ACENET
Microcredential in Advanced Computing
ISP Report

Project title: Using a Machine Learning Strategy to Predict Cancer Patient Outcome based on Clinical and Molecular Biomarkers

Participant name: Tom Belbin

Date: July 31, 2024

Abstract:

I would like to use a machine learning strategy with gene expression and clinical data to predict two critical parameters that would be useful in guiding treatment for head and neck cancer patients. The variables to predict would be: 1) presence of lymph node metastasis at diagnosis, and 2) overall patient survival.

1. Introduction

The decision of which treatment to pursue in head and neck cancer is often made based upon anatomic criteria only, certainly leading to inappropriate therapy in some patients. In order to provide effective treatment strategies, we will need to be able to predict, with high specificity and sensitivity, the tumour potential for metastasis and patient response to therapy. Development of such a prediction tool will advance the field of personalized medicine.

With this in mind, I would like to use a machine learning strategy with gene expression and DNA methylation data to predict two critical parameters that would be useful in guiding treatment. The variables to predict would be: 1) presence of lymph node metastasis at diagnosis, and 2) overall patient survival

2. Background

Head and neck squamous cell carcinomas (HNSCCs) rank among the 10 most common malignancies in men and women worldwide **(Figure 1)**. Overall, 5-year survival rates are

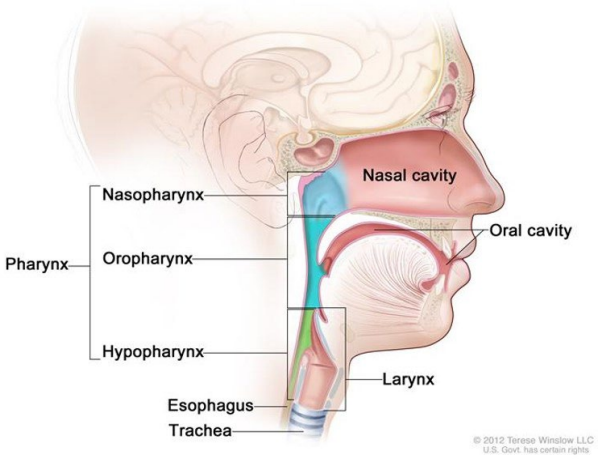

Figure 1. Anatomy of Head and Neck Squamous Cell Carcinoma.

approximately 50%, but there is substantial variability in response to treatment and long-term prognosis that cannot be predicted on the basis of standard histopathology. As a result, more than 4,300 Canadians will develop cancer of the head and neck this year; over 1,600 of them will die from this disease (1). Conventional treatment will usually employ surgery and adjuvant radiation therapy, with or without chemotherapy depending upon pathologic results. Any of these costly therapies can, and do, produce significant morbidities affecting speech, swallowing, and overall quality of life. Despite these interventions, recurrence of the disease is observed in about 50% of patients with high rates of associated mortality.

3. Analysis

The dataset used for this project included 530 head and neck cancer patients from the Cancer Genome Atlas (TCGA) (2). Of these, 11 patients had no clinical data available, so these were excluded from the analysis. In addition, 3 lip cancer samples were removed as these were difficult to categorize, leaving a total of 516 patients in the cohort. For each patient, clinical data was available on stage of disease, including lymph node metastasis (N stage), as well as overall patient survival. There was a total of 93 additional clinical variables, such as the primary site of the tumour. A summary table of the HNSCC patient cohort is included as **Supplementary Table 1**. The relevant features for training are "node_status" (Negative or Positive), "vital_status" (Dead or Alive), and "days_to_event" (the length of time from diagnosis to death or last followup). For molecular data, there were a total of 482,421 DNA methylation measurements and measurements of expression for 20,531 genes. For the purposes of this project, I chose to focus on the expression of 150 cancer-related genes identified by myself and others as being associated with poor prognosis in this disease (3). From this dataset, we identified an additional 8 patients for which no gene expression data was available, bringing the final cohort size to 508 patients. From this dataset, we plan to apply a random forest algorithm to test the hypothesis that the expression of these cancer-related genes can predict the presence of lymph node metastasis at diagnosis. In addition, we utilized Kaplan Meier survival analysis to identify features that might be significantly associated with patient survival time.

