Rasheed Alabi

Machine learning for personalized prognostication of tongue cancer

ACTA WASAENSIA 457

Vaasan yliopisto
UNIVERSITY OF VAASA

ACADEMIC DISSERTATION

To be presented, with the permission of the Board of the School of Technology and Innovations of the University of Vaasa, for public examination on the 15th of April, 2021, at noon.

Reviewers

Professor Jyrki Tapio Heinämäki
Pharmaceutical Nanotechnology
Programme Director, Faculty of Medicine
University of Tartu
Ülikooli 18
50090 TARTU
ESTONIA

Associate Professor Fabricio Passador-Santos
Faculdade São Leopoldo Mandic, Campinas
Department of Oral Pathology
Campinas - SP, 13045-755,
Brazil

Julkaisija	**Julkaisupäivämäärä**	
Vaasan yliopisto	Huhtikuu 2021	
Tekijä(t) Rasheed Omobolaji Alabi	**Julkaisun tyyppi** Artikkeliväitöskirja	
ORCID tunniste https://orcid.org/0000-0001-7655-5924	**Julkaisusarjan nimi, osan numero** Acta Wasaensia, 457	
Yhteystiedot Vaasan yliopisto Teknologian ja Innovaatiojohtamisen akateeminen yksikkö Tietoliikennetekniikka PL 700 FI-65101 VAASA	**ISBN** 978-952-476-944-0 (painettu) 978-952-476-945-7 (verkkoaineisto)	
	http://urn.fi/URN:ISBN:978-952-476-945-7	
	ISSN 0355-2667 (Acta Wasaensia 457, painettu) 2323-9123 (Acta Wasaensia 457, verkkoaineisto)	
	Sivumäärä 213	**Kieli** Suomi
Julkaisun nimike Koneoppimismenetelmä kielisyövän henkilökohtaisen ennusteen arviointiin		

Tiivistelmä

Kielisyöpä yleisin pään ja kaulan alueen pahanlaatuisista kasvaimista. Yhdysvaltain syöpäkomitean (AJCC) syöpäkasvainten luokitusjärjestelmä (TNM) on perinteisesti osoittautunut objektiiviseksi ja yleismaailmalliseksi työvälineeksi arvioida syöpäpotilaiden ennustetta. TNM-luokitusjärjestelmä on kuitenkin myös kritisoitu, koska sen ennustekyky yksittäisten potilaiden kohdalla on osoittautunut rajoitetuksi. Lisäksi varhaisvaiheen kielisyövän suhteen se ei ole osoittanut vakuuttavaa ennustekykyä. Tätä tarkoitusta varten työkalu, joka tarkastelee samanaikaisesti useita ennustekijöitä potilaan tilan ennustamiseksi tarkasti, olisi hyödyllinen tehokkaassa syövän hoidossa – ehkäisemään tehottoman hoitomuodon valintaa ja tarpeetonta ylihoitamista.

Tässä kansainvälisessä tutkimusyhteistyössä käytimme koneoppimistekniikoita, joissa otettiin huomioon TNM-luokituksen puutteet arvioitaessa ja ennustettaessa kielisyöpäpotilaiden tuloksia, kuten paikallisten ja alueellisten uusiutumisten esiintymistä sekä eloonjäämistä. Laajoja potilasaineistoja, joita käytettiin analyyseissä, saatiin viidestä opetussairaalasta Suomesta, A.C. Camargon syöpäkeskuksesta, Sao Paulosta, Brasiliasta ja Yhdysvaltain kansallisen terveysinstituutin (NIH) seuranta-, epidemiologia- ja lopputulokset (SEER) -ohjelmasta. Lisäksi arvioimme syöttöparametrien ennusteellista merkitystä käyttäen koneoppimistekniikoita. Useita eri koneoppimisalgoritmeja verrattiin parhaiten menestyvään malliin ja integroitiin sitten web-pohjaiseksi yksilöllisen hoidon työkaluksi. Vertailimme myös koneoppimistekniikoiden suorituskykyä nomogrammi-kaavioiden analysoimiseksi kielisyöpäpotilaiden ennusteen arvioinnissa (kokonaisselviytyminen). Lisäksi pohdittiin niitä eettisiä haasteita, jotka voivat vaikuttaa koneoppimismallien käyttämiseen päivittäisessä kliinisessä toiminnassa. Edellisten lisäksi ehdotettiin toimintaohjeita koneoppimisen sujuvaksi integroimiseksi päivittäisiin kliinisiin käytäntöihin.

Asiasanat

Koneoppiminen, kielisyöpä, ennuste, diagnoosi

Publisher
Vaasan yliopisto

Date of publication
April 2021

Author(s)
Rasheed Omobolaji Alabi

ORCID identifier
https://orcid.org/0000-0001-7655-5924

Contact information
University of Vaasa
School of Technology and Innovations
Telecommunications Engineering
P.O. Box 700
FI-65101 Vaasa
Finland

Type of publication
Doctoral thesis by publication

Name and number of series
Acta Wasaensia, 457

ISBN
978-952-476-944-0 (print)
978-952-476-945-7 (online)

http://urn.fi/URN:ISBN:978-952-476-945-7

ISSN
0355-2667 (Acta Wasaensia 457, print)
2323-9123 (Acta Wasaensia 457, online)

Number of pages
213

Language
English

Title of publication
Machine learning for personalized prognostication of tongue cancer

Abstract
Tongue cancer constitutes the majority of the malignancies of the head and neck region. Traditionally, the staging system of the American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) has been shown to be an objective and universal tool for predicting the prognosis for cancer patients. However, the TNM staging system has been criticized because it showed limited prognostic ability for individual patients. In addition, for early-stage tongue cancer, it has not shown convincing prognostic capabilities. To this end, a tool that considers many prognostic factors together to accurately predict patients' outcomes would be pertinent for effective cancer management – prevention of ineffective treatment and avoidance of unnecessary overtreatment.

In this international collaborative study, we applied machine learning techniques that considered the shortcomings of the TNM staging to estimate and predict tongue cancer patients' outcomes such as locoregional recurrences and overall survival. Large patient cohorts from five teaching hospitals in Finland, A.C Camargo Cancer Center, Sao Paulo, Brazil, and the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institute of Health (NIH), United States were used in the analyses. Moreover, we evaluated the prognostic significance of the input parameters using machine learning techniques. Several machine learning algorithms were compared for the best performing model and then integrated as a web-based tool for personalized medicine. Furthermore, we compared the performance of machine learning techniques to nomograms in the prognostication of outcomes (overall survival) for tongue cancer patients. Ethical challenges that can affect the implementation of machine learning models for daily clinical practice were highlighted and a framework for smooth integration of machine learning for daily clinical practices was proposed.

Keywords
Machine learning, Tongue cancer, Prognostication, Prediction

"There are six stages to knowledge: Firstly: Asking questions in a good manner. Secondly: Remaining quiet and listening attentively. Thirdly: Understanding well. Fourthly: Memorising. Fifthly: Teaching. Sixthly- and it is its fruit: Acting upon the knowledge and keeping to its limits."

--Ibn Qayyim Al-Jawziyya.

"If you know and believe in yourself, then you are ready to climb the ladder of success to greatness. The believe in yourself is the main ingredient to success and greatness"

--Alabi Rasheed

ACKNOWLEDGEMENT

This study was conducted during the years 2018-2020 at the School of Technology and Innovations, University of Vaasa; and was supported through a scholarship grant by the same department at the University of Vaasa. In addition, I received financial support for personal development by the Oral Cancer Research group, Institute of Biomedicine, Pathology, University of Turku, Turku, Finland.

I wish to thank the Dean of the School of Technology and Innovations, Dr. Harry Linnarinne for providing open scholarship opportunities for some of the PhD students. He has also provide an open and conducive atmosphere at the School of Technology and Innovations. Similarly, I thank the Head of the Department of Computer Science, Professor Tero Vartiainen, for the coordination of the activities of the team and providing an excellent research environment at the department during these years.

I am most grateful to my brilliant mentor and supervisor, Professor Mohammed Elmusrati, who has offered immense guidance, tutoring, support, endless enthusiasm and valuable comments since I started at the University of Vaasa as a Masters (MSc) Student. For every discussion, both formal and informal, I am indeed grateful for your time. You hold an important role in my life. I will never forget your positive contribution. I am fortunate to have met you, Sir.

Additionally, I appreciate my kind, gentle and approachable co-supervisor, Professor Timo Mantere. Your lectures during my MSc studies formed the foundation for my interest in research. Your attention and comments have improved me as a researcher. I sincerely appreciate your valuable comments and contributions to improve this dissertation.

My profound appreciation goes to my instructor, Docent Alhadi Almangush. I am privileged to have the opportunity to work with you. You taught me everything that I need to know about the art of research, most importantly, to be an independent researcher. Your continuous guidance, support, motivation, enthusiasm and criticism since I started with my MSc thesis make my research infinitely more valuable. Discussions with you regarding my research activities open up doors for improvements. I have learned greatly from your attentiveness to the slightest detail. You have influenced my life positively. I am most grateful.

I am grateful to all of my co-authors from the multi-institution (University Teaching Hospital in Finland) and from A.C. Camargo Hospital, Sao Paulo, Brazil. My sincere gratitude goes to Professor Antti A. Mäkitie and Professor Ilmo Leivo

for their continuous efforts to provide guidance and improvements to all the manuscripts. Your support and constructive discussions have strongly improved my research.

I wish to warmly show my profound appreciation to the official reviewers, Professor Jyrki Tapio Heinämäki and Associate Prof. Fabricio Passador-Santos who have worked tirelessly to review the manuscript and provided valuable suggestions and constructive comments to improve this thesis.

I would like to warmly acknowledge the appropriate agencies for granting the permission to use the data in this study.

I thank Deborah Kaska for the language review of this thesis.

I wish to express my warm appreciation and gratitude to all my friends who have helped during the period of my PhD studies. I appreciate your support and guidance. Your friendship meant a lot to me.

My deepest, heartfelt, and sincere gratitude to my wife (Ummu-Khayr: Atunrase Mistura Omolara) and my son (Mu'adh). Your understanding and support is unrivaled. To my wife, you have continuously offered me moral and emotional support. You are indeed a rare gem. I thank you with all my heart. This is only possible because you believe in me. I honestly find solace and tranquility in you. To my lovely son, thank you for your understanding. It has not been an easy experience travelling between Helsinki and Vaasa on a weekly basis. Thank you son for your understanding. I love you so much.

Additionally, I am forever indebted to my mother, who remain my source of joy. Thank you for your words of encouragement and prayers.

Lastly but the ultimate most, all thanks to Allah, the Lord of Incomparable Majesty. I am grateful for making this a reality.

Vaasa, April 15, 2021

Alabi Rasheed Omobolaji

Contents

Figures

Tables

Abbreviations

AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
ANN	Artificial Neural Network
AUC	Area Under Receiving Operating Characteristics Curve
BDT	Boosted Decision Tree
CNN	Convolution Neural Network
DF	Decision Forest
eHealth	Electronic Health
FP	False Positives
FN	False Negatives
IoT	Internet of Things
LHR	Lymphocytic Host Response
LR	Logistic Regression
LRNN	Layer Recurrent Neural Network
LSTM	Long-Short Term Memory
mHealth	Mobile Health
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MLT	Machine Learning Techniques
MSE	Mean Squared Error
NB	Naïve Bayes
NCI	National Cancer Institute
NIH	National Institute of Health
OTSCC	Oral Tongue Squamous Cell Carcinoma

OS	Overall Survival
PNI	Perineural Invasion
ReLU	Rectifier Linear Unit
RMSE	Root Mean Squared Error
SEER	Surveillance Epidemiology and End Results
SVM	Support Vector Machine
TN	True Negatives
TNM	Tumor Nodal Metastasis
TP	True Positives
TSCC	Tongue Squamous Cell Carcinoma
WPOI	Worst Pattern of Invasion
WHO	World Health Organization

Formulas

(1) $$sig(t) = \frac{1}{1+e^{-t}}$$ (6)

(2) $$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + ... + \beta_p x_p^{(i)}$$ (7)

(3) $$P(y^{(i)} = 1) = \frac{1}{1+e^{\left(-(\beta_0+\beta_1 x_1^{(i)}+...+\beta_p x_p^{(i)})\right)}}$$ (8)

(4) $$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$ (9)

(5) $$\frac{P(y=1)}{1-P(y=1)} = odds = e^{(\beta_0+\beta_1 x_1+...+\beta_p x_p)}$$ (10)

(6) $$\frac{odds_{x_j+1}}{odds} = \frac{e^{(\beta_0+\beta_1 x_1+...+\beta_j(x_j+1)+...+\beta_p x_p)}}{e^{(\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p)}}$$ (11)

(7) $$\frac{odds_{x_j+1}}{odds} = e^{(\beta_j(x_j+1) - \beta_j x_j)} = e^{(\beta_j)}$$ (12)

(8) $$\frac{d\sigma}{d\alpha} = \sigma(1-\sigma)$$ (13)

(9) $$p(z|w) = \prod_{t=1}^{T} y_t^{z_t} \{1-y_t\}^{1-z_t}$$ (14)

(10) $$E_{ntropy}(w) = -\ln p(z \mid w) = -\sum_{t=1}^{T}\left\{z_t \ln y_t + (1-z_t)\ln\left(1-y_t\right)\right\}$$ (15)

(11) $$\nabla E_{ntropy}(w) = \sum_{t=1}^{T}(y_t - z_t)\phi_t$$ (16)

(12) $$y(x,w) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right)$$ (17)

(13) $$a_j = \sum_{i=1}^{T} w_{ji}^{(1)} x_i + w_{j0}^{(1)}$$ (18)

(14) $$Z_j = h(a_j)$$ (19)

(15) $$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)}$$ (20)

(16) $$y_k = \sigma(a_k)$$ (21)

(17) $$y_k(x,w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{T} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$ (22)

(18) $$a_j = \sum_{i=0}^{T} w_{ji}^{(1)} x_i$$ (23)

(19) $$Y_k(x,w) = \sigma\left(\sum_{j=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{T} w_{ji}^{(i)} x_i\right)\right)$$ (24)

(20) $$y = \sigma(a) \equiv \frac{1}{1+e^{(-a)}}$$ (25)

(21) $$p(t \mid x,w) = y(x,w)^t \left\{1 - y\left(x,w\right)\right\}^{1-t}$$ (26)

(22) $$p(t \mid x,w) = \prod_{k=1}^{K} y_k\left(x,w\right)^{t_k} \left[1 - y_k\left(x,w\right)\right]^{1-t_k}$$ (28)

(23) $$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk} \ln y_{nk} + \left(1-t_{nk}\right)\ln\left(1-y_{nk}\right)\right\}$$ (29)

(24) $$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \ln y_k(x_n, w)$$ (30)

(25) $$W^{(\tau+1)} = W^{(\tau)} + \Delta W^{(\tau)}.$$

(26) $$W^{(\tau+1)} = W^{(\tau)} - \eta \nabla E(W^{(\tau)})$$ (27)

(27) $$E(w) = \sum_{n=1}^{N} E_n(w)$$ (28)

(28) $$W^{(\tau+1)} = W^{(\tau)} - \eta \nabla E_n(W^{(\tau)})$$ (29)

(29) $$y_k = \sum_i w_{ki} x_i$$ (30)

(30) $$E_n = \frac{1}{2}\sum_k \left(y_{nk} - t_{nk}\right)$$ (31)

(31) $$\frac{\partial E_n}{\partial w_{ji}} = \left(y_{nj} - t_{nj}\right) x_{ni}$$ (33)

(32) $$a_j = \sum_i w_{ji} z_i$$ (35)

(33) $$\delta_k = y_k - t_k$$ (38)

(34) $$\delta_j = h'\left(a_j\right)\sum_k w_{kj}\delta_k$$ (45)

(35) $$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$ (46)

(36) $$\frac{\partial E}{\partial w_{ji}} = \sum_n \frac{\partial E_n}{\partial w_{ji}}$$ (48)

(37) $$f(x)=\begin{Bmatrix} ax & x\geq 0 \\ 0 & x<0 \end{Bmatrix}$$ (54)

(38) $$y(t)=\int_{-\infty}^{\infty} x(\lambda)h(t-\lambda)d\lambda=\int_{-\infty}^{\infty} x(t-\lambda)h(\lambda)d\lambda$$ (55)

(39) $$\hat{y}[k]=\sum_{m=-\infty}^{\infty}\hat{h}[m]\hat{x}[k-m]=\sum_{m=-\infty}^{\infty}\hat{h}[k-m]\hat{x}[m]$$ (56)

(40) $$\hat{y}[k_1k_2]=\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\hat{h}[m,n]\hat{x}[k_1-m,k_2-n]=\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\hat{h}[k_1-m,k_2-n]\hat{x}[m,n]$$

(41) $$y(x)=W^T\,\emptyset\,(x)+b$$ (56)

(42) $$y\,(x_n)\;>0; for\; t_n=\;+1\;\text{And}\;\; y\,(x_n)\;<0; for\; t_n=\;-1$$ (58)

(43) $$\arg\max_{w,b}\left\{\frac{1}{\|w\|}\min_{n}\left[t_n\left(w^T\phi(x_n)+b\right]\right]\right\}$$ (60)

(44) $$\arg\min_{w,b}\frac{1}{2}\|w\|^2$$ (66)

(45) $$L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{n=1}^{N}a_n\left\{t_n(w^T\phi(x_n)+b)-1\right\}$$ (67)

(46) $$P(C_i|x_1x_2,...,x_M),\;\forall_i=1,...N$$ (70)

(47) $$P(C_i\mid x_1,x_2,...x_M)=\frac{P(x_1,x_2,...x_M\,|C_i)P(C_i)}{P(x_1,x_2,...,x_M)},\forall_i=1,...N$$ (77)

(48) $$P(C_i\mid x_1,x_2,...x_M)=\frac{P(x_1,x_2,...x_M\,|C_i)P(C_i)}{P(x_1)P(x_2)...P(x_M)},\forall_i=1,...N$$ (78)

(49) $$\hat{P}(C_i\mid x,m)=p_m^i=\frac{N_m^i}{N_m}$$ (78)

(50) $$I_m=-\sum_{i=1}^{K}p_m^i\log_2(p_m^i)$$ (78)

(51) $$\hat{f}_{boosting\,(adabost)}(y_{new})=sign\left(\sum_{b=1}^{b_{stop}}\alpha_b\hat{h}^{[b]}(y_{new})\right)$$ (78)

(52) $$\text{Accuracy}\;=\;\frac{TP+\;TN}{TP+FP+FN+TN}$$ (78)

(53) $$Precision=\frac{\text{TP}}{\text{TP + FP}}$$ (78)

(54) $$\text{Recall or sensitivity}=\frac{TP}{TP+FN}$$ (78)

(55) $\text{Specificity} = \frac{\text{TN}}{TN+FP}$ (78)

(56) $\text{F1 score} = \frac{2\,(Precision * Recall)}{(Precision + Recall)}$ (79)

(57) $\text{MSE} = \frac{1}{n}\,\Sigma\,(\text{y} - y_i)$ (78)

(58) $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y-y_i)^2}{N}}$ (79)

(59) $\text{MAE} = \frac{1}{n}\sum\left|(y-y_i)\right|$ (79)

(60) $R^2 = 1 - \frac{MSE(\text{mod}\,el)}{MSE(baseline)}$ (79)

(61) $\text{Adjusted R}^2 = R^2_{adjusted} = \left[\left(\frac{n-1}{n-k-1}\right)\times\left(1-R^2\right)\right]$ (80)

Publications

This doctoral thesis is based on the following peer-reviewed publications, which were subsequently referred to in the text by their Roman numerals (I-V).

