UNIVERSITY OF CALIFORNIA,
IRVINE

Machine Learning Models to Predict Diagnosis and Surgical Outcomes in Otolaryngology

THESIS

Submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Biomedical and Translational Science

by

Khodayar Goshtasbi

Thesis Committee:
Professor Hamid R. Djalilian, Chair
Assistant Professor Harrison W. Lin
Professor Sheldon Greenfield

2020

DEDICATION

To my family and mentors in recognition of their never-ending support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my committee chair and mentor Dr. Hamid Djalilian for his continuous guidance and support. Dr. Djalilian's mentorship has been crucial in cultivating my research skills and interests, and his vision will continue to greatly inspire my future. I would also like to thank Dr. Harrison Lin for being an incredible mentor and teacher. Dr. Lin has consistently encouraged me throughout my medical school journey while playing a key role in growing my passion for academics and research.

In addition, I thank Dr. Sheldon Greenfield for providing invaluable teachings regarding scientific conduct and data interpretation throughout the past year. I also thank Dr. Sherrie Kaplan for generously contributing to my clinical research training. Dr. Greenfield and Dr. Kaplan's teachings and emphasis on purposeful and methodological research will greatly serve my future.

Furthermore, I sincerely thank Dr. Mehdi Abouzari, Mr. Tyler Yasaka, Mr. Kotaro Tsutsumi, Mr. Pooya Khosravi, Dr. Edward Kuan, Mr. Brandon Lehrich, and Dr. Ronald Sahyouni for their significant contributions and continuous support.

I have no financial disclosures. All research was performed in affiliation with the University of California Irvine.

ABSTRACT OF THE THESIS

Machine Learning Models to Predict Diagnosis and Surgical Outcomes in Otolaryngology

By

Khodayar Goshtasbi

Master of Science in Biomedical and Translational Science

University of California, Irvine, 2020

Professor Hamid R. Djalilian, Chair

Background: Despite the rapid evolvement of machine learning (ML) applications within the medical literature, there exist a paucity of knowledge regarding diagnosis, decision-making, and predicting surgical outcomes in otolaryngology using ML models. This thesis aims to 1) Construct ML models that use pre-operative-only inputs to predict length of stay (LOS) and discharge to non-home facility (DNHF) following complex head and neck (HN) surgeries, 2) Utilize deep learning and a large number of input variables to predict short-term adverse events following vestibular schwannoma (VS) surgery, 3) Construct preliminary binary (normal vs. abnormal) and multiclass (normal vs. acute otitis media vs. otitis externa vs. chronic suppurative otitis media vs. cerumen) image classification models of otoscopic images, and 4) Publish the best-performing models as public web-based interfaces.

Methods: To develop novel ML models that predict various post-operative outcomes, the 2005-2017 National Surgical Quality Improvement Program database was utilized to extract subjects undergoing major HN surgery (N=2667) and VS surgery (N=1783). Datasets were randomly stratified into training and testing sets using an 80:20 ratio with k-fold cross-validation used for training. To develop the novel otoscopic image classification model utilizing Inception-Resnet-V2 networks, 400 publicly available and labeled otoscopic images were obtained.

Results: Four classification models for predicting DNHF were developed with high specificities (range 0.80-0.84), where the generalized linear model and gradient boosting machine models outperformed artificial neural network and random forest models with receiver operating characteristic (ROC), accuracy, and negative predictive value (NPV) of 0.72-0.73, 0.75-0.76, and 0.88-0.89, respectively. Four regression models were developed for predicting LOS in days, where all performed similarly with mean absolute error and root mean squared errors of 3.95-3.98 and 5.14-5.16, respectively. The DNHF and LOS models were developed into a web-based interface: https://uci-ent.shinyapps.io/head-neck/. Using pre-, peri-, and post-operative inputs, three deep-learning models to predict unplanned reoperation, surgical complications, and medical complications following VS surgery were developed with ROC of 0.74, 0.70, and 0.83, and accuracy of 0.91, 0.71, and 0.92, respectively. Lastly, a preliminary image classification model with ROC, accuracy, and NPV of 0.89, 0.77, and 0.74 for binary classification, and the capability to suggest the diagnosis among abnormal images, was developed: https://headneckml.com/tm/.

Conclusion: Novel ML models to predict DNHF or LOS following complex HN surgery, reoperation or complications following VS surgery, and abnormalities on otoscopic imaging were successfully developed. Publishing such models as interactive web-based interfaces will help advance this frontier by providing opportunities for examination/validation and practical benefit to clinicians.

CHAPTER 1: INTRODUCTION

Machine learning (ML) is an analytical application of artificial intelligence with the ability to "learn" from data, improve automatically through experience, and perform various tasks without explicit directions or programming.[1] In contrast to most current predictive models in clinical literature which are associative and reliant on direct coefficient-outcome relationships, ML models can potentially improve predictive outcomes by executing complex, hidden, and non-linear computations using big data.[2] The advent of big data has enabled ML to gain popularity in the medical literature, leading to landmark studies evaluating the application of ML techniques especially for disease diagnoses,[3] imaging classifications,[4] and outcome predictions.[5,6] An abundance of information is being collected from patients in the era of evidenced based medical and surgical practice. Compared to traditional statistical analyses with limited interpretation and real-life applicability, this opportunity can prove invaluable for training complex ML applications and building automated self-improving predictive models. The field of surgery is well-positioned to consider a limited embracement of ML applications in the healthcare system and investigate ways in which its utility can optimize workflow, decision-making, resourcefulness, and surgical outcomes.[7] Most recently in the otolaryngology literature, authors have reported ML algorithms for detecting pharyngeal cancer through imaging[8] and predicting complications following HN microvascular free tissue transfer.[9] Despite the clear need and interest, there remain a paucity of knowledge regarding ML applications for the vast majority of otolaryngology patients, and whether the development of such models can benefit patient care in a practical manner.

Important surgical outcomes to be investigated via ML models can include post-operative adverse events, length of stay (LOS), and discharge to non-home facility (DNHF). Adverse events can include medical or surgical complications or unplanned readmission/reoperation, all of which

can significantly alter the patents' overall health. Additionally, prolonged LOS and DNHF have been shown to associate with cost[10] and insurance barriers,[11] susceptibility to adverse events,[12] hospital performance,[13] and patient satisfaction.[14] Despite the clear need to incorporate such ML models and gain a better understanding of their predictive capabilities, there exist a significant gap of knowledge regarding these applications in the otolaryngology literature. Such models can hypothetically serve as helpful supplementary tools in improving care for patients undergoing major surgeries who may be at risk of certain outcomes, such as prolonged hospitalization or DNHF. Moreover, although ML models for predicting readmission, reoperation, and short-term complications have been recently published in the fields of orthopedic surgery[15] and neurosurgery,[16] similar attempts are yet to be made for otolaryngology patients. This can have important utilities if such models can identify at-risk patients and lead to optimizing pre-surgical planning and post-operative risk management. In addition to these literature gaps, the majority of clinical ML studies focus on common algorithmic models such as artificial neural network (ANN) and generalized linear model (GLM), but there exist other complex ML models that may provide more superior predictive capabilities[17,18] warranting investigation. Another complex application of ML is "deep learning" that can learn from large amounts of unstructured data in a more unsupervised fashion. That being said, more complexity is not always better,[19] and the performance of these models can heavily depend on a variety of factors.

Image classification and photographic diagnosis is another important application in ML. In a 2016 *JAMA* study, authors successfully developed and validated a deep learning algorithm that utilized retinal fundus photographs for detecting diabetic retinopathy with high confidence.[20] If appropriately incorporated within the ophthalmologic practice, this could certainly lead to improved care and outcome especially among regions with limited resources. Similar ML-assisted

image classification models have been reported for prostate[21] and breast[22] cancer ultrasound imaging. Very recently, the Google Cloud Vision AutoML platform (Google Inc.) was utilized to evaluate otoscopic images with promising results,[23] but manually constructed models as well as their publication as interactive interfaces for evaluation of external images are still needed.

This thesis aims to achieve three main objectives: 1) Construct different supervised ML models that predict LOS and DNHF following complex HN surgeries, compare these algorithms' performances, and publish the best-performing models in a public web-based interface, 2) Utilize deep learning on a large number of input variables to predict adverse events including readmission, reoperation, or surgical complications following vestibular schwannoma (VS) surgery, and 3) Construct a proof-of-concept image classification model that identifies otoscopic images with pathologies, and publish this as a public web-based interface where readers can upload images and view predictions. As such, the thesis is organized into four chapters subsequent to the introduction. Chapter 2 describes more background information regarding ML in medicine and specifically otolaryngology as well as current gaps in knowledge. Chapter 3 details the methodological approaches and statistical analyses, and chapter 4 summarizes the results. Finally, chapter 5 provides a thorough discussion of the presented results and related ML principals, novelty and limitations of the findings, and future directions.

CHAPTER 2: BACKGROUND

ML can be described as computational algorithms learning and improving from data with the ability to make accurate predictions through experience.[24] This learning is data-driven by design and combines fundamental principles from computer science, statistics, and model optimization. Given its rapid advancement and incorporation into the current technologies, it is likely that ML will change the medical and surgical landscape in the following decades. This chapter provides background information regarding ML terminology and basic concepts, relevant existing literature on ML in medicine and surgery, current gaps of knowledge, and areas of potential significance in the field of otolaryngology.