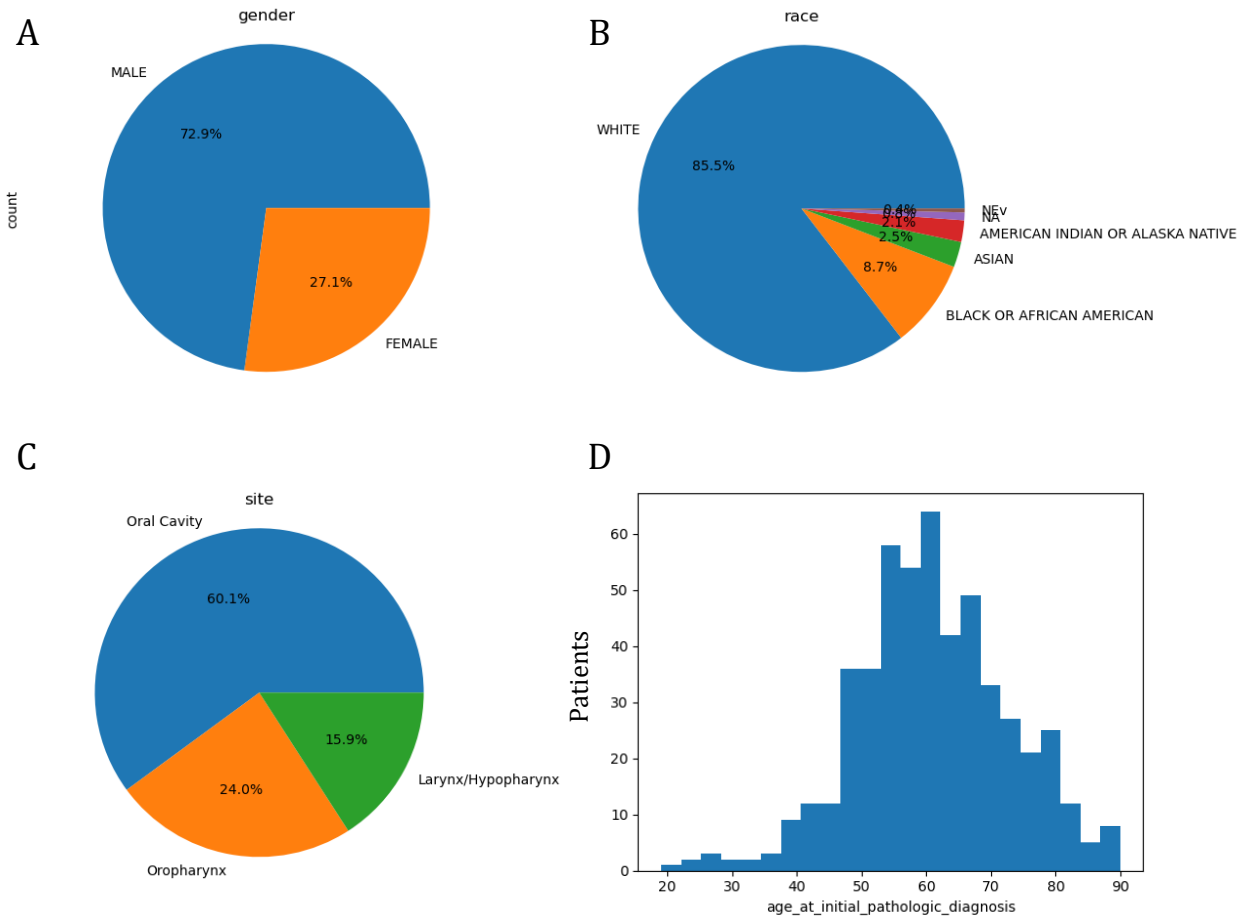

Figure 2. Breakdown of TCGA HNSCC patient population by **(A)** Sex, **(B)** Race, **(C)** Anatomic site, and **(D)** Patient age at diagnosis.

4. Results.

Initial exploratory analysis of the patient cohort indicated that the population was overwhelmingly a male Caucasian patient population **(Figure 2A-B)**. Most of the patients in this cohort (60%) had primary tumours that originated in the oral cavity, as opposed to the oropharynx or larynx **(Figure 2C)**. While there was a wide distribution of patient ages ranging from 20-90 years of age, most patients were between the ages of 50-70 years old, with a median age of 62. One of the most important features of HNSCC is the presence of tumour cells in the regional lymph nodes of the head and neck **(Figure 3A)**. In our patient cohort, almost half of our patients (46%) were positive for lymph node metastasis (node_status=1) **(Figure 3B)**. Unfortunately, approximately 100 patients (19.4%) had an unknown nodal status at the time of this analysis so these were removed prior to applying the machine learning algorithm.