(I) **Alabi**, R.O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L.O., Haglund, C., Coletta, R.D., Mäkitie, A.A., Salo, T., Ilmo, L., Almangush, A. Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. ***Virchows Arch*** 475, 489–497 (2019).

(II) **Alabi**, R.O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L.O., Haglund, C., Coletta, R.D., Mäkitie, A.A., Salo, T., Ilmo, L., Almangush, A. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. ***International Journal of Medical Informatics*** 136, 1-8 (2020).

(III) **Alabi**, R.O., Mäkitie, A.A., Pirinen, M., Elmusrati, M., Ilmo, L., Almangush, A. Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. ***International Journal of Medical Informatics*** 145, 1-9 (2021).

(IV) **Alabi**, R.O., Vartiainen, T., Elmusrati, M. Machine learning in oral tongue cancer: Addressing ethical challenges. Proceedings of the Conference on Technology Ethics (October 2020). ***CEUR-Workshop Proceedings*** 2737, 1-22 (2020).

(V) **Alabi**, R.O., Youssef, O., Pirinen, M., Elmusrati, M., Mäkitie, A.A., Ilmo, L., Almangush, A. Machine learning for oral squamous cell carcinoma: current status, clinical concerns and prospect for the future. ***Artificial Intelligence in Medicine,*** (2021). **Under review.**

Author's contribution

Publication I: "Machine learning application for prediction of locoregional recurrences in early oral tongue cancer"

The author conducted the machine learning analysis and evaluate the performance of the artificial neural network for the prediction of locoregional recurrences. Additionally, the web-based prognostic tool was designed and developed by the author. The author wrote the manuscript. The author's supervisor and instructor assisted the design of the study.

Publication II: "Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer"

The author had the main responsibility for the design, experiments and the preparation of the article. The author compared various machine learning algorithms and evaluate the top-performing algorithm to predict locoregional recurrences in early-stage oral cancer. The author generated the final model for predictions and participated in external evaluation of the developed model.

Publication III: "Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer"

In particular, the author obtained the permission to use the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institute of Health (NIH), United States. The author extracted the data used for this publication. The author had the main responsibility for the design and selection of various algorithms and nomogram for comparison. The author did the experiments and the preparation of the article.

Publication IV: "Machine learning for prognosis of oral cancer: What are the ethical challenges? Conference on Technology"

The author had the main responsibility for the design, experiments and the preparation of the article. The author performed the systematic review of the articles. The author prepared the article.

Publication V: “Machine learning for oral squamous cell carcinoma: current status, clinical concerns and prospect for the future”

The author had the main responsibility for the design, experiments and the preparation of the article. The author wrote the article and conducted the research on the clinical concerns of the machine learning model in actual daily clinical practices.

1 INTRODUCTION

Cancer is a dreadful disease that is capable of causing significant devastation in the life of individuals diagnosed with it. Cancer patients and their families are gripped by traumatic and emotionally overwhelming experiences due to the influence of this serious disease. In addition, the fact that it is the second leading cause of death globally makes it a source of great concern to the patients and their respective families. Globally, every sixth death was reported to be due to various forms of cancer (Roser & Ritchie, 2019). In 2018, an estimated of 9.6 million people were reported to have died from cancer worldwide (World Health Organization, 2018).

Cancer is characterized by abnormal cellular growth where normal cells disregard the regular pattern of tissue growth and differentiation, which is important for maintaining tissue physiology, function, and homeostasis (Jaiswal, 2018). In other words, these cancerous cells make more copies of themselves (Weinberg, 2014). Several terms have been used to depict this condition. These include malignant tumors and neoplasms (World Health Organization, 2018). However, the term cancer appeared as the most widely used.

This abnormal cellular growth can affect any part of the body (World Health Organization, 2018). These include lung, breast, colorectal, skin, and head and neck cancer to mention a few. Head and neck cancer are further categorized in accordance with the area of the head or neck where the cancerous growth begins. These can be the oral cavity, pharynx, larynx, paranasal sinuses and nasal cavity, or salivary glands as shown in Figure 1 (National Cancer Institute, 2017).

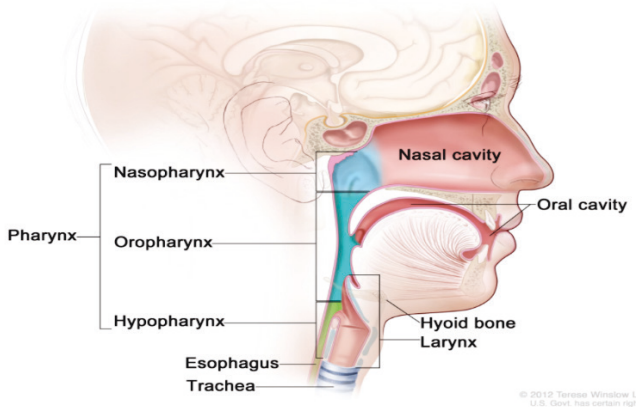

Figure 1. The head and neck cancer region (The Anatomy of the pharynx. National Cancer Institute, 2017, Credit: Terese Winslow).

The oral cavity represents the most common subtype of head and neck cancer. Globally, it is the eighth most common cancer (Ng et al., 2017) with a <60% chance of surviving above 5 years (Amit et al., 2013; R. Siegel et al., 2014). Thus, it represents a major threat to patients' health. Of note, oral tongue cancer constitutes the majority of cancers of the oral cavity (Almangush, 2015). Interestingly, it also has the worst prognosis (Listl et al., 2013). As shown in Figure 1, the anterior two-thirds of the tongue is a subsite that belongs to the oral cavity. This part can also be referred to as the oral tongue or mobile tongue. Similarly, the posterior third, also known as the base of the tongue is a subsite that belongs to the oropharynx (Almangush, 2015).

The oral tongue squamous cell carcinoma (OTSCC) has been reported to have a worse prognosis than squamous cell carcinomas arising from other subsites of the oral cavity (Rusthoven et al., 2008). Therefore, it is important to properly stratify cancer patients into risk groups for effective management and to alleviate the psychological, social, and economic burden caused by oral tongue cancer (Jaiswal, 2018).

Substantial progress has been made in terms of understanding the causes of oral cancer, prevention mechanisms, and treatment strategies. However, the main concern is in the effective and accurate stratification of the patients into risk groups. These stratifications can be in the form of prediction of locoregional recurrences, disease-specific survival, or overall survival of oral cancer patients. To this end, several approaches such as the use of the staging system of the American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) (Low et al., 2015), molecular markers (Almangush, 2015), and nomograms (Li et al., 2017) have been used in the risk stratification in oral cancer.

However, several shortcomings have been reported in the afore-mentioned approaches for the prognostication of oral tongue cancer. For example, the staging system of the American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (cTNM) has been shown to be an objective and accurate tool for predicting the prognosis for an entire population of cancer patients. Thereby making the cTNM risk stratification approach widely considered in the treatment planning for oral tongue cancer patients (American Joint Committee on Cancer, 2002; Low et al., 2015; Li et al., 2017).

In spite of this, it has been reported that the cTNM staging system showed limited prognostic ability for individual patients due to its inability to consider tumor- and patient-related risk factors (S. G. Patel & Lydiatt, 2008; Sobin, 2003). In addition, for early-stage oral cancer, the cTNM staging system has not shown convincing prognostic capabilities as it cannot properly access the biologic behavior of the

tumor (Piazza et al., 2014; Po Wing Yuen et al., 2002). Likewise, for molecular markers, lack of repeated validation for most of these markers have not provided reliability for their use in clinical practice (Søland & Brusevold, 2013). To this end, a tool that considers different prognostic factors together (i.e. staging system and clinicopathologic parameters) to accurately predict patients' outcomes would be pertinent for effective cancer management – prevention of ineffective treatment and avoidance of unnecessary overtreatment (Almangush, 2015; Li et al., 2017).

The goal of this thesis is to apply machine-learning techniques that consider the aforementioned shortcomings of the TNM staging to estimate and predict tongue cancer patients' outcomes such as locoregional recurrences and overall survival. Furthermore, this thesis is also aimed at developing a web-based prognostic tool for the stratification of tongue cancer patients into a low- or high-risk of locoregional recurrence. This is an important step towards personalized medicine. Additionally, this thesis is further aimed at comparing the performance of machine learning techniques to nomograms in the prognostication of outcomes for oral tongue cancer patients.

The prediction of oral cancer survival outcomes is of utmost interest to both clinicians and patients. This is because determining cancer outcomes may crucially contribute to personalized treatment planning, avoid unnecessary therapies, and offer effective management decision-making (Kudo, 2019). Also, early prediction of the possibility of cancer recurrence has been reported to decrease the mortality rates (Safi et al., 2017; Vázquez-Mahía et al., 2012). Therefore, with accurate risk stratification of oral cancer patients, realistic counselling can be offered to the patients while the clinicians are well posited to make informed decisions. Consequently, the overall survival rates of oral cancer patients may be improved.

A wide variety of machine learning techniques that involve supervised learning methods and algorithms were used to develop prognostic models for oral tongue cancer. These predictive models are expected to become important for the emerging concepts of personalized medicine and precision oncology. The prognostication of oral tongue cancer using machine learning as presented in this thesis was based on two different datasets. The first dataset contained clinicopathologic characteristics of early-stage oral tongue cancer patients treated at teaching hospitals between 1979 and 2009. These hospitals were University Hospitals of Helsinki, Oulu, Turku, Tampere, and Kuopio (all in Finland) and at the A.C. Camargo Cancer Center in Sao Paulo, Brazil. The second dataset was obtained from the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH).

2 REVIEW OF THE LITERATURE

2.1 Oral tongue squamous cell carcinoma

Oral cancer begins in the oral cavity (mouth) which includes the lips (upper, inside lining, and lower), buccal mucosa (cheeks), gums, retromolar trigone, frontal two-thirds part of the tongue, the floor of the mouth (below the tongue), and the hard palate (bony roof of the mouth) (American Cancer Society, 2018; Chang, 2013) as shown in Figure 2.

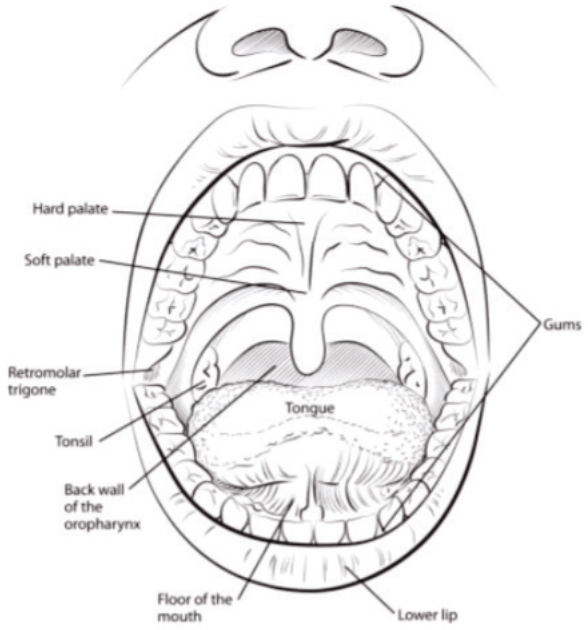

Figure 2. The oral cavity and oropharynx (American Cancer Society, 2018).

In the oral cavity, more than 90% of cancers are squamous cell carcinomas while less than 5% are verrucous carcinoma (American Cancer Society, 2018). As the oral tongue is the most common subsite in the oral cavity, oral tongue squamous cell carcinoma (OTSCC) arises from the anterior two-thirds part of the tongue. Globally, there were 354,864 new cases of oral cavity cancer with the inclusion of lip cancer diagnosed in 2018 (World Cancer Research Fund, 2018). In the United States, it has been estimated that 53,260 people will get oral cavity or oropharyngeal cancer with an estimated 10,750 death from this cancer in 2020 (American Cancer Society, 2020). Likewise, for oral tongue cancer, it has been

estimated that there were 17,060 new cases and 3,020 death in the United States in 2019 (Siegel et al., 2020). The high mortality rate is due to late diagnosis (Chang, 2013).

The most significant risk factors for OTSCC include alcohol, use of tobacco, and areca nut (betel quid) (Agnihotri & Gaur, 2014; Al-Amad et al., 2014; Scully, 2011). Other potential risk factors reported include potentially malignant lesions (Casparis et al., 2015; L. Sun et al., 2013), infection with oncogenic viruses such as human papilloma virus (Jalouli et al., 2012; Y. Zheng et al., 2010), dietary factors such as low consumption of vegetables and fruits (Meurman, 2010), poor oral hygiene (Oji & Chukwuneke, 2012), and genetic susceptibility (Hillbertz et al., 2012). Other possible risk factors include dental trauma that may be caused by several factors such as the sharp edge of a broken tooth (Bektas-Kayhan et al., 2014; Manoharan et al., 2014), allergies to dental restorations (Weber et al., 2012) and periodontal disease (Yao et al., 2014).

2.2 Diagnosis of oral tongue squamous cell carcinoma

The diagnosis of OTSCC is based on histology (Kudo, 2019). To determine the histology of OTSCC cancer, tissues are obtained from patients with excision or biopsy, cytological smears, and fine-needle aspiration (Kudo, 2019). This is most effective for lesions where malignancy is already suspected (Brinkmann et al., 2011). To this end, pathologists shoulder an immense responsibility to accurately diagnose OTSCC based on histology.

Biopsy is still considered as the gold standard for the diagnosis of OTSCC. Similarly, it has been reported that timely intervention in the carcinogenetic process and a quick response between the appearance of symptoms, small size of the tissue, and positive histological confirmation of OTSCC is capable of reducing cancer-specific mortality (Almangush, 2015; van der Waal et al., 2011). Thus, the early-diagnosis of OTSCC becomes important, as most cases of OTSCC are asymptomatic at the initial stage. Therefore, it is important to provide education aimed at self-examination and identification of oral lesions (Sarode et al., 2012). The possibility of self-identification is reasonable, as the subsite (tongue) is easily accessible for examination.

A delay in diagnosis and management of OTSCC may lead to poor management strategies, increased comorbidity, and reduced quality of health and chance of survival. For patients with oral lesions, the examination should include a clinical inspection. Likewise, in the case of patients with an established diagnosis of OTSCC, imaging techniques should be used to confirm the presence or absence of

metastasis. To this end, the presence or absence of metastases is used in the prediction of the biologic behavior of cancer (Kudo, 2019). Consequently, this forms the basis for determining the treatment plan and decision making regarding the patients. However, metastases are usually not accurately determined without the need for surgical exploration of neck lymph nodes. Hence, the need for predicting patients' outcomes becomes imperative.

2.3 Prediction of outcomes

In the medical parlance, the identification of a disease based on its signs and symptoms is known as diagnosis (Chang, 2013). Similarly, the prediction of the outcome of a disease and status of the patients such as overall survival, disease-specific survival, and locoregional recurrences is known as prognosis (Chang, 2013). The survival of patients from cancer is the most important outcome of interest to clinicians, oncologists, nurses, patients, and their families (Kudo, 2019). This is because it can significantly assist the patients in planning for their lives, and their families may be well-positioned on how best to take care of them. Similarly, the clinicians may also benefit from the accurate prediction of outcomes by making informed-decisions on the treatment strategies for the patients. In addition, the recurrence of cancer, which is the return of cancer after treatment as a result of incomplete resection of the tumor (Almangush, 2015) has also been touted as an important outcome of interest in the quest to properly manage cancer. It can be either local, regional, or the combination of both (locoregional) recurrences and has the unpleasant consequence of being the main cause of treatment failure and poor prognosis of oral tongue cancer (Peng et al., 2014; Yanamoto et al., 2013).

The accurate estimation of recurrences may guide daily clinical practice. With the proper prediction of recurrences in cancer patients, patients can be advised with realistic expectations. Also, the clinicians may be well equipped to make informed decisions about the patients through proper planning and offer personalized treatment and follow-up strategies such as postoperative adjuvant therapy. In this thesis, locoregional recurrences and overall survival were the outcomes of interest examined by machine learning techniques.

2.4 Approaches to predict outcomes in TSCC cancer

The early prediction of recurrences in tongue cancer patients can be beneficial for the identification of high-risk patients. Thus, corresponding multimodality treatment strategies can be planned for them. Admittedly, cancer diagnostics and

management have witnessed significant advancements in recent years. However, the 5-year relative overall survival (OS) for patients was reported to be 61% for patients treated with curative intent (Mroueh et al., 2017). With the advancement in technology, improved mathematical and statistical computations and analyses, and processing capacity of computer software, several technology-based advances such as graphical tools like nomograms and disruptive technologies like machine learning techniques have emerged. These technology-based tools have been touted for accurate diagnosis and prognosis prediction of cancer in patients. Such approaches ensure that the patients are treated on a case-by-case basis. Thus, this further supports the concept of personalized medicine, improved quality of care, and increased overall survival.

The basic goal of personalized medicine is to accurately identify individualized treatment therapies that maximize effectiveness aimed at improving the quality of care offered and increasing the chance of survival for the patient. Additionally, it ensures that unnecessary therapies for patients are avoided and suffering associated with the cancer is controlled. Furthermore, it provides a useful insight into effective management decision-making.

2.4.1 Nomograms

A nomogram can be said to be a graphical prognostic model where complex mathematical and statistical formulas are used to transform certain variables such as demographics, clinical, or treatment variables into an estimated outcome of a cancer patient (Balachandran et al., 2015; Grimes, 2008). The examples of estimated outcomes may include clinical events such as occult nodal metastases, recurrences, disease-specific survival, or overall survival for a given patient (Balachandran et al., 2015). Several articles have been reported that used nomograms in predicting survival in breast cancer (W. Sun et al., 2016), gastric cancer (J. Liu et al., 2016), and head and neck cancer (Gross et al., 2008; Li et al., 2017; Montero et al., 2014).

2.4.2 Machine learning techniques (MLT)

The application of machine learning techniques (MLT) in cancer research has been touted to facilitate the early diagnosis and prognosis of cancer to ensure proper management of patients (Kourou et al., 2015). Our medical hospitals and centers are reservoirs for large amounts of cancer data. These can be socio-demographic, clinical, pathologic, or genomic/microarray data. Recently, several studies have combined these data for diagnosis and prognosis purposes (Chang, 2013).

Clinical data consist of signs and symptoms such as the size of the primary lesion, clinical neck node, and other symptoms observed directly by the clinicians or physicians (Chang, 2013). Similarly, pathological data are obtained from laboratory examinations of the patient (Chang, 2013). Examples of pathological data include the number of neck nodes, tumor thickness and size, and other post-surgical pathological parameters. The clinical and pathological data may be combined to form clinicopathologic data (Chang, 2013). Considering the advancements in digitalization and data analysis, information regarding genomic markers of patients are now stored in the hospital databases (electronic health records).

Similarly, with the advancements in the internet of things (IoT), viz-a-viz in eHealth and mHealth, more medical-related data have become available. Interestingly, these data contain vital information that can assist in the proper management of cancer. Therefore, new technologies that are able to extract this information become imperative.

Machine learning, a subfield of artificial intelligence (AI), is a methodology that has become popular in medical research in recent years due to its ability to discover and identify patterns and complex relationships contained in these data (Kourou et al., 2015). These relationships were learned by MLT to be able to effectively estimate the possible future outcomes of cancer. Notably, the introduction of MLT to cancer diagnosis and prognosis significantly improved the accuracy of outcome prediction by 15% - 20% (Kourou et al., 2015).