Automated Learning Techniques and Tasks

Two types of learning techniques, namely supervised and unsupervised learning, are employed to construct most ML models. Supervised learning involves training ML models on data that contain both input variables and known outcomes. In other words, this technique utilizes a ground truth to learn an associative function, which following proper training and validation can lead to predicting outcomes associated with novel input information. Unsupervised learning on the other hand only requires inputs, which will use to independently find clustering or grouping of data points. In other words, predefined outcomes are not required, and the algorithm seeks patterns and infer natural structures in a more exploratory manner. Furthermore, two standard ML tasks that will be examined in this thesis are classification and regression. The former involves assigning discrete categories (also called classes) for an outcome, for instance presence versus absence of post-operative complications, determining one of several possible diagnoses based on imaging (e.g. glioblastoma, meningioma, or brain metastasis on brain MRI), or classifying the content and tone of patients' comments on physician rating websites (e.g. positive outcome, good bedside

manner, positive perception of office staff, etc.). The number of categories can certainly exceed hundreds in complex and large models especially for image classification.[25] Regression is the task of predicting a real value for a continuous outcome variable, for instance predicting length of post-operative hospitalization (e.g. day two, day three, etc.), duration of a procedure (e.g. four hours, six hours, etc.), or life expectancy following cancer diagnosis (e.g. ten years, twelve years, etc.).

Machine Learning Models

Four common types of supervised ML algorithms explored in parts of this thesis include generalized linear models (GLM), artificial neural networks (ANN), random forest (RF), and gradient bosting machine (GBM), all of which can be utilized to build classification or regression-based models. GLM algorithms which use traditional regression mathematical models have been most frequently utilized for constructing predictive models in the medical literature.[26] Specifically, features are considered linearly and additively for constructing line(s) of best fit while minimizing the distance between individual data points and line(s), making GLM relatively quick to develop and easy to visualize and interpret. ANNs are gross neuron-by-neuron simulations of the human brain with finite numbers of layers, nodes, interconnections, and weighted variables, usually able to identify meaningful connections within complex datasets.[27] Based on statistical learning theory introduced by Vapnik,[28] ANN's neurons receive customized weight, bias term, and sigmoid activation function, and model complexity can increase by more hidden layers and connection webs. Notably, GLM and ANN are the two most common models in the medical literature. Convolution neural network (CNN) is a type of deep learning neural network especially efficient for image[29] or sentence[30] classification/recognition. A CNN model is similarly trained on many input pictures (e.g. animal pictures) associated with an output class (e.g. wolf, dog, etc.), to be able to predict the class of a new and unlabeled picture with a certain probability (e.g. 80% probability

that this picture is a dog). RF is a type of decision tree algorithm utilizing different data sample bootstraps for creating each tree, where the number of trees and model predictors positively correlate.[31] In other words, decision trees go from observation (i.e. tree branches) to target conclusions (i.e. leaves), and RF is an ensemble of multiple decision trees where the final outcome takes into consideration the mean output of each individual tree. Another form of decision tree model is GBM, which fits weak learner decision tress to a regression model,[32] with the advantage of being highly adaptable and interpretable.[33]

Machine Learning in Medicine: A Brief Overview

The staggering increase in both clinical data volume and computational power/storage have led to the ability of large- and small-group researchers to develop ML applications in medicine.[19,34] Another important contributor is the emerging availability of publicly available packages and libraries of codes for MATLAB, R, and Python programming languages.[35-37] Regardless, success in this field benefits greatly from dedicated time and personnel, prior computer science experience, and availability of large and high-quality training datasets. In the field of medicine, disease diagnosis and outcome prediction are two common parameters to explore via ML applications. Examples of impactful works in the past decade include developing predictive models for morphological/functional echocardiography assessment,[38] progression to prediabetes and type 2 diabetes,[39] length of hospitalization among cardiac patients,[40] prognosis in oral[41] or lung[42] cancer, decision-making for septic patients,[43] and early detection of Alzheimer's disease.[44] It is important to mention that for the most part, these models are not meant to serve as stand-alone tools or replacements of existing resources, but to provide prompt assistance based on their learning of many past patients and physician decisions,[45] detect potential medical errors prompting second looks,[46] or address specialized needs in resource-burdened regions.[38] It is also possible that these

tools, if widely available and utilized, can identify more at-risk patients at earlier time points, thus presenting more opportunities for disease prevention or conservative management.[39,44]

Machine Learning in Surgery: An Evolving Landscape

Applications of ML can theoretically involve many aspects of the surgical field including diagnosis, decision-making, pre-operative planning, peri-operative guidance, post-operative care, short- and long-term prognosis, and early detection of adverse events or recurrence. When combined with a specialist physician, ML was shown to reduce human error from 3% to less than 1% for detecting positive lymph nodes in breast cancer,[46] which can play a significant role in long-term prognosis of breast cancer surgical patients. Among patients with high risk lesions on biopsy, ML can identifying those with lower risk of progressing to cancer and suggest close follow-up as opposed to surgical excision.[47] In urology, ML applications to predict prostate cancer extra-capsular extension,[48] Gleason score from MRI,[49] and readmission[50] or complications[51] following urologic operations have been recently reported. In ophthalmology, image classification models that predict diabetic retinopathy[20] or glaucoma[52] have been introduced with important surgical applications. The plastic surgery literature has also seen an increase in ML studies pertaining to post-operative healing time, tissue perfusion monitoring, and nerve grafting outcomes.[53] Using a large public database, a model for predicting short-term complications following liver, pancreatic, and colorectal surgery was also developed.[54]

Recently, utilizing big data to develop ML predictive models for post-operative outcomes has been actively explored in the fields of neurosurgery and orthopedic surgery. Namely, well-performing models for predicting readmission,[15] mortality,[18] discharge destination,[17,55] LOS,[56] and complications[57,58] after a variety of neurologic and orthopedic surgeries have been reported. In addition to a paucity of such developments for otolaryngology surgeries, another gap in the

surgical literature includes the publication of these models as interactive interfaces to 1) allow clinicians to explore the utility and real-life applicability of these models for hypothetical patients, and 2) allow external validation of the model with novel data, which will eventually lead to opportunities for improvement. Also, another important principle not always discussed in ML studies is whether it can provide practical value with a window of opportunity for change. For instance, if a model intended to predict surgical outcomes for pre-operative planning takes as inputs variables which are only known intra-operatively (e.g. length of operation, unexpected transfusion, etc.) or post-operatively (e.g. wound class, pain management, etc.), the model may fail to provide any practical benefit.

Machine Learning in Otolaryngology and Study Rationale

ML studies are beginning to gain popularity in the recent otolaryngology literature similar to the other aforementioned surgical fields. Two objectives of this thesis involve using big data to develop ML models for predicting various short-term outcomes following VS surgery or complex head and neck operations. The previous section mentioned this being actively investigated for several orthopedic and neurosurgery cohorts, yet there is a paucity of similar investigations in the otolaryngology literature. To our knowledge, the closest study which was published in early 2020 utilized 364 institutional patients to predict complications following head and neck microvascular free tissue transfer.[9] Besides the low number of subjects, there is yet again the need for such models to be openly published for practical usage. Although there exist big data studies investigating short-term adverse events in VS surgery using traditional statistical analyses,[59,60] incorporating ML algorithms to develop predictive models for adverse events has not yet been attempted. Similarly, there exist no ML studies predicting discharge destination or LOS following otolaryngology

operations. Accomplishing these objectives present two additional novelties beyond the development of these novel models for VS and complex HN surgeries. First, the majority of clinical ML studies predict categorical outcome variables, where even in the case of LOS which is a continuous value in nature, outcomes are grouped into "short" versus "long" LOS options.[40,61] On the contrary, our developed LOS model will regard LOS as a continuous variable and predict an actual number. Additionally, the best-performing models with all the respective input features will be published on a public interface, serving as a practical tool for users to test hypothetical patients or evaluate the external validity. Our last objective will be to construct a CNN model to identify healthy versus abnormal tympanic membranes on otoscopic imaging. Within the past year, two studies introduced similar prototypes for otoscopic imaging.[23,62] However, both groups discuss the preliminary nature of the work with limitations and room for improvements, and the algorithms were not published as a public interface for the readership's benefit or real-life usage. This justifies the third objective which is to not only develop a proof-of-concept otoscopic image classification algorithm, but to also publish this as an online platform where readers can upload images and receive diagnosis predictions. With the foundation in place, this model can be temporally improved with continuous collection of institutional images in the future.

CHAPTER 3: METHODS

The research activities were exempted from Institutional Review Board approval due to the de-identified patient information and publicly available nature of the databases.

DNHF and LOS Following Complex HN Surgery

The following methods are for developing different ML models that use pre-operative inputs to predict LOS and DNHF following complex HN surgeries. The 2005-2017 American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database, which collects 30-day morbidity and mortality information for various operations, was retrospectively reviewed. Complex HN surgeries were defined as laryngectomy or composite tissue excision followed by free tissue transfer. The following current procedural terminology (CPT) codes were used to collect patients undergoing laryngectomy: 31360, 31365, 31368, 31390, and 31395. For the other group of patients, inclusion criteria required two linked surgeries: 1) head and neck mucosal or composite resection with CPTs including 21034, 21044, 21045, 21047, 31230, 31255, 40814, 40816, 41116, 41120, 41130, 41135, 41140, 41145, 42120, 41150, 41153, 41155,42845, 42894, and 2) free tissue transfer with CPTs including 15756 15757, 15758, 20955, 20956, 20962, 20969, and 20970.

