing a function in which we call cost function. In AdaBoost algorithm that we talk here we are focusing on minimizing the exponential loss function which can be interpreted as our Loss function using gradient

descent method. This is the reason why it is important to understand gradient descent method in machine learning and AdaBoost algorithm.

Considering a linear relation of $h(x) = \theta_1 x + \theta_0$ where h(x) is our prediction value and x is our given data point and θ_1 , θ_0 are weights, our task is to find an optimal θ_1 , θ_0 in such a way that the Standard Squared Error (SSE) of our prediction from the actual value is at its lowest and the accuracy of model is at highest. Considering our SSE cost function of $J(\theta)$ we are going to find the minima point in this function using gradient descent. Since we have previously showed that this function is convex we can conclude that the local minima of $J(\theta)$ is also global minima of the whole graph.

The steps of gradient descent start with randomly choosing a starting point by assigning weights of (θ_1 , θ_0) and calculate the SSE. This is exactly where the convex properties of our loss function are taken into consideration. Therefore, because our function is convex we are assured that choosing different starting point does not impact the procedure of finding the minima in graph. In order to find the minima of our graph, assume we are dealing with m dimensional space, we randomly choose a vector of θ_m of weights on our cost ($J(\theta)$) function, we need to move this point toward the minima point in the graph in step-wise manner. The new adjusted θ is calculated using the formula below:

Equation 7

$$\theta_m = \theta - \eta \frac{\partial f(\theta)}{\partial \theta}$$

Where $\frac{\partial f(\theta)}{\partial \theta}$ is a gradient or slope of our cost ($J(\theta)$) function and η is a constant of *Learning rate* which defines the size of step at each iteration.

The higher the learning rate, the bigger the steps and more possible to overshoot minima point during our steps. Considering our convex cost function, we can say if a chosen point is less than our minima point then it means that the gradient or slope is negative, hence the overall value of $-\eta \frac{\partial f(\theta)}{\partial \theta}$ will be positive. Which means the new θ will be bigger and moving into right direction toward the minima.

Therefore, we can be sure that we are moving toward the right direction to our minima point. The role of gradient descent algorithm is to take us from a random initial point to the minima and since the function is a convex, the local minima is also global minima. Essentially finding the global minima of cost function means reducing the error.

The steps of gradient descent algorithm can be summarized as following:

Step 1: Initially assign random values for weights (θ_0 ... θ_m) and calculate sum of squared error (SSE). Figure 4 shows a cost function graph in a three-dimensional space with two weights. As the dot in the bottom of the graph shows the optimal θ with the lowest SSE. If $x \in R$ then we are going to represent the weights in a format of vector as $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$.

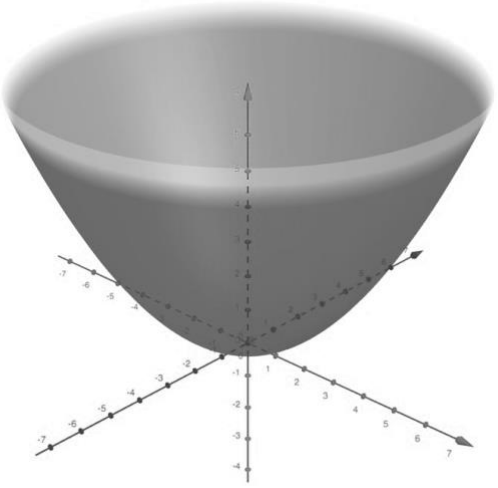

Figure 4. Three-dimensional convex cost function

Step 2: Calculate the gradient. After randomly assigning weights of (θ_1 , θ_0) we come back to the cost function as

Equation 8

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^i) - y^i)^2$$

We are going to calculate the gradient which is $\frac{\partial f(\theta)}{\partial \theta}$. As $x \in R$ therefore we have a vector of $\theta = \begin{bmatrix}\theta_0 \\ \theta_1\end{bmatrix}$ and our prediction formula as $h(x) = \theta_1 x + \theta_0$. The gradient descent of our model is calculated as below.

Equation 9

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{2}.2\sum_{i=1}^{n}(h_\theta(x^i) - y^i).\frac{\partial h(x^i)}{\partial \theta}$$

Since we have two θ then we calculate the derivatives with respect to each θ as

Equation 10

$$\frac{\partial J(\theta)}{\partial \theta_0} = \sum_{i=1}^{n} \left(h_\theta\left(x^i\right) - y^i\right).1$$

And Equation 11

$$\frac{\partial J(\theta)}{\partial \theta_1} = \sum_{i=1}^{n} \left(h_\theta\left(x^i\right) - y^i\right).x$$

Generally, in n-dimensional space the k-th gradient descent can be denoted as

Equation 12

$$\frac{\partial J(\theta)}{\partial \theta_k} = \sum_{i=1}^{n} \left(h_\theta\left(x^i\right) - y^i\right).x_k$$

Step 3: Adjust the weights with the gradients to reach the optimal values where SSE is minimized. In this step we are going to adjust θ with respect to the calculated gradient. The updated θ is denoted as:

Equation 13

$$\theta_k = \theta - \eta \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right).x_k^{(i)}$$

This formula can also be represented as:

Equation 14

$$\theta_k = \theta + \eta \sum_{i=1}^{n} \left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right).x_k^{(i)}$$

We will update the θ until our algorithm is converged meaning that until we are going further away from the optimal minima. Since we are dealing with convex function our stopping criteria is when our gradient or the slope of function is zero. This means that we reached to the minima point in our cost function. However, for variety of other functions there are multiple termination rules for stopping this iteration such as predefining maximum number of iterations, predefining maximum number of seconds in terms of timing out, stop when getting close enough to zero or stop when not adequate improvement is observed. Depending on the context of our algorithm and our specific aim we can choose between any of this criterion to terminate our iteration process. However sometimes one may not reach the optimal point, but meeting the criteria is good enough to stop the process (Bonnin, 2016).

3.2 AdaBoost Algorithm

3.2.1 Introduction

The AdaBoost algorithm, introduced in 1995 by Freund and Schapire, which is the focus of this paper, solved many of the difficulties of the earlier boosting algorithms. This learning algorithm has been attracted by many researchers not only in machine learning community but also in different areas of statistics such as game theory. AdaBoost algorithm were developed to find the possibility of boosting a weak algorithm which perform somewhat better that random guessing to achieve a high performance and particularly accurate algorithm.

Friedman et al. (Friedman et al., 2000) developed a new outlook of boosting algorithm just to say that boosting can be applied in many other contexts than binary classification. Further in time other researchers applied the boosting model in other machine learning perspectives just to find how well this algorithm can enhance the performance of models. Some of these areas that boosting was applied are boosting method for regression, density estimation for large, noisy and high dimensional datasets, survival analysis even with the presence of censored data, multivariate analysis in high dimensional datasets (Bühlmann & Yu, 2003; Yin, Huang, Li, on, 2012, n.d.). Pseudocode for AdaBoost is given in Figure 5.

Algorithm AdaBoost

1: Init data weights $\{w_n\}$ to $1/N$
2: **for** $m = 1$ to M **do**
3: fit a classifier $y_m(x)$ by minimizing weighted error function J_m:
4: $J_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x_n) \neq t_n]$
5: compute $\epsilon_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x_n) \neq t_n] / \sum_{n=1}^{N} w_n^{(m)}$
6: evaluate $\alpha_m = \log\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$
7: update the data weights: $w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m 1[y_m(x_n) \neq t_n]\}$
8: **end for**
9: Make predictions using the final model: $Y_M(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(x)\right)$

Figure 5. Pseudo code for AdaBoost algorithm

AdaBoost can also be applied for multiclass classification cases. Gradient boosting has been successful when applying on decision tree models such as REPTree, Random Forest, Decision Stump, Hoeffding Tree or J48. However, it has been advised to avoid complex base learners on AdaBoost in order to avoid overfitting problem while

maintaining a minimum complexity in order to avoid too weak classifier which will also lead to low margins and consequently overfitting again. Gradient Boosting has empirically been known to have a highly good performance in classification and regression problem. After its first introduction it has been winning over many algorithms in machine learning competitions (Nielsen, 2016).

The key factor for its success is its simplicity yet effective logic that boosts the performance of underlying classifier. AdaBoost calls a given weak classifier repeatedly in a series of rounds while applying revision weights to training dataset in each round. Initially all weights are equal in the first round, then the weights are adjusted in each iteration in a way that misclassified examples are given higher weights and correctly classified items are given a lower weight. This simple strategy improves most of classifiers' performance drastically. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier. One of the reasons why AdaBoost improves performance of classifiers is giving higher weight to misclassified examples. In this way we emphasize more on misclassified items and try to improve the classification model where it poorly performs. Therefore, this property makes AdaBoost more susceptible to uniform outliers in data. The final equation for AdaBoost classification that defines the class value for a binary classification problem is as follows:

Equation 15

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)),$$

where f_m stands for the $m\ th$ weak classifier and θ_m is the corresponding weight. The final function calculates the sign function of the weighted combination of M weak classifiers which values 0 or 1 for binary classification. The whole procedure of the AdaBoost algorithm can be summarized as follow:

Given a dataset containing n points, where

Equation 16

$$x_i \epsilon \mathbb{R}^d , y_i \epsilon \{ -1 , 1 \}$$

In the above formula -1 denotes negative class while 1 denotes positive class values.

The initial weight for each data point in the first run is calculated as:

Equation 17

$$w(x_i , y_i) = \frac{1}{n} , i = 1, \dots , n$$

For each iteration from $m = 1, \dots , M$ we need to

1- Apply our weak base learner to dataset and select the one which has lower weighted classification error:

Equation 18

$$\epsilon_m = \mathrm{E}_{w_m} [1_{y \neq f(x)}]$$

2- Calculate the weight or voting power for the m-th weak classifier

$$\theta_m = \frac{1}{2} ln(\frac{1 - \epsilon_m}{\epsilon_m}).$$

The weight is positive for any classifier with the accuracy of more than 50%. And the more accurate the classification result, the larger the weight. This value shows how good a classifier is doing in the final algorithm.

3- Update the weight for each data point as:

Equation 19

$$w_{m+1}(x_i, y_i) = \frac{w_m(x_i, y_i) exp[-\theta_m y_i f_m(x_i)]}{Z_m},$$

Where Z_m is a normalization factor that ensures the sum of all instances' weights is equal to 1.

If a misclassified case is from a positive weighted classifier, the exponential function in the numerator would be always larger than 1 (y*f is always -1, θ_m is positive). Therefore, misclassified cases would be updated with larger weights after an iteration. The same logic applies to the negative weighted classifiers. The only difference is that the original correct classifications would become misclassifications after flipping the sign.

After M iteration is completed we can get the final prediction by summing up the weighted predictions of each classifier. It has been proven that most of the times decision stumps perform better once chosen as a base learner in additive models and boosting (Friedman et al., 2000).

3.2.2 The Number of Iteration M

As the number of iterations M increase, the complexity of the model increases too, which may eventually result to overfitting of the model to dataset. On the other hand, lower

score of M can lead to more general model that does not fit dataset appropriately and hence will not give suitable results. Therefore, it is crucial to determine the appropriate number of iterations in order to avoid overfitting. We can check for this factor by monitoring the accuracy of the model on validation set or cross-validation or percentage split of data (Nielsen, 2016). There can be various criteria that can be met to stop the iterations. For example, one can decide to stop the iterations once a desirable accuracy measurement is reached or setting the limit for the run time. In our scenario, we stopped the iterations once we didn't see any substantial improvement in the overall accuracy of model by adding more iterations.