In this thesis, machine learning techniques are applied to clinicopathologic data. However, the limited amount of sample size is one of the main challenges with medical datasets (Chang, 2013). In addition, the extraction of the medical dataset is time-consuming. Also, the extracted sample cohorts usually need preprocessing to handle the inconsistencies, missing, and incomplete data (Chang, 2013). Despite these challenges, with preprocessed data of reasonable size, high-performance machine learning models with accurate and reliable risk estimation can be developed for prognostication in cancer.

MLT learn from the data samples with the aim of making informed and accurate deductions and inferences from these data. The learning process involves two distinct phases. Firstly, complex known and unknown relationships and dependencies between the variables contained in the datasets are estimated and established (Bishop, 2006; Kourou et al., 2015). Secondly, these estimated dependencies are consequently used to predict the outcomes of new cases, given that the new cases have the same parameters or variables for which the initial training was done (Bishop, 2006; Mitchell, 2006; Witten et al., 2011). The

schematic flow usually involves the extraction of data and their corresponding attributes from the database, training with machine learning, evaluation of the results obtained from the training, and decision-making based on the presented result (Figure 3).

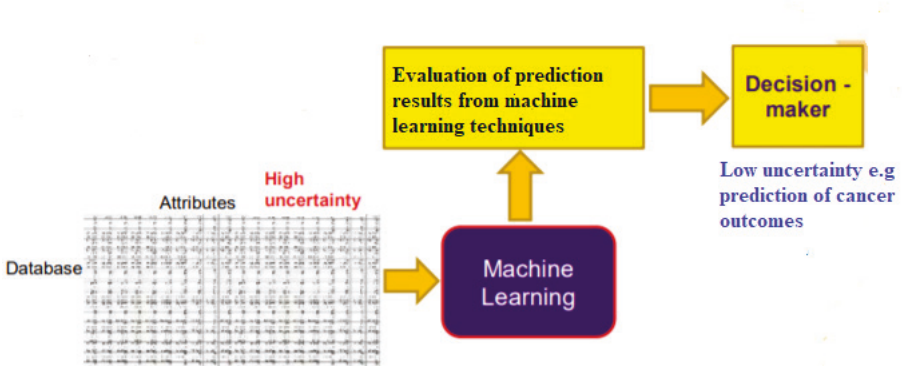

Figure 3. Schematic summary of machine learning techniques for decision-making (Elmusrati, 2020).

Interestingly, the learning process takes place automatically without the need for explicit programming (Expert System, 2020). The trained model can be re-trained with more data so that it can learn and improve from experience (Expert System, 2020). Hence, they are sometimes called data-driven systems (Elmusrati, 2020).

Despite the improved performances offered by the machine learning techniques in cancer diagnosis and prognosis, it is important to mention that the machine learning technique is not able to perfectly extract all the information contained in the data. This is due to noise, distortion, and possible corruption of some aspects of the data used to train the model (Elmusrati, 2020). In spite of this, the machine learning is usually capable of extracting reasonable amounts of information that are sufficient to understand the relationships between the variables and parameters contained in the date (descriptive) in order to provide valuable estimates or predictions of the outcomes of the patients (predictive) with a reasonable confidence level. However, to enhance better machine learning models, it is important to preprocess the data to remove missing and distorted data points (Elmusrati, 2020).

Several machine learning algorithms have been developed and used in the training phase (Bishop, 2006; Mitchell, 2006; Witten et al., 2011). The machine learning methodologies have been broadly divided into supervised and unsupervised learning methods (Kourou et al., 2015). In some other reports, machine learning methods have been divided into supervised, unsupervised, and reinforcement

(Expert System, 2020) while it includes semi-supervised in other reports (Elmusrati, 2020) as shown in Figure 4.

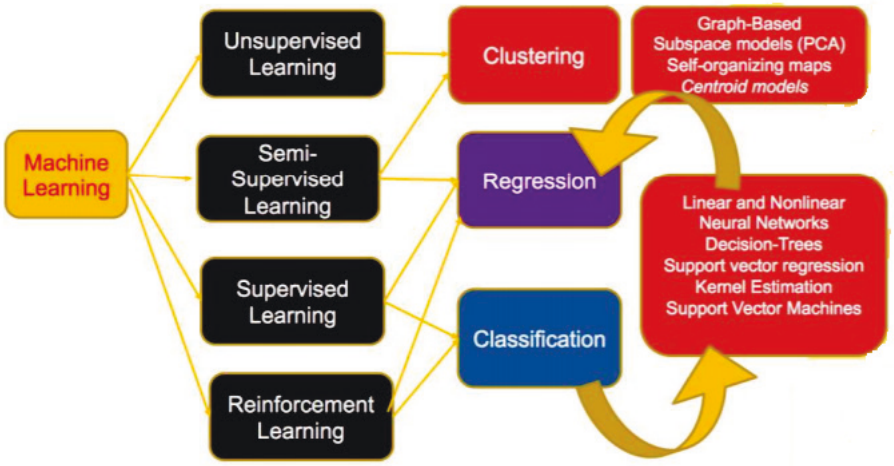

Figure 4. An overview of machine learning methodologies (Elmusrati, 2020)

In a supervised machine learning method, known training data are used to map the input data and the corresponding variables to the desired output (Kourou et al., 2015). In this case, the output produced after the thorough and sufficient training of the input training data is known as the predicted output while the expected or initially known output from the original data is called the target or desired output. Thus, the difference between the desired output and predicted output is known as the prediction error (Expert System, 2020). The prediction error usually informs the decision to further modify the model to increase the performance and accuracy accordingly. However, in some case, where not all the input data are labelled or only the statistical properties of the data are known without labels, then this type of machine learning technique is known as semi-supervised learning methodology (Elmusrati, 2020).

In contrast, the unsupervised learning method is a machine learning methodology where the input data to be used in the training phase are neither classified nor labelled and there is no notion of the output during the training or learning phase (Kourou et al., 2015). The idea is to classify or group the input data into clusters of similar attributes. Thus, the model does not figure out an output, rather the model explores the training data for relationships and patterns and forms clusters of similar attributes based on these patterns (Kourou et al., 2015).

The reinforcement machine learning methodology involves the possibility to interact and learn online from the environment by taking actions and discovering errors and rewards. These actions can be labelled as either right or wrong based on the response from the environment (Elmusrati, 2020; Expert System, 2020). Thus, trial and error become an integral part of reinforcement learning (Expert System, 2020). In this methodology, the model is given the liberty to automatically determine the ideal behavior that maximizes its performance within a given context. Usually, the reinforcement signal is required as reward feedback for the model to learn which action is best for its performance.

2.4.3 Tasks of machine learning

Based on the aforementioned definitions of machine learning methodologies, three common tasks can be inferred. These are classification, regression, and clustering (Figure 4). It is important that enough data are available to properly train and tune the model for better performance. This ensures the generalization of the model. However, it would be erroneous to have the notion that the more data that is available for training, the better the corresponding machine learning model generated will be (Elmusrati, 2020). The data available should be carefully divided to have enough for training, that is, the generalization of the model and not a memorization effect as shown in Figure 5.

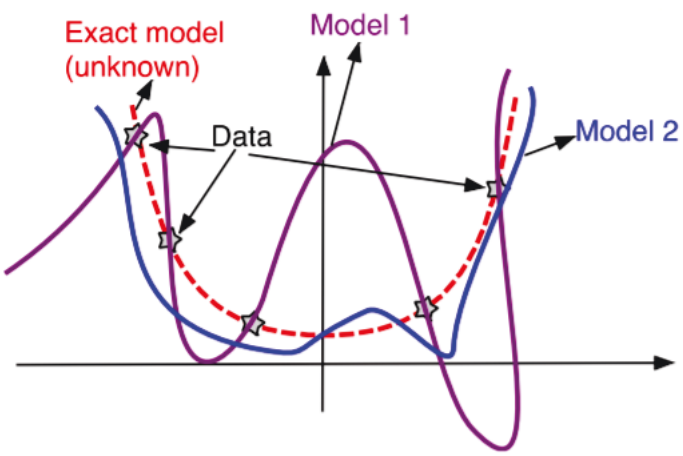

Figure 5. Memorization effect of ML training (Elmusrati, 2020)

By memorization effect, it means that the training process fails to capture the input/output relations between the available data. Instead, the model matches the available inputs with the output data (Elmusrati, 2020). As shown in Figure 5, the

model 1 appeared to have properly captured the relationships between the input data. However, it failed to learn the complex relationships between the input variables (Figure 5). Instead, it memorized and mapped the input variables. This is the reason why the shape of the model is different from the *exact model* as shown in Figure 5. Considering the *model 2*, it is evident that it does not just map the input data; rather, it learned the complex relationships and patterns between the input variables. Hence, *model 2* follows the same pattern and resemblance as the *exact model* and avoids the memorization of the relationships between the training data.

2.4.3.1 Classification

The most common machine learning task are classification tasks aimed at categorizing the data into a set of finite classes. The output variables are used to classify the input variables into one of the possible output classes. Hence, supervised and reinforcement machine learning methodologies can be thought of as a classification or regression problem (Figure 4). For instance, the prediction of whether a tumor is malignant or benign (Ayer et al., 2010) and stratification of the patients into finite classes of either low- or high-risk of recurrence (W. Kim et al., 2012) are all examples of classification problems. Likewise, classification of patients into positive (cN+) or negative (cN-) lymph nodes in the neck (Bur et al., 2019) and survival status as either dead or alive (Karadaghy et al., 2019) can be thought of as classification tasks.

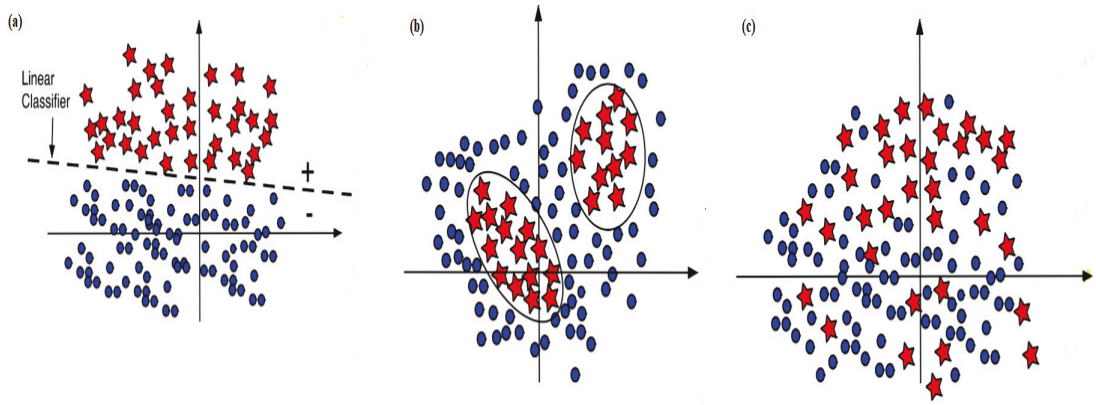

Figure 6. Class boundaries for classification classifiers (a) easy distinction (b) moderately distinguishable (c) extremely difficult to distinguish.

A linear classifier as shown in Figure 6a can easily distinguish the classes. Similarly, it can be a moderately distinguishable (Figure 6b) or extremely complex

to distinguish the classes (Figure 6c). Therefore, due to noise, corruption, and bias in the data, it becomes an important challenge to find an accurate boundary of the classifiers. In the process of developing a predictive model to classify the data into predefined classes, two errors may emerge. These are training and generalization errors (Kourou et al., 2015). The training error refers to the misclassification of the training data, while the misclassification of the testing data is known as the generalization error (Kourou et al., 2015).

2.4.3.2 Regression

The objective of the regression task is to learn the observed input-output relations to find an accurate model. That is, the training data are used to fine-tune the model for prediction. The resultant learned model after the training process can be used to test data that was not part of the training data (interpolation) or external data (extrapolation) as shown in Figure 7. This gives the actual performance and generalizability of the model. As shown in Figure 7, the best line could be used to fit the model.

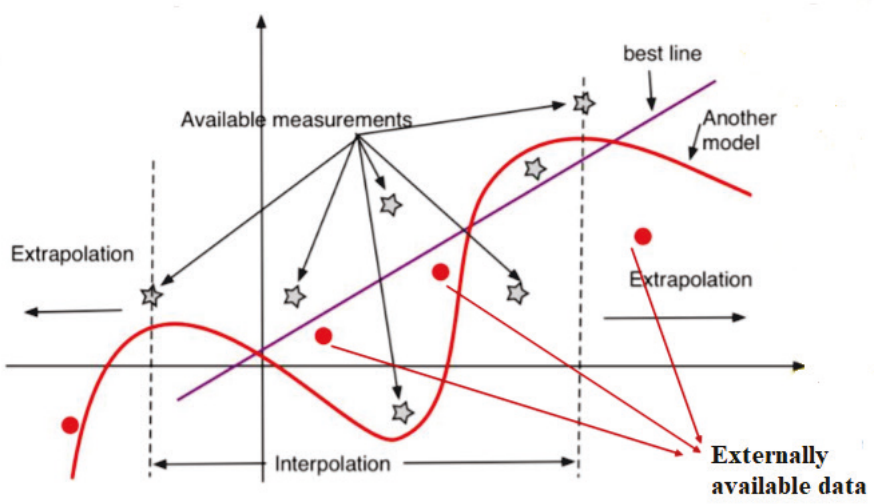

Figure 7. The concept of interpolation and extrapolation in regression (Elmusrati, 2020).

A good way to evaluate the regression model is to divide the available data into training and testing sets, i.e., one of the sets is used for training, while the other set is used for testing. The prediction of real-value variables, such as the prediction of survival time in cancer patients can be considered as an example of regression tasks (Bartholomai & Frieboes, 2018). Furthermore, semi-supervised machine learning methodology can also be thought of as a regression task (Figure 4).

2.4.3.3 Clustering

In clustering, input data are classified by finding the hidden patterns (Bishop, 2006; Kourou et al., 2015). It is one of the most common unsupervised machine learning methodologies (Bishop, 2006). In terms of similarity, it follows the same approach as the classification technique except that there is no label (target output) from which the model can learn. That is, there is no teacher for the model (Elmusrati, 2020). In clustering problems, the training of the model is aimed at finding clusters to describe the relationship between the data items. After the training, new samples can be assigned to any of the identified clusters. Thus, unsupervised and semi-supervised machine learning methodologies are used for clustering tasks (Figure 4) (Elmusrati, 2020).

Irrespective of the type of machine learning methodology used or the tasks (regression, classification, or clustering) to be performed, mathematical tools are critical for the design, analyses, and evaluation of the machine learning algorithms. Linear algebra, optimization theory, and probability and stochastic processes formed the foundation for a successful machine learning algorithm. Over the years, several machine learning algorithms have been developed. In this thesis, only the machine learning algorithms examined in the published articles of the author are discussed (Section 2.4.4).

2.4.4 Machine learning algorithms

The machine learning algorithm is a mathematical-based model that automatically learns from data without the need to explicitly program the algorithm (Koza et al., 1996). In order to learn hidden relationships in supervised learning, the available data are divided into training and testing data. The algorithms learn the relationships between the variables using the training data to make predictions (Koza et al., 1996). Likewise, the testing data are used to evaluate the performance and validity of the trained model.

Therefore, when applying machine learning algorithms, the data are positioned as the basic components (Kourou et al., 2015). These data are described by several features, with each feature consisting of different types of values. As such, it is important that the data are preprocessed to handle possible noise, outliers, missing, and duplicate values to improve the quality of the data (Kourou et al., 2015). Similarly, understanding the dependencies and relationships between the features could also assist to select the best features during the machine learning training process (Kourou et al., 2015).

Several techniques have been used for data processing to improve the quality of the data. These include dimensionality reduction, feature selection, and feature extraction. Dimensionality reduction is specifically important when the number of features contained in the dataset is very large. With lower dimensionality, the ML algorithms are poised to show improved performance (Tan et al., 2006). This is because it eliminates possible noise and irrelevant features in the data (Kourou et al., 2015; Tan et al., 2006). In the quest to improve the quality of the data, new features that are a subset of the old features can be selected. This process is known as feature selection. Conversely, it is possible to create a new set of features from the initial set that captures all the important details in the dataset. This process is known as feature extraction. It is aimed at manifesting the benefits of dimensionality reduction.

Once the data have been pre-processed, the type of task determined, inputs, and output feature defined, the next step is to determine the type of machine learning algorithm. Several machine learning algorithms have been reported in the literature. However, for this thesis, only the machine learning algorithms used in the published articles (I-III) have been highlighted in section 2.4.4.1

2.4.4.1 Logistic regression

Logistic regression is one of the machine learning algorithms that has been widely used in statistics. It is an extension of the linear regression algorithm due to inaccuracies of linear regression for classification problems. It derives its name from the function used at the core of the algorithm called the logistic function, also known as the sigmoid function (Swaminathan, 2018).

$$(1) \qquad sig(t) \equiv \sigma(a) = \frac{1}{1+e^{-t}}$$

Logistic regression is used for classification problems with two possible outcomes, hence it is sometimes known as two-class logistic regression. Thus, the logistic function is used to compress the output of a linear equation between 0 to 1, instead of fitting a straight line or hyperplane as with the linear regression.

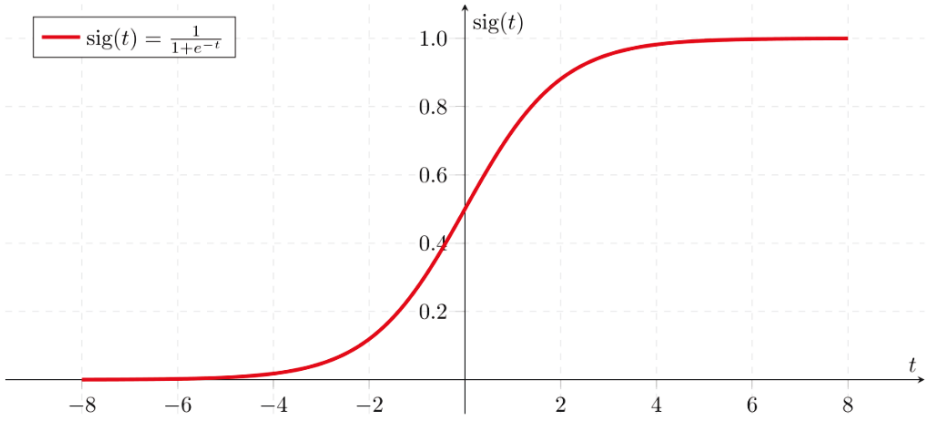

Figure 8. The logistic function for the logistic regression algorithm (Swaminathan, 2018).

The process of transformation of linear regression to logistic regression is straightforward. With the linear equation, the relationship between features and outcome is represented by:

(2) $$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

Considering a two-class outcome of high-risk of recurrence (denoted as 1) or low-risk of recurrence (denoted as 0) with input features $(x_1^{(i)} \ldots\ldots x_p^{(i)})$ and corresponding slopes and shift as $(\beta_1 \ldots\ldots \beta_p)$ and (β_0). As the aim of the logistic regression is to squeeze the output between 0 to 1 for classification tasks, the right-hand side of the Equation (2) into logistic function presented in Equation (3). Therefore, Equation (2) is modified to Equation (3).

(3) $$P(y^{(i)} = 1) = \frac{1}{1 + e^{\left(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)})\right)}}$$

Equation (3) ensures that the output assumes only values between 0 and 1. In cancer management, logistic regression can be used to predict the outcome of a patient as having either a benign or a malignant tumor. In training, the algorithm with a dataset that contains features such as tumor size, the expected outcome can be either of the two classes, as shown in Figure 8. The threshold is usually set to 0.5, thus, the inclusion of additional points does not affect the estimated curve shown in Figure 9.