The detailed description of the NSQIP variables are found in the user guide for the 2017 ACS-NSQIP Participant Use Data File. The American Society of Anesthesiologists (ASA) score which measures pre-operative comorbidities and overall health was binarized as low (class 1-2) and high (class 3-4) ASA class. Discharge to non-home facility included skilled care or unskilled facility, rehabilitation center, separate acute care center, or hospice. Furthermore, LOS was defined as days from operation to discharge. The NSQIP data, similar to other national databases, suffers from a considerable amount of missing values. As such, input variables with more than 25%

missing values (e.g., albumin, PT, and PTT levels) were excluded from this study. Due to the importance of properly handling missing data, we used the *missForest* package in the R statistical programming language,[63] which imputed missing values in a manner less prone to bias compared to alternative methods of handling missing values.[64] To determine which factors would be included for DNHF or LOS models, univariate analysis (e.g., chi-square, independent *t*-test, and Pearson correlation) was performed to evaluate the association between the pre-operative input and outcome variable, and those with p value <0.2 or deemed clinically important were included.

The dataset was randomly stratified into a training and testing set using an 80:20 ratio separately for the DNHF and LOS models. The training set was used to train the algorithms, and free parameters were adjusted according to results from the training set's cross-validation. Hyperparametric optimization was achieved using random search methods.[65] Our classification models were configured to output probabilities for each prediction, with predictions above a given probability (the "threshold") classified as positive outcomes.[66] After optimizing the classification models for the receiver operating characteristic area under the curve (ROC-AUC), the classification thresholds were adjusted to target specificities of approximately 80%. The models were trained using k-fold cross-validation.[67] Models were evaluated via internal validation, where each model was trained on the training partition and predictions on the test set were evaluated using multiple performance metrics. The reported performance metrics are based on averages of twenty trials to control for bias introduced by randomness in the models. All statistical analyses, including ML development/testing and figure generation, were performed using R version 3.6.3 (The R Foundation for Statistical Computing, Vienna, Austria) via RStudio version 1.1.463 (RStudio, Boston, MA).

The two models had slightly different cohorts depending on exclusion criteria: For DNHF classification models, patients with unknown discharge information were excluded (N=386), and for LOS regression models, patients with unknown or >30-day discharge were excluded (N=119). A total of four classification models were trained and tested to predict DNHF: GLM, ANN, RF, and GBM. Since DNHF was a binary outcome variable, the performance of these models was assessed using sensitivity, specificity, ROC-AUC, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Four regression models were trained and tested for predicting days of post-operative LOS: GLM, ANN, RF, and GBM. Since LOS was treated as a continuous outcome variable, the performance of the regression models was assessed using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and root mean squared logarithmic error (RMSLE). To evaluate how the regression models performed according to the extent of LOS, patients were divided into quartiles and the relative prediction performances (assessed by RMSE) were analyzed accordingly.

Short-term Adverse Events Following Vestibular Schwannoma Surgery

This section aimed to develop classification models that predict short-term adverse events following VS surgery. Unless specified in the following sentences, the methodology was similar to the previous section. The 2005-2017 ACS-NSQIP database was queried for identifying patients with VS undergoing resection using ICD codes 225-1 and D33.3, and CPT codes appropriate for skull base surgery. The post-operative outcome variables assessed by these models included 30-day unplanned reoperation, surgical complication, and medical complication. Surgical complications included superficial or deep surgical site infection (SSI), organ/space SSI, wound disruption, and blood transfusion. Medical complications included pneumonia, unplanned reintubation, urinary tract infection, deep vein thrombosis, renal insufficiency, acute renal failure,

pulmonary embolism, ventilation >48 hours, cerebrovascular accident or stroke, cardiac arrest, myocardial infarction, sepsis, and septic shock. To distinguish from the previous model, this time deep learning on all available variables except those related to the outcome variables were included (N=115, most of which had unknown values).

Otoscopic Image Classification

A tympanic membrane (TM) image database was developed by utilizing publicly available online images and open databases. A total of 400 appropriate images from the Van Akdamar Hospital eardrum database[68] and Google Images (Google Inc., Mountain View, CA) using the terms "tympanic membrane," "otoscopic image," "normal," "acute otitis media," "otitis externa," and "chronic suppurative otitis media" were included. The database was bipartitely organized into normal (196) and abnormal (204) classes. The abnormal class contained images representing 4 different pathologies: acute otitis media (AOM; 116), otitis externa (OE; 44), chronic suppurative otitis media (CSOM; 23), and cerumen (21). To develop and test the algorithm, 60% of the database was used for training, 15% for cross-validation, and 25% for testing of the model.

Since our dataset was relatively small for training a deep learning model from scratch, we employed a technique referred to as transfer learning, in which pretrained networks are fine-tuned on new datasets. Networks pretrained on large datasets such as the ImageNet database, consisting of millions of labelled images belonging to around 1000 different categories, already have learnt general features present among such images. Hence, fine tuning such models for classification of external image categories is possible even with small datasets. For this study, we utilized three publicly available models pretrained on the ImageNet database including the ResNet-50, Inception-V3 and the Inception-Resnet-V2 networks. The pretrained networks were loaded through the keras library written in the Python programming language. All layers of the loaded

models were frozen excluding BatchNormalization layers. The following five layers were added at the ends of the models: GlobalAveragePooling2D, Dense (256, activation = 'relu'), Dropout (0.25), and BatchNormalization. A fully connected layer with two output nodes with a softmax activation function was added as the last layer. Hyperparameters for the training process were as follows: batch size 32, number of epochs 20, learning rate 0.001, optimizer root mean square propagation (RMSprop). We performed data augmentation with rotation range 45, sheering range 0.2, zoom range 0.2, and random horizontal and vertical flips. The study was conducted via Google Colaboratory notebook ran on its Graphics Processing Unit. Representation of model training for otitis externa and acute otitis media images are demonstrated in **Figure 1**. We evaluated algorithms through their classification accuracy and ROC-AUC on novel images not used for training the models.

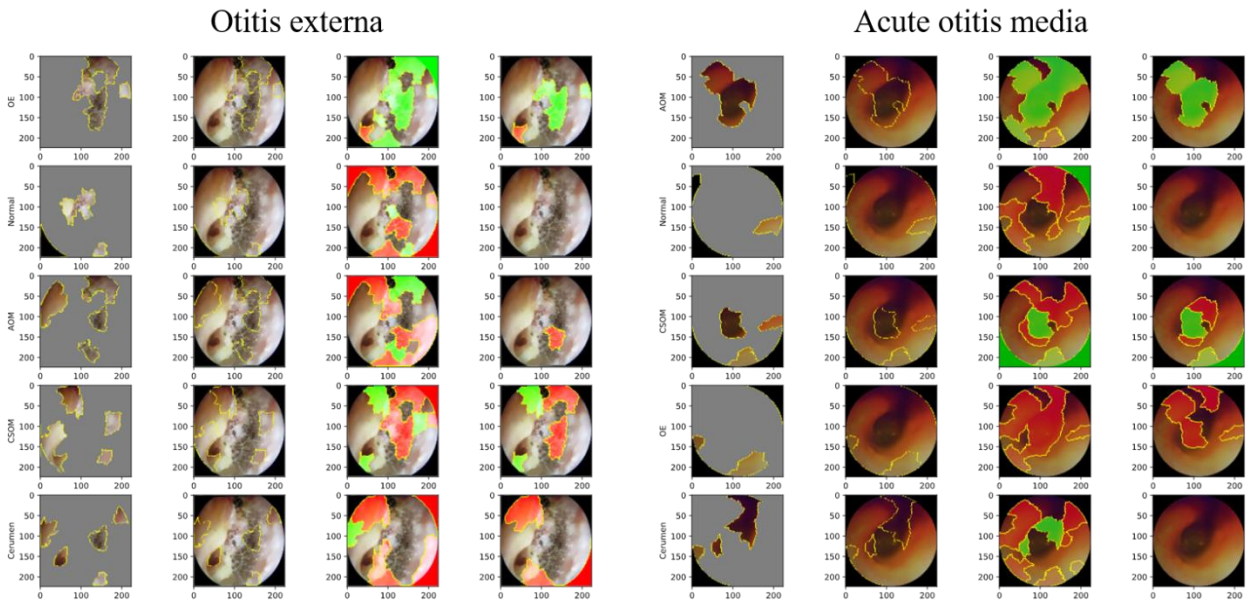


Figure 1. Representation of image classification model training for otitis externa (left panel) and acute otitis media (right panel). The labels are presented from top to bottom based on their probability of positive outcome. Green and red regions of interest had the greatest influence on assigning positive or negative labels, respectively.

CHAPTER 4: RESULTS

Discharge to Non-home Facility in Complex HN Surgery

A total of 2,400 patients (27.6% female) with a mean age of 61.9 ± 11.7 years and including 421 (15.1%) DNHF patients were included. Variables utilized for the DNHF predictive models are demonstrated in **Table 1**.

Table 1. Variables included in the DNHF machine learning predictive models compared via univariate analysis (chi-square or independent *t*-test).