3.2.3 Base Learner

As it has been mentioned above, researchers proved that Decision stump algorithm provide the better results when chosen as a base learner in AdaBoost classification technique (Friedman et al., 2000) (Technol2015, n.d.). Therefore, we also decided to choose Decision Stump algorithm as a base algorithm for our model. Decision stump is a tree-based machine learning model with only one node. That is, a decision stump makes prediction model based on only one root decision tree which is immediately connected to the node. Therefore, due its nature, the prediction model generated by decision stump is consisting of only one feature or attribute. The attribute that is chosen as a node to split is being chosen by some sort of purity measures like entropy or Gini index. These measurements define how well this attribute is doing in predicting the results. Since decision stump in its nature is only one single tree with one node, therefore it does not perform well in producing a good

predictive model. But using ensemble method by applying decision stump multiple times on training set boosts the model significantly.

Decision stump algorithm can handle any type of attributes and classifications. For binary attributes the decision tree model consists of two leaves, whereas a missing value can be treated as another category for classification too. For multiclass models it can either generates two leaves whereas if the classification equals to one of the categories and the other leaf for all other categories. Or it can generate a leaf for each possible values of that category. Also, for numeric classifications usually a threshold value is selected and the tree is generated with two leaves – one for values higher than threshold and the other for values lower than the specified threshold. Usually in all different types of classification in decision stump missing value is treated as a separate category.

In a forest of stumps, where the aggregation of multiple stumps defines the model, each of the stumps has an equal say on the final classification. However, in AdaBoost the weight of stumps differs and therefore some stumps get the more say in the final classification than other stumps. The other difference is that in forest of trees the order of trees does not make any difference in the final classification, while in AdaBoost the order of stumps is an important part of model which makes a huge difference in final classification. As mentioned before AdaBoost makes prediction by combining multiple weak learners, in this case decision stumps, to produce more accurate final results. In order to do that sumps of one node is produced based on each attribute in the model and all the attributes has the same amount of weight. The sum of weights always equal to 1 therefore the weights for each attribute is defined as 1/n initially for all stumps.

In this step we need to find the variable that does better than others in classification to choose as the starting first stump. In order to decide which stump does better than others we need to have some sort of measurement such as Gini index or entropy. Gini index is an impurity-based criterion that represent the level of heterogeneity and is broadly used in decision trees (Rokach & Maimon, 2008). Therefore, we are looking for a stump that has the lower Gini index value to branch the tree based on it. It is calculated by the formula:

Equation 20

$$Gini\ (D) = 1 - \sum_{i=1}^{n} p_i^2$$

where P is the estimated probability of that the item is actually in a particular class. The Gini index ranges from 0 to 1 where Gini index of 0 means all elements are belonging to one class which means the impurity is zero, while Gini index of 1 means that all elements are equally distributed to all classes. While choosing a node fir decision tree we would choose the variable that has the lowest Gini index as the root node.

The other impurity measurement is entropy which is broadly used in decision tree algorithms such as ID3, C4.5 (Rokach & Maimon, 2008).

Equation 21

$$E(T) = \sum_{i \epsilon T} -p_i * \ log_2(p_i)$$

Where p is the probability of each class in a set and T is the number of labels. In our case of binary classification, the n equals to 2. The entropy -same as Gini index- also ranges between 0 and 1 where entropy of 0 denotes that the sample is completely

homogenous and if entropy is 1 it means that the sample is equally distributed. The entropy itself does not provide comprehensive outlook of prediction ability of each node, yet we need another measure based on entropy to select the best node. This measure is called information gain which takes into account entropy. Unlike entropy the information gain measures the purity of attribute, hence the more value of information gain, the better partitioning. In order to calculate Information Gain we first need to calculate the entropy based on two attributes (Rokach & Maimon, 2008).

$$E(T,X) = \sum_{c \epsilon X} P(c)E(c)$$

Later the information gain of each node is calculated using:

$$Gain(T,X) = Entropy\ (T) - Entropy\ (T,X)$$

After deciding which impurity measure we want to choose, we calculate that measure for all possible nodes and choose the one that has the lower value of impurity for example Gini index or higher value of purity such as Information Gain. Then the first node of our weak classifier is selected in first iteration. Then in the next iteration all the adjusted weights are calculated based on Decision stump algorithm, is simple and fast yet effective model when chosen as a base learner for boosting or bagging algorithms. This method, when combined with AdaBoost, had been employed in some state-of-the-art models for face detection and other functional models (Sewaiwar, of, 2015, n.d.).

3.2.4 Accuracy

It has been shown in many researches and artificial intelligence competitions that applying the boosting method to any weak classifier can drastically enhance the accuracy

of classification model. The accuracy of applying the base learner alone on our dataset was 67.8% which shows a significant improve by applying boosting method to the base learner. However, with the ensemble method we could boost this accuracy by 9% to 76.38% which demonstrate a superiority of boosting method.

Some of the major benefits of ensemble learning over other conventional machine learning algorithms are:

1- More reliable and stable models due to diversification*.

2- Better predictions

3- It is fast, simple and easy to program

4- It is flexible as it can combine any learning algorithm to achieve better algorithm which consists of a linear combination of these algorithms

5- It is demonstrably effective and can be extended to any datatypes that is textual, numeric and even beyond binary classification

6- It proposed a new mindset which caused a shift in traditional mentality by setting a goal of finding a classifier that does just slightly better than a rule of thumb or random guessing or chance. While other learning algorithms focus more on finding a classifier with the highest accuracy.

*Diversification is a term used in economics, however in this context it refers to having a mixed portfolio of multiple models rather than just one model. This can significantly boost the model accuracy and stability and reduce the bias in data.

Chapter Four: Method

4.1 Research Design

For this study, we focus on two major objectives:

1- Perform descriptive analysis on the database of all patients who has been treated in SMH over the past two decades: For this study, our main goal is to look at changing trends over time in terms of stone size, location, and patient BMI, site of origin, gender and age over 20 years period at the only center in Toronto. We also intend to perform a retrospective cohort study to analyze and compare the effectiveness of the lithotripter currently in use with the two old lithotripters that have been used previously based on clinical retreatment rates. For this matter, we performed bivariate analysis to evaluate and compare the effectiveness of each lithotripter in regard to treatment outcome.

2- Implementing machine learning approach to predict treatment outcome for candidates of ESWL: In order to construct a rigorous machine learning model that can be confidently applied to assist in decision making process, we built our model based on the whole dataset of patients ages over 18 for year 1998 to 2016. Our objective is to build a classification model to predict treatment outcome using SWL prior to making any decision on treatment modality. The success or failure was based on having retreatment plan for the same patient within less than 90 days of initial treatment. We also compared six machine learning algorithms' performance

on dataset in terms of their accuracy and positive predictive value using t-test with 95% confidence interval the results of this assessment are provided in section 7.3.

We performed classification techniques to predict whether the patient is likely to come for a retreatment in less than 3 months or not. For this matter, we selected several classification algorithms and applied them on dataset to build a model. In order to find the best algorithm that fits perfectly on dataset, we performed statistical experiment between these algorithms to discover the best option. Following is the list of algorithms that has been applied:

1. J48- Class for generating a pruned or unpruned C4.5 decision tree. Developed by Ross Quinlan.
2. NaiveBayes- Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data.
3. BayesNetwork- Bayes Network learning using various search algorithms and quality measures.
4. Lazy IBK- K-nearest neighbor's classifier. Which not only can select appropriate value of K based on cross-validation, but can also do distance weighting.
5. MultilayerPerceptron- Neural Network Algorithm is a Classifier that uses backpropagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid except the time for when the class is numeric.

A paired sample T-test was carried out these algorithms with the significance level of 0.05% to compare these algorithms and find the best one that fits the data and gives higher percentage of accuracy.

4.2 Experimental Setting and Dataset

Dataset includes all patient visits for ESWL to the St. Michael hospital from 1990 to 2016. Some of patients had their preoperative and postoperative follow ups conducted at other centers. 58349 ESWL procedures has been performed completely on 31569 patients during this period. Some of the booked procedures were canceled due to patients' complications, unavailable staff, machine malfunction or unknown reasons, therefore we excluded these from the study.

Only patients of 18 years of age or older were included in the study. Patients whose stone size exceeded 25mm in diameter were also excluded from the study. This is because based on urology guidelines patients whose stone exceeds 20mm in diameter are not a good candidate for SWL and even though of noninvasiveness and desirability of this treatment, they may not benefit from SWL and hence may require further auxiliary therapies(Bozzini et al., 2017). So, the exclusion criteria were:

1- Under 18 years of age
2- Stone size 25mm in diameter or more
3- Treatment were identified as canceled or partially completed during one session
4- Patients with special genetic conditions including Staghorne, horseshoe, caliceal, Duplex, Solitary or MSK conditions

Initially among these procedures, 73.06% (38042 procedures) were determined as successful and 26.93% (14027 procedures) as Failure, where failure was defined as "having a re-treatment for the same stone, same side within less than 3 months of initial therapy in a same center". For this matter all data from different sources was aggregated into one single sheet that includes all the data with over 25 columns of information.

Data were coming from 6 different separate sheets where each datasheet contained different information and was inputted by various staff before, on or after the treatment session. Then this data was sorted out in the order of patient, side of the kidney date of diagnosis, date of treatment and the stone. After that the incomplete sessions of treatment that could be because of machine malfunction, patient intolerance or complications were removed from the datasheet. Later, the number of days difference between each two consecutive treatment sessions were calculated and if the difference was less than 90 days then the initial treatment session was labeled as *Failure* and if not then *Successful*.

According to our data since we were unable to collect all follow ups comprehensively we denote failure as retreatment. Some patients may have done their subsequent SWL treatment in other locations across Canada or abroad, albeit due to shortage of SWL resources in Canada and specially in Ontario this case is unlikely. On the other hand, some patients may opt for other auxiliary treatments such as URS after failing to reach the stone-free state with SWL. Since the initial dataset is an aggregation of all SWL visits at St. Michael hospital, it lacks the information about auxiliary treatments that may have been performed on patients. In order to eliminate the effect of this absence of information, after building the rigorous model according to the whole dataset, we tested it against a subset of

patients whose preoperative and post-operative follow-ups were conducted at St. Michael Hospital in order to further validate the results of our classification model.

The success and failure of treatment for these patients were determined based on follow up CT Scan after 3 months. In this way, not only we were able to make a best use out of all data we had, but also, we ensured the validity of our results by testing it against well-known and sure data. The follow ups for these patients had been conducted in SMH and the stone free state were evaluated using preoperative and post-operative CT Scan. With this method we ensure that lack of a follow up information have the least effect on our prediction model.

4.2.1 Protection of human subjects

All the data were collected from SMH patient health record system repository retrospectively. All subjects of study were anonymous and partial Postal code and MRN numbers were provided. Patients were aware of partial data collection during the course of treatment. Data were collected retrospectively since the initial introduction of SWL technology at SMH. No names, personal or family information were captured during the study.

Data were only accessible from local intranet server in SMH and no offsite access was possible which ensured the security of dataset. Also any format of data, analysis and model were saved and maintained on local server for the duration of study and were removed after the official study period ended in 2018.

Biomedical safety, data privacy and confidentiality courses and exams were performed by myself and other staff involved in the study prior to access the dataset. The whole study and data collection procedure were fully complied with HIPAA standards and guidelines. In order to use patients demographic information only a postal code was used. A rationale of the reason for using the postal code were given and was confirmed by the ethics committee of SMH before getting access to them. The rationale for using the postal code was to being able to locate patient geographical pattern to SMH during the 20 years period and to find out which areas of the province has higher referrals to SMH in order to be able to optimize the waiting time for patients in future studies.

4.3 Attributes

The list of attributes that has been investigated and implemented in the final model and their values are shown below in table 2. Initially more variables were assessed prior to the modeling however only these ones were included in further analysis as they have contributed more in the final result of model.

4.4 Preprocessing

The preprocessing of our data includes data cleaning, data integration, data transformation, data reduction and feature selection in order to improve our evaluations and essentially enhance the model accuracy. The comprehensive explanation of each step in preprocessing this large dataset is discussed in upcoming sections.