In order to predict the presence of lymph node metastasis at diagnosis, the patient dataset was split into training set (70%) and test (30%) using the Scikit-learn package in Python. I applied a random forest machine learning algorithm using 100 trees in order to train a binary classifier on patient nodal status (positive versus negative). Unfortunately, the resulting classifier only performed with an accuracy of 62% (precision: 63%, recall: 82%) **(Figure 3C)**. The results of a

high recall were encouraging, as this was an indication that the algorithm was able to successfully detect most of the node-positive cancers in the cohort, albeit with a high false positive rate. I tried to represent the balance between precision and recall using a receiver operator (ROC) curve **(Figure 3D)**. As shown in the final presentation, parameter tuning by increasing the number of trees used to train the random forest classifier did not increase the accuracy of the model.

In addition to predicting nodal metastasis at diagnosis, my project utilized survival analysis using the lifelines module in Python to look features associated with overall patient survival in

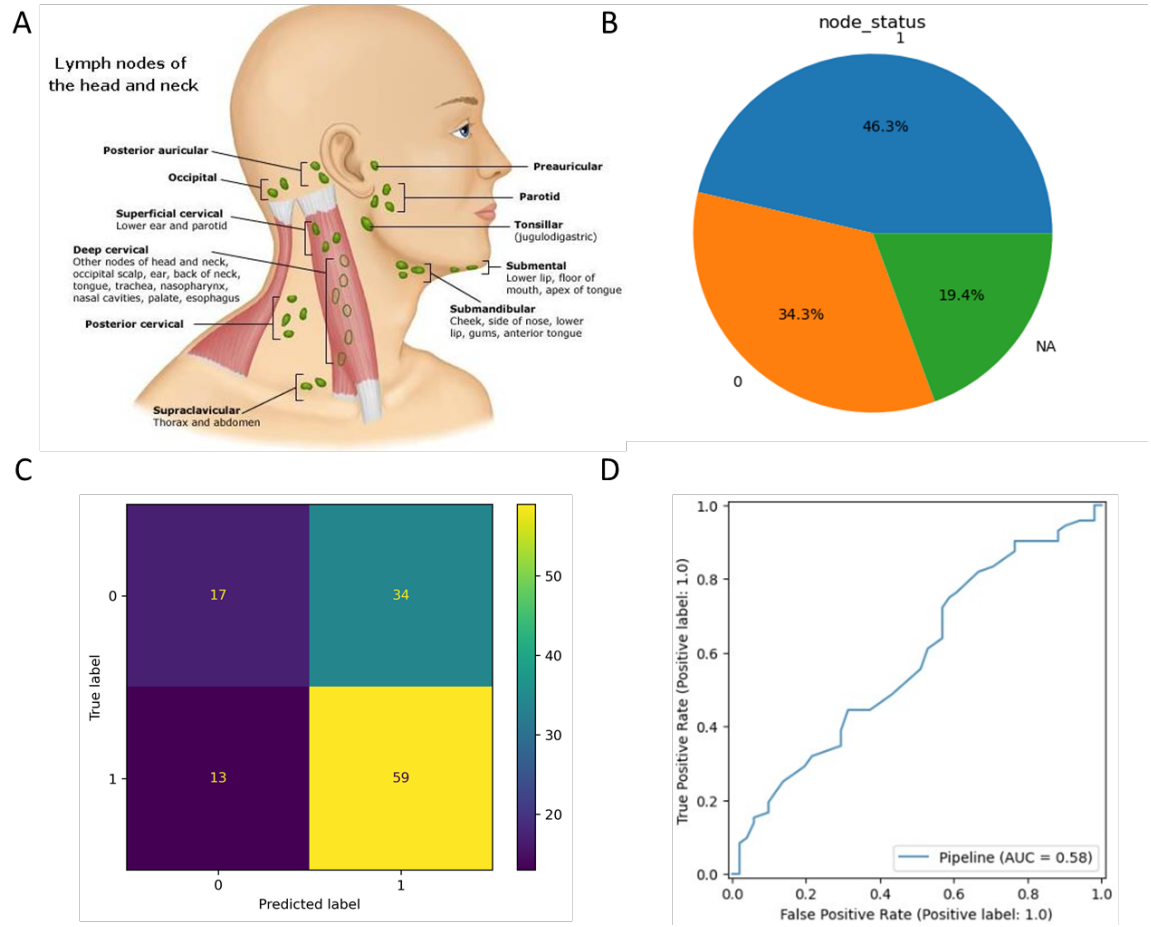

Figure 3. (A) Overview of lymph nodes of the head and neck. **(B)** Breakdown of TCGA HNSCC patient population by nodal status. **(C)** Confusion matrix showing the performance of the random forest classifier at predicting node status in the test set of HNSCC patients. **(D)** Performance of the same random classifier using a ROC curve.

this cohort. A Kaplan-Meier (KM) plot summarizing patient survival for this cohort is shown in **Figure 4A**. Plotting survival data over time also allows us to stratify the population by specific parameters and compare how the survival compares across different patient groups. I utilized Cox proportional hazards modelling to identify features that were significantly associated with