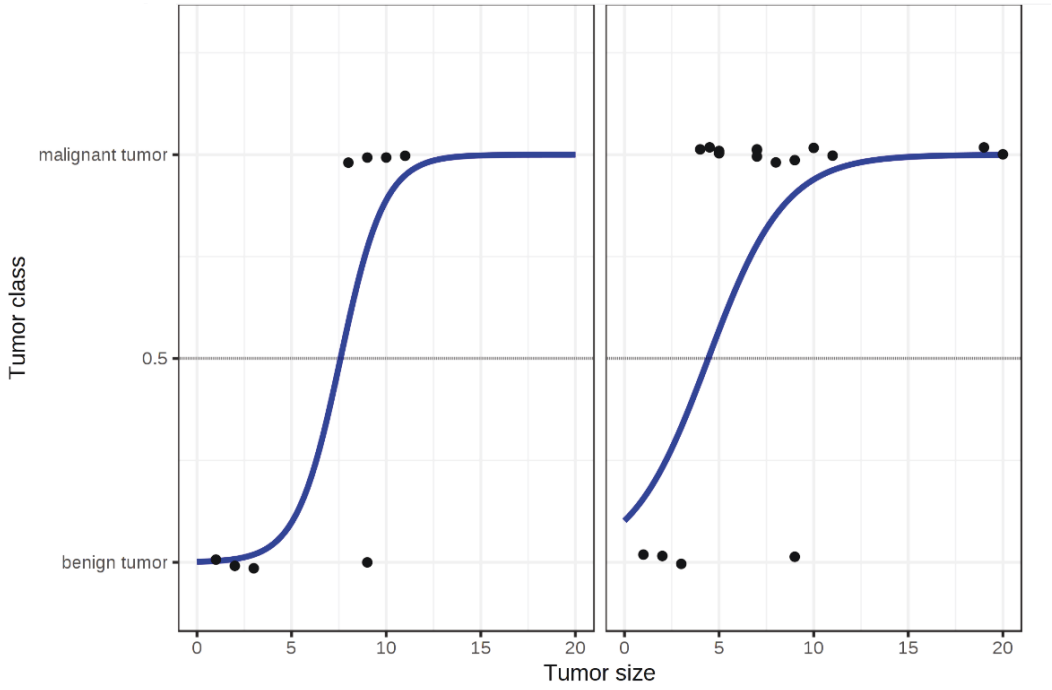

Figure 9. The use of logistic regression for classification problems (Swaminathan, 2018).

In logistic regression, the weights do not influence linearly. Rather, the weighted sum is changed to a probability using the logistic function. In this case, equation 3 is modified so that the linear term is on the right-hand side of the formula as shown in equation 4.

$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right)=\log\left(\frac{P(y=1)}{P(y=0)}\right)=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p \quad (4)$$

The term in the $\log()$ function is called the odds, that is, the probability of an event divided by the probability of no event. From Equation (4), it can be said that the logistic regression gives a linear model for the log odds. However, when one of the features x_j changes by 1 unit, the prediction changes. This change can be accommodated by taking the exponential function of the equation 4. The new equation is:

$$\frac{P(y=1)}{1-P(y=1)}=odds=e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)} \quad (5)$$

Equation (5) can simply be used to examine the effect of increasing one of the feature values by 1. Increasing one of the feature values by 1 is best described by the ratio of the odds. This is given by Equation (6)

$$\frac{odds_{x_j+1}}{odds} = \frac{e^{(\beta_0+\beta_1 x_1+\ldots+\beta_j(x_j+1)+\ldots+\beta_p x_p)}}{e^{(\beta_0+\beta_1 x_1+\ldots+\beta_j x_j+\ldots+\beta_p x_p)}} \quad (6)$$

Then, applying the exponential rule: $\frac{e^{(a)}}{e^{(b)}} = e^{(a-b)}$, Equation (6) is further modified to Equation (7).

$$\frac{odds_{x_j+1}}{odds} = e^{(\beta_j(x_j+1) - \beta_j x_j)} = e^{(\beta_j)} \quad (7)$$

An increase in the value of feature x_j by one unit means that the estimated odd change by a factor of $e^{(\beta_j)}$. For a binary classification problem, changing the feature x_j from one category (usually the reference category) to another category changes the estimated odds by a factor of $e^{(\beta_j)}$. For a categorical feature with more than two categories; each category is expected to have its own column (one-hot-encoding). To avoid over-parameterization in the case of more than two categories, $M-1$ columns are used for features with M categories. In this case, the M-th category is considered as the reference category. It is important to mention that when all numerical features are zero and the categorical feature are at the reference category, then the estimated odds are e^{β_0}.

To determine the parameters of the logistic regression, maximum likelihood estimation can be used. To achieve this, the derivative of logistic regression may be used which may be expressed conveniently considering the sigmoid function as shown below:

$$\frac{d\sigma}{d\alpha} = \sigma(1-\sigma) \quad (8)$$

For a given dataset $\{\phi_t, z_t\}$, where $z_t \in \{0,1\}$ = output parameter and $\phi_t = \phi(x_t)$. Also, t ranges as $t = 1, \ldots\ldots T$ with likelihood of the observation is given as $\prod_{t=1}^{T} y_t^{z_t}$.

Thus, the likelihood function can therefore be written as:

$$p(z|w) = \prod_{t=1}^{T} y_t^{z_t} \{1-y_t\}^{1-z_t} \quad (9)$$

From Equation (9), $z = (z_{1,\ldots\ldots\ldots,} z_T)^T$ and $y_t = p(C_1|\phi_t)$. Therefore, taking the negative logarithm of the likelihood equation given in Equation (9) gives the cross-entropy function. The cross-entropy function is given as (Bishop, 2006):

$$(10) \qquad E_{ntropy}(w) = -\ln p(z \mid w) = -\sum_{t=1}^{T} \left\{ z_t \ln y_t + (1 - z_t) \ln \left(1 - y_t \right) \right\}$$

From Equation (10), $y_t = \sigma(a_t)$ and $a_t = w^T \phi_t$. Taking the gradient of the cross-entropy error function with respect to w gives Equation (11) where Equation (8) was used. From Equation (11), the factor that involved the derivative of the logistic sigmoid function has cancelled out. This leads to a simplified version for the gradient of the logarithm likelihood as depicted in Equation (11).

$$(11) \qquad \nabla E_{ntropy}(w) = \sum_{t=1}^{T} (y_t - z_t) \phi_t$$

Of note, the contribution to the gradient from the dataset point of view is expressed by $(y_t - z_t)$ which is the error between the target value and the predicted value of the model multiplied by the basis function vector ϕ_t. Interestingly, Equation (11) takes precisely the same form as the gradient of the sum-of-squares error function for the linear regression model (Bishop, 2006).

There are specific concerns with the maximum likelihood used in the logistic regression. Notably, it can exhibit significant overfitting, especially with the dataset that is linearly separable. The reason is that the solution for maximum likelihood occurs when the hyperplane ($\sigma = 0.5$) separates two classes and the magnitude of w tends to infinity. Additionally, the logistic sigmoid function increases infinitely into the feature space so that each class m is assigned a posterior probability $p(C_m|\mathrm{x}) = 1$. Furthermore, choosing another separating hyperplane would give rise to the same posterior probabilities (Bishop, 2006). Therefore, the maximum likelihood does not provide a clear means of favoring one solution over the other. Also, the solution that is found depends on several factors such as parameter initialization and choice of optimization algorithm. Irrespective of the size of the data, the problem exists as long as the data is linearly separable. This singularity problem can be addressed by the inclusion of a prior and finding a maximum posterior (MAP) solution for w. Alternatively, adding a regularization term to the cross-entropy error function may be posed as an intuitive solution. In summary, logistic regression (for classification problems) as well as its counterpart linear regression (for regression problems) consist of fixed basis functions. This makes them useful in analytical and computational related tasks. However, their practical application is limited due to dimensionality issues. This is because their application to large data involves the adaptation of the basis functions to the data.

To address this problem, it is important to define all the basis functions that are directed at the training data points and subset of these functions during training.

The artificial neural network and support vector machine are examples of algorithms that can use these techniques to address the singularity problems encountered by the logistic regression. Therefore, these two algorithms are discussed in detail in sub-sections 2.4.4.2 and 2.4.4.3, respectively.

The artificial neural network fixes the number of basis functions in advance while allowing these functions to be adaptive. That is, parametric forms for the basis function are used where the parameter values are adapted in the training phase. The most common type of artificial neural network used for this purpose is called the feedforward neural network, also termed the multilayer perceptron (Bishop, 2006). The details of the artificial neural network are presented in sub-section 2.4.4.2.

2.4.4.2 Artificial neural network

The artificial neural network is a subfield of artificial intelligence that works in a way that is inspired by the human brain (A. Biglarian et al., 2011; Akbar Biglarian et al., 2010; M.-H. Zheng et al., 2013). It is an adapted general model or mapping that is capable of learning the relationships between input and output variables which can be used in the prognostication of various cancers (A. Biglarian et al., 2011; Akbar Biglarian et al., 2010; Lisboa, 2002; J. Patel & Goyal, 2007; M.-H. Zheng et al., 2013). Thus, it has found applications in several cancer types (Keogan et al., 2002; Selaru et al., 2002; Spelt et al., 2013). Structurally, it is composed of the input, hidden, and output layers (Figure 10).

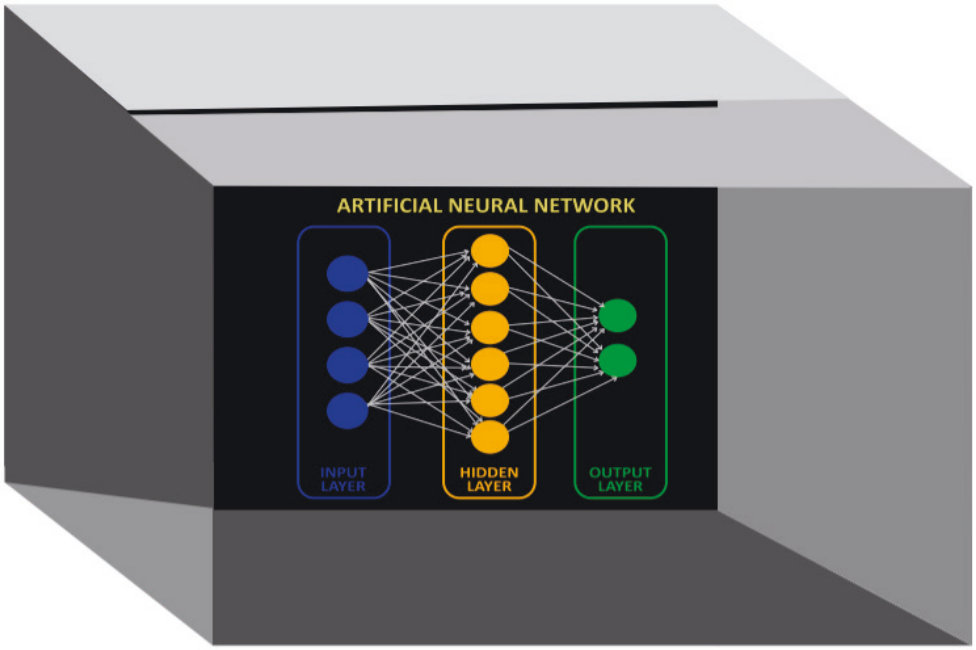

Figure 10. The structure of an artificial neural network with an interconnected group of nodes (Kourou et al., 2015).

It is a non-linear algorithm that achieves better generalization than the conventional linear/nonlinear regression algorithms. In terms of the application of ANN, it can be a feedforward neural network (Figure 10) or a backward propagation algorithm.

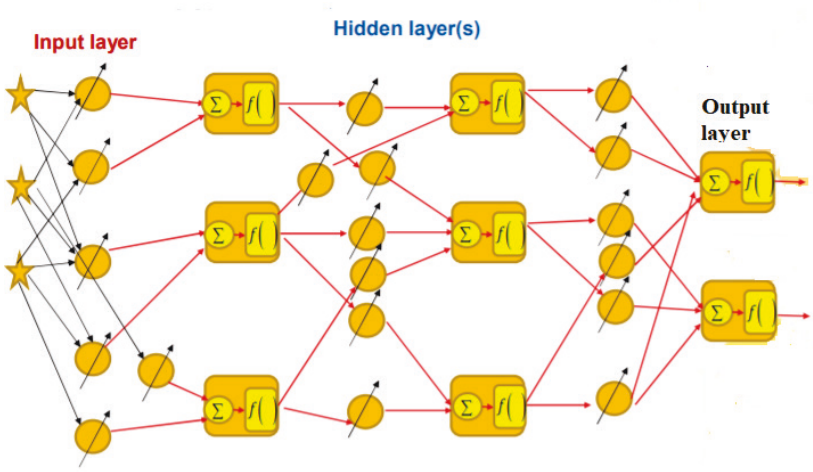

Figure 11. Feedforward neural network.

The feedforward neural network is also called the multilayer perceptron (Bishop, 2006). Of note, the term multilayer perceptron can be misleading because in this case the multilayer perceptron consists of multilayers of logistic regression with continuous nonlinearities (Bishop, 2006). Therefore, it is not a case of multiple perceptrons, as suggested by its name. The feedforward neural network (multilayer perceptron) results in a model that is more compact and faster to evaluate. However, the trade-off is that the likelihood function, which formed the basis function for the network during the training, is no longer a convex function of the model parameters. In addition, the compactness requires a significant amount of computational resources during the training.

The general mathematical models for regression as well as classification is given as:

$$y(x,w) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right) \tag{12}$$

It is based on linear combinations of fixed linear or nonlinear basis functions $\phi_j(x)$ as shown in Equation 12. From Equation (12), the function *f (.)* can be a linear or nonlinear activation function for regression/classification problems. Extending Equation (12) ensures that the basis function ($\phi_j(x)$) depends on parameters that is adjustable along with the coefficient $\{w_j\}$ in the training phase. The neural network itself uses parametric nonlinear basis functions. Therefore, it follows the same form as Equation (12). That is, each basis function is a nonlinear function that consist of combinations of inputs linearly with adaptive coefficient parameters.

The basic neural network model is derived from a series of functional transformations. Considering M linear combinations of input variables that ranged between $x_1, \ldots, x_T$, the activations are given as Equation (13):

$$a_j = \sum_{i=1}^{T} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{13}$$

From equation (13), *j* ranged between *1,……,M* while the superscript (1) means that the parameters are in the first layer of the neural network. Also, $w_{ji}^{(1)}$ represents the *weights* while $w_{j0}^{(1)}$ indicates the *biases*. Each parameter from Equation (13) is transformed using an activation function, $h(.)$ that is differentiable and nonlinear as shown in Equation (14).

(14) $Z_j = h(a_j)$

These quantities that are given in Equation (14) correspond to the output of the basis function given in equation (12). However, in the context of neural networks, these are known as the hidden units. Of note, sigmoid functions such as the *logistic sigmoid* or *tanh* are usually selected as the nonlinear functions $h(.)$. Linearly combining these values gives the output unit activation given in equation 15.

(15) $$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)}$$

From equation (15), K represents the total number of outputs, and k is between 1,.....,K. The equation (15) above corresponds to the second layer of the network where the bias parameter is represented by $w_{k0}^{(2)}$. To get a set of network outputs (Y_k), an appropriate activation function is used to transform all the output unit activation functions. It is essential to mention that the choice of activation function to be selected for this purpose depends mostly on the nature of the data and also the magnitude distribution of the target variables. As a rule of thumb, for a standard regression problem, the activation function is aimed at having $Y_k = a_k$. Similarly, the logistic sigmoid function has been the widely used activation function for a two-class (binary) classification problem to transform each output unit activation. Hence for classification, it is given as:

(16) $y_k = \sigma(a_k)$

Where: $\sigma(a) = \dfrac{1}{1+e^{(-a)}}$

However, for a multi-class classification problem, the softmax activation function is the preferred choice (Bishop, 2006). Therefore, combining the derivations to give the overall network function for a sigmoidal output unit activation function as shown in Equation (17). The set of all weights and biases have been combined together into vector w.

(17) $$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{T} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} \right)$$

Thus, it can be deduced from the above that the neural network model is basically a nonlinear function that is made up of a set of input variables $\{x_i\}$ controlled by an adjustable vector parameter w, to form a set of output variables $\{Y_k\}$.

The bias parameter present in equation (13) can be absorbed by defining an additional input variable $x_0 = 1$. It is absorbed into a set of weight parameters. Thus, Equation (13) can be rewritten as:

$$a_j = \sum_{i=0}^{T} w_{ji}^{(1)} x_i \tag{18}$$

Similarly, the second-layer bias can be absorbed into second-layer weights. As a result, the overall network function is given as Equation (19).

$$y_k(x,w) = \sigma\left(\sum_{j=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{T} w_{ji}^{(i)} x_i\right)\right) \tag{19}$$

As seen from the derivation stages of Equation (19), the neural network consists of stages of processing, hence, the name multilayer perceptron (MLP). Of note, the multilayer perceptron of the neural network differs as the hidden units of the neural network uses the continuous sigmoidal nonlinearities unlike the ordinary perceptron that uses the step-function nonlinearities (Bishop, 2006). For the neural network, it implies that it is differentiable with respect to the network parameters. The differentiability characteristic is essential for backpropagation network training.

In the training of the network, for example, in a classification task with a target variable tagged as $\boldsymbol{t}$ in such a way that it has two-classes denoted as $\mathbf{C_1}$ where $t = 1$ and $\mathbf{C_2}$ where $t = 2$ and a logistic sigmoid activation function shown in Equation (16) given that $0 \leq y(x,w) \leq 1$. From this, $y(x,w)$ represents conditional probability $p(C_1|x)$ while $1 - y(x,w)$ gives the probability $p(C_2|x)$.

$$y = \sigma(a) \equiv \frac{1}{1+e^{(-a)}} \tag{20}$$

Therefore, the conditional distribution of the target variable is given by a Bernouli distribution as given in Equation 17.

$$p(t \mid x,w) = y(x,w)^t \left\{1 - y\left(x,w\right)\right\}^{1-t} \tag{21}$$

Assuming that the variables contained in the training dataset are independent to avoid collinearity issues, then the error function is given by the cross-entropy function already presented in Equation (10).

(Recall: 10): $$E_{ntropy}(w) = -\ln p(z \mid w) = -\sum_{t=1}^{T}\left\{ z_t \ln y_t + (1 - z_t)\ln\left(1 - y_t \right)\right\}$$

Where y_t denotes $y(x_n,$ w). Of note, cross-entropy is preferred instead of sum-squares to achieve faster training and generalization of the model (Simard et al., 2003).

Therefore, in more complex problems that consider K separate binary classification tasks, the network, in this case, has K outputs, each with a logistic sigmoid activation function. Each of these outputs has a target label denoted as $t_k \in \{0,1\}$, where the value of $k = 1, \ldots, K$. Considering the independence of the training variables as well as the class labels, then, the conditional distribution of the target variable is given as:

(22) $$p(t \mid x, w) = \prod_{k=1}^{K} y_k\left(x, w\right)^{t_k}\left[1 - y_k\left(x, w\right)\right]^{1-t_k}$$

Hence, taking the negative logarithm of the likelihood function (Equation 10) gives the error function given in Equation 23.

(23) $$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\left\{ t_{nk} \ln y_{nk} + \left(1 - t_{nk}\right)\ln\left(1 - y_{nk} \right)\right\}$$

Similarly, y_{nk} denotes $y_k\ (x_n, w)$.