Table 2- Values of final investigated attribute in the model

Attribute	***Value***
Side	Left / Right
Electrode	integer
Stone Treatment Number	integer
Shocks	integer
Location	LC,LU,MC,MU,P,UC,UU,UVJ,RP
Area	integer
Gender	Female /Male
BMI	Real
Age	18,…,95
Number of Stones	integer
Family History	True, False
Asymptomatic	True, False
Stent	True, False
Frequency	120, 90, 60
Antibiotic	True, False
Maximum Voltage	integer
Lithotripter	Dornier MFL 5000, Philips LithoTron , Storz Modulith SLX-F2
Result	Success, Failure

4.4.1 Data cleaning

Data included 6 independent and heterogeneous datasets including demographic information dataset, screening dataset, stone information dataset, treatment information dataset, treatment details information, and medical record dataset. The first step in preprocessing these miscellaneous datasets was to combine them together in order to

generate a comprehensive and homogeneous dataset that can be further used as a complete repository of patients who underwent SWL at St. Michael hospital.

The next step of preprocessing data included finding outliers and correct the ones that was possible. As one of our missions to conducting this research was to make the best use out of this data, we tried to fix the errors as much as we could instead of just removing the inconsistent ones. This means that some of the outliers were just mistakenly entered by one staff on one dataset, however in other parallel dataset the correct input was entered. By defining certain constraints on some attributes, we could locate and find the outliers and replace them with the correct example from the other dataset. This process itself took a lot of time since data were entered manually by different people and also it included some text data on some of the attributes which required to alter manually one after the other. For example, the complication attribute was handwritten information by staff. This was also the case for older data from 1990s when predefined instances were not constructed by the electronic patient record system, hence staff had to manually input most of the data by themselves which resulted in many mistakes in dataset.

4.4.2 Data Integration

In this step data with different representations were put together and conflicts in the data were resolved. Preprocessing of data including the part of combining datasets and removing duplicate and outliers were performed using VB.NET platform in Excel. Since the data came from multiple different segregated datasheets, and each datasheet came from different origin and was inputted by different people, hence the format is slightly different

between them and they do not talk to each other in a context of data integration. Therefore, the first step was to aggregate all data into one unified datasheet with a uniform format.

Our mindset was to make use of as much as data as possible and since each of datasets were generated by different people in different sections of hospital, they were many outliers and mistakes in the data collection which needed to be fixed by assigning and introducing appropriate restrictions on attributes. For example, the treatment information dataset was inputted by general nurse prior to treatment, however the treatment detail dataset was inputted by doctor or operation room nurses after the treatment has been fully completed. Therefore, it was crucial to find and correctly input the adjustments that had been done during the operation due to patient's condition.

4.4.3 Data Transformation

In this step of data preprocessing data is normalized, aggregated and generalized. For this matter, after combining relational databases we normalized the aggregated dataset in order to remove redundancy in data and attributes. Furthermore, in the aspect of data mining technique, some of the attributes that contained numeric values were centered to have zero mean and some of the attributes were normalized to reduce their range to [0,1] to compute normalization intervals. As such patients' body mass index (BMI) were calculated from height and weight with the calculation of: Weight / (Height^2) where weight and height are measured in kilogram and meter respectively.

4.4.4 Data Reduction

This step aims to present a reduced representation of the data in a data warehouse. For this step first, we applied our inclusion and exclusion criteria to the data. This process included:

1- Removing incomplete or canceled treatments
2- Removing patients under 18 years
3- Removing patients with special stone type or congenital disorders of kidney including Staghorn kidney stones, horseshoe kidney, caliceal stones, Duplex kidney, Solitary kidney or medullary sponge kidney conditions in order to achieve the most coherent and unbiased model
4- Removing duplicate treatments for a patient in a same day
5- Removing patients whose stone were not located in kidney or ureter (some patients had stones in gallbladder, bladder or other parts)

After the inclusion and exclusion criteria were determined and errors were fixed per the previous section, we removed the inconsistent data rows with extreme number of blank data or the ones that were not possible to fix.

4.4.5 Feature Selection

In addition to data reduction, we also performed data reduction in the context of feature selection in order to reduce number of features that are useless to the modeling or have many missing values. Some of the attributes for example *patient's complications, treatment complications, symptoms, treatment plan, max voltage, electrode, fluoroscope*

included more than 50% of missing values and hence were removed from dataset. In addition to that, some attributes that contained information about the post treatment were also removed due to nature of this research which we only require attributes prior to getting to treatment in order to predict the outcome. This technique is further explained in detail and is well executed in this paper (An, Huang, Huang, & Cercone, 2005).

For the remaining attributes we performed feature selection method using Pearson's correlation coefficient method. With this method we can calculate the correlation between each attribute and the output variable and select only those attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) and drop those attributes with a low correlation (value close to zero). In Weka, the *CorrelationAttributeEval* does the work by using *Ranker* search method. This technique evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal values are considered on a value by value basis by treating each value as an indicator. An overall correlation of nominal attributes is then generated by averaging the weighted average of its values. We performed this task by using 10 fold cross validation of training set with a threshold of -1.7976931348623157E308 in which attributes can be discarded if exceeding this value in a Ranker method.

The results of attributes selection forced us to remove attributes of *Family history, complications* and *frequency* since they did not provide any useful information to the model while adding complexity and unnecessary dimensions to the model. Therefore, we also removed these attributes and ended up with 11 attributes overall. The result of attribute selection feature is attached to the index of this paper for further reference.

Chapter Five: Statistical Analysis Result

5.1 Descriptive statistics

37013 patients had ESWL treatment at SMH lithotripsy unit from 1980 to 2016. Among those 19546 (52.8%) were males and 11176 (30.19%) were females while the rest of 17.01% had missing value for gender. Majority of patients were located in greater Toronto area (22102 patients, 59.6%), however there were referrals from other provinces including Manitoba, Yukon and even Quebec. Some patients had multiple visits to SMH for retreatments, however this data is not included in this part of analysis as we wanted to focus on the number of unique patients who are referred to SMH center. Figure 7 represent the distribution of patients' geographical locations. City of Toronto compromised the largest number of referrals among small municipalities followed by Mississauga and Scarborough. Figure 6 demonstrate the top 15 municipalities that had higher patients' referral to St. Michael Hospital. Inclusion criteria for this figure included patients over 18 and renal or ureteral stones less than 20 mm.

Table 4 illustrate the average stone size, number of stones, age and BMI dividend by the ten-year periods. The result of nonparametric statistical analysis represent that the distribution of stone area is statistically significant among three lithotripters with (p-value = 0.000). The p-values from the Kruskal Wallis test of statistics is demonstrate in table 4.

By running ANOVA test to determine if there is any statistically significance difference among 3 decades we were able to find that the average number of stones has significantly

increased during the past three decades from average of 1.35 in first decade compared to 2.28 in the later decade (p-value= 0.000). However, on the other hand the stone area has significantly decreased from 75.87 mm^2 to 60.49^2 (p-value= 0.000). The average age of patients also escaladed from 49.80 to 54.36 (p-value = 0.000). BMI of patients however did not differ that much in number value as the average BMI in all 3 decades were ranging in the 26. However, even this small difference was still statistically significant (p-value=0.046).

Table 3. Comparison time frames in terms of stone size, age and BMI by ANOVA

Duration	Number of Stones	Stone Area	Age	BMI
1993-2001	1.35	75.87	49.80	26.30
2001-2010	1.99	67.86	51.63	26.57
2010-2016	2.28	60.49	54.36	26.52
P-Value	0.000	0.000	0.000	0.046

The post-hoc analysis that has been done by pairwise comparison of each time span in terms of the stone area demonstrates that average stone area was significantly Lower in timespan 3 compared to timespan 1 and 2 (60.49 vs 67.86 and 75.87, p-values of 0.000 and 0.000). However, this analysis suggest that the average stone area was not statistically significant between time span 1 and 2 (75.87 vs 67.86, p-Value=0.426). The post-hoc analysis for number of stones shows that the number of stones in the third time span is significantly higher than in first and second timespans. (2.28 vs. 1.35 and 1.99, p-

values=0.000 and 0.0000). Also, the number of stones in second time span was significantly higher than of first time span (1.99 vs 1.35, p-value=0.000).

The average BMI during the whole period is 26.49. Post-hoc pairwise comparison of BMI among three timespans shows that the average BMI in second time span is significantly higher than first timespan (26.57 vs 26.30, p-value=0.043). However, comparing first to third (26.30 vs 26.52, p-value = 0.146) and second to third timespan (26.57 vs 26.52, p-value= 1.000) shows no statistical significance of average BMI.

The average age and number of stones are significantly higher in third timespan compared to first and second (2.28 vs 1.35 and 1.99 , p-values = 0.000, 0.000 for age and 54.36 vs 49.80 and 51.63, p-values= 0.001 and 0.000 for number of stones). The same result was observed when comparing average age and number of stones of second timespan to first timespan (51.63 vs. 49.80 with p-value= 0.000 for age, and 1.99 vs 1.35 with p-value= 0.000 for number of stones).

Table 5 demonstrates the gender frequency in each time span. We can see that the number of male patients is significantly higher than females in each three time periods (p-value= 0.001). This finding can be justified by the findings from other researchers that suggest the prevalence of kidney stone is higher among men compared to women (Stamatelou et al. 2003). The overall ratio of men to women in our database is 1.88 which suggest that the number of men treated with SWL in Toronto is almost two times higher than of women.

Table 4. Gender frequency among three timespans

		Timespan			
		1990-2001	2001-2010	2010-2016	Total
SEX	Male	10640	12259	5232	28131
	Female	5384	6755	2784	14923
Total		16024	19014	8016	43054

In terms of geographical referral pattern, we could see that the majority of patients resided in Toronto downtown area. City of Toronto has the highest number of patients referred to SMH followed by Mississauga and Scarborough. Very few patients were also referred from out of province locations including Quebec and Manitoba, however this case was very rare. The below figure 6 demonstrate the top 15 municipalities that had higher number of referrals to SMH for SWL. Figure 7 demonstrate the visual density of patients' referral pattern based on location.