Pre-Operative Feature	**Discharge to Home (N=1978)**	**Discharge to Non-home Facility (N=421)**	***P* Value**
Age (mean ± SD)	60.7 ± 11.4	67.7 ± 11.3	**<0.001**
BMI (mean ± SD)	25.6 ± 6.3	24.5 ± 5.8	**<0.001**
Gender: Female	536 (27.1)	126 (29.9)	0.236
Procedure: Laryngectomy	1108 (56.0)	242 (57.5)	0.575
Diabetes	265 (13.4)	66 (15.7)	0.217
Smoking	819 (41.4)	163 (38.7)	0.312
Dyspnea	250 (12.6)	98 (23.3)	**<0.001**
Functional status: Dependent	36 (1.8)	46 (10.9)	**<0.001**
Ventilator dependent	9 (0.5)	6 (1.4)	**0.022**
History of COPD	213 (10.8)	81 (19.2)	**<0.001**
History of CHF	18 (0.9)	11 (2.6)	**0.004**
History of wound infection	52 (2.6)	21 (5.0)	**0.010**
Hypertension	925 (46.8)	241 (57.2)	**<0.001**
ASA class: High	1651 (83.5)	389 (92.4)	**<0.001**
Chronic steroid use	75 (3.8)	30 (7.1)	**0.002**
Emergency surgery	15 (0.8)	5 (1.2)	0.378
Elective surgery	1836 (92.8)	352 (83.6)	**<0.001**
Pre-operative sepsis	31 (1.6)	19 (4.5)	**<0.001**
Pre-operative transfusion	11 (0.6)	6 (1.4)	0.053
Pre-operative Sodium (mean ± SD)	138.4 ± 3.7	137.7 ± 4.0	**0.001**
Pre-operative BUN (mean ± SD)	15.9 ± 9.0	17.6 ± 9.9	**0.002**
Pre-operative Creatinine (mean ± SD)	0.90 ± 0.53	0.86 ± 0.36	0.114
Pre-operative WBC (mean ± SD)	8.0 ± 3.0	8.5 ± 3.5	**0.003**
Pre-operative HCT (mean ± SD)	38.7 ± 5.2	36.8 ± 5.3	**<0.001**
Pre-operative Platelets (mean ± SD)	265.7 ± 96.6	279.7 ± 98.4	**0.008**

Five models were constructed to predict DNHF, and their performance on the testing set are demonstrated in **Table 2**. Overall, with the optimized specificity of approximately 0.80-0.84 for all models, accuracy and sensitivity ranged between 0.73-0.76 and 0.39-0.53, respectively. GLM and GBM had the highest ROC's of 0.72-0.73 (**Figure 2**).

Table 2. Performance of different classification machine learning models predictive of discharge to nonhome facility.

ML Model	Accuracy	Sensitivity	Specificity	ROC-AUC	PPV	NPV
GLM	0.7542	0.5294	0.8025	0.7253	0.3659	0.8880
ANN	0.7316	0.4200	0.7986	0.6719	0.3419	0.8735
RF	0.7603	0.3894	0.8401	0.6639	0.3451	0.8648
GBM	0.7623	0.4712	0.8249	0.7181	0.3666	0.8788

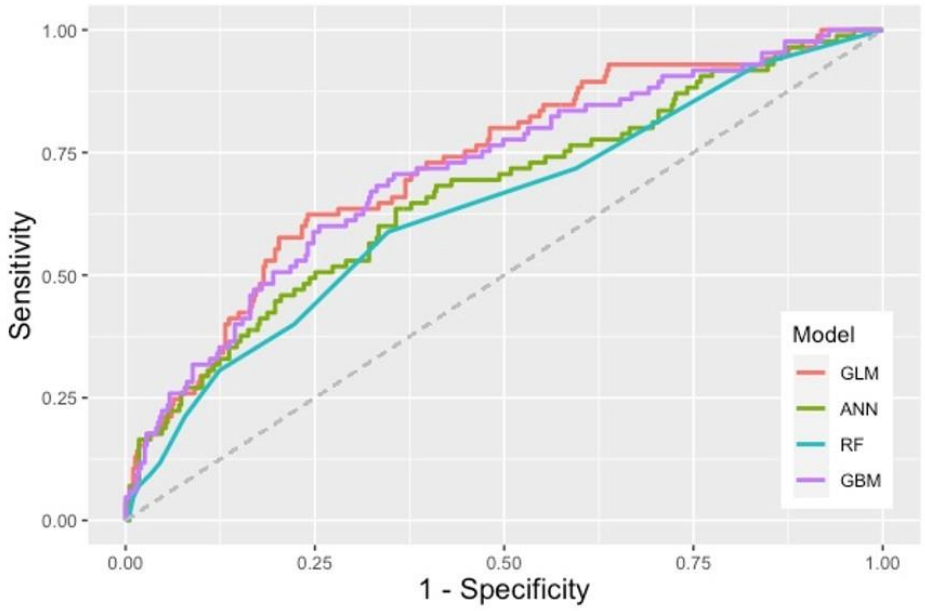


Figure 2. Receiver operating characteristic of the GLM, ANN, RF, and GBM models that used pre-operative features to predict DNHF (ROC-AUC range=0.66-0.73)

Since the algorithms' designated classification thresholds are directly related to the resulting sensitivity and specificity, the associations between these two metrics and the threshold are depicted in **Figure 3**. This demonstrates that although the majority of our reported metrics

were dependent on the optimization of ~0.80 specificity, adjusting the threshold to a value corresponding to lower specificity may raise the sensitivity; for instance, in the GBM model we developed, a balance of both sensitivity and specificity at around 0.65 could be achieved simply by using a lower classification threshold (**Figure 3B**).

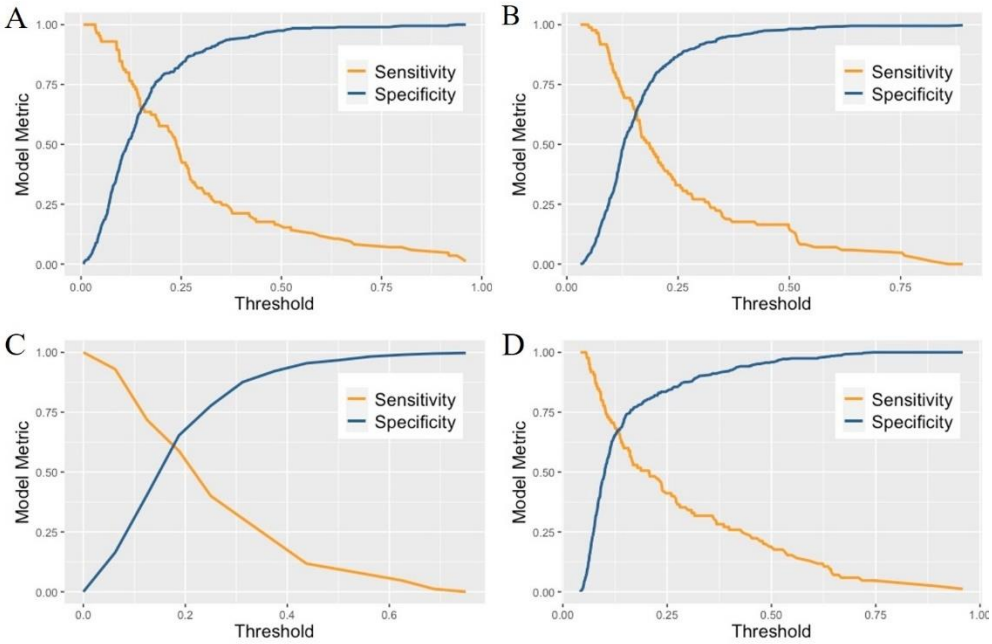


Figure 3. The trade-off relationship between sensitivity and specificity of the constructed models, directly dependent on the designated classification threshold of a given model, are demonstrated for (A) GLM, (B) ANN, (C) RF, and (D) GBM models.

Length of Stay in Complex HN Surgery

A total of 2,667 patients (27.1% female) were included for these regression-based (continuous) LOS predictive models/ The mean age and post-operative LOS were 61.9 ± 11.7 years and 10.4 ± 5.5 days, respectively. The input variables included age (p<0.01), BMI (p=0.02), gender (p=0.52), race (p<0.01), procedure type (p=0.08), elective surgery (p<0.01), diabetes (p=0.05), smoking (p=0.02), dyspnea (p=0.01), hypertension (p=0.08), dependency functional status(p<0.01) , history of congestive heart failure (p=0.05), disseminated cancer(p<0.01), history

of wound infection (p=0.02), pre-operative sepsis (p=0.09) or transfusion (p=0.10), ASA class (p=0.06), and pre-operative sodium (p<0.01), WBC (p<0.01), and HCT (p<0.01), with the p-values representing their association with LOS on independent t-test (for categorical variables, e.g., gender) or correlation analysis (for continuous variables, e.g., age). Four models were constructed to predict LOS, and their performance on the testing set are demonstrated in **Table 3**.

Table 3. Performance of different regression ML models predictive of LOS in days.

ML Model (Regression)	MAE	MSE	RMSE	RMSLE
GLM	3.9559	26.5832	5.1559	0.4545
ANN	3.9456	26.5878	5.1563	0.4536
RF	3.9770	26.4624	5.1442	0.4590
GBM	3.9783	26.6121	5.1587	0.4610

The performances of these models were compared according to different extents of LOS, demonstrating better predictive ability when the actual LOS was >8 days (**Figure 4)**.

