Received: 10 June 2023 | Revised: 8 January 2024 | Accepted: 23 January 2024
DOI: 10.1002/hed.27674

ORIGINAL ARTICLE

WILEY

A supervised machine learning model for identifying predictive factors for recommending head and neck cancer surgery

Max L. Jiam BS[1] | **Kevin Z. Xin MD**[2] | **Patrick K. Ha MD**[3] | **Nicole T. Jiam MD**[3,4]

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[2]Department of Radiology, University of California – Irvine, Irvine, California, USA

[3]Department of Otolaryngology – Head & Neck Surgery, University of California – San Francisco, San Francisco, California, USA

[4]Department of Otolaryngology – Head & Neck Surgery, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, USA

Correspondence
Nicole T. Jiam, Department of Otolaryngology – Head and Neck Surgery, University of California – San Francisco, 505 Parnassus Avenue, San Francisco, CA 94143, USA.
Email: nicole.jiam@ucsf.edu

Section Editor: Dennis H. Kraus

Abstract

Background: New patient referrals are often processed by practice coordinators with little-to-no medical background. Treatment delays due to incorrect referral processing, however, have detrimental consequences. Identifying variables that are associated with a higher likelihood of surgical oncological resection may improve patient referral processing and expedite the time to treatment. The study objective is to develop a supervised machine learning (ML) platform that identifies relevant variables associated with head and neck surgical resection.

Methods: A retrospective cohort study was conducted on 64 222 patient datapoints from the SEER database.

Results: The random forest ML model correctly classified patients who were offered head and neck surgery with an 81% accuracy rate. The sensitivity and specificity rates were 86% and 71%. The positive and negative predictive values were 85% and 73%.

Conclusions: ML modeling accurately predicts head and neck cancer surgery recommendations based on patient and cancer information from a large population-based dataset. ML adjuncts for referral processing may decrease the time to treatment for patients with cancer.

KEYWORDS
carcinoma, decision tree, head and neck cancer, machine learning, random forest, surgery

1 | INTRODUCTION

Treatment delays are one of the most significant, yet modifiable, variables affecting cancer patient morbidity and mortality. The consequences of delayed treatment are not benign. There is approximately a 1.2%–3.2% absolute increased risk of mortality per week of delayed treatment for cancers such as lung, kidney, and pancreas.[1] Another study by Cone et al. reported higher all-cause 5- and 10-year predicted mortality with delays in treatment initiation across the most prevalent cancers in the United States of America: breast, prostate, non-small cell lung, and colon cancer.[2]

Unfortunately, institutional infrastructure and personnel failures to recognize and process urgent cancer

---

This manuscript was presented as an oral presentation at the AHNS 11th International Conference on Head and Neck Cancer on July 12, 2023 in Montreal, Canada.

Max L. Jiam and Kevin Z. Xin contributed equally to this study.

patient referrals contribute to this delay. A study of over 100 000 scheduling attempts highlighted significant heterogeneity in the amount of time between the original referral and the scheduled appointment date, ranging from 8 to 73 days, with an average wait time of 22 days.[3] The same study concluded that the longer the wait time, the less likely an appointment was to be completed. For individuals with critical conditions, 22 days is not an acceptable wait time; these patients need to be seen as soon as possible for physicians to make a proper clinical diagnosis and management before the patient's condition worsens. This is especially important for patients with cancer, where days spent waiting for a referral to process or being referred to the incorrect provider (i.e., the incorrect subspecialty clinic) may adversely affect the patient's treatment options and prognosis.

Currently, the responsibility of processing and triaging referrals often falls upon the shoulders of practice coordinators, who may not be medical experts and often have competing responsibilities. Practice coordinators and even referring providers are unlikely to distinguish between benign versus potentially serious findings, compared to an medical or radiation oncologist, or oncological surgeon. Even if a referral is flagged as urgent by a referring provider, practice coordinators often rely on internal nurse navigators or the specialists themselves to review referral documents and relay how soon the patient should be seen. Processing bottlenecks may be amplified by factors such as pandemic backlogs, underfunding, understaffing, and supply chain constraints.