Figure 6. Distribution of patients according to site of origin top 15 municipalities

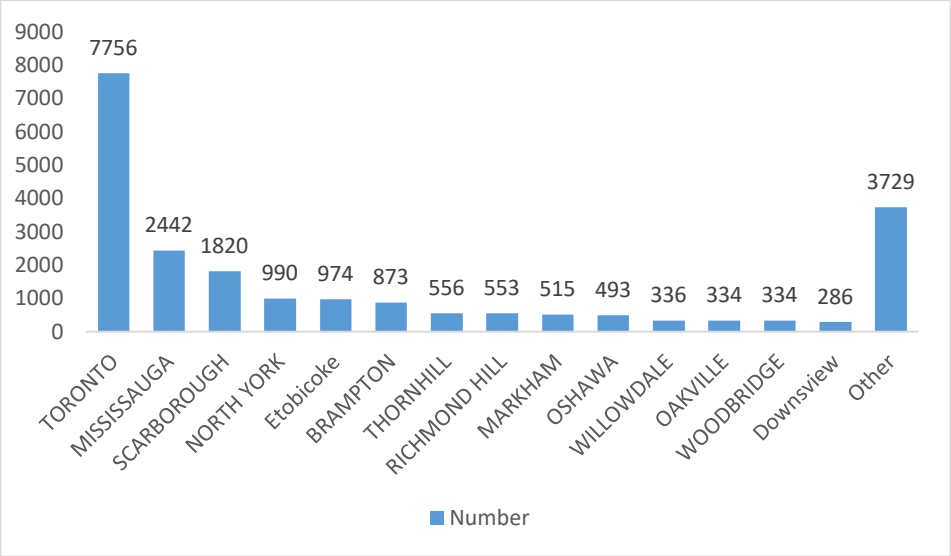

Figure 7. Density of patients' distribution according to site of origin

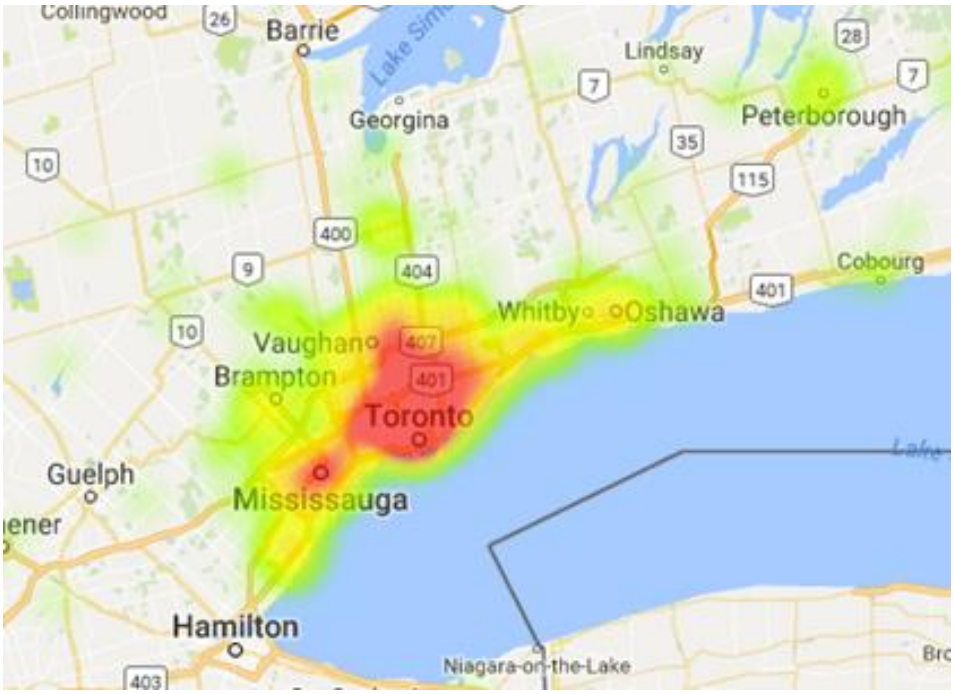

Table 5. Distribution of renal and ureteral stones based on location and diameter

	size	Stone size (mm)	Number of Stones	Retreatment rate
Kidney				
Unspecified	4290	65.19	1.68	25.43%
Lower Calix	17612	65.01	1.53	23.51%
Median Calix	4954	66.34	1.72	23.29%
Upper Calix	2526	69.65	1.85	33.29%
Ureter				
Lower ureter	4333	65.09	2.22	22.45%
Median ureter	1877	67.81	1.89	24.82%
Upper ureter	8523	65.99	1.82	28.15

Table 6 illustrate the distribution of renal and ureteral stones based on location of stone in ureter or kidney. Renal calyx is flower shaped or funnel shaped chambers of kidney through which the urine passes. It is located at the end of pelvis area which is the lower part of torso between abdomen and legs. As we can see in the table the retreatment rate for the stones in upper calix is higher than other locations in urinary tract system.

5.2 Bivariate Analysis

In a bivariate analysis of dataset Chi square test of homogeneity along with fisher exact test were performed on nominal attributes with confidence interval of 95%. Statistical tests were performed to ascertain the effect of each attribute in constructing the boosting model. The statistical analysis was performed on the subset of dataset that has been used to validate the result of model.

This means that all patients whose preoperative and postoperative follow-ups were administered in SMH and had a follow up CT scan to assess the stone free status. The results of statistical analysis show there is significant difference in treatment success rate between men and women (P-value = 0.032) whereas women showed better response to SWL treatment than men (74.85% vs 71.82%).

Studies has shown that the management of kidney stones for people with extreme ages, more precisely elderly and pediatric patients, does not differ compared to standard adult population when treating with either ureteroscopy or percutaneous nephrolithotomy. However pediatric patients should be chosen more attentively for SWL treatment. On the other hand, these studies show that elderly may benefit less from SWL treatment when

compared to standard adults (Ng, 2009). This can be further corroborated from our results which shows that age significantly influence the result of SWL treatment (P-value = 0.013).

Having Calyceal stone, Duplex, Solitary, MSK Stones did not defer in treatment outcome. However, the SWL treatment outcome on Radiolucent stones which are stones that do not usually appear on KUB imaging is significantly different than non-radiolucent stones with p-value of 0.000. Furthermore, having Staghorn stones which spread and branch through the renal pelvis and may fill the entire area have significant effect on treatment outcome with p-value of 0.000. Patients who are suffering from Horseshoe kidneys have shown less success to SWL treatment (p-value = 0.007).

Horseshoe kidney is a congenital disorder that affect 1 in 500 people, with no symptoms and studies showed that these patients are more prone to get stone disease. Duplex kidneys showed no significant effect on the SWL treatment outcome (P-value = 0.240). Duplex kidney is a condition where two ureters are connected to a kidney, whereas in normal urinary tract system each kidney is connected to only one ureter. Duplex kidney is considered a safe condition meaning that they are a normal variant and may have few to zero medical conditions associated with them. This condition happens in one percent of the population. Although removed from our classification model but these rare congenital disorders were still studied in statistical analysis part.

Another interesting finding was that having solitary kidney or only one kidney does not affect in treatment results (p-value = 0.414). Patients with solitary kidney may be born with only one kidney or have two kidneys but only one of them is functional due to some disease or cancer. Having medullary sponge kidney (MSK) also does not make any difference in the results of SWL treatment (p-value=0.095). medullary sponge kidney, also

known as Cacchi Ricci disease is a special and rare congenital disorder in which the collecting tubules or ducts in one or two kidneys are dilatated into cysts. This means that small cystic malformations are formed in tubules of the fetus kidney which cause the appearance of kidney look like a sponge. Although the prevalence of this disorder is 1 in 5000 of the population (Garfield & Leslie, 2019), urologist and nephrologist specializing in kidney stone usually diagnose this condition in adulthood as a result of having kidney stone. Patients suffering from MSK are more prone to having recurrent kidney stones. This disorder is less common in children, however the severe cases of MSK which engage bones can happen in children.

In addition to that, although having a family history of kidney stones significantly increases the chance of stone formation in patients (Garfield & Leslie, 2019), however, our results shows that this factor does not influence the treatment outcome of SWL (p-value= 0.266). This result also suggests that having previous SWL treatments does not differentiate the treatment results of subsequent therapies (P-Value = 0.686). Furthermore, SWL treatment performs equally same on kidney sides of left and right with p-value 0.466.

The electrode that has been used for each patient may have been used once or twice or never before. This factor did not significantly affect the chance of retreatment for patients whose electrode were used before (P-value = 0.93). Furthermore, patients who had stent tubes installed were shown no difference in succession from SWL treatment with P-values of 0.128. Due to an extremely low ratio of patients who had nephrostomy tubes to patients who did not, we removed this factor for further analysis and also from classification prediction.

We also compared the probable effect of lithotripter in treatment outcome. The number of patients who had been treated with the first-generation lithotripter with only 45 patients with Dornier MFL 5000 compared to 859 for Philips LithoTron and 189 for Storz Modulith SLX-F2. The result of bivariate analysis of lithotripter to treatment outcome suggests that lithotripter does not influence treatment outcome (p-value= 0.85).

As it is illustrated in the table below, the success rate for each of lithotripters are 60% for Dornier MFL 5000, 62.3% for Philips LithoTron and 64% Storz Modulith SLX-F2. Although the results show improvement in success rate for the latter technology however our findings do not suggest that this difference is statistically significant. This finding can be justified by the lack of balanced data in all three different lithotripters.

Table 6- Chi-Squared test of homogeneity of lithotripters

RESULT * LITHO Crosstabulation

			LITHOTRIPTOR			Total
			1	2	3	
RESULT	Failure	Count	18	324	68	410
		% within LITHO	40.0%	37.7%	36.0%	37.5%
	Success	Count	27	535	121	683
		% within LITHO	60.0%	62.3%	64.0%	62.5%
Total		Count	45	859	189	1093
		% within LITHO	100.0%	100.0%	100.0%	100.0%

5.3 Multiple Regression Model

A binomial logistic regression was performed to ascertain the effects of Side, Electrode, Stone treatment number, gender, age, BMI, shocks, area, family history, frequency, asymptomatic, number of stones, antibiotic, maximum voltage and location of

the stone on the likelihood of patients ending up with a retreatment of SWL when considering all factors at the same time. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure in the next paragraph. Based on this assessment, all continuous independent variables were found to be linearly related to the logit of the dependent variable. 790 cases (71.4%) out of 1106 cases were initially included in the analysis and the remaining 28.6% had some missing values that were excluded from further analysis. Furthermore, there were 10 standardized residual outliers with values more than 2 times of standard deviation that ranged from 2.569 to 8.698 of the standard deviations, which were removed from analysis.

In the multiple regression analysis using binomial logistic regression model, one of the assumptions is that there should be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable. In order to test this assumption, we used Box Tidwell approach. For this purpose, each numeric independent variable was multiplied to its natural log transformed variable and the combination of all variables were then tested against the dependent variable to find any linearity between them. After running a binary logistic regression with added interaction terms, we will only focus on the p-values obtained from these interaction terms in order to find any linearity.

The null hypothesis is “There is a linear relationship between continuous independent variable and a log transformation of dependent variable”. If the interaction term is statistically significant, the original continuous independent variable is not linearly related to the logit of the dependent variable. Based on this assessment, all continuous

independent variables were found to be linearly related to the logit of the dependent variable, therefore all variables were kept in the analysis.

Table 7- Box-Tidwell Analysis of Linearity

Box Tidwell Analysis						
	B	S.E.	Wald	df	Sig.	Exp(B)
Ln_StTxNo by StTxNo	-20.017	12243.820	.000	1	.999	.000
Ln_Shocks by shocks	.000	.001	.218	1	.640	1.000
AREA by Ln_AREA	.001	.004	.162	1	.687	1.001
Ln_BMI by bmi	.093	.085	1.190	1	.275	1.098
Ln_age by age	-.045	.041	1.201	1	.273	.956
Ln_Nstone by Nstones	.070	.400	.031	1	.860	1.073
Ln_Voltage by MAXVOLTAGE	2.115	4.536	.217	1	.641	8.292

a. Variables entered on step 1: side 1=left 2=right, StTxNo, shocks, AREA, GenderM0, bmi, age, Nstones, FamilyHx, ASYMP, stent 0=no 1=yes, FREQ, ANTIBIOT, MAXVOLTAGE, LITHO, LOCATION, Ln_StTxNo * StTxNo , Ln_Shocks * shocks , AREA * Ln_AREA , Ln_BMI * bmi , Ln_age * age , Ln_Nstone * Nstones , Ln_Voltage * MAXVOLTAGE .

The logistic regression model was found to be statistically significant with $\chi_2(21) = 274.353$, p-value < 0.0005. From a binomial logistic regression analysis, we can conclude that 37.9% of variation in the dependent variable can be explained by the model. This conclusion is based on pseudo R_2 value that was obtained from Cox and Snell R_2 method. The model is able to correctly classify 78.4% of cases. Sensitivity of model is 92.4% whereas specificity of model is 52.9%. Of the initial 24 variables two of the variables (Maximum voltage and antibiotic) were removed from analysis due to the high number of missing values. The final binomial logistic regression model overview is presented below in the table 8.

From the 22 predictor variables, 10 variables were found to be statistically significant. These variables are stone area, patient's BMI, age, location of stone, number of shocks and frequency. Among the different locations of stone, middle calyx, middle ureter and ureterovesical junction stones also known as UVJ were found to be statistically significant with likelihood ratios of 4.628, 2.228 and 3.281 respectively. Also, patients who received less amount of frequency shocks were found to be 32.78 times more likely for a retreatment of SWL. However, as we have discussed before due to lack of resources and staff in SMH, very few number of patients received 60 or 90 shocks per minute and majority of patients were treated using 120 shocks per minute frequency.