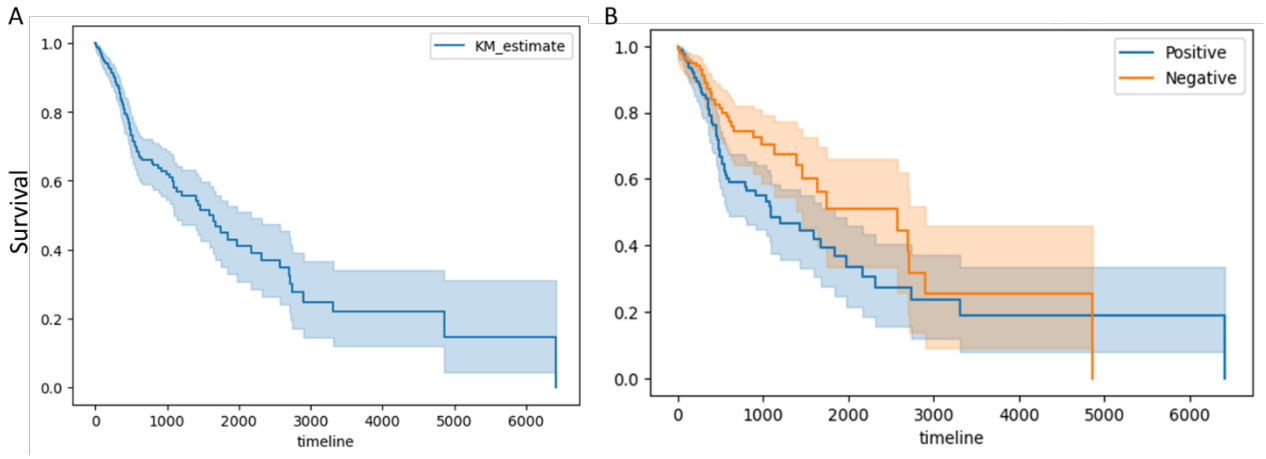

Figure 4. Kaplan Meier plot of overall survival over the timeline of days for (A) the entire cohort and (B) when patients are stratified by positivity for nodal metastasis. Significance of the difference in survival was determined by Log rank statistic.

overall patient survival in this disease. Based on Cox proportional hazards modelling, I found that node status, patient age, and patient history of a prior malignancy were all significantly associated with decreased overall survival. An example of the KM plot stratifying patients by nodal status in shown in **Figure 4B**, and shows a significantly decreased survival in patients with node-positive disease.

5. Discussion

The attempts to build a classifier to predict nodal metastasis at diagnosis were somewhat disappointing, with an accuracy of 62%. However, the fact that the classifier had a relatively high recall of 82% were encouraging, as this was an indication that the algorithm was able to successfully detect most of the node-positive cancers in the cohort. This would be important in the design of a screening tool that would pick up a possible node-positive case that other screening tools might miss. Another interesting hypothesis is that the "false positives" identified in this analysis might not actually be false-positives, but might instead represent a micro-metastasis that was not detectable by the pathologist. Indeed, it might be interesting to follow eventual patient outcomes for patients who were pathologically node-negative but were predicted by the classifier to be node-positive (n=34 in the confusion matrix). Do these patients eventually develop lymph node metastasis and have a worse outcome than those predicted to be node-negative (n=17 in the confusion matrix)?