Machine learning has played an increasingly important role in science and medicine over the past decade.[4] The adoption of data-intensive machine learning models into clinical workflow has the potential to reduce operational inefficiencies and promote evidence-based decision-making—even among non-medical personnel. Decision tree algorithms offer classification (categorical outcomes) and regression (continuous outcomes) modeling in a rich multidimensional clinical setting.[5] We, therefore, evaluated if a tree-based machine learning model could predict whether a patient with cancer would be recommended for head and neck surgery based on the medical information provided at the time of a new patient referral. In addition, we sought to determine which patient variables carried the most weight in driving the decision for recommending surgery. Findings from this study may inform operational directors and practice coordinators about the best practices when it comes to processing head and neck new patient referrals—ultimately reducing inappropriate clinic visits and freeing up specialists' time so that they can "operate at the top their license."

2 | MATERIALS AND METHODS

2.1 | Training and testing dataset

This study was reviewed and approved by an institutional review board. The details of the machine learning model development are outlined in Figure 1. We used the Surveillance, Epidemiology, and End Results (SEER) dataset. The SEER database is a cancer registry funded by the National Institute of Health's (NIH) National Cancer Institute and has been in existence since 1975. Each year, the National Cancer Institute collects cancer incidence information on approximately ~800 000 new patients from 16 different cancer registries nationwide. The SEER database has been reported to represent approximately 47.9% of the United States patient population.[6] We isolated a subset population of 64 222 patients with head and neck cancer from the SEER database and split this dataset into a training and testing set, with 70% of data points assigned to the former and 30% of data points assigned to the latter (Table 1).

2.2 | Preprocessing

From the SEER dataset, we meticulously identified and isolated 21 out of a total of 263 patient variables that would be pertinent and plausibly available at the time of a new patient referral. We also removed other proxies for patient outcomes (i.e., survival months) to prevent the algorithm from exploiting information it would not generally have in the real-world setting. The specific data values utilized for this study are listed in Table S1, Supporting Information.

We compressed the parameters for cancer TMN classifications from multiple American Joint Committee on Cancer (AJCC) editions into one, as the discrepancies between tumor (T), nodal (N), and metastasis (M) stages between different AJCC cancer staging editions resulted in sparse, low-fidelity data. We recognize that compressing staging systems from multiple editions into one column could distort the actual morbidity attached to each staging value so we kept these values separated during the data analyses. Patients with metastatic disease were kept within the testing and validation dataset because in the real world, it is unlikely for a referring general practitioner or non-medical practice coordinator to recognize that patients with metastatic disease are unlikely to be recommended curative surgery.

The isolated data underwent one-hot encoding, wherein categorical variables were split into subcategories that could only have values of one or zero, before being inputted into the model (Table S1). This approach