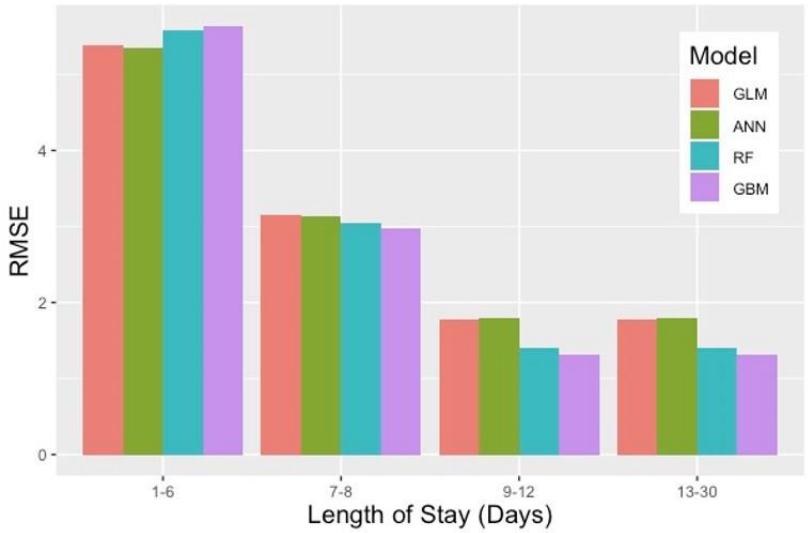


Figure 4. Comparing RMSE of the regression models stratified based on LOS quartiles: 1-6 days (N=105), 7-8 days (N=154), 9-12 days (N=131), and 13-30 days (N=144). The stratification is according to the actual LOS, and the bars represent the root mean squared error of the respective predicted LOS.

Predictive Modeling Interface for DNHF and LOS

The best-performing ML algorithm overall for each model was developed into an encrypted web-based interface (**Figure 5**), which can be accessed at the following link: https://uci-ent.shinyapps.io/head-neck/.

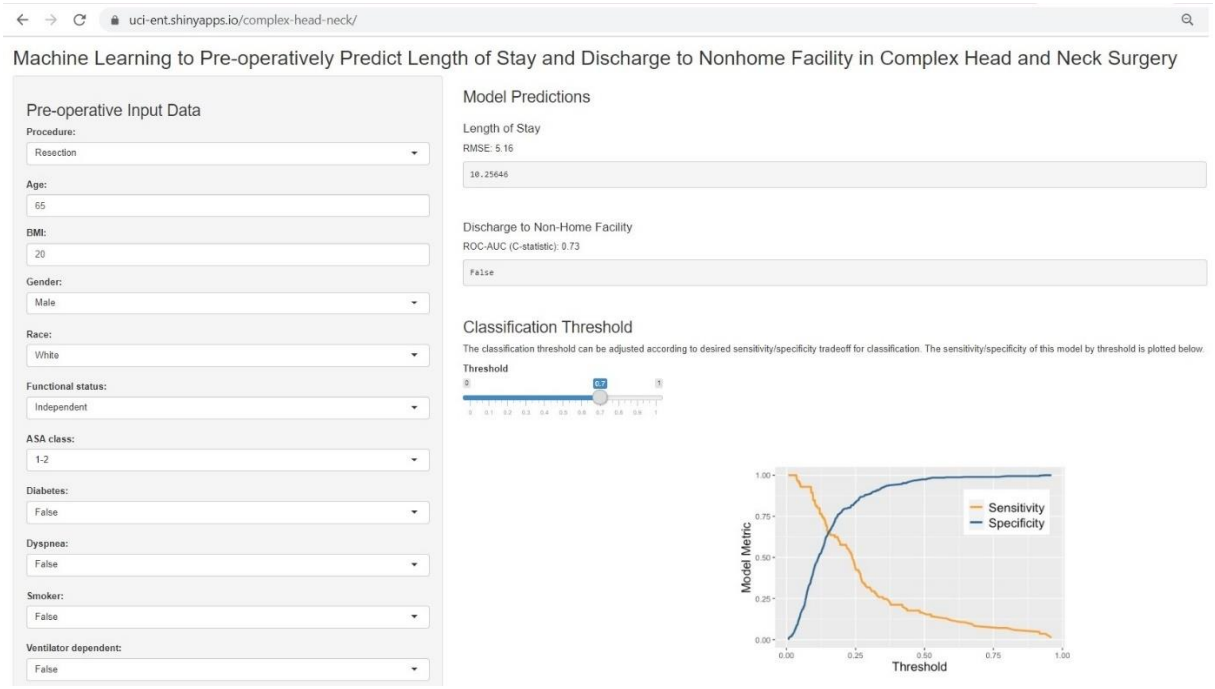


Figure 5. Display of the developed Complex HN Surgery DNHF and LOS on a publicly available interface (https://uci-ent.shinyapps.io/head-neck/). Eleven of the total 26 possible input variables are shown in the figure. The online model is capable of predicting DNHF and LOS, and the threshold (directly related to sensitivity/specificity) can be adjusted by changing the "classification threshold".

Short-term Adverse Events Following Vestibular Schwannoma Surgery

A total of 1,783 patients (57.1% female) with a mean age of 50.4 ± 13.9 years undergoing VS surgery were included. Mean operation time and LOS were 403.1 ± 163.1 minutes and 5.0 ± 4.4 days, respectively. Surgical complication, medical complications, and unplanned reoperation were observed in 92 (5.2%), 111 (6.2%), and 151 (8.5%) of patients, respectively. The best performing model for each outcome variable, optimized for AUROC, are described in **Table 4**. The AUROC's are compared in **Figure 6**.

Table 4. Performance of the three deep learning models predicting short-term reoperation, surgical complication, and medical complication following VS surgery.						
Outcome	ROC-AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Reoperation	0.74	0.91	0.38	0.95	0.38	0.95
Surgical complication	0.70	0.71	0.56	0.72	0.14	0.95
Medical complication	0.83	0.92	0.48	0.95	0.37	0.97

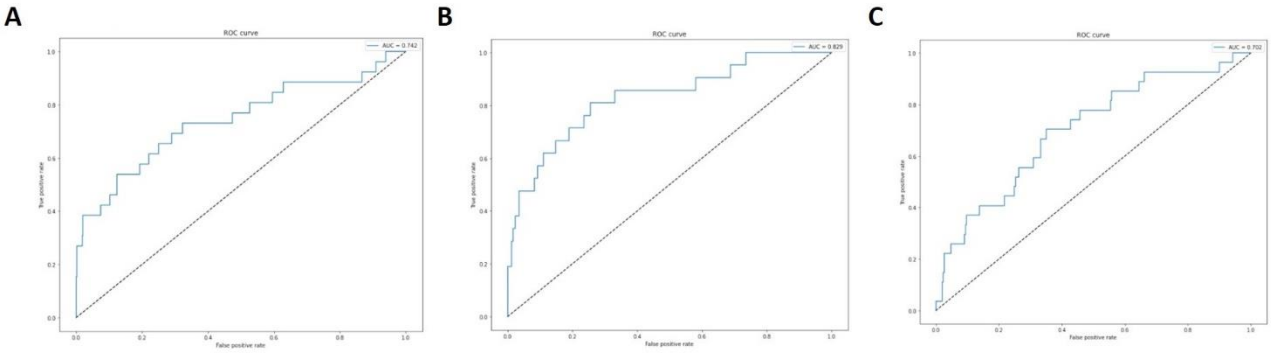


Figure 6. Comparing ROC-AUC of models predicting reoperation (0.74), surgical complication (0.83), and medical complications (0.70).

Due to the much larger number of inputs, some of which are difficult to estimate depending on the time of using the model, a web-based model was deemed less practical.

Otoscopic Image Classification

The highest AUROC and accuracy were yielded by the Inception-Resnet-V2 network. Two models were constructed, a binary model to differentiate between normal and abnormal images, and a multiclass model to differentiate between normal, AOM, OE, CSOM, and cerumen. The ROC-AUC of the two models are depicted in **Figure 7**. The binary classification model had the following performance for predicting abnormal otoscopic images: 70% sensitivity, 84% specificity, 81% PPV, 74% NPV, and 77% accuracy. **Table 5** shows the performance of the multiclass model for predicting not only abnormal images but also suggest a pathology, which should be interpreted with caution due to the low number of pathologic images for training/testing. **Figures 8** compares the confusion matrices of the binary and multiclass models.

Table 5. Performance of the binary image classification model for predicting normal versus abnormal otoscopic imaging.						
Outcome	ROC-AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
Abnormal	0.89	0.77	0.70	0.84	0.81	0.74

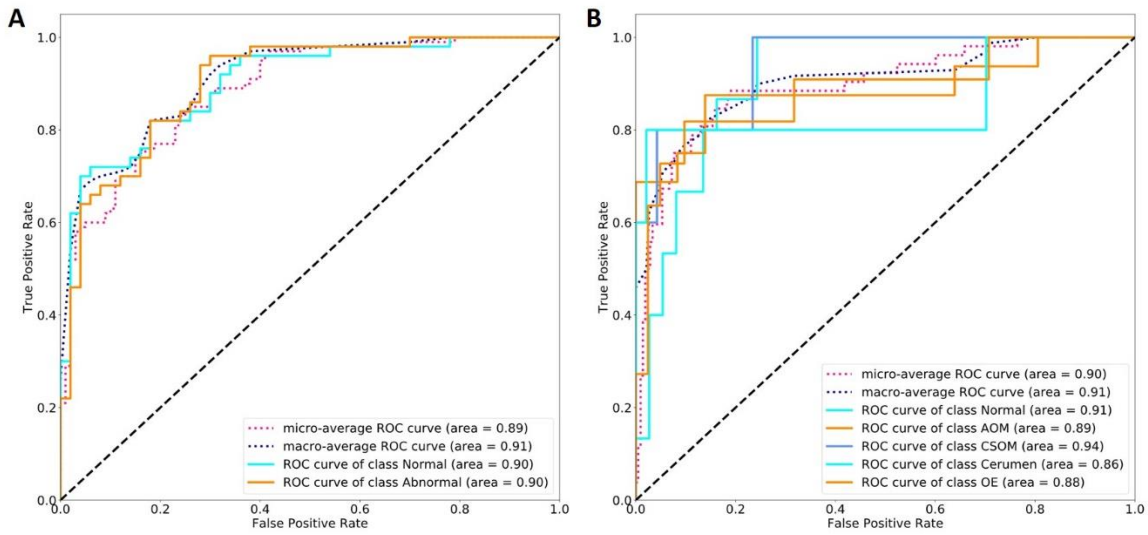


Figure 7. ROC-AUC for A) binary (normal versus abnormal), and A) multiclass image classification (healthy versus AOM vs. CSOM vs. cerumen).