We conclude from the logistic regression model that side of the kidney, Electrode, Stone treatment number, gender, number of stones, family history, asymptomatic, lithotripter and stent insertion do not have significant effect on predicting the treatment outcome. As we can see, the traditional statistical approach has many limitations which to name a few, it is not able to handle missing values, outliers have significant effect on the analysis and rigorous amount of assumptions that need to be met preliminary to running any test. For these reasons we were not able to use all of the data for multiple regression statistical analysis part due to high number of missing values and not meeting assumptions.

Furthermore, unlike machine learning approach where we can build a model based on all data and test it against some other dataset, traditional statistical approach does not provide such solution therefore we could only take a use of a very small subset of our database in which all patients had follow up information. This is the same database that we used in order to validate our machine learning model performance.

Table 8- Binomial Logistic Regression Analysis

Variables in the Equation						
	B	**SE**	**Wald**	**Df**	**Sig.**	**Exp(B)**
Side1=left 2=right(1)	-.259	.200	1.672	1	.196	.772
Electrod	.086	.129	.442	1	.506	1.089
StTxNo	-40.694	3380.36	.000	1	.990	.000
AREA	-.013	.002	36.356	1	.000	.987
GenderM0(1)	-.247	.208	1.405	1	.236	.781
bmi	-.039	.018	4.844	1	.028	.962
age	-.023	.008	8.606	1	.003	.977
Nstones	.080	.120	.440	1	.507	1.083
FamilyHx(1)	.365	.256	2.033	1	.154	1.441
ASYMP(1)	-.089	.224	.156	1	.693	.915
LITHO(1)	.273	1.191	.052	1	.819	1.314
Location			25.962	7	.001	
Location(1)	1.532	.506	9.161	1	.002	4.628
Location(2)	.801	.349	5.261	1	.022	2.228
Location(3)	.907	.703	1.663	1	.197	2.476
Location(4)	.785	.480	2.678	1	.102	2.193
Location(5)	.011	.288	.001	1	.970	1.011
Location(6)	-.782	.700	1.248	1	.264	.458
Location(7)	1.188	.368	10.413	1	.001	3.281
FREQ			13.029	2	.001	
FREQ(1)	3.489	1.152	9.168	1	.002	32.78
FREQ(2)	1.461	.980	2.225	1	.136	4.312
shocks	-.002	.000	25.317	1	.000	.998
stent 0=no 1=yes(1)	.469	.327	2.064	1	.151	1.599
Constant	48.760	3380.36	.000	1	.988	150.000

a. Variable(s) entered on step 1: side 1=left 2=right, Electrod, StTxNo, AREA, GenderM0, bmi, age, Nstones, FamilyHx, ASYMP, LITHO, Location, FREQ, shocks, stent 0=no 1=yes.

In addition to that, the specificity of our model using machine learning approach with AdaBoost algorithm significantly outperformed the statistical analysis 65.6% vs 52.9%. The specificity of the model is very important in this study because it defines the ability of model to detect percentage of patients who fail the treatment so that can be offered an alternative treatment instead of SWL.

Indeed, one of the main goals of this study is to identify patients who fail the SWL treatment, as such we focused more on finding the classification model with higher negative predictive value or recall. For this reason, the AdaBoost algorithm significantly outperforms other classification algorithms in correctly classifying patients in failure class. Machine learning approach and statistical approach has different views of same problem in a way that machine learning approach is more liberal in terms of data distribution and assumptions however statistical approach is more conservative.

In general our findings suggest that statistical approach on this specific dataset is heavily restricted by the assumptions about the data and its distribution, making it unfeasible to take the advantage of majority of data rows, while on the other hand machine learning approach follows the abundance ideology which encourage the more data the better and hence is more pragmatic as it encompasses the more data rows in the final model (Yang, 1999). In this thesis we used both techniques to further ascertain and validate the results of one another.

Chapter Six: Machine Model Result

6.1 Model Overview

The AdaBoost model that was produced on this dataset consisted of 30 iterations. Summary and pseudocode of AdaBoost iteration can be illustrated as:

First initialize the weights of data points then

FOR each t=1 to N, do

Select a weak classifier that does slightly better than a chance 50% on data

Compute the classification error for the chosen weak classifier for each data point

Increase the weights of item sets that are misclassified

Normalize weights

END

Output strong classifier as a linear combination of N weak classifiers

On the above code N represents the number of iterations. It is analogous to the number of weak classifiers in the final strong classifier. We used 30 number of iterations for our model. This means that the base learner was applied to the dataset 30 times iteratively and in each iteration the misclassified itemset were given higher weights so that in the next iteration focus will be more on these misclassified examples. Although increasing the number of iterations usually increase the accuracy of model, we ceased adding more iterations to the model as we wanted to avoid overfitting the model. The result of final model can be found in appendix. The first leaf of our model on the node is based on stone treatment number with the weight of 1.12. This finding suggests that patients who have had SWL treatment for their first time are more likely to benefit from SWL at first

place. However, patients who have had previous SWL treatments are classified into failing this treatment.

The second node of model is constructed based on number of stones with weight of 0.62. This node classifies patients who has more than one stones in each kidney as failing the treatment. The third node is constructed based on area of the stone which is calculated by multiplying two diameters of stone. This node classifies stones less than 64 mm^2 as a failure with the weight of 0.49. The further nodes of this model are generated based on misclassified examples in the previous iteration which can be found in appendix.

6.2 Performance Measurements for Unbalanced Data

The overall accuracy of model is 76.38% with mean absolute error of 0.2996 and root mean squared error of 0.3912. The area under ROC curve is 0.843 and PRC area is 0.835 for overall classification. The sensitivity of model is 0.875 which means that 87.5% of all patients who succeeded in SWL treatment were correctly identified by our model. On the other hand, the specificity is 0.6528 which measures the proportion of correctly classified patients who actually failed the treatment. This measure is one of the most important factors in constructing our model as such the aim was to correctly identifying patients who are going to fail the treatment.

Furthermore, the probability that subjects with a failure class truly failed the treatment or in other words the negative predictive value of the model is 0.839. Likewise, the positive predictive value or precision which is the probability that subjects with a success result truly succeeded in the treatment is 0.7159. These two factors describe the performance of model and are related to the prevalence in addition to the accuracy of model

itself. The F1 score and Matthews Correlation Coefficient of our model are 0.7875 and 0.5413 respectively. F1 is a function that takes into consideration both precision and recall. The formula for F1 score is as follows:

Equation 22

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

F1 score is more desirable when we have unbalanced data and are seeking to make a balance between precision and recall. This is usually the case when we have tangible or intangible costs associated with False Negative or False Positive values. For our model as we assumed the cost of False Positive is higher than False Negative, this measure can be a better reference to assess the accuracy of model.

Data imbalance is frequently observed in data mining applications specially in biomedical field. When the class distributions are of a very unequal and unbalanced sizes, some measures such as accuracy is not a good measurement to assess the performance of model as it does not take into account the effect of unbalanced data. Sometimes resampling techniques or oversampling such as Synthetic Minority Oversampling Technique (SMOTE) can be applied to tackle with the issue of imbalanced dataset. However, each of these techniques have their own issues and struggles. For example, resampling technique to reduce number of instances in higher class may reduce the overall accuracy as we limit ourselves from taking a benefit from useful data. Also, oversampling techniques may shift results further away from their actual value and as such decrease performance measurements.

One of the more reliable measures that can be used in these scenarios is Matthews Correlation Coefficient (MCC). Mathews Correlation Coefficient is an evaluation technique in machine learning that asses the quality of binary classifications and as it involves all of the four quadrants of a confusion matrix, it is considered as a balanced measure. This measurement is first introduced by Brian W. Matthews in 1975 (Matthews, 1975) and ranges from -1 to 1 whereas coefficient of 1 demonstrate a perfect prediction, 0 shows prediction no better than flipping a coin and -1 represents total disagreement between model prediction and the real value. MCC can be calculated using the following formula:

Equation 23

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

As such we can see the MCC of our model is 0.5413 which shows a pretty decent prediction performance on our unbalanced data (Liu, An, & Huang, 2006; Liu, Yu, Huang, & An, 2011).

6.3 Ensemble Learning Vs Other Classification Models

In this section we are going to compare the performance of our model to some of the other well-known machine learning models. For this purpose, we built various prediction models based on 5 other learning algorithms and compared their performance results to our base algorithm of AdaBoost. The results were then tested pairwise against our model using student t-test to identify whether the results are statistical significant or not. The performance measurements that has been compared are Accuracy, Mathew's

Correlation Coefficient, F1 measure, Area under ROC curve, Root Mean Squared Error. The table below illustrates the result of this comparisons. Each of these 5 classifiers' performance were tested against AdaBoost algorithm. The classifiers are C4.5 decision tree classifier, Naïve Bayes, neural network Multilayer Perceptron, Bayesian Network and k nearest neighborhood lazy IBK which are some of the top-notch algorithms that are extensively used in classification problems specially in biomedical datasets and information retrieval processes (X. Huang & Hu, 2009).

As we have done multiple comparisons and performed multiple t-tests on dataset, in order to eliminate type I error, we used Bonferroni correction technique and reduced significance level to 0.01. The star on top of numbers represent that the results of t-test was statistically significant.

Table 9. Classifier Performance Comparison

Classifier	**AdaBoost M1**	trees. J48	Naïve Bayes	Multilayer Perceptron	Bayesian Network	Lazy.IBk
Accuracy	**77.59**	75.26*	75.82*	69.11*	76.49*	57.52*
MCC	**0.53**	0.46*	0.47*	0.34*	0.49*	0.09*
F1 score	**0.84**	0.82*	0.83*	0.76*	0.83*	0.66*
Area under ROC	**0.80**	0.74*	0.75*	0.74*	0.78*	0.54*
Root mean squared error	**0.39**	0.43*	0.43*	0.51*	0.40*	0.65*

As we can see from the table all the performance measurements for AdaBoost algorithm were significantly higher than all other classifiers. The * beside the number shows that the difference is statistically significant in a two by two comparison of AdaBoost to each of other algorithms.

Surprisingly AdaBoost has higher values of performance measurement compared to all other algorithms that we tested and these differences are statistically significant even though of its small proportion. This is how we came up to choose AdaBoost algorithm at a first place. We initially compared many algorithms on our dataset and essentially chose the one that had highest performance on the data. This statistical comparison affirms the superiority of ensemble learning method over other classification techniques.

The comprehensive process of comparing the algorithms on this dataset is included in the appendix 2. For this purpose, we used the experimenter environment of WEKA platform that provides a user-friendly environment to compare and test multiple algorithms on the same or multiple datasets at the same time. In order to reduce the run time, we cross validated each algorithm on the dataset with 3 folds. Later the final models and their accuracy measurements were statistically compared pairwise with student t-test in multiple comparisons against the AdaBoost algorithm.

Chapter Seven: Analysis and Discussion

7.1 Importance of Treatment Prediction for Management of Urinary Stone

ESWL failure and repeated procedures can increase the risk for potential side effects. This also will result in delay in receiving other definitive endourological treatments, increasing morbidity and higher cost for patients and healthcare system (Joseph et al. 2002). Therefore, it is crucial to establish a model to specifically target patients who will most benefit most from this treatment and provide other definitive options for those who won't. In this regard our machine learning approach not only can benefit patients to find the most effective treatment for them, but also help physicians to select most effective treatment with highest possible success in achieving stone free status with minimal morbidities based on patient's characteristics.

In regard to statistical analysis, there are number of other studies which found similar results to ours. However, some of these studies collected different attributes which resulted in more significant variables in their equations. Abad. et.al identified stone size, volume, mean density, BMI and skin to stone distance as predictive variables in determining success of ESWL for urinary lithiasis (Muratori et al. 2017).