For multi-class target variables, meaning, each input is assigned to one of the possible K classes that are mutually exclusive, the error function follows the same principles as given by the binary. The error function for a multiclass is given by:

(24) $$E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \ln y_k(x_n, w)$$

To improve the speed of solving the problem by finding the location of the minimum point of the error function, it is crucial to evaluate the gradient of an error function. This can be achieved using the backpropagation procedure. The gradient descent optimization algorithm is one of the most widely used. Given the weight update in Equation (25)

(25) $W^{(\tau+1)} = W^{(\tau)} + \Delta W^{(\tau)}$

Choose the weight update presented in Equation (25) to comprise a small step in the direction of the negative gradient as shown in Equation (26)

(26) $W^{(\tau+1)} = W^{(\tau)} - \eta \nabla E(W^{(\tau)})$

Here $1 > \eta > 0$ is the adaptation rate, and setting its value correctly is essential to avoid the weight's divergence. After every update, the gradient is re-evaluated for the new weight vector and the process is repeated. To evaluate ∇E, the entire training set is processed. This technique (that requires the entire training set) is known as the batch training technique. Notably in the batch training technique, at each step, the weight vector is moved in the direction of the greatest rate of decrease in the error function. Hence, it is called steepest descent or gradient descent. Unfortunately, this algorithm does not perform reasonably well using this technique even though it seems intuitive (Bishop & Nabney, 2008). Therefore, for the batch technique, there are other methods that have been reported to be more efficient, robust, and faster (Nocedal & Wright, 1999). Examples of such methods include the conjugate gradient and quasi-Newton (Bishop, 2006; Nocedal & Wright, 1999). These algorithms performed better than the gradient descent algorithm because their error functions decrease at every iteration unless the weight vector has arrived at a local or global minimum.

Despite this, the gradient descent is one of the widely used algorithms. The problem mentioned previously is usually addressed by running a gradient descent algorithm multiple times. At each run, a different initial weight vector is chosen randomly. The resulting performance is compared with a validation set in order to determine the best performing run. In addition, there is an on-line version of the gradient descent algorithm that has been reported to be vibrant and showed good performance on large datasets. The on-line gradient descent algorithm is known as stochastic gradient descent or sequential gradient descent.

In the on-line gradient descent, the error functions based on the maximum likelihood for the input variables that are independent are given in Equation (27). It consists of a sum of terms, one for each data point.

(27) $$E(w) = \sum_{n=1}^{N} E_n(w)$$

It is called the stochastic gradient descent because it updates the weight vector based on a single data point at a time as given in Equation (28). The update is done

repeatedly in cycling through the data sequentially or using a random selection of points with replacement.

$$W^{(\tau+1)} = W^{(\tau)} - \eta \nabla E_n(W^{(\tau)}) \quad (28)$$

A significant advantage of on-line gradient descent is that it handles redundancy in data more effectively and efficiently than the ordinary or traditional gradient descent. Also, the computational intensiveness in the case of an online-gradient descent is reasonably low. Moreover, the on-line gradient descent is capable of escaping from the local minima.

It should be noted that the term backpropagation has been used to mean a variety of concepts in neural network analyses. For instance, the multilayer perceptron is otherwise known as a backpropagation network. Similarly, where the training of a multilayer perceptron is done using a gradient descent algorithm that is applied to a sum-of-squares function, it is called backpropagation. Hence, it is important to summarize the training process of a neural network to ensure that the meaning of backpropagation is not confused.

In general, for any type of neural network, the training algorithms minimize the error function through an iterative process. Secondly, the weights are adjusted in a sequence of steps. In minimizing the error function (the gradient of the error function with respect to the network weights is evaluated). It can be shown that the weights adaptation depends on the error at the neuron output. Although it is straightforward to compute the errors in the output layer as the difference between the actual output and the desired output, it is not the case for the hidden layers. The desired values of the outputs of the neurons in the hidden layers are unknown. Hence, the backpropagation algorithm provide a simple method to compute the errors in the hidden layers by propagating and projecting of the output errors backward. This process involves two distinct steps. In the first step, the errors are propagated backward through the network in order to evaluate the derivatives. This step is peculiar to any other network (not limited only to the multilayer perceptron).

In the case of a feedforward neural network, the inputs go after weighting and processing in a layer by layer till the output. It is essential to have the input and output layer with any desirable number of hidden layers (between the input and output). To produce a neural network with a good performance, the weights of each neuron should be adjusted to minimize the local error at the output of each neuron. This essential process is achieved by computing the error at the output layer by evaluating the difference between the actual output and the desired output. The error could be reduced by repeatedly adjusting the connection weights of all

neurons in all layers. The gradient descent may be used to achieve this adaption. This error could be propagated backward in order to compute the local error at all neuron's outputs. The backpropagation algorithm is one of the most widely used supervised algorithms. The number of layers or neurons should be carefully selected to avoid overfitting or underfitting the network.

The learning concept is based on adapting weights to minimize some cost functions that are related to the average error. Therefore, the goal in the learning process is to minimize the average error between the target output and the predicted outputs considering all available learning samples. This is usually achieved using the stochastic gradient algorithm. By removing the expectation from the cost function, this algorithm can be modified and hence is called the steepest descent algorithm. As shown in the equation, the step size is vital for the convergence speed as well as the stability of the algorithm. With a small step size, the stability of the algorithm could be guaranteed. However, this makes the algorithm very slow to converge. Conversely, a large step sizes might enhance the convergence speed but higher risk of divergence and instability. To address this problem, several adaptive algorithms have been developed that self-adapt to the appropriate value of the step size. Adaptive and optimized step-size algorithms are used in almost all implementations of backpropagation algorithms in computer packages. However, the discussion is beyond the scope of this thesis. As a rule of thumb, the step size must be less than 1.

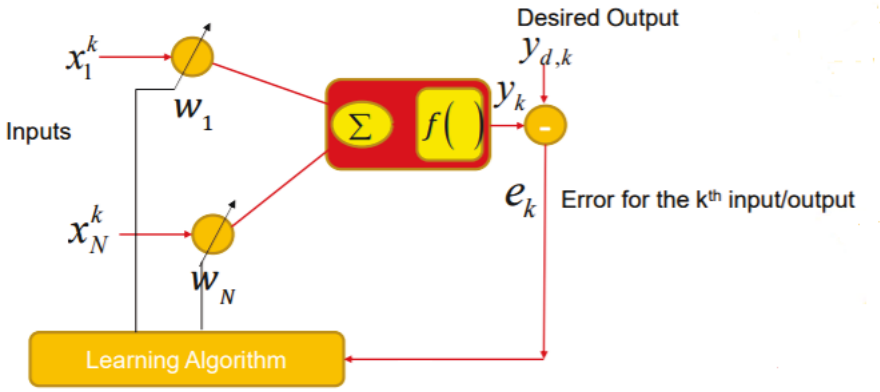

Figure 12. A single neuron neural network

To evaluate the error function derivatives of a feedforward topology with differentiable nonlinear functions that use a layer of sigmoidal hidden units with sum-of-squares error, the error function was given in equation (27).

(Recall. 27) $$E(w) = \sum_{n=1}^{N} E_n(w)$$

Considering a linear combination of input variables (x_i) with outputs y_k, the output in this case is given by Equation (29).

(29) $$y_k = \sum_i w_{ki} x_i$$

While the error function for an input pattern n takes the form:

(30) $$E_n = \frac{1}{2} \sum_k \left(y_{nk} - t_{nk} \right)$$

Where: $y_{nk} = y_k\ (x_n, w)$. With respect to the weight w_{ji}, the gradient of the error function is then given as:

(31) $$\frac{\partial E_n}{\partial w_{ji}} = \left(y_{nj} - t_{nj} \right) x_{ni}$$

It is the product of the error parameter ($y_{nj} - t_{nj}$) and the variable x_{ni}.

In general, for a feed-forward neural network, each unit calculates the weighted sum of the input using the formula, where z_i correspond to the input (or activation of a unit) which connects to unit j with an associated weight of w_{ji}.

(32) $$a_j = \sum_i w_{ji} z_i$$

The above equation can be transformed by using a nonlinear activation function $h(.)$ to obtain an activation in the form of equation (14)

(Recall: equation 14) $Z_j = h(a_j)$.

Therefore, for a neural network, all the inputs or the inputs of interest are supplied to the network, and the activation of all the hidden and output units in the network are calculated using Equations (14) and (27). This process is known as forward propagation (forward flow of information through the network).

Conversely, for error backpropagation, the process involves applying an input vector x_n to the network and forward propagation using the derivatives presented

in Equations (14) and (27). Therefore, the output units are evaluated in such a way that:

(33) $$\delta_k = y_k - t_k$$

This is followed by the backpropagation of each hidden units in the network using:

(34) $$\delta_j = h'\left(a_j\right)\sum_k w_{kj}\delta_k$$

Thereafter, the required derivatives are evaluated using:

(35) $$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$

The process is similar in the case of the batch training method. However, the derivative of the total error is given in Equation (36) by summing over all the training errors:

(36) $$\frac{\partial E}{\partial w_{ji}} = \sum_n \frac{\partial E_n}{\partial w_{ji}}$$

The above Equation (36) works with the assumption that each hidden or output unit in the network in the case of batch training has the same activation function.

The weight updating algorithm for the multi-neurons and multi-outputs neural network is similar to that of a single neuron, as discussed earlier. In some instances, a more complicated structure can be designed where the output of the first neuron is connected as the input for the second neuron. This type of design is known as a recurrent neural network. As expected, the weight adapting algorithm would be different from the conventional backpropagation algorithms (Figure 12).

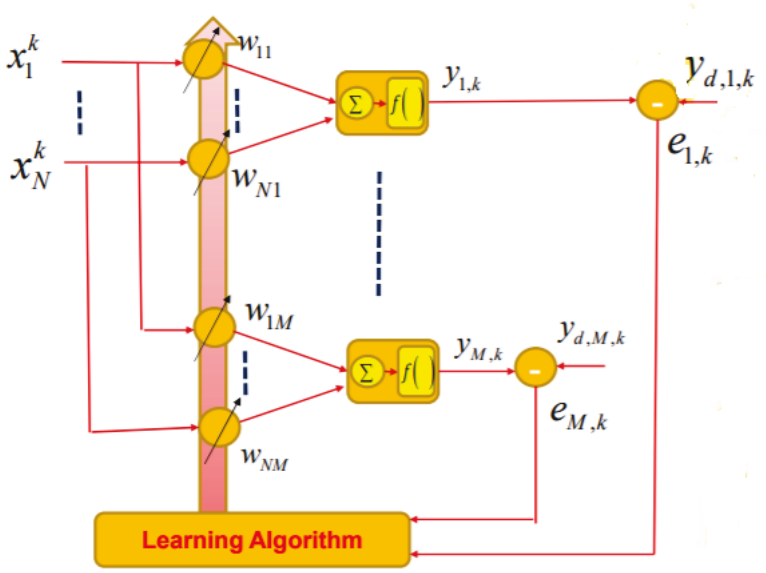

Figure 13. A multi-neurons and multi-outputs neural network

Of note, the weight adaptation is the same as the single neuron neural network. The difference is mainly with each neuron to update its weight. Conversely, a multilayer neural network is a powerful neural network construction that is able to learn complex relationships between inputs and outputs. The weight adaptation is usually performed with the backpropagation algorithm. Figure 13 shows a simple sketch for multiple layer neural network.

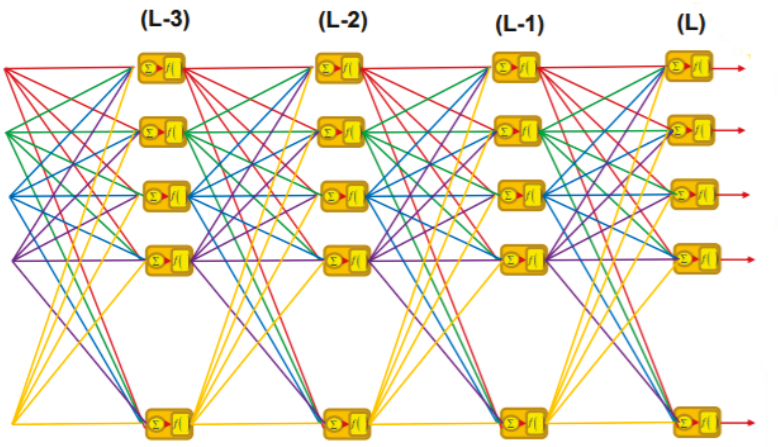

Figure 14. The multilayers neural network

Given the multi-layered neural network (Figure 14) with L number of layers, the output layer as the layer number L, each layer k has M_k neurons and N inputs to

the first layer. In this case, the number of outputs in each layer is the same number of neurons. This is because each neuron has only one output. At each layer k, there is a matrix of weights with dimension $M_{k-1} \times M_k$ where W_{ijk} is the weight at layer number k connecting the i^{th} input (which is the output of the i^{th} neuron at layer $k-1$ except for the input layer) to the j^{th} neuron at layer k. The main goal is to update the weight of all layers to minimize the overall average error between the actual output and the predicted output. The weight of the layers may be updated with the steepest descent algorithm. However, with the hidden layer, it presents a great challenge to update the weights as the desired output of the neurons is unknown. To address this, it is crucial to backpropagate the error of the output layer to the hidden layers. Although, backpropagation showed improved performance, there are certain problems with backpropagation neural networks.

For the layer recurrent neural network (LRNN), the output will be a recursive function in its input. Figure 15b showed a single neuron with recursive output. The weights can be updated by different methods such as unfolding the process in time and applying the stochastic gradient. This is similar to the backpropagation algorithm, but with a little modification by unfolding the LRNN in time. However, in backpropagation, we have the problem of gradient vanishing. The LRNN is effective only for short time sequence memory systems.

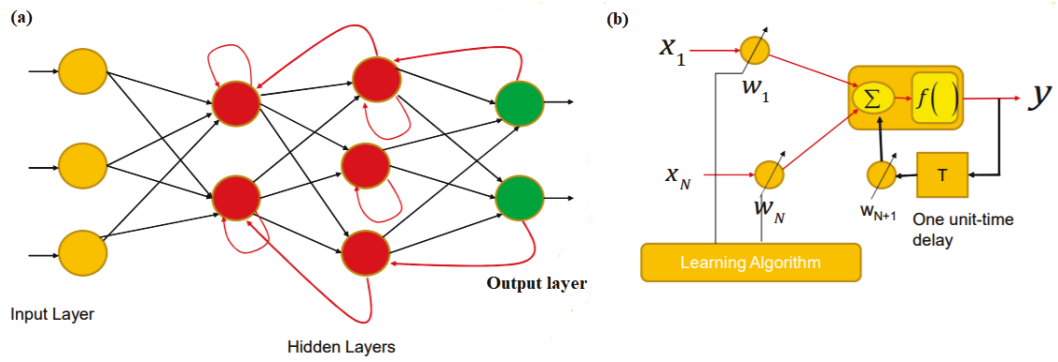

Figure 15. Layer recurrent neural network

The LRNN captures the time series properties in many applications such as the stock market, climate predictions, or any applications where the desired output depends not only on the current input but also on the previous history.

For the long-short time memory system (LSTM), it has been specifically designed to handle the problem of vanishing gradients that affect the backpropagated

network. That is, it allows the network to retain information for longer periods compared to the traditional LRNN.

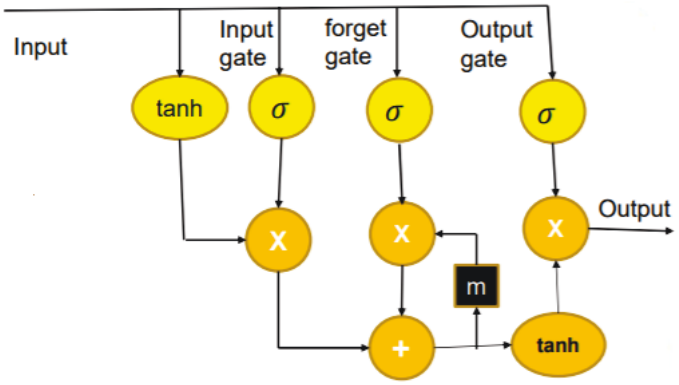

Figure 16. The structure of the long-short time memory system (LSTM)

Of note, the LSTMs can maintain a constant error (Figure 16). This allows the LSTMs to continue learning over numerous time-steps and backpropagate through time and layers. The idea is based on adding a memory cell for the hidden neurons in the LRNN. Each memory cell is connected to three gates: forget, input, and output gates. The input gate basically adds information to the cells. The forget gate, as the name implies, deletes information that is no longer necessary. The output gate selects and outputs the necessary information.

In recent years, modifications have been made to the traditional artificial neural network. These modifications gave rise to the concept of deep learning algorithms. Interestingly, the multilayer neural networks with many hidden layers are a practical example of the models with deep architectures. Considering the backpropagation learning algorithm, it works exceedingly well for multi-layer neural networks with only a few hidden layers. Likewise, it is a gradient descent based algorithm and thus, it has the challenge that it may be trapped in poor local minima. This becomes more evident, and the severity increases significantly as the number of hidden layers increases. The closer to the local minimum, the absolute value of the weights using gradient descent becomes much smaller than one, and with successive multiplications, the value can almost be zero. Thus, the proper adaptation of the weight ceases.

Alternatively, the learning algorithm for the multi-hidden layers neural network is called the deep belief networks (DBNs). It provides a faster and more effective algorithm for neural networks with many hidden layers. It overcomes several limitations posed by the standard backpropagation learning algorithms. In this case, the learning is performed layer by layer. It offers one smart way of initializing

the network weights instead of starting with completely random weights. It is most appropriate for many inputs such as the case of image recognition and huge data applications.

Similarly, the rectifier linear unit (ReLu) has received special attention in deep learning and large sized neural networks. The rectifier linear unit is presented in Equation (37). It is linear for positive x and zeroes for negatives; that is, it introduces some nonlinearities.

$$f(x)=\begin{cases} ax & x\geq 0 \\ 0 & x<0 \end{cases} \tag{37}$$

The ReLu algorithm is useful for applications require some nonlinearity. Convolution neural networks (CNN) are very important in the applications of deep learning in image recognition and classifications. The convolution is the mathematical process to compute the output of a linear systems represented by its impulse response as shown in Figure 17.

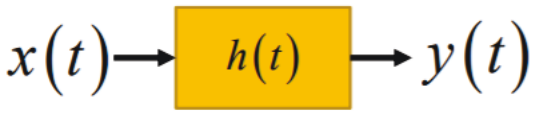

Figure 17. Schematic of convolution

Assume that, a signal $x(t)$ is applied as an input to a linear system with an impulse response given as $h(t)$, and output $y(t)$ is the convolution between $x(t)$ and $h(t)$. Mathematically this is given as:

$$y(t)=\int_{-\infty}^{\infty} x(\lambda)h(t-\lambda)d\lambda=\int_{-\infty}^{\infty} x(t-\lambda)h(\lambda)d\lambda \tag{38}$$

From Equation (38), it is evident that the convolution is the mathematical title of the filtering process in the time-domain. In discrete time, the same concept of convolution can be expressed as:

$$\hat{y}[k]=\sum_{m=-\infty}^{\infty}\hat{h}[m]\hat{x}[k-m]=\sum_{m=-\infty}^{\infty}\hat{h}[k-m]\hat{x}[m] \tag{39}$$

The Equation (39) clearly defines the convolution for a one dimensional signal. The same concept is applied for two dimensional signals like images. For discrete convolution, it is given as:

$$(40) \qquad \hat{y}[k_1 k_2] = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{h}[m,n]\hat{x}[k_1 - m, k_2 - n] = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{h}[k_1 - m, k_2 - n]\hat{x}[m,n]$$

The output depends on the elements of the $\hat{h}$ matrix. Notably, the convolutional neural networks make the feature extraction part of the learning process of the algorithm. The feature extraction is done by the linear filtering process which is performed by the convolution. Each layer of the neural network extracts certain features which are done automatically to achieve the required performance in the learning phase. Therefore, a deep neural network which involves deep learning, implies that enough features must have been extracted and highly sophisticated learning is achieved for an image recognition system.