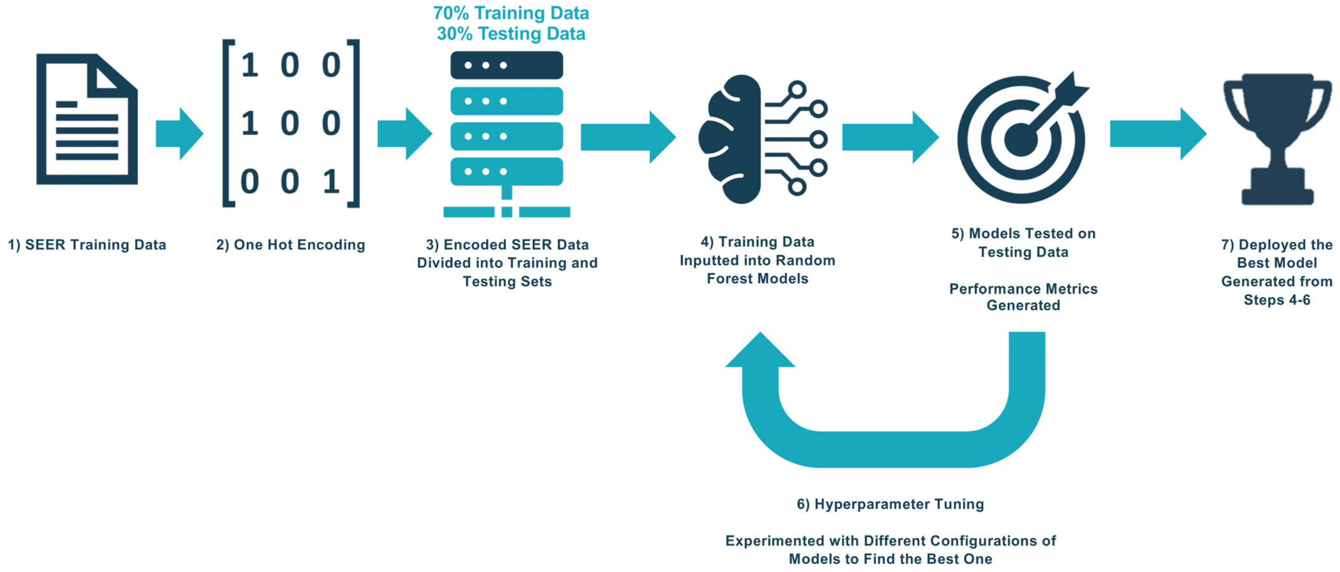

FIGURE 1 Overview schematic of the machine learning decision tree model development. SEER, Surveillance, Epidemiology, and End Results Database. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Characteristics of overall SEER data, training subset, and testing subset.

Characteristic	SEER overall	SEER training	SEER testing
Number of cases	63 553	44 487	19 066
Sex			
Male	43 696 (69%)	30 414 (68%)	13 282 (70%)
Female	19 857 (31%)	14 073 (32%)	5784 (30%)
Primary site			
Nasopharynx	3786 (6%)	2890 (6%)	896 (5%)
Oral cavity	24 984 (39%)	17 938 (40%)	7046 (37%)
Oropharynx	23 069 (36%)	15 381 (35%)	7688 (40%)
Hypopharynx	3467 (5%)	2587 (6%)	880 (5%)
Salivary gland	7148 (11%)	4999 (11%)	2149 (11%)
Other	1099 (2%)	692 (2%)	407 (2%)
Cancer grade			
Grade I	8067 (13%)	5989 (13%)	2078 (11%)
Grade II	20 474 (32%)	14 320 (32%)	6154 (32%)
Grade III	14 193 (22%)	10 153 (23%)	4040 (21%)
Grade IV	2187 (3%)	1672 (4%)	515 (3%)
Grade unknown	18 632 (29%)	12 353 (28%)	6279 (33%)
AJCC stage			
Stage I	6976 (11%)	4865 (11%)	2111 (11%)
Stage II	3857 (6%)	2745 (6%)	1112 (6%)
Stage III	4633 (7%)	3217 (7%)	1416 (7%)
Stage IV	13 988 (22%)	9225 (21%)	4763 (25%)
Unknown	34 099 (54%)	24 435 (55%)	9664 (51%)
Outcomes			
Surgery recommended	41 242 (65%)	29 830 (67%)	11 412 (60%)
Not recommended	22 311 (35%)	14 657 (33%)	7654 (40%)

Abbreviations: AJCC, American Joint Committee on Cancer; SEER, Surveillance, Epidemiology, and End Results.

helped convert non-numerical variables (i.e., sex, race, cancer classification) into a form that could be interpreted by the machine learning model.

2.3 | Random forest model development

Our model predicted whether a patient was recommended for head and neck surgery using a set of demographic and cancer variables that would be pertinent and plausibly available at the time of a new patient referral. SEER treatment outcomes were as either “Surgery performed,” “Not recommended,” “Recommended, unknown if performed,” “Recommended but not performed, unknown reason,” “Recommended but not performed, patient refused,” “Not recommended, contraindicated due to other conditions; autopsy only (1973–2002),” “Unknown; death certificate; or autopsy only (2003+),” or “Not performed, patient died prior to recommended surgery.” We removed the last two of the aforementioned outcomes from the dataset because they are the equivalent of missing or unknown variables. In addition, the scope of this study does not extend to elucidating the reasons why a patient failed to follow through with their treatment recommendation. Thus, we grouped all “recommended” categories together and “not recommended” categories together to transform our random forest model into a binary classifier.

Our random forest model classifies patients on whether they were recommended surgery, based on the isolated 21 patient variables. The model utilizes an ensemble of several decision trees trained independently, and the outcomes of each of the decision trees are aggregated to determine the final result. These “bagging” methods improve the algorithm's generalizability and accuracy when interfacing with different datasets.

The random forest model was sourced from existing libraries in the “sklearn” Python library (Scikit-Learn; https://scikit-learn.org). We performed hyperparameter tuning via grid search to extensively search for possible model configurations. These configurations are defined by variances in the model's hyperparameters, which define the learning process. The hyperparameters include how the forest's decision trees determine the quality of each split, how the forest distributes the trees, and how many trees to train. One of the hyperparameters determined when an individual tree was randomly assigned a subset of patient variables to train upon. It tweaked the number of features each individual decision tree was allowed to consider. Hyperparameters such as this one determined the balance between the model's interpretability versus overfitting. The full search space of specific hyperparameter configurations and how we split the configuration space are specified in Table S2. We tested a total of 336 hyperparameter combinations and selected the best-performing combination for the machine learning model.