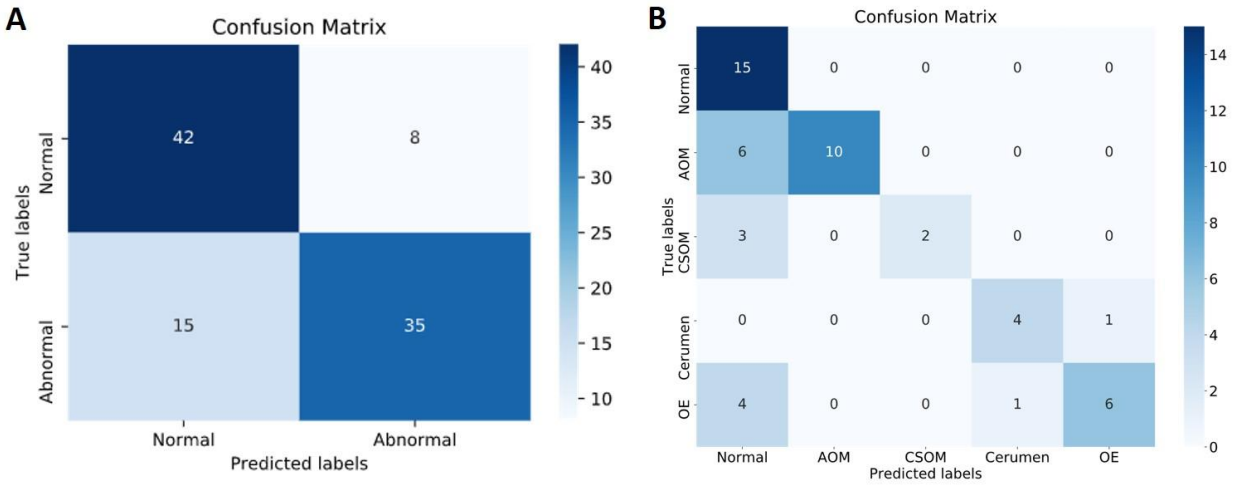


Figure 8. Confusion matrices for A) binary (normal versus abnormal), and A) multiclass image classification (healthy versus AOM vs. CSOM vs. cerumen). True and predicted values are on Y- and X-axis, respectively.

The best-performing ML algorithm for classifying otoscopic images was developed into an encrypted web-based interface (**Figure 9**), which can be accessed at the following link: https://headneckml.com/tm/.

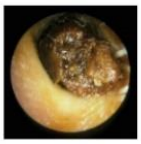
Image Preview	Classification Label	Class probabilities (Normal, Abnormal)	Multi Classification Label	Multi Class probabilities (Normal, AOM, CSOM, Earwax, OE)
	Normal	0.778,0.222	Normal	1.000,0.000,0.000,0.000,0.000
	Abnormal	0.040,0.960	CSOM	0.000,0.000,1.000,0.000,0.000
	Abnormal	0.257,0.743	Earwax	0.021,0.002,0.000,0.977,0.000

Figure 9. Image classification interface with the built-in ML algorithm for evaluating otoscopic images.

CHAPTER 5: DISCUSSION

This thesis demonstrated the successful development of novel ML models to predict DNHF and LOS following complex HN surgeries, short-term complications following VS surgeries, and possible pathologies among otoscopic images. Furthermore, one of the major novelties was the publications of some of these ML algorithms as public interfaces for simulation and further examination, thus providing practical value and opportunities for external validation. Additional important aspects of the performance and applicability of these models included the nature of the original database, appropriate selection of input features, tradeoff between performance parameters of the model, and plans for continuous improvement.

Utilizing the NSQIP dataset and only pre-operative features, our DNHF classification models demonstrated that GLM and GBM models performed well with ROC-AUC of 0.72 and 73, followed by ANN and RF models with ROC-AUC of 0.67 and 0.66, respectively. These models were classified to optimize specificity to approximately 0.80, resulting in high NPVs ranging 0.86-0.89. Moreover, we successfully built four regression ML models that performed similarly in predicting LOS (in days) with MAE and RMSE of 3.9-4.0 and 5.1-5.2, respectively. Although these results lacked external validation, we published both models via an encrypted open-access interface for the readership as a means of potential future external validation. The different performance of our predictive models compared to each other, or to other DNHF models published in the non-head and neck literature,[17,55] stem from several reasons[19]: 1) feature selection and the algorithm's treatment (e.g., attributed weight) of each input variable; 2) finding optimal free parameter values used for feature transformation and class-based prediction; 3) balancing the trade-off between model complexity and generalizability to novel cases, while remaining cognizant of the potential for overfitting;[69] and finally 4) the inherit associative strengths of the

features in predicting the outcome of interest. To achieve optimal predictive performance, it is imperative that multiple models and parametric configurations be developed and evaluated for a given ML application, and that a robust training method (such as k-fold cross-validation) be utilized to minimize the risk of overfitting. Further, it is worth noting that the performance and validity of a ML model depends not only on the characteristics of the model itself, but also on the nature of the dataset and the degree of underlying direct or hidden relationships between the features and outcome.[70-72]

All of the developed ML models were supervised ML constructs. Since the existence of missing values can diminish ML model optimization, this can be addressed in several ways including subject elimination (which can eliminate a large number of subjects overall, considering all the available input variables), substitution of missing variables with that input's median value (e.g. assuming age 59 for all subjects with missing age), or using ML applications to substitute missing values with predicted estimation relative to each subject's other characteristics. To our knowledge, the last option is one of the more sophisticated and robust methods to address this, which was utilized for our outcome models. Namely, we used the *missForest* package in the R statistical programming language,[63] which imputed missing values in a manner less prone to bias compared to alternative methods of handling missing values.[64] GLM algorithms have been most frequently utilized for constructing predictive models in the medical literature,[26] and it is possible for them to perform similarly or better than more complex or non-linear models.[17,73] The fact that GLM and GBM models were the most ideal in predicting DNHF in this cohort suggests that model performance is not always correlated with its complexity, but that it is also heavily dependent on the dataset and nature of input-output relationships. Compared to GLM, ANN can detect more complex and non-linear relationships and detect implicit interactions, but they can also be

associated with more computational time, difficult interpretation, and prone to overfitting.[74] Likewise, all other ML models present their own unique advantages and disadvantages, which should be considered when approaching a specific scientific question as well as the availability and quality of resources. Regarding our regression ML models for post-operative LOS (in days), GLM, GBM, ANN, and RF all performed similarly. Although previous papers in orthopedic and cardiology literature have developed classification models for predicting LOS (e.g., short *vs.* long LOS),[40,75] our proof-of-concept regression model with numeric outputs was more novel to the literature, paving path for future studies to develop such algorithms.

In applying classification models, adjusting the classification threshold values determined the balance between sensitivity and specificity of the predicted outcomes. As we demonstrated in Figure 2, adjusting the threshold can significantly modify the respective sensitivities and specificities, which is an important theme when considering a model's practicality and intended application. Of note, changing the threshold does not change the underlying model and is not a model parameter; rather, the model outputs a probability for each prediction, after which the threshold is applied such that probabilities above the threshold will be classified as positive (e.g., DNHF). Our DNHF results prioritizing specificity over sensitivity expressed a preference for having high confidence in positive predictions, as opposed to preferring to detect a greater proportion of positive outcomes (which would be prone to more false alarms). Our VS models were optimized for ROC-AUC, and many variables including those known at various pre- or post-operation timelines may have resulted in better performance but lower practical value and lack of a useful online interface. It is also important to emphasize the time-point at which a model is intended to be utilized. Our designed *pre-operative* models precluded the use of intra- and post-operative variables which could have strong associations with the outcome. This distinction is

further demonstrated in a study predicting readmission following spine surgery, where the aggregate ML models' ROC decreased from 0.81 to 0.58 when only including pre-operative characteristics.[15] When constructing a ML model, it is crucial to consider the applicability of such a model. If, for example, a model intended to predict surgical outcomes for pre-operative planning takes as inputs variables which are only known intra-operatively or post-operatively, the model may fail to provide any practical benefit.