Despite the fact that the artificial neural network has performed reasonably well in several classification problems (Ayer et al., 2010), it suffers from certain criticisms. An example of such a criticism is that the structured nature of artificial neural network could be time-consuming during training and lead to poor performance. Also, the black-box nature of the algorithm is another concern. By black-box, it means the lack of details of the exact functions that achieved the learning and mapping between the inputs and outputs.

2.4.4.3 Support vector machine

The support vector machine (SVM) is one of the robust algorithms in machine learning. It is highly preferred as it produces reasonable accuracy with less computational power (Gandhi, 2018). It is used for both regression and classification problems, but mostly for classification problems (Witten et al., 2011). The SVMs seek to address the dimensionality challenge in the shallow neural network. Therefore, the SVM uses basic functions that are directed at the training dataset. Subsequently, a subset from these training datasets would be selected. Of note, the training process using SVM involves nonlinear optimization. However, the objective function is convex, thus, any local solution is also considered as a global optimum (Bishop, 2006).

Therefore, the SVMs map the input vector into a feature space, that is, N-dimensional space (where N – number features) of higher dimensionality and identify the hyperplane that distinguishes the data points into two classes distinctly as shown in Figure 18 (Gandhi, 2018; Kourou et al., 2015). Interestingly, the SVM does not provide posterior probabilities. However, it uses an extensive concept of Lagrange multipliers.

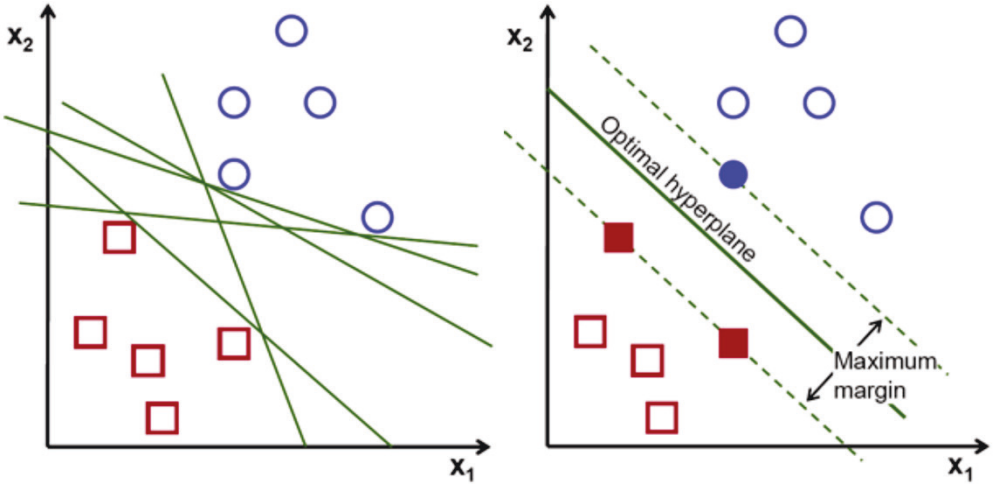

Figure 18. The support vector machine showing possible hyperplanes.

As shown in Figure 18, it is possible that more than one hyperplane separates the two classes. The most important concern in the case of more than one possible hyperplane is to find a plane that has the maximum margin (Gandhi, 2018). The maximum margin is the maximum distance between the data points of the classes (Figure 18). This ensures that the best hyperplane is chosen that classify the classes with more confidence (Gandhi, 2018).

The data points that fall on either side of the hyperplane are thought of as different classes. The nature of the hyperplane in terms of the dimension depends on the number of feature variables. A straight-line hyperplane is obtained if the number of features is two (Figure 18), and showed a two-dimensional plane if the feature is three. For more than 3 features, it becomes difficult to visualize the separating hyperplane (Gandhi, 2018). The SVMs have been used extensively in the prognostication of cancer. As shown in Figure 19 where it is important to classify cancer tumor size according to the age of the patient.

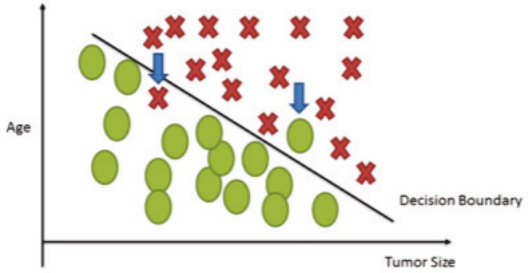

Figure 19. A simple illustration of linear SVM with two input features to classify cancer according to tumor size (Adam, 2012; Kourou et al., 2015).

As shown in Figure 19, the hyperplane classifies cancer according to tumor size as benign (green) and malignant (red) based on their size and patients' age. The hyperplane is the decision boundary between the classes of cancer according to the tumor size. As shown in Figure 19, based on the hyperplane that formed at the boundary, some of these classes (benign or malignant) can be said to be misclassified. Of note, the SVMs can also give probabilistic outputs (Platt et al., 1999).

Consider a two-class classification problem that follows a linear model given by:

(41) $y(x) = W^T \emptyset\,(x) + b$

In the above Equation (41), $\emptyset\,(x)$ corresponds to a fixed space transformation while b denotes the bias parameter. The training data consist of N input vectors that vary as $x_1, \ldots., x_N$, and corresponding targets as $t_1, \ldots \ldots, t_N$ where $t_n \in \{-1, 1\}$. Similarly, the training dataset were assumed to be linearly separable in feature space so that Equation (41) can conveniently hold true as:

(42) $y\,(x_n) > 0; for\ t_n = +1$ And $y\,(x_n) < 0; for\ t_n = -1$

Unlike the neural network that depends on initial values (arbitrary) for the weight and bias, the SVM uses the concept of margin (Figure 18). The margin is said to be the smallest distance between the decision boundary and any of the input features (samples) as shown in Figure 18. Therefore, to ensure distinct classification, the decision boundary is chosen to maximize the margin. An example of an intuitive approach to maximize the margin is the statistical learning theory (computational learning theory). The maximum margin solution can be derived using:

(43) $$\arg\max_{w,b}\left\{\frac{1}{\|w\|}\min_{n}\left[t_n\left(w^T\phi(x_n)+b\right)\right]\right\}$$

Of note, the above equation (43) appears complex to solve. Thus, the equation is further optimized as:

(44) $$\arg\min_{w,b}\frac{1}{2}\|w\|^2$$

The above Equation (44) is positioned as an example of quadratic programming. Interestingly, the bias parameter is not present in Equation (44) as seen in Equation (43). This constrained optimization problem is addressed with Lagrange

multipliers $a_n \geq 0$; with a single multiplier denoted as a_n. The Lagrangian function is thus given below where a $=(a_1, \ldots . a_N)^T$:

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n (w^T \phi(x_n) + b) - 1 \right\} \quad (45)$$

The negative (minus) that precedes the Lagrange multiplier term is due to the minimizing with respect to weight and bias and maximizing the term a.

2.4.4.4 Naïve Bayes

This classifier is based on Bayes theorem (Bishop, 2006) and is one of the widely known and efficient probabilistic algorithms. It has been prefixed with naive as the algorithm naively assumes that the predictors are conditionally independent, although this assumption is not obeyed in practice. In summary, this classifier applies density estimation to the data and approximates the optimal Bayesian average (theoretical) of the generalization performance by choosing one average classifier (Bishop, 2014). Hence, the name Bayes Point Machine (Microsoft Azure Machine Learning Studio, 2018). The Naïve Bayes classifier has the advantage that it is not prone to overfitting to the training data, and it is robust to class density estimates that are biased (Hastie et al., 2009). The algorithm works by estimating the densities of the predictors within each class. Using Equation (12), the algorithm models posterior probabilities. Finally, it classifies an estimation by estimating the posterior probability for each class. This is followed by the assignment of the observation to the class yielding the maximum posterior probability (*maximum a posteriori decision rule*) (Manning et al., 2008).

Assume N possible classes for M available features or attributes, then the probability of being in a certain class given certain specified attribute is given as:

$$P(C_i | x_1 x_2, \ldots, x_M), \ \forall_i = 1, \ldots N \quad (46)$$

This gives the probability of being in class i. Using Bayesian theorem and assuming discrete-value attributes, then:

$$P(C_i \mid x_1, x_2, \ldots x_M) = \frac{P(x_1, x_2, \ldots x_M | C_i) P(C_i)}{P(x_1, x_2, \ldots, x_M)}, \forall_i = 1, \ldots N \quad (47)$$

Where:

$P(C_i|x_1,x_2,..,x_M)$ = Posterior Probability

$P(x_1,x_2,...x_M|C_i)$ = Likelihood;

$P(C_i)$ = Class Priori Probability;

$P(x_1,x_2,...,x_M)$ = Evidence.

The Equation (47) is usually challenging to compute because of the unknown interrelationships between attributes or features. To address this concern, it is important to assume that the attributes are independent of each other. Thus, the equation (47) can be simplified as:

$$(48) \quad P(C_i \mid x_1, x_2,...x_M) = \frac{P(x_1,x_2,...x_M|C_i)P(C_i)}{P(x_1)P(x_2)...P(x_M)}, \forall_i = 1,...N$$

With this simplification, the probabilities could be easily computed. However, due to the assumption of independence between attributes, it is termed naïve Bayesian as previously mentioned. However, this assumption has been reported to work well in various machine learning applications.

2.4.4.5 Decision trees

The decision tree is one of the earliest and widely used algorithms for classification problems. It follows a true-structured classification that is made up of nodes and leaves. The general structure of the decision tree is given in Figure 20. The internal decision nodes correspond to the input feature (X, Y, Z in Figure 20), branches (X<T1; X>=T2; X<T2; X>=T2; Z<T3; Z>=T3) and the terminal leaves (Class A, Class B in Figure 20) represent the outcomes. Therefore, it is easy to understand and interpret decision tree architecture.

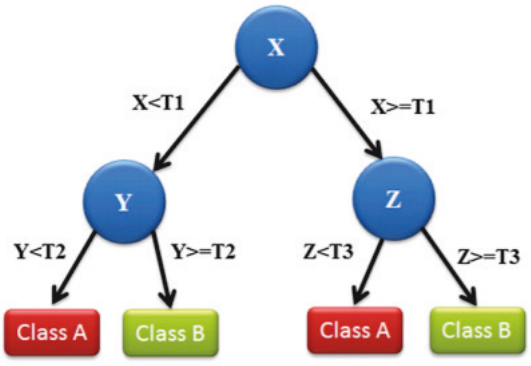

Figure 20. The structure of a decision tree

Each internal decision node m implements a test function $f_m(x)$ with discrete outcomes labelling the branches. Practically, a test is applied to the input node, and one of the branches is taken depending on the outcome. Usually, the process starts at the root and continues repeatedly and recursively until a leaf node is hit. At this point, the output is written to the leaf. The decision tree has the advantage of fast localization and interpretability (Alpaydin, 2014). The decision tree may be thought of as a nonparametric model in the sense that there is no assumption for the parametric form for the class densities. Additionally, the tree structure is not fixed a priori. Rather, the tree grows and leaves are added during the learning process depending on the complexity of the problem and also the nature of the data.

In the classification tree, the effectiveness of a split from the root is quantified by an impurity measure.

A split can either be a pure or impure split. A split is considered pure if after the split, all the instances that chose a branch belong to the same class. For instance, taking N_m to be the number of training instances that reached node m and the root node is N. In this case, N_m^i of N_m belongs to class C_i, with $\sum_i N_m^i = N_m$.

The estimate for the probability of class is given as:

$$\hat{P}(C_i \mid x, m) = p_m^i = \frac{N_m^i}{N_m} \tag{49}$$

Given that an instance reaches node m. In terms of purity, the node m is considered pure if p_m^i for all the values of i are either 0 or 1. Simply put, it is 1 if all such instances are of class C_i, and 0 when none of the instances that reaches node m are of class C_i. In the case of a pure split, there is no need to split any further. A good function to measure impurity is entropy, which is given as:

$$I_m = -\sum_{i=1}^{K} p_m^i \log_2(p_m^i) \tag{50}$$

It is a measure that specifies the minimum number of bits needed to encode the class code of an instance. An impure node implies that a further split is needed to decrease the impurity. In all, it is important to calculate impurity and chose the one that has the minimum entropy. For branches that are impure, the tree construction continues recursively and repeatedly in parallel until the branches are

pure. This forms the foundation for the classification and regression trees (CART) algorithm. Based on this foundation, decision trees have evolved over the years.

In recent years, the decision tree algorithm has evolved to boosting and bagging decision tree algorithms. The basic idea is to iteratively apply simple decision tree algorithms and combine their predictive performances to obtain a better model with improved prediction ability (Freund, 1995; Schapire, 1990; Schapire & Freund, 2012). Therefore, the boosting was aimed at transforming the weak algorithm to a better algorithm through combination with stronger algorithms. In this way, a weak base-learner can be improved to become strong learning. This improvement is known as boosting. Boosting is achieved by allowing the algorithm to develop new strategies needed to handle problematic observations while rewarding the base-learner in the final aggregation. The concept of boosting is not about the combination with stronger algorithms but manipulation of the training data by re-weighting the observations iteratively. As such, the base learner b will find a new solution $\hat{h}^{[b]}(.)$ from the available data.

By the repeated application of the weak base-algorithm on observations that are weighted based on the base-learner's success in the previous rounds, the algorithm focusses on objects that are difficult to classify by assigning higher weights to them.

In each iteration, $b = 1,.....,b_{stop}$, the weight vector $w^{[b]} = (w_1^{[b]},...,w_n^{[b]})$ contains the individual weights of all observations. In the iteration cycle, the attention is directed towards observations that happened to be misclassified up to the current iteration b.

In the final step, the previous results of the base-learner are combined into a more accurate prediction. Using an iteration-specific coefficient, there is an increase to the weight of better performing solutions of the base-learner.

The early boosting algorithm developed (Freund, 1995; Schapire, 1990) gave rise to a more concrete algorithm called adaptive boosting, also known as Adaboost (Freund & Schapire, 1996). It is called the adaptive boosting algorithm as it automatically adjusts its parameters to the data. It is usually used with simple classification trees as base-learners and has been found to show significant improvement in performance (Bauer & Kohavi, 2003; Meir & Rätsch, 2003; Ridgeway, 1999). The overview of the boosting algorithm is given below:

(a) Initialization: The iteration counter is initialized

$b = 0$; w_i for individual weights; observations given as $i = 1,....,n$ to $w_i^{[0]} = \frac{1}{n}$

(b) Base-learner: compute the base-learner for the weighted data set

Set b : $b+1$

$$\text{Then, re-weight with } w_1^{[b-1]},\ldots,w_n^{[b-1]} \xrightarrow{base-learner} \hat{h}^{[b]}(.)$$

(c) Update the weights by computing the error rate and update the iteration-specific coefficient α_b

(d) Iterate steps (b) and (c) until $b = b_{stop}$

(e) The final aggregation for new observation y_{new} is computed. It is given as:

$$\hat{f}_{boosting(adabost)}(y_{new}) = sign\left(\sum_{b=1}^{b_{stop}} \alpha_b \hat{h}^{[b]}(y_{new})\right) \quad (51)$$

The performance of the boosting algorithm has been compared to another approach known as the bagging algorithm (Breiman, 1996). In bagging, there is no need to rely on the misclassification rate of earlier iterations as seen in the boosting algorithm, rather, bootstrap generated samples are used to modify the training data (Breiman, 1998). It was concluded that boosting is poised to outperform the bagging approach due to the ability of the boosting approach to have a reduced bias-variance (Breiman, 1998).

This accounts for why the boosting decision tree is still relevant in present-day application of machine learning to other fields such as medicine. Therefore, the boosted decision tree can be said to be a (Microsoft Azure Machine Learning Studio, 2018) machine learning method that is made up of trees where the second tree corrects the errors in the first tree. Similarly, the third tree corrects the errors in the second tree and the sequence goes on until the final tree is reached. All these trees are combined together to make the prediction. Hence, it is known as the ensemble machine learning method.

Several modifications have been made to the basic decision tree algorithm. A modification to the boosted-decision tree algorithm gave rise to the decision forest algorithm. In decision forest, several decision trees are created to train the available dataset, but from different starting points. The process of creating many individualized classification trees is usually randomized. Each tree in the decision forest gives a frequency histogram of labels that is non-normalized as the outputs. These histograms are summed together and then normalizes to get the probabilities for each label (Criminisi & Shotton, 2013; Microsoft Azure Machine

Learning Studio, 2018). The trees with high prediction confidence are voted as the popular output class. It is another type of fast-supervised ensemble method that is used for large datasets.

By extension of the decision forest algorithm, the decision jungle was developed. It is aimed at achieving better generalization performance than the traditional decision trees and decision forest (Microsoft Azure Machine Learning Studio, 2018). It works in a similar way as to the decision forest, but with directed acyclic graphs (DAGs) (Criminisi & Shotton, 2013). Although, they are a non-parametric ensemble method, the longer training time is one of the major concerns about using this algorithm.

In general, decision trees have become popular in the machine learning process, as they are fast with efficient computation and memory usage. Of note, they have the capacity to capture both linear and non-linear decision boundaries. Similarly, they appear as one of the widely preferred algorithms as feature selection is integrated into the training and classification process. Additionally, the trees can accommodate noisy data. Moreover, they are able to handle data with varied distributions, thus, they are non-parametric algorithms.

2.4.5 Data division

In supervised machine learning such as in regression and classification tasks, the data are usually divided into three types. These are **training, validation, and testing sets**. In some cases, the available data is divided into the **training and testing sets**, respectively. The training set is otherwise known as the learning set. This set of data is used as the observation data input/output that is mainly used for training. With the aid of the machine learning algorithm, the input relationship between the input parameters can be captured with acceptable accuracy. The learning process may be smooth and straightforward especially if the data have been preprocessed. However, with noisy, distorted, and biased data, the learning process may be challenging.

The size of the data and the algorithm are critical factors that determine the speed of the learning process. The validation set are a part of the available data set that was not used during the training phase. This set of data is used to validate the trained model. The performance of the model after validation provides an insight into how the model is likely to perform in real-time. However, to have a more convincing insight into the actual performance of the model and also enhance the generalizability of the model, it is important to further evaluate the model with the test data which may be from the same source (externally validated with same

source), but used neither in the training nor validation phase. Likewise, the data may come from an entirely different source, in this case, this is known as external validation of the model (externally validated with different source).

There are different ways to assign the percentage between the training and validation and how to select the data into these categories. Several studies have proposed 80% for training and 20% validation. Similarly, 70% training and 30% validation have been proposed in other studies. Likewise, 50% training and 50% validation is beginning to gain relevance in recent studies. The amount of available data, the nature of the data, the training algorithm, and types of machine learning tasks may be determining factors in selecting the appropriate data division ratio. Irrespective of the data division ratio, it is important to avoid overfitting or underfitting scenarios as these have detrimental effects on the model produced. To avoid these concerns, it is usually a good practice to start with 30% for the training and then check the performance of the model on the validation set. If the performance is not good, the division ratio may be increased until a model with reasonable performance metrics is obtained. It is important to mention that the quality of the data is extremely important. Low-quality data sets will lead to biased models. By low quality, it means that the data does not represent the real problem due to high biases, missing values, not a number (NaN), and missing important input parameters that contains vital information. Also, it is important to avoid input features that are highly linearly correlated as this may lead to erroneous conclusions. The training and validation should be selected carefully to span all the dimensions of the problem as shown in Figure 21.