To assess the model performance, we performed fivefold cross-validation. The fivefold cross-validation approach involves rotating training and testing data points (i.e., during training, leaving out a fifth of the dataset for testing). Rotating the training set reduced the likelihood of the model overfitting on any subsection of the dataset. The results from the cross-validation process are averaged to calculate the overall performance of the model. In summary, all 336 hyperparameter configurations were trained five different ways on the same training data and produced a total of 1680 fits.

A total of 63 553 adult patients with head and neck cancer were included in the study; 41 242 of these patients (64.9%) were recommended surgery and 22 311 (35.1%) were not recommended surgery. The decision trees trained on 70% of the available data with fivefold cross-validation. The remaining 30% of the dataset was kept unseen during the entirety of the training process and reserved as the test set. Our hyperparameter tuning analyses determined the best-performing combination to be when the maximum features were limited to two-tenths of the full set of variables, when the minimum samples threshold was increased to 50, when entropy was used as the criterion for choosing a separator, and no pruning was performed.

3 | RESULTS

The accuracy rate of this model on unseen test data was 81%. The sensitivity rate was 86% and the specificity rate was 71% (Figure 2A). The positive predictive value was 85% and the negative predictive value was 73%. A receiving operator characteristic (ROC) curve was constructed by plotting the algorithm's true positive rate (TPR) against the false positive rate (FPR); the area under the curve (AUC) was 0.88 (Figure 2B).

We also performed hyperparameter tuning based on the size of the training set and the number of forests, and we found diminishing returns on the number of decision trees after 10 estimators. With that in mind, we decided to use 100 estimators because there did not seem to be any performance benefit when the number of decision trees was greater than 100 (Figure 3A). Similarly, we found that the training and cross-validation metric converged as the number of training set samples increased, with visible signs of non-decreasing variance (Figure 3B). There was no benefit in raising the number of samples within the training set beyond the 70–30 split.