However same conventional approach of multiple logistic regression was performed to obtain this result. Another study by researchers from Germany demonstrate that age, sex, BMI, stone size and location were not significant predictive factors of treatment outcome in univariate analysis. This data included 68 patients with a mean age

of 42 years. However the only significant predictor they found was stone density (Ben Khalifa et al. 2016). Other authors verify these findings that skin to stone distance (SSD), stone density and BMI as predictive factors of success rate of SWL (El-Nahas et al. 2007; Joseph et al. 2002) ((Lee et al. 2015). Nevertheless determining absolute prognostic factors in predicting the success of ESWL is a controversial subject as there are some studies suggest other factors such as stone location, stone number as most significant predictive factors while others oppose that (Chongruksut et al. 2011). Also the absence of ureteral stent, right sided stones were found to associate with higher stone free rate in a study by the group from McGill university, albeit we did not find any significance for these variables in our study (Elkoushy et al. 2011).

In regard to the type of lithotripter machine, we found no significant difference of success or failure according to the model of the lithotripter used. This finding was also confirmed by the group from university of Minnesota showing the same result in comparison of Medstone STS™ and the Modulith® SLX machines (Alanee, Ugarte, and Monga 2010). In terms of location of the stone, authors found lower pole renal *calculi* present worse results than mid-pole and upper pole *calculi* and less likely to achieve stone free status. Calyceal *calculi* have worse outcomes compared to renal pelvic and ureteral stones(Chongruksut et al. 2011; Kanao et al. 2006).

Besides the statistical analysis part, we could not find any other research in the literature using machine learning approach to predict treatment outcome or success of ESWL. Thus, it makes this research a novel approach to improve management of the nephrolithiasis.

For the machine learning approach, the limitation of our study includes lack of complete follow up data for some of the patients enrolled.

As a result of that the failure of treatment were defined only based on having retreatment of stone in a same center (SMH). However, in order to confirm the accuracy of our model, we tested our model on a sample of comprehensive subset of dataset that included only patients who had complete preoperative and post-operative follow ups which were conducted at SMH. The follow up data and the stone free rate and successfulness of treatment for these patients were assessed based on the follow up CT-scan conducted at SMH after 3 months of initial SWL. This way we mitigated the effect of raw data and the possibility of incomplete follow ups that were captured in SMH.

In regard to statistical analysis, the limitation of our study includes lack of sufficient attributes such as SSD and stone density, which are found to have significant impact on treatment outcome in other studies.

Chapter Eight: Conclusion and Remarks

8.1 Ensemble Learning Technique for Binary Classification problems

There are many applications for classification tasks in real world datasets including bioinformatics, medicine, commerce and etc. many of these classification tasks involve more than two categories for the classification problem. Usually building a classifier to distinguish between two classes is an easier task due to low margins and simpler boundaries. Therefore, some researchers prefer to break the original complex multiclass problem into smaller binary classification tasks in order to build a simpler yet more effective model on complex multiclass datasets. This process is called binarization technique which is now widely used in variety of classification tasks in different fields("An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," 2011).

In this paper we built a rigorous machine learning model to assist physicians and decision makers to choose a best treatment option for SWL candidates based on their demographics and stone characteristics, which can result in more efficient retreatment. The model is generated based on ensemble learning technique using AdaBoost algorithm. The proposed algorithm exploits a very simple one branch decision tree for binary classification in multiple iterations in a way that in each iteration it produces a model by adding more weight to misclassified examples. After a predefined number of iterations has been performed, the final model is generated by averaging the weights from all previous runs. This model can help physicians and care takers to reduce unnecessary overtreatment by predicting which stones are likely to fail SWL. Due to the lack of comprehensive follow

ups in the database, the model was validated against a comprehensive supplementary dataset whose patients had their follow ups conducted at same center which included 862 patients. Also, a bivariate and multiple regression statistical analysis were performed on this supplementary dataset acknowledging the superiority of AdaBoost learning model over conventional statistical analysis and also to ascertain the effect of each attribute on treatment outcome. As a result of bivariate statistical analysis some of the attributes that did not contribute enough to the treatment result were removed from further analysis. Furthermore, pairwise comparison has been performed between AdaBoost classifier and 5 other classification techniques in regard to their accuracy, MCC, area under ROC curve and root mean squared error. The findings of these comparisons suggest the privilege of AdaBoost to those algorithms. In order to develop a more robust model, the same ensemble learning approach can be applied on a more comprehensive database to ascertain the applicability of model.

8.2 Recommendation for Further Study

The proposed technique in this paper is able to identify patients who are likely to fail the SWL treatment. This study was performed with the hope that patients who benefit most, would undergo this treatment modality. However, considering the limited resources of SWL in Canada, one of the areas that can be improved significantly is the patients' referral pattern. This means that taking into account the geographical location of patients, population of patients within designated area, availability dates of each lithotripter center, number of resources and staff, financial restrictions, laws and regulations one can develop a model with data envelopment analysis to identify the geographical borders for each lithotripter center in which only patients who are residing within that border are going to

be referred to that specific location. This way not only we eliminate unnecessary long travels for patients and reduce the traffic, we also ensure that the reasonable number of patients are treated in each location which may essentially reduce the waiting times and more satisfaction for both patients and staff.

Aside from optimization in the management of patients' referral, some enhancements can be applied on current guidelines for managing nephrolithiasis using SWL to make better use of this technology for best candidates. In other words, considering the reported failure rate of SWL after first session in recent research that ranges from around 30% to 60% (Altok et al. 2016; Javanmard et al. 2016; Yamashita et al. 2017).

This number can be reduced significantly by identifying candidates who would benefit most from this treatment and provides alternative therapy for those who won't. This task requires a more robust and detailed database which captures more attributes that are proved scientifically to have an effect on treatment outcome of SWL such as stone density or skin to stone distance.

We also plan to evaluate our proposed ideas and methods on more data sets, including some document collections in real-world applications (for instance, Liu, Huang, An, & Yu, 2007; Liu, Huang, An & Yu, 2008; Feng, Zhang, Hu, & Huang, 2014; Yu, Liu, Huang, & An, 2012; Yin, Huang, Li, & Zhou, 2013).

Bibliography

Altok, M., Güneş, M., Umul, M., Şahin, A. F., Baş, E., Oksay, T., & Soyupek, S. (2016). Comparison of shockwave frequencies of 30 and 60 shocks per minute for kidney stones: a prospective randomized study. Scandinavian Journal of Urology, 50(6), 477–482. http://doi.org/10.1080/21681805.2016.1235609

Satapathy, S. C., Bhateja, V., Somanah, R., Yang, X.-S., & Senkerik, R. (2019) An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. (2018) Pattern Recognition, 44(8), 1761–1776. http://doi.org/10.1016/j.patcog.2011.01.017 Springer

An, A., Huang, Y., Huang, X., & Cercone, N. (2005). Feature Selection with Rough Sets for Web Page Classification. In Transactions on Rough Sets II (Vol. 3135, pp. 1–13). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-27778-1_1

Bonnin, R. (2016). Building Machine Learning Projects with TensorFlow. Birmingham, Mumbai. Packt Publishing Limited.

Bozzini, G., Verze, P., Arcaniolo, D., Dal Piaz, O., Buffi, N. M., Guazzoni, G., et al. (2017). A prospective randomized comparison among SWL, PCNL and RIRS for lower calyceal stones less than 2 cm: a multicenter experience: A better understanding on the treatment options for lower pole stones. World Journal of Urology, 35(12), 1967–1975. http://doi.org/10.1007/s00345-017-2084-7

El-Assmy, A., El-Nahas, A. R., Abou-El-Ghar, M. E., Awad, B. A., & Sheir, K. Z. (2013). Kidney stone size and hounsfield units predict successful shockwave lithotripsy in children. Urology, 81(4), 880–884. http://doi.org/10.1016/j.urology.2012.12.012

Feng, W., Zhang, Q., Hu, G., & Huang, J. X. (2014). Mining network data for intrusion detection through combining SVMs with ant colony networks. Future Generation Computer Systems, 37, 127–140. http://doi.org/10.1016/j.future.2013.06.027

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337–407. http://doi.org/10.1214/aos/1016218223

Gambaro, G., Fabris, A., Puliatta, D., & Lupo, A. (2006). Lithiasis in cystic kidney disease and malformations of the urinary tract. Urological Research, 34(2), 102–107. http://doi.org/10.1007/s00240-005-0019-z

Garfield, K., & Leslie, S. W. (2019). Medullary Sponge Kidney. StatPearls Publishing, Treasure Island, FL. https://www-ncbi-nlm-nih-gov.ezproxy.library.yorku.ca/books/NBK470220/

Gücük, A., & Uyetürk, U. (2014). Usefulness of hounsfield unit and density in the assessment and treatment of urinary stones. World Journal of Nephrology, 3(4), 282–286. http://doi.org/10.5527/wjn.v3.i4.282

Huang, X., & Hu, Q. (2009). A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval (307-314). Presented at the the 32nd international ACM SIGIR conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/1571941.1571995

Huang, X., Peng, F., Schuurmans, D., Cercone, N., & Robertson, S. E. (2003). Applying Machine Learning to Text Segmentation for Information Retrieval. Information Retrieval, 6(3-4), 333–362. http://doi.org/10.1023/A:1026028229881

Huang, X., Huang, Y. R., Wen, W., An, A., Liu, Y., Poon, J. (2006) Applying Data Mining to Pseudo-Relevance Feedback for High Performance Text Retrieval. In Proceedings of IEEE ICDM 2006: 295-306

Huang, X., Zhong, M., Si, L. (2005) York University at TREC 2005: Genomics Track. In proceedings of TREC 2005.

Kalmegh, S., (2015). Analysis of WEKA data mining algorithm REPTree, Simple CART and RandomTree for classification of Indian news. International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2. Pdfs.Semanticscholar.org.

Liu, Y., An, A., & Huang, X. (2006). Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. In Advances in Knowledge Discovery and Data Mining (Vol. 3918, pp. 107–118). Berlin, Heidelberg: Springer, Berlin, Heidelberg. http://doi.org/10.1007/11731139_15

Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing & Management, 47(4), 617–631. http://doi.org/10.1016/j.ipm.2010.11.007

Liu, Y, Huang, X, An, Aijun, Yu, X. (2007) ARSA: a sentiment-aware model for predicting sales performance using blogs. In Proceedings of SIGIR 2007: 607-614

Liu, Y, Huang, X, An, Aijun, Yu, X. (2008) Modeling and predicting the helpfulness of online reviews. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 443-452.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica Et Biophysica Acta (BBA) - Protein Structure, 405(2), 442–451. http://doi.org/10.1016/0005-2795(75)90109-9

Nakasato, T., Morita, J., & Ogawa, Y. (2015). Evaluation of Hounsfield Units as a predictive factor for the outcome of extracorporeal shock wave lithotripsy and stone composition. Urolithiasis, 43(1), 69–75. http://doi.org/10.1007/s00240-014-0712-x

Ng, C.-F. (2009). The effect of age on outcomes in patients undergoing treatment for renal stones. Current Opinion in Urology, 19(2), 211–214.

Nielsen, D. (2016). Tree Boosting With XGBoost Why Does XGBoost Win "Every" Machine Learning Competition?.Norwegian University of Science and Technology.

Okada, A., Yasui, T., Taguchi, K., Niimi, K., Hirose, Y., Hamamoto, S., et al. (2013). Impact of official technical training for urologists on the efficacy of shock wave lithotripsy. Urolithiasis, 41(6), 487–492. http://doi.org/10.1007/s00240-013-0586-3

Robert, C. (2014). Machine Learning, a Probabilistic Perspective. Chance. http://doi.org/10.1080/09332480.2014.914768

Rokach, L., & Maimon, O. Z. (2008). Data Mining with Decision Trees. World Scientific.