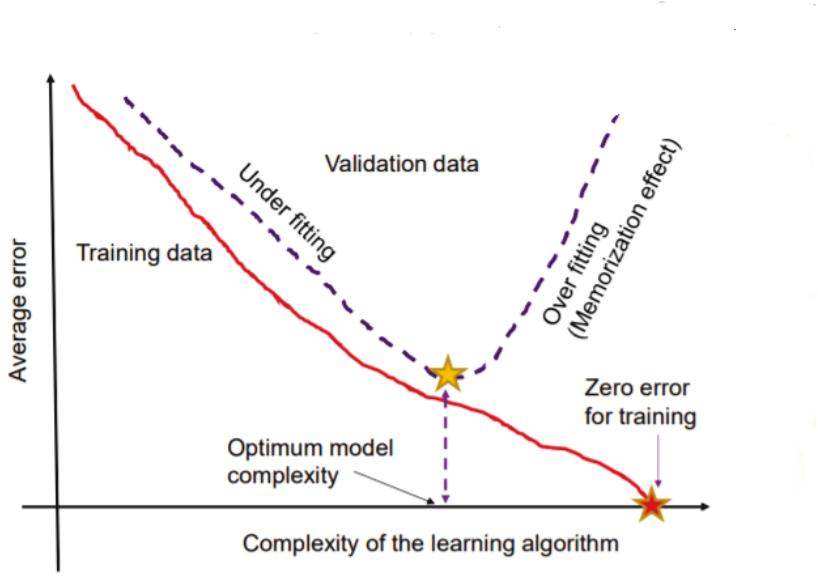

Figure 21. The training and validation phases against algorithm complexity (Elmusrati, 2020).

Due to the increasing application of machine learning techniques in various fields, new algorithms continue to emerge. Interestingly, the complexity of the algorithms increases reasonably. Therefore, as shown in Figure 21, it is important that the algorithm is trained with a reasonable amount of data to obtain an optimum model. The learning data should also be informative. Interestingly, even if the learning algorithm and the model complexity are both perfect to have an optimum model (Figure 21), this will not help if the algorithm was not exposed to enough information during the training process.

2.4.5.1 Data division methods

Based on the data division explanations presented in the aforementioned, (1) holdout, (ii) random sampling, (iii) cross-validation, and (iv) bootstrap methodologies emerged (Kourou et al., 2015). In the holdout method, the available data are divided broadly into training, validation, and testing sets. While the model is developed based on the training set, the performance of the model presented through the validation set, and the generalizability of the model with the aid of the test sets. However, this data division is done manually before the training process begins (Kourou et al., 2015). Although, the random sampling data division methodology follows an approach similar to the holdout method, the significant difference between the two is that the data division approach, in this case, is done randomly. That is, the holdout method is repeated several times the training and testing sets are randomly selected. Thus, this approach produces training and testing sets that are representative of the population and reduces the possibility of a biased dataset (Kourou et al., 2015). In cross-validation, each sample in the training set is used the same number of times during the training process and just once during the testing phase. This is to ensure that all the available data are covered in the machine learning training and testing process. For the bootstrap data division method, the data used for the training are returned back to the available dataset (Kourou et al., 2015). In this case, they may be re-used during the testing phase. In this type of approach, the performance produced by the model after the training and testing phases are less reliable. The model requires a significant amount of external validation for the model to be considered reliable.

2.4.5.2 Machine learning performance metrics

After a successful machine learning training process, the performance metrics produced by the learning algorithm varies, depending on the type of tasks. For classification problems, the performance metrics include log-loss, accuracy, area under receiving operating characteristics (ROC) curve (AUC), precision, recall,

and F1 score. The majority of these performance metrics have been defined with the aid of the confusion matrix.

The confusion matrix is one of the most intuitive and self-explanatory approaches to understanding the performance matrix in classification problems. It is generally used for classification tasks where the expected outcomes can be two-class (two outputs) or multiclass (more than two outputs). Technically speaking, the confusion matrix may not be considered as a performance metric. However, it gives insightful information about four main threshold parameters of the performance of the model. These threshold parameters are true positives, true negatives, false positives, and false negatives as shown in Figure 22.

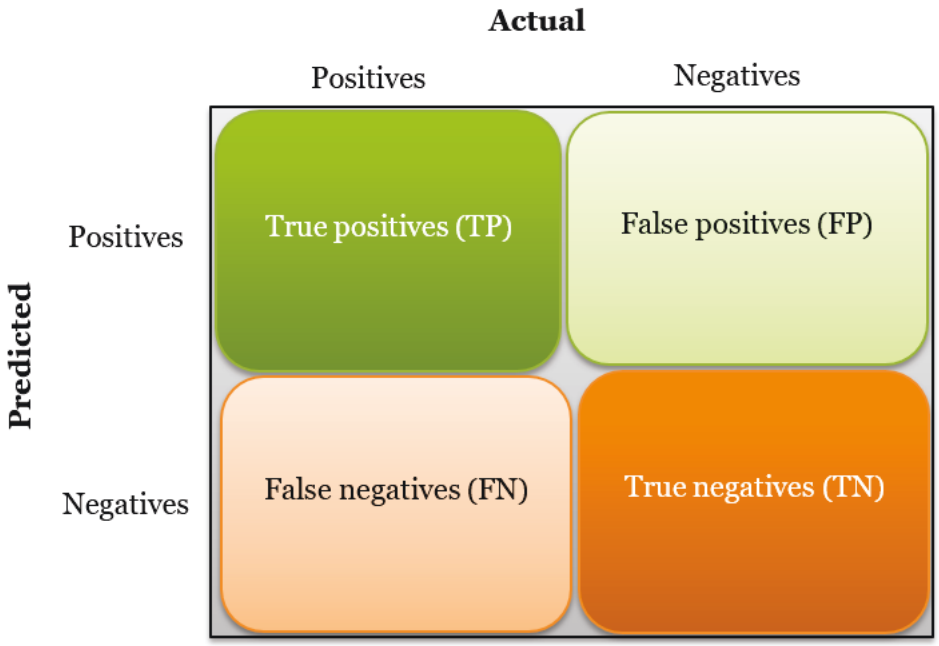

Figure 22. Confusion matrix for machine learning classification problems.

True positives (TP) refer to the cases when the actual class of the data was actually true and the machine learning model also predicted the cases as true. An example is when the patient actually had a locoregional recurrence and the machine learning correctly predicted the individual as having a locoregional recurrence.

True negatives (TN) refer to the instances where the actual class of the data was false and the machine learning model also predicted the class as false. An example is when the actual outcome of the patient's tumor is non-malignant and the machine learning model rightly predicted the tumor as such, that is, non-malignant.

False positives (FP) implies the cases when the actual class of the data was false and the machine learning model incorrectly predicted the class as true. An instance is when the machine learning model predicted the case as having a locoregional recurrence of cancer whereas the patient actually had no locoregional recurrence of cancer.

False negatives (FN) indicate the cases when the actual class of the data was true while the machine learning model incorrectly predicted the patient as false. For instance, when the machine learning model a patient to have 5-year survival whereas the patient actually died of cancer.

These four basic threshold parameters have formed the basic parameters that define the performance metrics in classification problems.

Accuracy in classification problems refers to the number of predictions that the machine learning model made correctly over all the possible predictions made by the model (Sunasra, 2017). Mathematically, it is given as:

$$\text{Accuracy} = \frac{TP + TN}{TP+FP+FN+TN} \tag{52}$$

Accuracy has shown to be a good performance metric when the target variable has classes that are nearly balanced. Thus, accuracy is not advisable in an imbalanced dataset where the target variable has classes with the majority of one class (Sunasra, 2017).

On the other hand, precision is a measure that defines the proportion of the patients that had been classified as true (TP and FP) by the machine learning model, and actually were true based on the expected outcomes (TP) (Sunasra, 2017). It is given as:

$$Precision = \frac{\text{TP}}{\text{TP + FP}} \tag{53}$$

Similarly, recall (sensitivity) is a measure that indicates the proportion of patients that was actually positive and was mainly identified by the algorithm as positive (Sunasra, 2017). It is presented as:

$$\text{(54)} \quad \text{Recall or sensitivity} = \frac{TP}{TP+FN}$$

It is essential to mention that recall gives information about the algorithm's performance regarding how many cases were wrongly or incorrectly classified. Conversely, precision gives information about the algorithm's performance regarding how many cases were correctly classified. As the name implies, it centered on being precise.

Likewise, specificity is a measure that gives an account of the proportion of the patients that was not positive and were predicted by the machine learning model as non-positive (negative). Thus, it is the exact opposite of recall. Hence, specificity is given as:

$$\text{(55)} \quad \text{Specificity} = \frac{\text{TN}}{TN+FP}$$

A very good performance measure that combines both precision and recall is the F1 score. It is the harmonic mean of precision and recall.

$$\text{(56)} \quad \text{F1 score} = \frac{2\,(Precision * Recall)}{(Precision + Recall)}$$

Therefore, the F1 score seeks to address some of the concerns with using precision and recall. It gives the model an appropriate score rather than just an arithmetic mean. However, for regression problems, the performance metrics are mean squared error (MSE), root-mean-squared error (RMAE), mean-absolute-error (MAE), R^2 or coefficient of determination, and adjusted R^2.

Mean squared error (MSE): This is one of the most widely used metrics for regression tasks because it is differentiable and provide better optimization. It is the average of the squared difference between the target class and the predicted class by the regression model (Mishra, 2019). The MSE is given by:

$$(57) \quad \text{MSE} = \frac{1}{n} \Sigma (\text{y} - y_i)^2$$

Where y is the actual value; y_i is the predicted value and $(y - y_i)^2$ is the squared difference between the actual and the predicted value.

The root mean squared error (RMSE) is actually the most preferred performance metrics for regression problems. It is the squared root of the averaged squared difference between the target classes and the predicted classes (Mishra, 2019). The RMSE is given as:

$$(58) \quad \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y - y_i)^2}{N}}$$

Of note, RMSE is mostly preferred because the errors are first squared prior to averaging. Thus, this penalizes large errors (Mishra, 2019).

The mean absolute error (MAE) means the difference between the target class and the predicted class by the machine learning model. Therefore, the MAE is robust to outliers. However, it does not penalizes errors compared to MSE.

$$(59) \quad \text{MAE} = \frac{1}{n}\sum\left|(y - y_i)\right|$$

For MAE, the absolute value of the difference between the actual and predicted classes is taken.

Considering R^2 error, also known as the coefficient of determination (CoD) is another widely used metric for analyzing the performance of a regression machine learning model. It evaluates and informs how better the model is when the regression model is compared with a constant baseline. Notably, the CoD is a scale-free score that implies that the metric takes into consideration how large or small the value is as CoD will always be less than or equal to one (Mishra, 2019).

$$(60) \quad R^2 = 1 - \frac{MSE(\text{mod}\,el)}{MSE(baseline)}$$

The adjusted R^2 follows the same interpretation as the ordinary R^2. However, it is an improvement over the ordinary R^2 as given below:

$$(61) \quad \text{Adjusted R}^2 = R^2_{adjusted} = \left[\left(\frac{n-1}{n-k-1}\right) \times \left(1 - R^2\right)\right]$$

Where :
n = number of observation
k = number of independent variables

The adjustments to the ordinary R^2 are needed because the ordinary suffers from an increasing scores without improved performance of the model when the terms are increased. This may lead to misjudgments of the actual performance of the model. Therefore, the score of the adjusted R^2 is usually lower than the ordinary R^2 because the model tries to adjust for the increasing input variables. In this case, the model will only show improvement if the increasing predictors contribute to a significant improvement.

In general, it is often expected that the values of R^2 ranges from 0 to 1. Thus, positive value of R^2 is expected. However, the reality is, it ranges from $-\infty$ to 1, meaning, the R^2 can be negative. The reasons for a negative value of R^2 could be that the chosen regression algorithm does not follow the trend of the data. In addition, the presence of outliers may be the cause of a negative R^2. Moreover, the failure to properly consider the intercept in the regressor may also be the cause of a negative value of R^2 (Mishra, 2019).

2.4.6 Errors in machine learning methodology: overfitting and underfitting

Errors usually occur during the training and/or in the testing phase. The misclassification error that occurs during the training phase is known as the training error. Likewise, the misclassification error on the testing data is termed the generalization error. In some cases, the model might fit the training data reasonably well, thereby decreasing the training error rates. However, the same model may experience increased testing error rates. This phenomenon is known as overfitting. It is a form of increased model complexity where the testing error rates increases. In the case of underfitting, the training error increases while the testing error also increases. The best model for classification, regression or clustering tasks is the ideal model that is devoid of underfitting or overfitting. It is a model that produces the lowest generalization error. To have an ideal model, it is important to balance the bias-variance decomposition. The bias components represent the error rate of the algorithm in use, while the variance represents the error over the training and testing sets. From these, the overall error of the model is the sum of bias and variance errors. Hence, the name bias-variance decomposition.

2.4.7 Machine learning in cancer prognostication

Several articles have been published that employed machine learning techniques in the prognostications of cancer (Bur et al., 2019; Cruz & Wishart, 2007; Exarchos et al., 2012; Karadaghy et al., 2019; K. Park et al., 2013; Y. Sun et al., 2007). These studies have examined the use of machine learning techniques to predict cancer susceptibility (risk assessment), recurrence, and survival (Ayer et al., 2010; Bur et al., 2019; Exarchos et al., 2012; Karadaghy et al., 2019; K. Park et al., 2013; Urbanowicz et al., 2013). The prognostication of outcomes has been reported to be beneficial for proper planning of treatment and to improve the prognostication of overall survival of the patients (Cruz & Wishart, 2007). For nearly three decades, artificial neural network and decision trees have been widely used algorithms in various research to employ machine learning for cancer prognostication (Bottaci et al., 1997; Maclin et al., 1991; Simes, 1985).

A growing trend is noted in recently published articles where several machine learning algorithms were considered (Bur et al., 2019; Karadaghy et al., 2019). Similarly, other types of machine learning algorithms that differ from the artificial neural network and decision tree have been used in recently published studies (Bur et al., 2019; Chao et al., 2014; Exarchos et al., 2012; Lynch et al., 2017; Montazeri et al., 2016; Tapak et al., 2019). These algorithms have been used with clinicopathologic, histologic, molecular, genomic, image, demographical, epidemiological, and combinations of any of these data input types (heterogeneous sources of data), and labelled, unlabelled, and pseudo-labelled patient data (Aubreville et al., 2017; Chang et al., 2013; Exarchos et al., 2012; J. Kim & Shin, 2013; Lu et al., 2017; Shams & Htike, 2017; Sharma & Om, 2013; Tseng et al., 2015).

Most of these studies reported that machine learning techniques showed significant prognostic ability for cancer outcome (Ahmad & Eshlaghy, 2013; Ayer et al., 2010; Chang et al., 2013; Chen et al., 2014; Delen et al., 2005; Exarchos et al., 2012; Gevaert et al., 2006; W. Kim et al., 2012; K. Park et al., 2013; Rosado et al., 2013; Xu et al., 2019). Despite this reported accuracy in machine learning models, they have not been widely used in actual daily clinical practice (Arambula & Bur, 2020; Kourou et al., 2015). Of note, several limitations and concerns have been raised on the integration of these models into daily clinical practice.

Therefore, it is important to identify the possible drawbacks in the development of machine learning models. Some of these potential drawbacks include the collection of appropriate data samples. These data should be pre-processed to remove noise and distortions (Kourou et al., 2015). Likewise, sufficiently large data samples should be collected. Additionally, the training and testing phases of the

machine learning model development should be mapped out with a scientific-based experimental design (Kourou et al., 2015). That is, the model should be devoid of overfitting or underfitting and properly validated internally or externally.

3 AIMS AND OBJECTIVES

3.1 Aims of the study

The aims of this study were to offer accurate prognostication that can aid in personalized medicine of tongue cancer patients (including early-stage) using machine learning techniques. Specifically, the objectives of this study were:

A. To apply machine learning techniques that consider the shortcomings of the TNM staging to predict tongue cancer patients' outcomes such as locoregional recurrences and overall survival.

B. To evaluate the prognostic significance of some input parameters using machine learning techniques.

C. To provide a web-based prognostic tool for stratification of oral tongue cancer patients into a low- or high-risk for the occurrence of locoregional recurrence. Such an online tool can be an important step towards personalized medicine for oral tongue cancer patients.

D. To compare the performance of machine learning techniques to nomograms in the prognostication of outcomes (overall survival) for tongue cancer patients.

E. To consider ethical challenges and other factors that affect the implementation of machine learning models into daily clinical practice.

4 METHODS

4.1 Dataset for the study

Two different datasets were used for this study. The first dataset was a multi-institution dataset. The second dataset was obtained from the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH).

4.1.1 Multi-institution data

The multi-institution data consist of retrospective data from the five (5) Finnish University Hospitals (Helsinki, Oulu, Turku, Tampere, and Kuopio) and at the A.C. Camargo Cancer Center, Sao Paulo, Brazil. From these multi-institutions, a total of 311 cases were obtained.

4.1.2 Surveillance, Epidemiology, and End Results (SEER) Program Data

The Surveillance, Epidemiology, and End Results (SEER) Program is one of the largest cancer database that is available to the public (SEER, 2012). In addition, it gives non-identifiable information on cancer statistics of the United States population (SEER, 2012). These important features makes it good choice for this study. The database is managed by the National Cancer Institute through the National Institutes of Health (NIH).

4.2 Ethical permission

The ethical permissions to use this multi-institution data was granted by the Finnish National Supervisory Authority for Welfare and Health (VALVIRA) and by the Brazilian Human Research Ethics Committee. Similarly, the use of the SEER database was granted via the user identification numbers of 10455-Nov2108 and 11522-Nov2019, respectively.

4.3 Selection of attributes

For the analysis that involved the use of the multi-institution data, the following attributes were selected and used; age at diagnosis, gender (sex), grade, World Health Organization (WHO) grade, the American Joint Committee on Cancer (AJCC) Tumor staging scheme, tumor budding, tumor depth, worst pattern of

invasion (WPOI), lymphocytic host response (LHR), perineural invasion (PNI), treatment (adjuvant [chemo] radiotherapy and/or surgery), neck treatments, survival times (in months), and overall survival status [Study I, II, and III] (Table 1). The parameters and their respective definitions is presented in Table 1.

Table 1. The histopathological parameters and their definitions

Variable	**Categories**	**Definitions**
WHO grade	Grade I	Well-differentiated tumor
	Grade II	Moderately-differentiated tumor
	Grade III	Poorly-differentiated tumor
Tumor budding	None	No tumor budding
	Low	Tumor has < 5 buds
	High	Tumor has >= 5 buds
Depth of Invasion (DOI)	Superficial	Tumor is < 4mm in depth
	Deep	Tumor with >= 4mm in depth
Worst pattern of invasion (WPOI)	Type 1-3	Pushing border; Finger-like growth; Large tumor islands
	Type 4	Small tumor islands
	Type 5	Tumor satellites
Lymphocytic host response	Type 1	Strong
	Type 2	Intermediates
	Type 3	Weak
Perineural invasion (PNI)	Absent	PNI: Not observed
	Present	PNI: Observed

Likewise, for analysis that involved the SEER database, the attributes to be extracted from the database depends on the attributes used in the literature reviews, that is, in the published articles. For example, in comparing machine learning models to nomogram, the extracted attributes were based on the nomogram that was used for comparison (Li et al., 2017). These attributes were age at diagnosis, race, marital status, grade, Tumor Nodal Metastasis (TNM) status according to the American Joint Committee on Cancer 7th edition, treatment (surgery, and radiotherapy) (Li et al., 2017). The survival period (in months) and overall survival status of the patients were also extracted. The detailed definitions of these attributes can be found in the SEER attribute documentation.