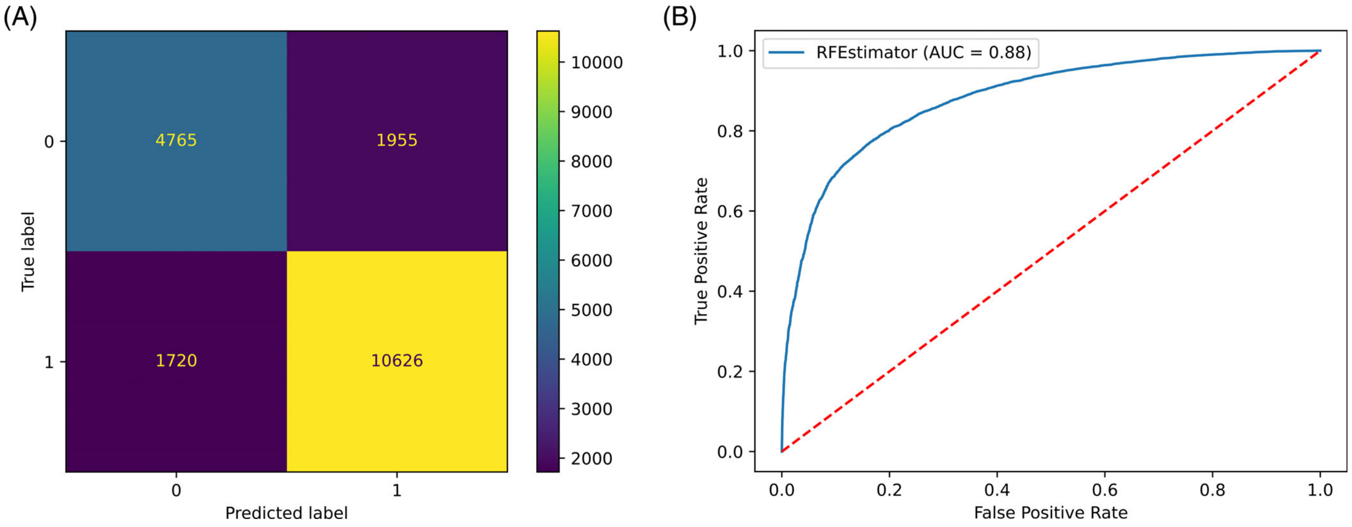

FIGURE 2 Performance metrics for the random forest model. (A) The confusion matrix demonstrates an unseen test data prediction accuracy of 81%. The "0" label refers to cases where the patient is recommended for surgery and "1" refers to cases when they are not. (B) The receiver operating characteristic (ROC) curve of the decision tree model. The closer the area under the curve (AUC) is to one, the better the algorithm is performing. The red line depicts the naive, baseline approach of randomly recommending patients regardless of their characteristics. The further an estimator deviates from this line, the better it is. AUC, area under the curve; RFEstimator, random forest estimator. [Color figure can be viewed at wileyonlinelibrary.com]

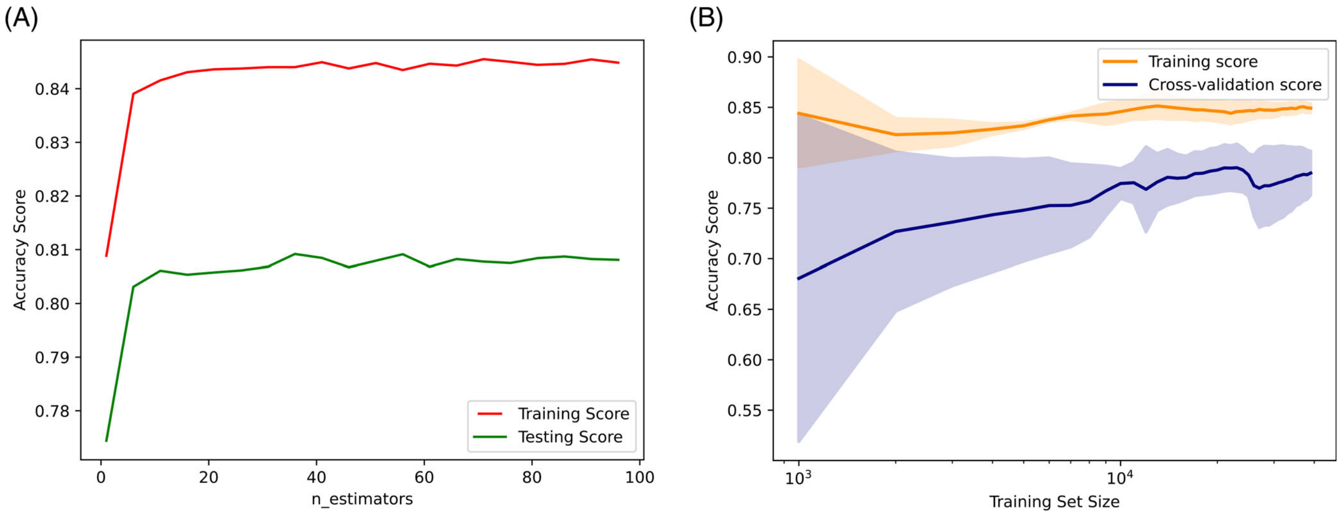

FIGURE 3 Model performance analysis: the impact of decision trees and training set size on the accuracy of the random forest model. These two figures demonstrate the model's performance relative to the information provided or resources it was allowed to use. The accuracy of the test set was plotted against (A) the number of decision trees in the ensemble of the random forest and (B) the number of samples provided within the training set. The two curves in (B) illustrate the mean and standard deviation of performances across the fivefold cross-validation runs. The ideal training set size is determined at the closest convergence point between the training and the cross-validation scores while minimizing variance. Note, performance plateaus when the training set size exceeds 10 000. [Color figure can be viewed at wileyonlinelibrary.com]

Sixty-seven percent of the decision splits ($n = 47\ 571$) were based on information about the lesion and 33% of the decision splits ($n = 96\ 143$) were made on the basis of patient demographic variables. However, a variable's number of splits within a random forest does not necessarily reflect the importance of the variable due to the structure of decision trees, where there are plenty more nodes in the lower depths. Decisions made further down the tree often translate to less important variables. Thus, we ran two different analyses to interpret how the random forest prioritizes these variables: mean minimal variable depth and permutation importance.