Schapire, R. E., & Freund, Y. (2014). Boosting: Foundations and Algorithms. Cambridge, MA: MIT Press.

Sewaiwar, P., Verma, K., (2015). Comparative study of various decision tree classification algorithm using WEKA. IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2. Maharashtra. India

Takahara, K., Ibuki, N., Inamoto, T., Nomi, H., Ubai, T., & Azuma, H. (2012). Predictors of success for stone fragmentation and stone-free rate after extracorporeal shockwave lithotripsy in the treatment of upper urinary tract stones. Urology Journal, 9(3), 549–552.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1/2), 69–90. http://doi.org/10.1023/A:1009982220290

Yilmaz, E., Batislam, E., Basar, M., Tuglu, D., Mert, C., & Basar, H. (2005). Optimal frequency in extracorporeal shock wave lithotripsy: prospective randomized study. Urology, 66(6), 1160–1164. http://doi.org/10.1016/j.urology.2005.06.111

Yin X., Huang J.X., Li, Z., Zhou, X. (2013) A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia. IEEE Trans. Knowl. Data Eng. 25(6): 1201-1212

Yu, X., Liu, Y., Huang, X., An, A. (2012) Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Transactions on Knowledge and Data engineering 24 (4), 720-734.

Appendices

Appendix A.

AdaBoost Model

=== Run information ===

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 50 -W weka.classifiers.trees.DecisionStump
Relation: validation2-weka.filters.supervised.instance.SMOTE-C0-K5-P100.0-S1-weka.filters.supervised.instance.SpreadSubsample-M0.0-X432.0-S1
Instances: 864
Attributes: 18
SIDE
ELECTROD
STTXNO
SHOCKS
LOC
AREA
SEX
BMI
AGE
Nstones
FAMILYHX
ASYMP
STENT
FREQ
ANTIBIOT
MAXVOLTAGE
LITHO
RESULTNoSide
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

AdaBoostM1: Base classifiers and their weights:

Decision Stump

Classifications

STTXNO <= 1.0028183968529663 : SUCCESS
STTXNO > 1.0028183968529663 : FAILURE

STTXNO is missing : FAILURE

Class distributions

STTXNO <= 1.0028183968529663
SUCCESS FAILURE
0.6708268330733229 0.3291731669266771
STTXNO > 1.0028183968529663
SUCCESS FAILURE
0.00904977375565611 0.9909502262443439
STTXNO is missing
SUCCESS FAILURE
0.0 1.0

Weight: 1.12

Decision Stump

Classifications

Nstones <= 1.5 : SUCCESS
Nstones > 1.5 : FAILURE
Nstones is missing : SUCCESS

Class distributions

Nstones <= 1.5
SUCCESS FAILURE
0.7003163006192364 0.29968369938076356
Nstones > 1.5
SUCCESS FAILURE
0.43446298610600603 0.565537013893994
Nstones is missing
SUCCESS FAILURE
0.6 0.4

Weight: 0.62

Decision Stump

Classifications

AREA <= 64.5 : SUCCESS

AREA > 64.5 : FAILURE
AREA is missing : SUCCESS

Class distributions

AREA <= 64.5
SUCCESS FAILURE
0.6491667631541844 0.35083323684581563
AREA > 64.5
SUCCESS FAILURE
0.43265935616490236 0.5673406438350976
AREA is missing
SUCCESS FAILURE
0.5695480747173117 0.43045192528268833

Weight: 0.49

Decision Stump

Classifications

AREA <= 49.013786945296374 : SUCCESS
AREA > 49.013786945296374 : FAILURE
AREA is missing : FAILURE

Class distributions

AREA <= 49.013786945296374
SUCCESS FAILURE
0.5620766090438759 0.43792339095612415
AREA > 49.013786945296374
SUCCESS FAILURE
0.34301420686672385 0.6569857931332761
AREA is missing
SUCCESS FAILURE
0.43809792434124967 0.5619020756587503

Weight: 0.47

Decision Stump

Classifications

LITHO = 3 : SUCCESS

LITHO != 3 : FAILURE
LITHO is missing : SUCCESS

Class distributions

LITHO = 3
SUCCESS FAILURE
0.7285553389742082 0.27144466102579173
LITHO != 3
SUCCESS FAILURE
0.4077405139971664 0.5922594860028336
LITHO is missing
SUCCESS FAILURE
0.7242446164927261 0.27575538350727385

Weight: 0.45

Decision Stump

Classifications

STTXNO <= 1.0028183968529663 : SUCCESS
STTXNO > 1.0028183968529663 : FAILURE
STTXNO is missing : FAILURE

Class distributions

STTXNO <= 1.0028183968529663
SUCCESS FAILURE
0.5597126293161357 0.4402873706838643
STTXNO > 1.0028183968529663
SUCCESS FAILURE
0.2572447813680046 0.7427552186319955
STTXNO is missing
SUCCESS FAILURE
0.0 1.0

Weight: 0.31

Decision Stump

Classifications

ELECTROD <= 2.99379294277576 : FAILURE

ELECTROD > 2.99379294277576 : SUCCESS
ELECTROD is missing : FAILURE

Class distributions

ELECTROD <= 2.99379294277576
SUCCESS FAILURE
0.44069673113973695 0.5593032688602629
ELECTROD > 2.99379294277576
SUCCESS FAILURE
0.6933033011465868 0.30669669885341333
ELECTROD is missing
SUCCESS FAILURE
0.47273937930366944 0.5272606206963305

Weight: 0.26

Decision Stump

Classifications

AGE <= 49.43425504833003 : SUCCESS
AGE > 49.43425504833003 : FAILURE
AGE is missing : SUCCESS

Class distributions

AGE <= 49.43425504833003
SUCCESS FAILURE
0.6008249023749316 0.39917509762506836
AGE > 49.43425504833003
SUCCESS FAILURE
0.462270845332327 0.5377291546676731
AGE is missing
SUCCESS FAILURE
0.5238146394576895 0.47618536054231064

Weight: 0.26

Decision Stump

Classifications

LOC = RP : SUCCESS

LOC != RP : SUCCESS
LOC is missing : SUCCESS

Class distributions

LOC = RP
SUCCESS FAILURE
0.681112746731448 0.3188872532685521
LOC != RP
SUCCESS FAILURE
0.5010549305689898 0.4989450694310103
LOC is missing
SUCCESS FAILURE
0.5317030716908591 0.46829692830914105

Weight: 0.13

Decision Stump

Classifications

LOC = RP : SUCCESS
LOC != RP : FAILURE
LOC is missing : SUCCESS

Class distributions

LOC = RP
SUCCESS FAILURE
0.6529218810587228 0.34707811894127727
LOC != RP
SUCCESS FAILURE
0.46934775827770125 0.5306522417222987
LOC is missing
SUCCESS FAILURE
0.5000000000000014 0.49999999999999856

Weight: 0.2

Decision Stump

Classifications

ASYMP = 0 : FAILURE

ASYMP != 0 : SUCCESS
ASYMP is missing : SUCCESS

Class distributions

ASYMP = 0
SUCCESS FAILURE
0.48997007778851775 0.5100299222114824
ASYMP != 0
SUCCESS FAILURE
0.6293209859365355 0.37067901406346443
ASYMP is missing
SUCCESS FAILURE
0.6703770848081017 0.3296229151918984

Weight: 0.19

Decision Stump

Classifications

AREA <= 297.0 : SUCCESS
AREA > 297.0 : FAILURE
AREA is missing : SUCCESS

Class distributions

AREA <= 297.0
SUCCESS FAILURE
0.5602764739791787 0.4397235260208214
AREA > 297.0
SUCCESS FAILURE
0.08946556551373253 0.9105344344862674
AREA is missing
SUCCESS FAILURE
0.5530355171587016 0.44696448284129847

Weight: 0.26

Decision Stump

Classifications

MAXVOLTAGE <= 22.010774445935965 : SUCCESS

MAXVOLTAGE > 22.010774445935965 : FAILURE
MAXVOLTAGE is missing : FAILURE

Class distributions

MAXVOLTAGE <= 22.010774445935965
SUCCESS FAILURE
0.6027184973626503 0.3972815026373498
MAXVOLTAGE > 22.010774445935965
SUCCESS FAILURE
0.45358826604367475 0.5464117339563254
MAXVOLTAGE is missing
SUCCESS FAILURE
0.47642453695646003 0.5235754630435401

Weight: 0.22

Decision Stump

Classifications

BMI <= 35.236841820980715 : SUCCESS
BMI > 35.236841820980715 : FAILURE
BMI is missing : SUCCESS

Class distributions

BMI <= 35.236841820980715
SUCCESS FAILURE
0.5343488288903837 0.4656511711096164
BMI > 35.236841820980715
SUCCESS FAILURE
0.2966048587393557 0.7033951412606443
BMI is missing
SUCCESS FAILURE
0.5772188248723802 0.4227811751276198

Weight: 0.18

Decision Stump

Classifications

Nstones <= 1.0038109793486236 : SUCCESS
Nstones > 1.0038109793486236 : FAILURE
Nstones is missing : SUCCESS

Class distributions

Nstones <= 1.0038109793486236
SUCCESS FAILURE
0.5073826282389798 0.49261737176102016
Nstones > 1.0038109793486236
SUCCESS FAILURE
0.38496873630478184 0.6150312636952181
Nstones is missing
SUCCESS FAILURE
0.529063841069812 0.4709361589301879

Weight: 0.15

Decision Stump

Classifications

LOC = LC : SUCCESS
LOC != LC : SUCCESS
LOC is missing : SUCCESS

Class distributions

LOC = LC
SUCCESS FAILURE
0.6062233263819654 0.3937766736180346
LOC != LC
SUCCESS FAILURE
0.5068230467432168 0.4931769532567833
LOC is missing
SUCCESS FAILURE
0.5785500250192703 0.4214499749807296

Weight: 0.16

LOC = LC : SUCCESS
LOC != LC : FAILURE
LOC is missing : SUCCESS

Class distributions

LOC = LC
SUCCESS FAILURE
0.5670340271787128 0.4329659728212873
LOC != LC
SUCCESS FAILURE
0.4664478835933662 0.5335521164066338
LOC is missing
SUCCESS FAILURE
0.538702323881284 0.46129767611871597

Weight: 0.18

Decision Stump

Classifications

Nstones <= 3.5 : SUCCESS
Nstones > 3.5 : FAILURE
Nstones is missing : SUCCESS

Class distributions

Nstones <= 3.5
SUCCESS FAILURE
0.529454573119081 0.470545426880919
Nstones > 3.5
SUCCESS FAILURE
0.3881811487136028 0.6118188512863972
Nstones is missing
SUCCESS FAILURE
0.5149232635849601 0.48507673641503996

Weight: 0.15

Decision Stump

Classifications

SHOCKS <= 2697.0 : SUCCESS
SHOCKS > 2697.0 : FAILURE
SHOCKS is missing : FAILURE

Class distributions

SHOCKS <= 2697.0
SUCCESS FAILURE
0.5533253385999132 0.44667466140008694
SHOCKS > 2697.0
SUCCESS FAILURE
0.4571206770143557 0.5428793229856443
SHOCKS is missing
SUCCESS FAILURE
0.41429293249109145 0.5857070675089084

Weight: 0.18

Decision Stump

Classifications

BMI <= 20.200000000000003 : SUCCESS
BMI > 20.200000000000003 : FAILURE
BMI is missing : SUCCESS

Class distributions

BMI <= 20.200000000000003
SUCCESS FAILURE
0.7488473670152196 0.2511526329847804
BMI > 20.200000000000003
SUCCESS FAILURE
0.4821089051249881 0.5178910948750118
BMI is missing
SUCCESS FAILURE
0.5384099301610757 0.4615900698389243