Constructing non-classification regression models are much more rare in the clinical ML literature,[56] and their comparison metrics are potentially more difficult to interpret. MAE and RMSE, which are easier to interpret since they can be thought in the same units of the outcome variable, was 4-5 days for our models. This was notably lower than the cohort's averaged LOS of around 10 days. Of note, MSE and RMSE are more sensitive to the predictive ability of outliers, which can be optimized depending on the application. Furthermore, classification ML studies may use a variety of metrics (e.g., sensitivity, accuracy, ROC, PPV, etc.) to evaluate and compare models, each of which individually can be misleading to compare on face-value while ignoring the other respective metrics.[76] As such, the important trade-off relationship between specificity/sensitivity and PPV/NPV as it relates to a model's classification threshold was demonstrated graphically in this paper. Another example of a potentially misleading metric when applied to an imbalanced dataset is accuracy. A ML model applied to a dataset with 90% negative outcomes and 10% positive outcomes could trivially achieve 90% accuracy simply by classifying every data point as negative. Yet, few would consider this a useful model in practice.[76]

Our image classification model was a proof-of concept development using publicly available images, in which InceptionResnetV2 had the best relative performance. The additional novelty of this work included the development of a novel website incorporating the algorithm,

where novel otoscopic images can be uploaded for a diagnosis prediction. Misdiagnoses of otologic disorders by healthcare workers in various roles or stages of training are not uncommon,[77-79] thus efforts to aid in objective diagnosis of otologic disorders using ML have emerged.[80,81] A recent study has advanced this frontier by utilizing otologic images using Google's AutoML platform with results that on average performed better than physician evaluations.[23] We believe that the development of multiple models by various groups will encourage advancement within this field, and that our publication of the algorithm as an online interface is a novel and important step for future utility of ML image classification models for actual benefit of patients and clinicians. Although the constructed model especially when classifying between different pathologies was limited in performance, additional institutional images for continuous training and improvement of the model will hopefully lead to a more reliable end-model.

The future of ML in the medical field including otolaryngology can be associated with both promises and perils. One central challenge is assembling large, accurate, diverse, and representative datasets for developing reliable ML applications, while another is the trade-off between model complexity and generalizability to unseen cases.[19,45] The advent of big data has allowed development of ML using national databases, but this can also bring disadvantages including heterogeneity in quality or reliability, various biases, and challenges regarding certain intnerpretations such as causal inferences.[82] As previously elucidated to, ML is especially useful for identifying complex or hidden patters in large databases that would be indiscernible to human eye or traditional statistical inferences. However, it is important to keep in mind that complexity does not always equate to improvement, and sometimes ML can perform similarly or worse than traditional analytical methods.[83,84] Besides being heavily dependent on a sizable and high-quality dataset for training, this learning can be prone to systematic biases since supervised learning is

predicated upon manually labeled inputs, feature inclusion, model selection, and adjustments for under- or over-fitting.[69,85-87] On the other hand, a shift of reliance on ML for certain applications can lead to automaton bias[88] and diminished vigilance for provider errors.[45] Also, although the Food and Drug Administration's guidelines for incorporation of ML into clinical care remain indefinite, regulation/liability and protecting patient privacy are important aspects of future real-life applications of these technologies.[89-91] Despite the hurdles, ML can potentially have the ability to transform surgical care in areas of diagnosis, prognosis, surgical planning, and risk-management.[2,6-9,38,47,51,55] This warrants investigating time and resources into developing reliable and practical ML applications for otolaryngology patients.

Limitations

Discussing the limitations of this research will be important for future progress of ML in the field of otolaryngology. First, the post-operative prediction models utilized retrospective data from a de-identified national database, which is prone to missing values, miscoding, or inherent biases. For the DNHF and LOS models, we chose not to include intra- and post-operative variables (e.g., operation time, transfusion, post-operative complications, etc.) even though they may strongly associate with DNHF or LOS. This was because these models intended to guide clinicians at a pre-operative standpoint to provide a window of opportunity for appropriate planning or risk adjustment. It is important for a practical model to incorporate a clear point in time at which all the input data could be known, for providing insight into the likelihood of a given outcome, yet this may be more difficult at earlier time points (e.g., pre-operatively *vs.* post-operatively). Additionally, the databases did not include socioeconomic variables which may potentially associate with DNHF or LOS, such as insurance, income, or employment status, or clinical variables which may potentially associate with post-operative adverse events such as the institution

type, experience of the surgeon or residents involved, and specific intra-operative complications not captured by the database. This can be addressed by performing future multi-institutional studies that benefit from more comprehensive input data. Furthermore, although the models contained internal validation by partitioning of the cohort into training and testing cohorts, there was no external validation which would require testing outside patients. To address this, we published our algorithms via an encrypted online interface and encourage clinicians to test hypothetical patients/situations. Finally, while our outcome models were carefully tuned to maximize predictive performance, they were limited by the extent of inherent relationships between the features and outcomes.

There exist several important limitations with the image classification model as well, as this is still at a preliminary and proof-of-concept stage. The number of images available for training was extremely low compared to some of the well-established models. For instance, the referenced diabetic retinopathy detection study used a dataset of 128,175 total retinal images.[20] Also, the classification labels were according to the original sources, and the reliability could not be guaranteed. In addition to the low number, the images were extracted from different resources with varying sizes and qualities. These limitations can be addressed with future large and multi-institutional studies to collect a much larger cohort of otoscopic images, including healthy and pathologic cases, with reliable physician diagnoses and homogenous image size/quality. Despite these limitations, developing these supplementary tools that may assist in diagnosis or post-operative outcome predictions have the potential to improve medical management. This can also lead to better resource allocation and universal high-quality care especially at disadvantaged or understaffed regions. As such, future studies to continue developing ML applications for different otolaryngology diagnoses, surgeries, and outcomes are warranted.

Future Directions

This thesis presented multiple applications of ML in the field of otolaryngology that could theoretically improve the provided care and patient outcomes. Future studies will need to improve these models and demonstrate reliable external validity using novel patients independent from the training/testing datasets. Although building these models was not feasible using institutional data due to a much lower number of available cases, we will plan identifying ML applications with enough institutional data so more comprehensive and reliable inputs could be used for model training. This is similar to the image classification model, and we are currently applying for Institutional Review Board approval to prospectively acquire institutional otoscopic images using hand-held otoscopes available at our tertiary care center clinic. Although it will likely take 2-3 years to acquire enough images of the various pathologies, we will plan to collaborate with other institutions for quicker data collection. If future models can predict various pathologies with high reliability, the current proof-of-concept website can transition into a publicly available smartphone which would be more user-friendly and portable in a clinical setting. We will continue investigating and developing additional ML applications of other post-operative outcomes in otolaryngology, especially ones with the potential to provide clinical value and feasibility to be incorporated in real-life patient care situations.

Conclusions

We have successfully developed novel otolaryngology-specific ML models with the ability to predict DNHF or LOS following complex HN surgery, reoperation or complications following VS surgery, and otologic abnormalities on otoscopic imaging. Furthermore, we have published several of the models as novel and publicly available interfaces for practical use as well as external examination/validation. We have further explored important factors that play significant roles in

performance and applicability of such ML models, including the nature of the original database, inherent relationship between input features and outcomes of interest, tradeoff between performance parameters of the model, and what clinical questions can be feasibly approached via ML for practical applications.

REFERENCES

1. Mitchell TM. Machine learning: McGraw-hill New York. 1997.
2. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*. 2016; 375:1216.
3. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*. 2001; 23:89-109.
4. Nguyen LD, Lin D, Lin Z, Cao J. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. *IEEE International Symposium on Circuits and Systems (ISCAS)*. 2018:1-5.
5. Shipp MA, Ross KN, Tamayo P et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*. 2002; 8:68-74.
6. Bihorac A, Ozrazgat-Baslanti T, Ebadi A et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Annals of surgery*. 2019; 269:652-62.
7. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Annals of surgery*. 2018; 268:70.
8. Tamashiro A, Yoshio T, Ishiyama A et al. Artificial-intelligence-based detection of pharyngeal cancer using convolutional neural networks. *Digestive Endoscopy*. 2020. [Online Ahead of Print]
9. Formeister EJ, Baum R, Knott PD et al. Machine Learning for Predicting Complications in Head and Neck Microvascular Free Tissue Transfer. *The Laryngoscope*. 2020. [Online Ahead of Print]
10. Missios S, Bekelis K. Drivers of hospitalization cost after craniotomy for tumor resection: creation and validation of a predictive model. *BMC health services research*. 2015; 15:85.
11. Carey MR, Sheth H, Braithwaite RS. A prospective study of reasons for prolonged hospitalizations on a general medicine teaching service. *Journal of general internal medicine*. 2005; 20:108-15.

12. Rosman M, Rachminov O, Segal O, Segal G. Prolonged patients' In-Hospital Waiting Period after discharge eligibility is associated with increased risk of infection, morbidity and mortality: a retrospective cohort analysis. *BMC health services research*. 2015; 15:246.
13. Kulinskaya E, Kornbrot D, Gao H. Length of stay as a performance indicator: robust statistical methodology. *IMA Journal of Management Mathematics.* 2005; 16:369-81.
14. Tokunaga J, Imanaka Y. Influence of length of stay on patient satisfaction with hospital care in Japan. *International Journal for Quality in Health Care*. 2002; 14:493-502.
15. Hopkins BS, Yamaguchi JT, Garcia Ret al. Using machine learning to predict 30-day readmissions after posterior lumbar fusion: an NSQIP study involving 23,264 patients. *Journal of Neurosurgery: Spine*. 2019; 1:1-8.
16. Van Niftrik CH, Van Der Wouden F, Staartjes VE et al. Machine learning algorithm identifies patients at high risk for early complications after intracranial tumor surgery: registry-based cohort study. *Neurosurgery*. 2019; 85:E756-64.
17. Goyal A, Ngufor C, Kerezoudis P, McCutcheon B, Storlie C, Bydon M. Can machine learning algorithms accurately predict discharge to nonhome facility and early unplanned readmissions following spinal fusion? Analysis of a national surgical registry. *Journal of Neurosurgery: Spine*. 2019; 31:568-78.
18. Karhade AV, Thio QC, Ogink PT et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. *Neurosurgery*. 2019; 85:E83-91.
19. Deo RC. Machine learning in medicine. *Circulation*. 2015; 132:1920-30.
20. Gulshan V, Peng L, Coram M et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*. 2016; 316:2402-10.
21. Moradi M, Abolmaesumi P, Siemens DR, Sauerbrei EE, Boag AH, Mousavi P. Augmenting detection of prostate cancer in transrectal ultrasound images using SVM and RF time series. *IEEE Transactions on Biomedical Engineering*. 2008; 56:2214-24.
22. Shan J, Alam SK, Garra B, Zhang Y, Ahmed T. Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. *Ultrasound in medicine & biology*. 2016; 42:980-8.

23. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: An automated machine learning approach. *The Laryngoscope*. 2019. [Online Ahead of Print]
24. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. MIT press, 2018.
25. Li Y, Crandall DJ, Huttenlocher DP. Landmark classification in large-scale image collections. *2009 IEEE 12th international conference on computer vision*. 2009:1957-64.
26. Hasan O, Meltzer DO, Shaykevich SA et al. Hospital readmission in general medicine patients: a prediction model. *Journal of general internal medicine*. 2010; 25:211-9.
27. Daniel G. Principles of artificial neural networks. *World Scientific*. 2013.
28. Vapnik VN. The nature of statistical learning. *Theory*. 1995.
29. Moeskops P, Viergever MA, Mendrik AM, De Vries LS, Benders MJ, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE transactions on medical imaging*. 2016; 35:1252-61.
30. Hu B, Lu Z, Li H, Chen Q. Convolutional neural network architectures for matching natural language sentences. *Advances in neural information processing systems*. 2014:2042-50.
31. Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002; 2:18-22.
32. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001:1189-232.
33. Ye J, Chow J-H, Chen J, Zheng Z. Stochastic gradient boosted distributed decision trees. *Proceedings of the 18th ACM conference on Information and knowledge management.* 2009:2061-4.
34. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC medical research methodology*. 2019; 19:64.
35. Vedaldi A, Lenc K. Matconvnet: Convolutional neural networks for matlab. *Proceedings of the 23rd ACM international conference on Multimedia*. 2015:689-92.
36. Bischl B, Lang M, Kotthoff L et al. mlr: Machine Learning in R. *The Journal of Machine Learning Research*. 2016; 17:5938-42.
37. Pedregosa F, Varoquaux G, Gramfort A et al. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*. 2011; 12:2825-30.

38. Narula S, Shameer K, Omar AMS, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. *Journal of the American College of Cardiology*. 2016; 68:2287-95.
39. Anderson JP, Parikh JR, Shenfeld DK et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of diabetes science and technology*. 2016; 10:6-18.
40. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International journal of cardiology*. 2019; 288:140-7.
41. Chang S-W, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*. 2013; 14:170.
42. Yu K-H, Zhang C, Berry GJ et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*. 2016; 7:12474.
43. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association*. 2014; 21:315-25.
44. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Initiative AsDN. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*. 2015; 104:398-412.
45. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019; 380:1347-58.
46. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv*. 2016: 160605718.
47. Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology*. 2018; 286:810-8.

48. Kim JK, Yook IH, Choi MJ et al. A Performance Comparison on the Machine Learning Classifiers in Predictive Pathology Staging of Prostate Cancer. *Studies in health technology and informatics*. 2017; 245:1273.
49. Fehr D, Veeraraghavan H, Wibmer A et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proceedings of the National Academy of Sciences*. 2015; 112:E6265-73.
50. Kirk PS, Liu X, Borza T et al. Dynamic readmission prediction using routine postoperative laboratory results after radical cystectomy. *Urologic Oncology.* 2020; 28:255-61.
51. Bhandari M, Nallabasannagari AR, Reddiboina M et al. Predicting intraoperative and postoperative consequential events using machine learning techniques in patients undergoing robotic partial nephrectomy (RPN): Vattikuti Collective Quality Initiative (VCQI) database study. *BJU International*. 2020 [Online Ahead of Print]
52. Burgansky-Eliash Z, Wollstein G, Chu T et al. Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. *Investigative ophthalmology & visual science.* 2005; 46:4147-52.
53. Kanevsky J, Corban J, Gaster R, Kanevsky A, Lin S, Gilardino M. Big data and machine learning in plastic surgery: a new frontier in surgical innovation. *Plastic and reconstructive surgery*. 2016; 137:890-7.
54. Merath K, Hyer JM, Mehta R et al. Use of machine learning for prediction of patient risk of postoperative complications after liver, pancreatic, and colorectal surgery. *Journal of Gastrointestinal Surgery*. 2019. [Online Ahead of Print]
55. Ogink PT, Karhade AV, Thio QC et al. Predicting discharge placement after elective surgery for lumbar spinal stenosis using machine learning methods. *European Spine Journal*. 2019; 28:1433-40.
56. Muhlestein WE, Akagi DS, Davies JM, Chambless LB. Predicting inpatient length of stay after brain tumor surgery: Developing machine learning ensembles to improve predictive performance. *Neurosurgery*. 2019; 85:384-93.
57. Kim JS, Arvind V, Oermann EKet al. Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. *Spine deformity*. 2018; 6:762-70.

58. Gowd AK, Agarwalla A, Amin NH et al. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. *Journal of shoulder and elbow surgery*. 2019; 28:410-21.
59. Patel VA, Dunklebarger M, Banerjee K, Shokri T, Zhan X, Isildak H. Surgical Management of Vestibular Schwannoma: Practice Pattern Analysis via NSQIP. *Annals of Otology, Rhinology & Laryngology*. 2020; 129:230-7.
60. Mahboubi H, Haidar YM, Moshtaghi O et al. Postoperative complications and readmission rates following surgery for cerebellopontine angle schwannomas. *Otology & neurotology*. 2016; 37:1423.
61. Biron DR, Sinha I, Kleiner JE et al. A Novel Machine Learning Model Developed to Assist in Patient Selection for Outpatient Total Shoulder Arthroplasty. *The Journal of the American Academy of Orthopaedic Surgeons*. 2019. [Online Ahead of Print]
62. Livingstone D, Talai AS, Chau J, Forkert ND. Building an Otoscopic screening prototype tool using deep learning. *Journal of Otolaryngology-Head & Neck Surgery*. 2019; 48:1-5.
63. Stekhoven DJ. Using the missForest package. *R package*. 2011:1-11.
64. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012; 28:112-8.
65. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *Journal of machine learning research*. 2012; 13:281-305.
66. Zou Q, Xie S, Lin Z, Wu M, Ju Y. Finding the best classification threshold in imbalanced classification. *Big Data Research*. 2016; 5:2-8.
67. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statistics surveys*. 2010; 4:40-79.
68. Başaran E, Cömert Z, Çelik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification. *Biomedical Signal Processing and Control*. 2020; 56:101734.
69. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*. 2010; 11:2079-107.
70. Riley P. Three pitfalls to avoid in machine learning. *Nature*. 2019; 527:27-9.
71. Brynjolfsson E, Mitchell T. What can machine learning do? Workforce implications. *Science*. 2017; 358:1530-4.

72. Rowe M. An introduction to machine learning for clinicians. *Academic Medicine*. 2019; 94:1433-6.
73. Mohammadinia A, Saeidian B, Pradhan B, Ghaemi Z. Prediction mapping of human leptospirosis using ANN, GWR, SVM and GLM approaches. *BMC infectious diseases*. 2019; 19:1-18.
74. Ayer T, Chhatwal J, Alagoz O, Kahn Jr CE, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics*. 2010; 30:13-22.
75. Ramkumar PN, Navarro SM, Haeberle HS et al. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. *The Journal of arthroplasty*. 2019; 34:632-7.
76. Baştanlar Y, Özuysal M. Introduction to machine learning. *miRNomics: MicroRNA Biology and Computational Analysis*. Springer, 2014:105-28.
77. Pichichero ME, Poole MD. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. *International journal of pediatric otorhinolaryngology*. 2005; 69:361-6.
78. Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. *Archives of pediatrics & adolescent medicine*. 2001; 155:1137-42.
79. Steinbach WJ, Sectish TC, Benjamin DK, Chang KW, Messner AH. Pediatric residents' clinical diagnostic accuracy of otitis media. *Pediatrics*. 2002; 109:993-8.
80. Viscaino M, Maass JC, Delano PH, Torrente M, Stott C, Auat Cheein F. Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. *Plos one.* 2020; 15:0229226.
81. Monroy GL, Won J, Dsouza R et al. Automated classification platform for the identification of otitis media using optical coherence tomography. *NPJ digital medicine*. 2019; 2:1-11.
82. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*. 2018; 39:95-112.

83. Austin PC, Tu JV, Lee DS. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *Journal of clinical epidemiology*. 2010; 63:1145-55.
84. Puddu PE, Menotti A. Artificial neural networks versus proportional hazards Cox models to predict 45-year all-cause mortality in the Italian Rural Areas of the Seven Countries Study. *BMC medical research methodology*. 2012; 12:100.
85. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*. 2019; 116:15849-54.
86. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*. 2018; 178:1544-7.
87. Schaffer C. Overfitting avoidance as bias. *Machine learning*. 1993; 10:153-78.
88. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*. 2017; 24:423-31.
89. Mattessich S, Tassavor M, Swetter SM, Grant-Kels JM. How I learned to stop worrying and love machine learning. *Clinics in dermatology*. 2018; 36:777-8.
90. Deist TM, Jochems A, van Soest J et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and translational radiation oncology*. 2017; 4:24-31.
91. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. *Proceedings of the IEEE*. 2016; 104:444-66.