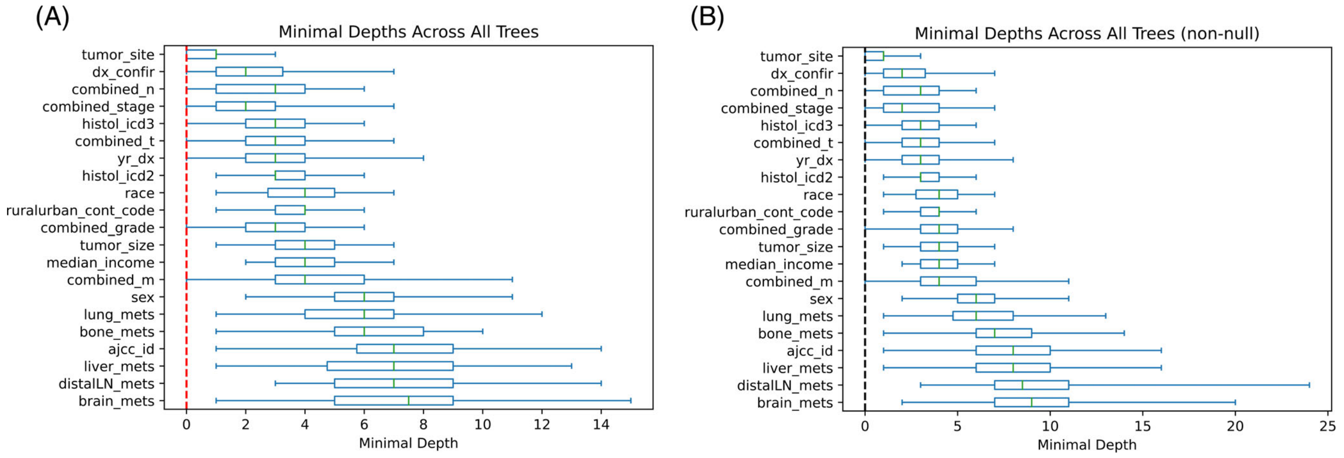

FIGURE 4 Minimal variable depths across all decision trees. The variables are sorted by mean minimal depth from smallest (most important) to largest (least important). (A) illustrates the minimum depths for each of the variables. (B) Depicts the minimum depths for each of the variables after accounting for unknown or null decisions. If the variable is not included within the model, the tree's maximum depth is assigned by default. The green line within each box indicates the median value. The box covers the interquartile interval (Q1–Q3), where 50% of the data is found. [Color figure can be viewed at wileyonlinelibrary.com]

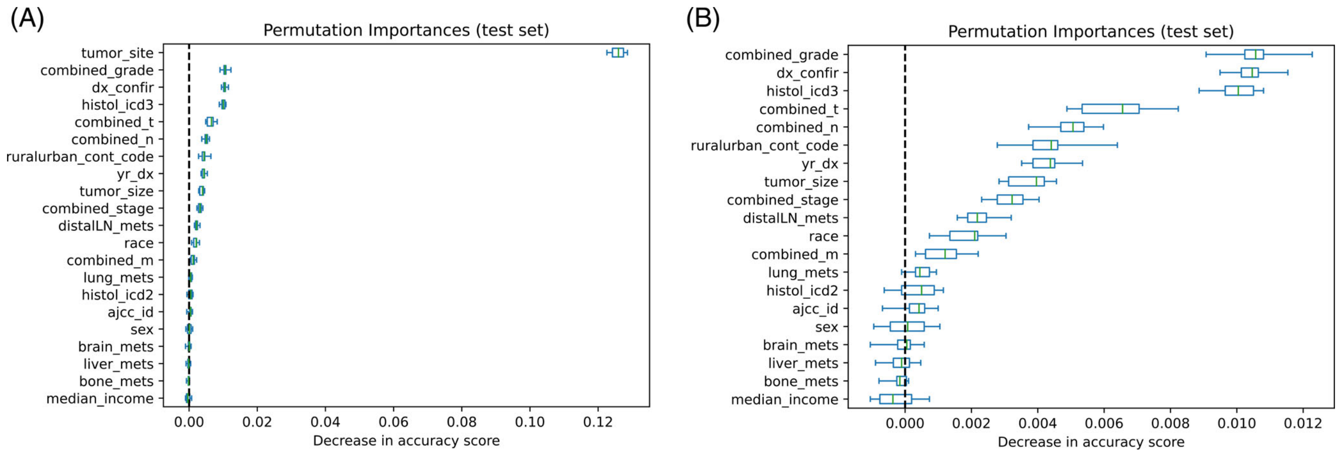

FIGURE 5 Permutation feature importance of patient demographic and cancer characteristics. This figure illustrates how much the model's accuracy score decreases when a particular variable was shuffled. In (A), all of the variables are included. In (B), the tumor site variable is omitted to provide detailed visualization of the permutation importance for the other variables. The box illustrates the Q1–Q3 range. The green line within each box symbolizes the median. The whisker ends indicates the range of scores. [Color figure can be viewed at wileyonlinelibrary.com]