Weight: 0.11

Decision Stump

Classifications

STENT = TRUE : FAILURE
STENT != TRUE : SUCCESS
STENT is missing : SUCCESS

Class distributions

STENT = TRUE
SUCCESS FAILURE
0.4472655409666456 0.5527344590333544
STENT != TRUE
SUCCESS FAILURE
0.5270429251971401 0.47295707480285987
STENT is missing
SUCCESS FAILURE
0.5042460262406591 0.4957539737593408

Weight: 0.14

Decision Stump

Classifications

AREA <= 297.0 : SUCCESS
AREA > 297.0 : FAILURE
AREA is missing : FAILURE

Class distributions

AREA <= 297.0
SUCCESS FAILURE
0.5054723420663707 0.49452765793362924
AREA > 297.0
SUCCESS FAILURE
0.11673032199510863 0.8832696780048914
AREA is missing
SUCCESS FAILURE
0.4999999999999998 0.5000000000000002

Weight: 0.04

Decision Stump

Classifications

Nstones <= 2.5 : SUCCESS
Nstones > 2.5 : FAILURE

Nstones is missing : FAILURE

Class distributions

Nstones <= 2.5
SUCCESS FAILURE
0.5055735886087134 0.4944264113912867
Nstones > 2.5
SUCCESS FAILURE
0.425773951924875 0.574226048075125
Nstones is missing
SUCCESS FAILURE
0.489462900472792 0.510537099527208

Weight: 0.08

Decision Stump

Classifications

AREA <= 64.5 : FAILURE
AREA > 64.5 : SUCCESS
AREA is missing : FAILURE

Class distributions

AREA <= 64.5
SUCCESS FAILURE
0.47038548106638883 0.5296145189336111
AREA > 64.5
SUCCESS FAILURE
0.542894199774129 0.45710580022587105
AREA is missing
SUCCESS FAILURE
0.49788277999239977 0.5021172200076003

Weight: 0.14

Decision Stump

Classifications

AGE <= 62.0090448752195 : SUCCESS
AGE > 62.0090448752195 : FAILURE
AGE is missing : FAILURE

Class distributions

AGE <= 62.0090448752195
SUCCESS FAILURE
0.5126784822439339 0.48732151775606614
AGE > 62.0090448752195
SUCCESS FAILURE
0.3980089690864512 0.6019910309135488
AGE is missing
SUCCESS FAILURE
0.4900849125340395 0.5099150874659604

Weight: 0.12

Decision Stump

Classifications

ELECTROD <= 2.99379294277576 : FAILURE
ELECTROD > 2.99379294277576 : SUCCESS
ELECTROD is missing : FAILURE

Class distributions

ELECTROD <= 2.99379294277576
SUCCESS FAILURE
0.45064954072376895 0.549350459276231
ELECTROD > 2.99379294277576
SUCCESS FAILURE
0.616479235338828 0.383520764661172
ELECTROD is missing
SUCCESS FAILURE
0.481376694461467 0.5186233055385329

Weight: 0.2

Decision Stump

Classifications

BMI <= 23.79 : SUCCESS
BMI > 23.79 : FAILURE
BMI is missing : SUCCESS

Class distributions

BMI <= 23.79
SUCCESS FAILURE
0.5957993167396594 0.4042006832603406
BMI > 23.79
SUCCESS FAILURE
0.4864002102610243 0.5135997897389757
BMI is missing
SUCCESS FAILURE
0.5002949886688962 0.4997050113311038

Weight: 0.13

Decision Stump

Classifications

Nstones <= 5.5 : FAILURE
Nstones > 5.5 : SUCCESS
Nstones is missing : SUCCESS

Class distributions

Nstones <= 5.5
SUCCESS FAILURE
0.4905303664132027 0.5094696335867973
Nstones > 5.5
SUCCESS FAILURE
0.9880500996061856 0.011949900393814423
Nstones is missing
SUCCESS FAILURE
0.5432281390938166 0.4567718609061834

Weight: 0.05

Decision Stump

Classifications

BMI <= 17.6 : SUCCESS
BMI > 17.6 : FAILURE
BMI is missing : FAILURE

Class distributions

BMI <= 17.6
SUCCESS FAILURE
1.0 0.0
BMI > 17.6
SUCCESS FAILURE
0.48993855094868854 0.5100614490513116
BMI is missing
SUCCESS FAILURE
0.4303695243726611 0.5696304756273388

Weight: 0.06

Decision Stump

Classifications

Nstones <= 5.5 : FAILURE
Nstones > 5.5 : SUCCESS
Nstones is missing : SUCCESS

Class distributions

Nstones <= 5.5
SUCCESS FAILURE
0.49647224421168457 0.5035277557883154
Nstones > 5.5
SUCCESS FAILURE
0.9870249959789161 0.012975004021083963
Nstones is missing
SUCCESS FAILURE
0.5224861933351789 0.4775138066648212

Weight: 0.03

Decision Stump

Classifications

SHOCKS <= 3630.0 : SUCCESS
SHOCKS > 3630.0 : FAILURE
SHOCKS is missing : FAILURE

Class distributions

SHOCKS <= 3630.0
SUCCESS FAILURE
0.5128743958482833 0.48712560415171674
SHOCKS > 3630.0
SUCCESS FAILURE
0.3798318097738239 0.6201681902261761
SHOCKS is missing
SUCCESS FAILURE
0.4985216489614424 0.5014783510385575

Weight: 0.1

Number of performed Iterations: 30

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

A.1 Summary of performance measurements

Correctly Classified Instances	660	76.3889 %
Incorrectly Classified Instances	204	23.6111 %
Kappa statistic	0.5278	
Mean absolute error	0.2996	
Root mean squared error	0.3912	
Relative absolute error	59.9097 %	
Root relative squared error	78.2426 %	
Total Number of Instances	864	

A.2 Detailed Accuracy by Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
SUCCESS	0.875	0.347	0.716	0.875	0.788	0.541	0.843	0.803
FAILURE	0.653	0.125	0.839	0.653	0.734	0.541	0.843	0.866
Weighted Avg.	0.764	0.236	0.778	0.764	0.761	0.541	0.843	0.835

A.3 Confusion Matrix

Classified as →	A	B
A= SUCCESS	378	54
B = FAILURE	150	282

Appendix B. Feature Selection Results

=== Run information ===

Evaluator: weka.attributeSelection.CorrelationAttributeEval

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: validation-weka.filters.unsupervised.attribute.Remove-R1-2-
weka.filters.unsupervised.instance.RemoveWithValues-S10.0-C9-Lfirst-last-
weka.filters.unsupervised.instance.RemoveWithValues-S12.0-C9-Lfirst-last-
weka.filters.unsupervised.instance.RemoveWithValues-S13.0-C9-L13-
weka.filters.unsupervised.instance.RemoveWithValues-S14.0-C9-
weka.filters.unsupervised.instance.RemoveWithValues-S15.0-C9-
weka.filters.unsupervised.instance.RemoveWithValues-S16.0-C9-

weka.filters.unsupervised.instance.RemoveWithValues-S17.0-C9-

weka.filters.unsupervised.instance.RemoveWithValues-S18.0-C9-

weka.filters.supervised.instance.SpreadSubsample-M1.5-X0.0-S1

Instances: 36050

Attributes: 18

SIDE

ELECTROD

STTXNO

SHOCKS

LOC

AREA

SEX

BMI

AGE

Nstones

FAMILYHX

ASYMP

STENT

FREQ

ANTIBIOT

MAXVOLTAGE

LITHO

RESULTNoSide

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 18 RESULTNoSide):

Correlation Ranking Filter

Ranked attributes:

0.24445 10 Nstones

0.16758 6 AREA

0.09027 13 STENT

0.06525 9 AGE

0.04611 17 LITHO

0.04206 5 LOC

0.03645 4 SHOCKS

0.03353 12 ASYMP

0.03274 2 ELECTROD

0.03151 7 SEX

0.01767 8 BMI

0.01255 14 FREQ

0.01232 3 STTXNO

0.01166 11 FAMILYHX

0.0043 16 MAXVOLTAGE

0.00428 15 ANTIBIOT

0.00297 1 SIDE

Selected attributes: 10,6,13,9,17,5,4,12,2,7,8,14,3,11,16,15,1 : 17

Appendix C. Pairwise Comparison of Classification Models

C.1 Accuracy

Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 0 -stddev-width 0 -sig-width 0 -count-width 5 -print-col-names -print-row-names -enum-col-names"

Analysing: Percent_correct

Datasets: 1

Resultsets: 6

Confidence: 0.05 (two tailed)

Sorted by: -

Dataset (1) meta.Ada | (2) trees (3) funct (4) bayes (5) bayes (6) lazy.

---

validation2 (30) 77.59 | 75.00 * 69.11 * 76.49 75.82 * 57.52 *

(v/ /*) | (0/0/1) (0/0/1) (0/1/0) (0/0/1) (0/0/1)

Key:

(1) meta.AdaBoostM1 '-P 100 -S 1 -I 50 -W trees.DecisionStump' -1178107808933117974

(2) trees.J48 '-C 0.25 -M 2' -217733168393644444

(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779

(4) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954

(5) bayes.NaiveBayes " 5995231201785697655

(6) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -3080186098777067172

C.2 Mathews Correlation Coefficient

Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-

Analysing: Matthews_correlation

Datasets: 1

Resultsets: 6

Confidence: 0.05 (two tailed)

Sorted by: -

Dataset (1) meta.Ad | (2) tree (3) func (4) baye (5) baye (6) lazy

---

validation2 (30) 0.53 | 0.45 * 0.34 * 0.49 0.47 * 0.09 *

(v/ /*) | (0/0/1) (0/0/1) (0/1/0) (0/0/1) (0/0/1)

Key:

(1) meta.AdaBoostM1 '-P 100 -S 1 -I 50 -W trees.DecisionStump' -1178107808933117974

(2) trees.J48 '-C 0.25 -M 2' -217733168393644444

(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779

(4) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954

(5) bayes.NaiveBayes " 5995231201785697655

(6) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -3080186098777067172

C.3 F1 Score

Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 2 -stddev-width 2 -sig-width 1 -count-width 5 -print-col-names -print-row-names -enum-col-names"

Analysing: F_measure

Datasets: 1

Resultsets: 6

Confidence: 0.05 (two tailed)

Sorted by: -

Dataset (1) meta.Ad | (2) tree (3) func (4) baye (5) baye (6) lazy

---

validation2 (30) 0.84 | 0.82 * 0.76 * 0.83 0.83 * 0.66 *

(v/ /*) | (0/0/1) (0/0/1) (0/1/0) (0/0/1) (0/0/1)

Key:

(1) meta.AdaBoostM1 '-P 100 -S 1 -I 50 -W trees.DecisionStump' -1178107808933117974

(2) trees.J48 '-C 0.25 -M 2' -217733168393644444

(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779

(4) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954

(5) bayes.NaiveBayes " 5995231201785697655

(6) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -3080186098777067172

C.4 Area Under ROC Curve

Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2 -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 2 -stddev-width 2 -sig-width 1 -count-width 5 -print-col-names -print-row-names -enum-col-names"

Analysing: Area_under_ROC

Datasets: 1

Resultsets: 6

Confidence: 0.05 (two tailed)

Sorted by: -

Dataset (1) meta.Ad | (2) tree (3) func (4) baye (5) baye (6) lazy

---

validation2 (30) 0.80 | 0.73 * 0.74 * 0.78 0.75 * 0.54 *

(v/ /*) | (0/0/1) (0/0/1) (0/1/0) (0/0/1) (0/0/1)

Key:

(1) meta.AdaBoostM1 '-P 100 -S 1 -I 50 -W trees.DecisionStump' -1178107808933117974

(2) trees.J48 '-C 0.25 -M 2' -217733168393644444

(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779

(4) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954

(5) bayes.NaiveBayes " 5995231201785697655

(6) lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -308018609877706717