The KM plots and Cox proportional hazards model demonstrated that it was possible to identify parameters associated with decreased overall patient survival. Not surprisingly, node-positive disease was significantly associated with a decreased overall survival. Similarly, patient age was also significantly associated with a decreased overall survival. This is not surprising as older patients can have a variety of comorbid conditions that can significantly affect their survival. It is important to keep in mind that overall survival is what is being measured in this analysis, and not death due to cancer, which is a more specific variable. It was interesting that a history of prior malignancy was also significantly associated with a decreased overall survival. This might

suggest an existing susceptibility to cancer development, or a patient's reduced ability to successfully respond to treatment due to prior susceptibility, or prior cancer treatments in earlier years.

There were some challenges and limitations in this project. Although the anatomic site of the tumour was included, it would also have been advantageous to have anatomic subsite of the tumour for each patient. That would have revealed the proximity of the tumour to regional lymph nodes, and this proximity would likely be a significant contributor to nodal metastasis. It was also unfortunate that nodal metastasis data was missing for 100 patients, which significantly reduced the size of the training and test sets.

Conclusion

In conclusion, the training of a random forest classifier was able to recall node-positive patients with an 82% recall, suggesting it might be useful as a screening tool. In addition, KM analysis revealed that nodal metastasis, patient age and history of prior malignancy were significantly associated with decreased patient survival in this cohort. Prediction tools such as these could be useful in guiding treatment for HNSCC patients while avoiding unnecessary surgeries if nodal disease is not present. In future work, I would like to expand the list of features utilized in the dataset, and include more specific parameters such as anatomic subsite to improve the performance of the classifier. Also, I would try to expand the number of patients used in the cohort, and include a more diverse population in order to make any classifier more applicable to a wide range of patient populations.

References

1. Statistics., C.C.S.s.A.C.o.C. *Canadian Cancer Statistics* 2016.
2. The Cancer Genome Atlas Network, Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517: 576–582 (2015).
3. Hier, J *et al.* Portrait of DNA methylated genes predictive of poor prognosis in head and neck cancer and the implication for targeted therapy. *Sci Rep,* 11(1): 10012 (2021).

Supplementary Materials

TCGA dataset - https://portal.gdc.cancer.gov/projects/TCGA-HNSC

Github - https://github.com/TomBelbin/ACENET_ISP

Table 1. HNSCC Patient charateristics by tumor site						
	Oral Cavity N=310		Oropharynx N=82		Larynx N=124	
	N	%	N	%	N	%
Gender						
Male	206	66.5	69	84.1	101	81.5
Female	104	33.5	13	15.9	23	18.5
Race						
White	266	85.8	76	92.7	99	79.8
Black or African American	20	6.5	6	7.3	19	15.3
Asian	10	3.2	0	0	1	0.8
American Indian or Alaska Native	1	0.3	0	0	1	0.8
Information not available	13	4.2	0	0	4	3.2
Ethnicity						
Hispanic/Latino	16	5.2	3	3.7	5	4
Non-Hispanic/Latino	269	86.8	76	92.7	109	87.9
Information not available	25	8.1	3	3.7	10	8.1
Smoking						
Ever smoker	215	69.4	56	68.3	113	91.1
Lifelong non-smoker	86	27.7	25	30.5	8	6.5
Information not available	9	2.9	1	1.2	3	2.4
HPV Status						
HPV +	31	10	56	68.3	11	8.9
HPV -	278	89.7	26	31.7	112	90.3
Indeterminate	1	0.3	0	0	1	0.8
Vital Status						
Alive	197	63.5	69	84.1	82	66.1
Deceased	113	36.5	13	15.9	42	33.9
Nodal Status						
Positive	147	47.4	34	41.5	58	46.8
Negative	119	38.4	16	19.5	42	33.9
Information not available	44	14.2	32	39	24	19.4
Pathologic Tumor Stage						
Stage I	19	6.1	4	4.9	2	1.6
Stage II	56	18.1	10	12.2	12	9.7
Stage III	52	16.8	9	11.0	14	11.3
Stage IV	160	51.6	28	34.1	79	63.7
Information not available	23	7.4	31	37.8	17	13.7