We calculated the mean minimal depth for each variable to determine its relative importance for the random forest model. The minimal depth is the earliest point at which a variable is considered in a decision tree. Variables with lower minimal depths (as opposed to larger minimal depths) are considered to be more important for the decision tree. We averaged the minimal variable depth across all decision trees and then sorted the variables in descending order of mean minimal depth (Figure 4A). Variables that were uninvolved in any given tree were assigned the maximum depth for that tree. To adjust for unknown values within the SEER dataset, we created a secondary graph that selected the minimal depth of the first node that did not use a threshold based on an unknown value (Figure 4B). For example, we did not count queries on whether the sex was "Unknown." Tumor site was the most important variable for the predictive algorithm; more than half of the trees queried tumor site at the root of the tree (first decision made) with an average minimum depth of 0.8 ± 0.81 (mean ± standard deviation). The order of variable importance was preserved whether or not unknown or null thresholds were excluded from the analysis.

We also calculated permutation feature importance to measure the relative weight of each variable within the random forest (Figure 5A).[7] Permutation importance is measured by shuffling the column of a dataset to create synthetic data. The model is then tested on the synthetic data with the expectation that the trained model's performance will suffer. The decrease in accuracy corresponds to the degree of importance for that particular variable. Among all the variables, tumor site permutations (13.2% decrease in the model's accuracy score) had the greatest impact on the model's accuracy rate than any other variable (<2% decrease in the model's accuracy score) (Figure 5B). After tumor site, diagnostic confirmation, histology, T stages, and N stage had the greatest effect on the model's accuracy rate. These findings are relatively consistent with the other study measure of variable importance (minimum variable depth) (Figure 4).

4 | DISCUSSION

In this study, we evaluated the clinical use case of a machine learning model for processing new patient referrals. Specifically, we investigated how well a random forest could predict how likely a patient with head and neck cancer would be recommended surgery at the time a new patient referral is placed. Machine learning models within this domain have the potential of assisting practice coordinators with processing referrals more accurately and expeditiously, thus improving patient outcomes.

Decision tree models are a form of non-parametric supervised learning that uses classification and regression questions to make predictions. These models are particularly effective with multidimensional datasets, like the SEER database. Limitations related to overfitting can be partially mitigated by setting constraints on model parameters and the use of pruning nodes. In addition, ensemble learning methods are often employed to address the trade-off between bias and variance. Random forest is an ensemble learning technique with excellent multi-dimensionality reduction capacity and within-training-set internal validation.[8]

Black-box models are algorithms whose inner workings are difficult to interpret. Although there are concerns surrounding how random forest models do not provide insight into their decision-making process (aka black-box model), we applied several methods to understand our final model's decision-making process. Specifically, we employed two different analyses to reveal the overall importance the model assigns to each variable. We also ran local interpretative model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) techniques to provide a local, interpretable model to explain each individual prediction.

Our results helped us understand which pieces of medical information have the greatest valence in affecting cancer treatment decisions. These findings may help guide clinical practice and workflows at cancer centers, particularly at the time of a new patient referral. This is especially important because referrals are primarily handled by practice coordinators with little-to-no clinical background. Our supervised machine learning algorithm was able to predict a head and neck cancer provider's decision to recommend surgery to a patient with an 81% accuracy rate.

The variable that could best predict a decision to offer oncologic resection was tumor site. Our study findings are consistent with the existing literature. Prior studies suggested a potential relationship between tumor site and treatment modality.[9,10] Johnson and his team report that surgery is frequently selected for oral cavity cancers, whereas radiation is more commonly recommended to patients with pharyngeal and laryngeal cancers.[10] This is not particularly surprising given the clinician's role in weighing treatment efficacy rates against patient morbidity and side effects. Interestingly, all of the cancer characteristics and patient demographic variables were more-or-less similar in its relative importance in determining whether or not a patient was likely to be recommended surgery—with a less 2% or less effect on the model's accuracy rate.

The implications of these findings could be significant. Head and neck cancer practice coordinators may be able to apply a supervised machine learning tool to determine whether the patient should be first seen by a head and neck cancer surgeon as opposed to starting with a medical oncologist or radiation oncologist. This has the potential of saving patients weeks to months of treatment delay as they are referred from one specialist to another. The financial and time cost-savings are also significant. Practice coordinators, nurse navigators, and physician specialists could spend less time directly screening these referrals on their off-time and improve surgical conversion rates in the clinic.

There are important study limitations to note. One of the study's strengths is the use of a large national cancer database rather than a single-institutional cohort; this approach improves the generalizability of the random forest model and decreases the risk of overfitting on a restricted population of patients. However, SEER does not collect data on the entire U.S. population and therefore may not be representative of certain regions of the United States. We also recognize that there are other variables affecting provider treatment recommendations that may not be captured by the SEER database. These include the physical exam, assessment of frailty, patient body cues, and patient-centered discussions that occur on the day of the visit. In addition, we combined the different AJCC editions into one, which could have diluted the relative importance of cancer staging on treatment recommendations. This variability was further compounded by the prevalence of unknown variables in the SEER database and eliminated the training resources the model had access to by half. As such, the individual importance of non-tumor site variables (i.e., race compared to M-stage) should be interpreted with caution.

5 | CONCLUSIONS

Using a supervised learning machine learning model, we trained an ensemble classifier that could predict which factors drive the decision to offer a head and neck cancer patient surgery. The random forest model accuracy rate was 81%. The most important variables were the tumor site (by far the most influential variable), followed by diagnostic confirmation, cancer histology, T stage, and N stage. These results suggest that machine learning algorithms could be a useful adjunct tool for screening new patient referrals, especially since practice coordinators have little-to-no medical background in subspecialty oncology. Faster and more accurate referral processing times could have a meaningful impact on institutional expenditures, personnel costs, and patient outcomes.

CONFLICT OF INTEREST STATEMENT

Max L. Jiam is the Co-Chief Technology Officer of IIAM corporation. Kevin Z. Xin is the Co-Chief Technology Officer and Chief Financial Officer of IIAM corporation. Patrick K. Ha is a consultant for Atos Medical and Checkpoint Surgical. He received educational funding from Stryker, Ethicon, and Medtronic. He receives royalties from Wolters-Kluwer and Wiley. Nicole T. Jiam is the Chief Executive Officer of IIAM corporation and a consultant for Oticon Medical.

DATA AVAILABILITY STATEMENT

The SEER dataset is publicly available via the National Cancer Institute website (https://seer.cancer.gov/). Interested parties may directly email the corresponding author regarding the data analysis and machine learning model discussed in this manuscript. Max L. Jiam, Kevin Z. Xin, Patrick K. Ha, and Nicole T. Jiam had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

ORCID

Nicole T. Jiam https://orcid.org/0000-0001-9838-5065

REFERENCES

1. Khorana AA, Tullio K, Elson P, et al. Time to initial cancer treatment in the United States and association with survival over time: an observational study. *PloS One*. 2019;14(3):e0213209.
2. Cone EB, Marchese M, Paciotti M, et al. Assessment of time-to-treatment initiation and survival in a cohort of patients with common cancers. *JAMA Netw Open*. 2020;3(12):e2030072.
3. Patel MP, Schettini P, O'Leary CP, Bosworth HB, Anderson JB, Shah KP. Closing the referral loop: an analysis of primary care referrals to specialists in a large health system. *J Gen Intern Med*. 2018;33(5):715-721.
4. Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920-1930.
5. Banerjee M, Reynolds E, Andersson HB, Nallamothu BK. Tree-based analysis. *Circ Cardiovasc Qual Outcomes*. 2019;12(5):e004879.
6. Doll KM, Rademaker A, Sosa JA. Practical guide to surgical data sets: Surveillance, Epidemiology, and End Results (SEER) database. *JAMA Surg*. 2018;153(6):588-589.
7. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
8. Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform*. 2013;14(3):315-326.
9. Karpathiou G, Giroult JB, Forest F, et al. Clinical and histologic predictive factors of response to induction chemotherapy in head and neck squamous cell carcinoma. *Am J Clin Pathol*. 2016;146(5):546-553.
10. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. *Nat Rev Dis Primers*. 2020;6(1):92.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jiam ML, Xin KZ, Ha PK, Jiam NT. A supervised machine learning model for identifying predictive factors for recommending head and neck cancer surgery. *Head & Neck*. 2024;46(5):1001-1008. doi:10.1002/hed.27674