4.4 Machine learning techniques

In this thesis, a supervised machine learning method was used. Both MATLAB version R2015b (Study I) and Microsoft Azure Machine learning studio (Study II & III) was used.

A total of 311 cases (224 Finnish, 87 Brazilian) were included in study I to predict locoregional recurrences in early-stage oral tongue squamous cell carcinoma (OTSCC) as shown in Table 1. These histopathological parameters were used in the training of the artificial neural network.

Table 2. The summary of histopathological parameters

Variable	Categories	Total
WHO grade	Grade I	105
	Grade II	131
	Grade III	75
Tumor budding+	None	114
	Low	102
	High	75
Depth of Invasion (DOI)	Superficial	116
	Deep	195

Worst pattern of invasion (WPOI)	Type 1-3*	78
	Type 4	190
	Type 5	43
Lymphocytic host response	Type 1	53
	Type 2	116
	Type 3	142
Perineural invasion (PNI)	Absent	269
	Present	42

+ Tumor budding is considered as a single cancer cell or cancer cluster of four cancer cells or less

* Type 1-3 of worst pattern of invasion were considered as one group.

The process of building a reliable predictive machine learning model for precision and personalized medicine begins with the selection of appropriate data and the corresponding attributes. Of note, the selected attributes were mentioned in **Section 4.2** of this thesis (for multi-institution data). In terms of the output, locoregional recurrences was considered as the output parameter. Furthermore, the selected attributes were pre-processed to remove missing, corrupted, or not a number (NaN) entries in the data. Then, the preprocessed data can now be safely used in the machine learning analysis to produce reliable machine learning models. The machine learning process is given in Figure 23. The pre-processed data are considered as the data (available data) to be used in the machine learning process.

As a rule of thumb, the available data were divided into training, validation, and testing datasets (sub-section 2.4.5). The data division ratio depends on many factors, in this case, the data division ratio was 70% training, 15% validation, and 15% testing sets (Jeong et al., 2013; Puri et al., 2016). Considering the selection of the machine learning algorithm, the artificial neural network was selected as the algorithm of choice. MATLAB has an inbuilt function for the training of the artificial neural network. Therefore, the network was trained using the *patternnet* function.

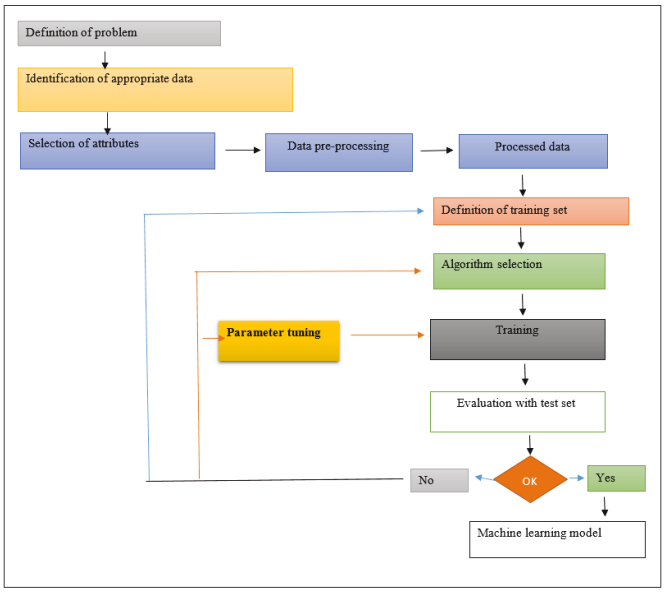

Figure 23. Machine learning process.

It creates a standard solution neural network that classifies inputs into a target. The final process involves training the neural network algorithm for prediction. For training of the neural network, scaled conjugate gradient backpropagation was used. In addition, the performance of the network was evaluated cross-entropy. The accuracy and area under the receiving characteristic curve gives the overall performance of the network on the test set (Figure 22). In case the performance of the trained network is not convincing enough, the data division ratio of the training data can be modified, the algorithm may be changed, or the training parameters may be tuned to produce a better model.

4.5 Comparison of machine learning algorithms

In the quest to produce an effective predictive model for precision and personalized medicine, several machine learning algorithms were compared to predict locoregional recurrences. Of note, there are numerous machine learning algorithms. However, only the most widely used algorithms were used for comparison. Thus, the performance of the artificial neural network model was compared with logistic regression in terms of accuracy (Study I). Similarly, the performance of four supervised machine learning algorithms were compared to predict locoregional recurrences. These algorithms were support vector machine,

Naïve Bayes, Boosted Decision Tree (BDT), and Decision Forest (DF) algorithms (Study II).

4.6 Comparison of machine learning algorithms with a nomogram

The nomogram used for comparison was constructed in a study that was previously published (Li et al., 2017). It was used to predict 5- and 8-year overall survival in OTSCC. The nomogram is presented in Figure 24-25, respectively.

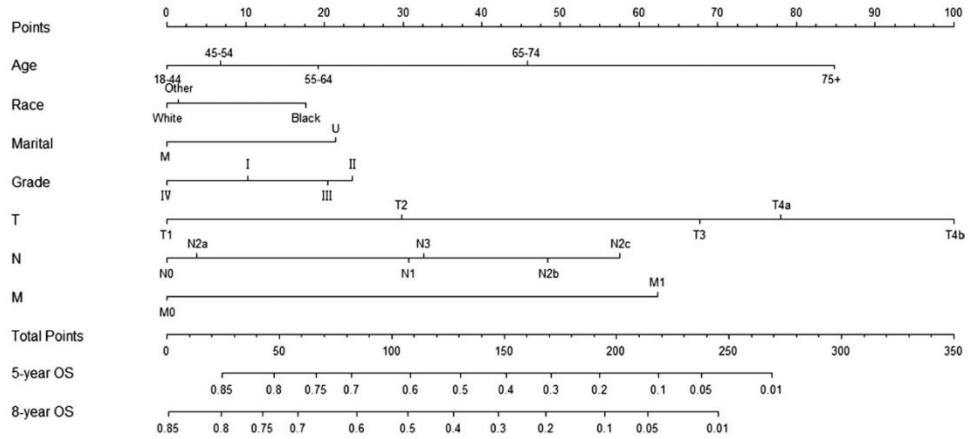

Figure 24. Nomogram to predict 5- and 8-year overall survival with surgical treatment (Li et al., 2017)

The nomogram presented in Figure 24 was well calibrated and validated to predicting the 5- and 8-year overall survival for patients that received surgery.

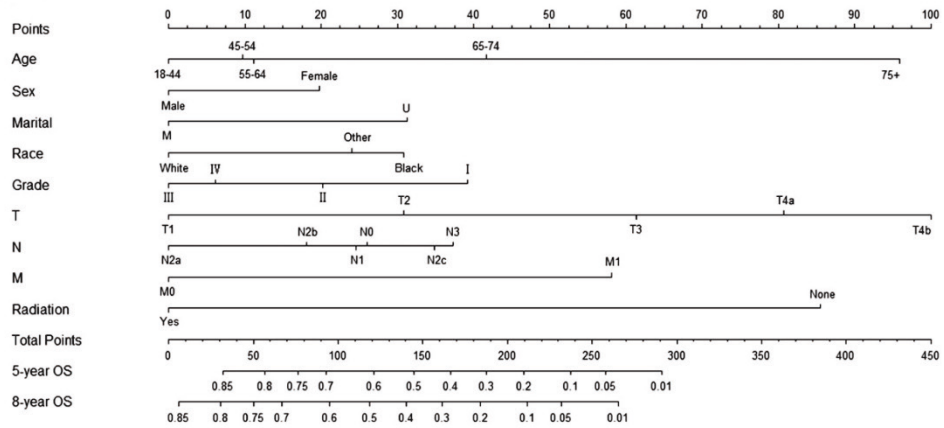

Figure 25. Nomogram to predict 5- and 8-year overall survival with radiotherapy (Li et al., 2017).

Interestingly, the nomogram was constructed in such a way that two distinct nomograms were constructed to differentiate between the TSCC patients that received surgical treatments and those that received radiotherapy.

The machine learning algorithms used for comparison with the nomogram were logistic regression, Naïve Bayes, support vector machine, neural network, boosted decision tree, decision forest, and decision jungle (**Study III**). In addition, the SEER data were used in the development of the machine learning model and construction of the nomogram. Of note, the November 2015 submission between 1973 and 2013 was used because the nomogram was built using the data of the same year.

A total of 7649 cases were extracted from the database. Out of these cases, the last 53 cases were reserved to further provide external validation for the trained machine learning model (**Study III**). Thus, a total of 7596 cases were used in the machine learning analysis. The baseline demography and tumor characteristics of the TSCC patients extracted from the SEER database are given in Table 3.

Table 3. Baseline demographic and tumor characteristics of patients extracted from the SEER database

Variables	**Overall survival, N = 7596 Training and testing cohort**	**Overall survival, N = 53 External validation cohort**
Age at diagnosis (years): Age of the patients at the time of diagnosis		
1 – 18	5 (0.1%)	0 (0.0%)
19 – 44	515 (6.8%)	2 (3.8%)
45 – 54	1412 (18.6%)	14 (26.4%)
55 – 64	2497 (32.9%)	18 (34.0%)
65 – 74	1877 (24.7%)	13 (24.5%)
75+	1290 (16.9%)	6 (11.3%)
Ethnic origin: This implies race and ethnicity of the patient. Other includes American Indian (native), Asian/Pacific Islander		
White	6597 (86.8%)	44 (83.0%)
Black	516 (6.8%)	9 (17.0%)
Other*	483 (6.4%)	
Sex: The biological sex		
Male	5322 (70.0%)	38 (71.7%)
Female	2274 (30.0%)	15 (28.3%)
Marital status: The marital status of the patient at the time of diagnosis of OTSCC		
Married	4430 (58.0%)	29 (54.7%)
Unmarried	3166 (42.0%)	24 (15.3%)
Grade: The differentiation of the cancer cell		
Grade I	1215 (16.0%)	8 (15.1%)
Grade II	3768 (49.6%)	29 (54.7%)
Grade III	2543 (33.5%)	15 (28.3%)
Grade IV	70 (0.9%)	1 (1.9%)
T stage (2010+): The measurement of the dimension of the tumor		
T1 (< 2cm or less)	2942 (38.7%)	19 (35.8%)
T2 (>2cm & ≤ 4cm)	2492 (32.8%)	16 (30.2%)

T3 (> 4cm)	1159 (15.3%)	6 (11.3%)
T4a (Moderately advanced)	920 (12.1%)	11 (20.8%)
T4b (Significantly advanced)	83 (1.1%)	1 (1.9%)
N Stage (2010+): Lymph node metastasis		
N0 (No regional lymph node metastasis)	3485 (45.9%)	14 (26.4%)
N1 (Single node)	1220 (16.1%)	10 (18.9%)
N2a (Cancer has spread)	327 (4.3%)	5 (9.4%)
N2b (Multiple node)	1498 (19.7%)	14 (26.4%)
N2c (Lymph node in the neck)	880 (11.6%)	9 (17.0%)
N3 (Spread to one or more neck lymph nodes)	186 (2.4%)	1 (1.9%)
M stage (2010+): The presence of distant metastasis		
M0 (No distant metastasis)	7425 (97.7%)	50 (94.3%)
M1 (Distant metastasis)	171 (2.3%)	3 (5.7%)
Surgery performed: This describes if surgery was performed		
Yes	4654 (61.3%)	22 (41.5%)
None	2942 (38.7%)	31 (58.5%)
Radiotherapy: This is the indication of whether the patient has received radiation		
Yes	4489 (59.1%)	37 (69.8%)
None	3107 (40.9%)	16 (30.2%)
Overall survival status: This indicate of the patient was alive or died		
Alive	5743 (75.6%)	47 (88.7%)
Dead	1853 (24.4%)	6 (11.3%)

The Microsoft Azure machine learning studio was used to build the machine learning model. The process was similar to the training phase described in Section 4.3 and Figure 23. First, all the available data were converted to numeric data, for easy machine learning training, and uploaded into Azure machine learning studio. The data were divided into training and testing sets. In addition, the synthetic minority oversampling technique was used to handle possible bias in the target output (Blagus & Lusa, 2013) and hyperparameters were fine-tuned to maximize the performance of the model. Each algorithm of interest was configured and the training was done using cross-validation. The performance of each of the algorithms was noted (Microsoft Azure Machine Learning Studio, 2018). The algorithm that produced the best performance in terms of accuracy and area under the receiving operating characteristic curve was used for comparison with the nomogram (**Study III**). This comparison was carried out using the external validation test (53 cases reserved for external validation) [**Study III**].

4.7 Systematic review of studies that applied machine learning in oral cancer (study V)

The keywords used were “oral cancer” AND “machine learning”. Additionally, the search word was extended to include [(“oral cancer”) AND (‘artificial neural network’ OR ‘ensemble method’)]. These words were searched in Scopus, PubMed, Web of Science, OvidMedline, and Institute of Electrical and Electronic Engineers (IEEE) databases. The flowchart for the search process is presented in Figure 26.

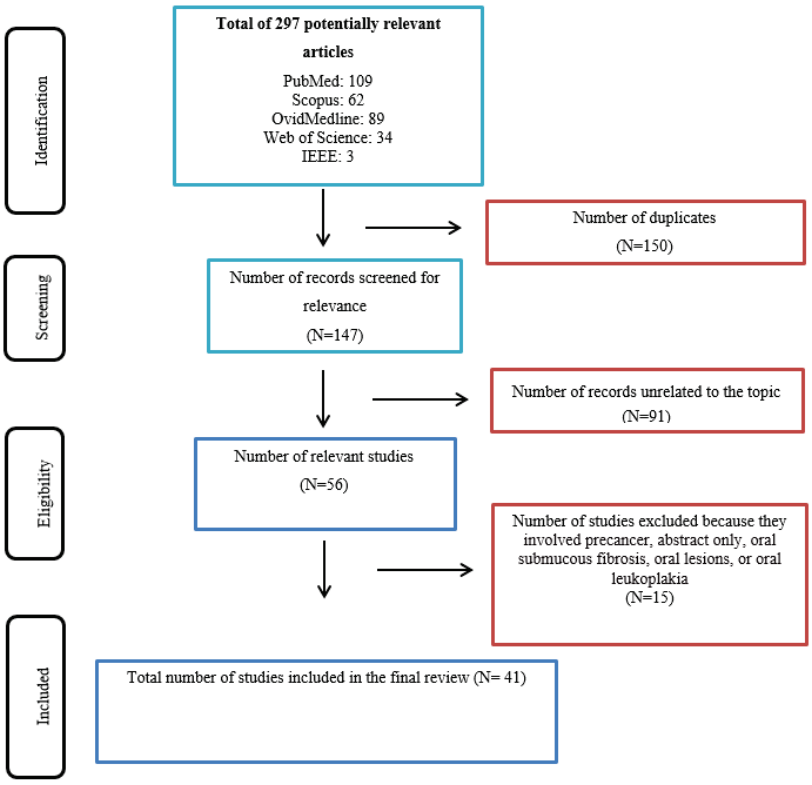

Figure 26. Flowchart of database search (**study V**)

As shown in the flowchart, the search was limited to articles written in the English language. Additionally, the search date for the databases were from inception until the end of February 2020. The details of the inclusion and exclusion criteria are as shown in Figure 26. The study aimed at examining the machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for the future. It seeks to examine why the machine learning models have made few contributions in actual daily clinical practice despite the touted benefits of artificial intelligence, or its subfield, machine learning in medicine.

4.8 Addressing ethical challenges related to the application of machine learning in oral tongue cancer: (**study IV**)

The application of machine learning techniques for personalized medicine have been touted to offer benefits such as precision medicine, improved prognostication and the overall survival of OTSCC patients. Despite these benefits, ethical issues have been raised in some quarters. The literature was systematically reviewed to examine these concerns and how they can be addressed.

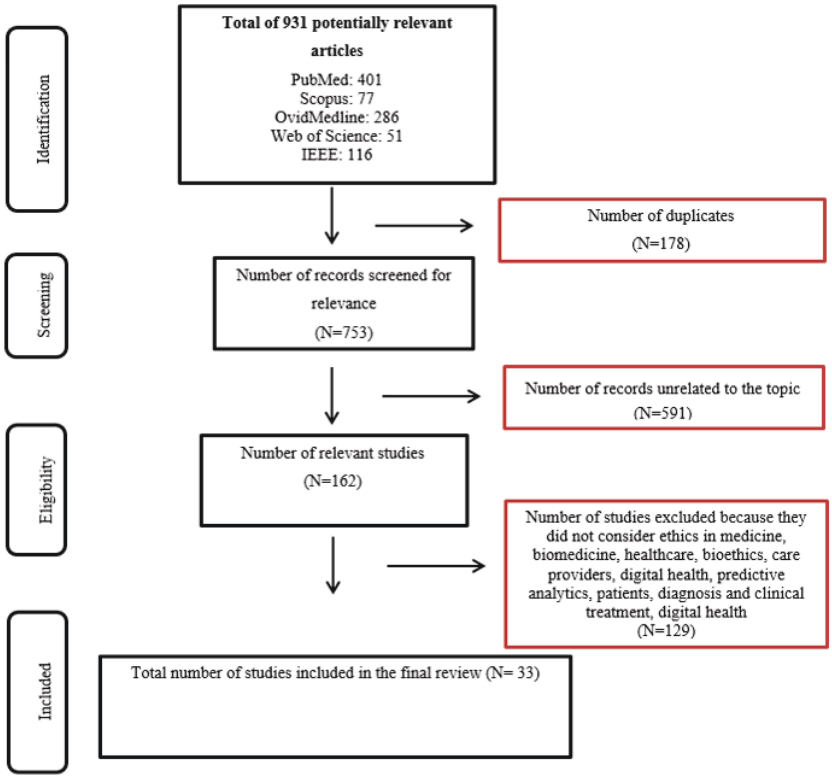

Figure 27. Flowchart for the database search on ethical challenges of the machine learning model in medicine (**study IV**).

The following databases; OvidMedline, PubMed, Scopus, Web of Science, Cochrane Library, and Institute of Electrical and Electronic Engineers (IEEE) databases were searched with keywords [('machine learning OR artificial intelligence') AND ('ethics')]. The search word was general. However, the ethical challenges mentioned were analyzed to examine how they relates to oral tongue cancer. The objective is to examine the ethical challenges to the integration of machine learning model into daily clinical practice of oral management, as well as

how these ethical challenges can be addressed. Moreover, flowchart towards a successful integration of machine learning in daily clinical practice was proposed.

5 RESULTS

The artificial neural network showed promising results in predicting locoregional recurrences in early-stage oral tongue squamous cell carcinoma (OTSCC). The overall accuracy of 92.7%. Other performance metrics were also evaluated: recall (sensitivity) was 71.2%, specificity (98.9%), and positive and negative predictive values were 97.7% and 84.5%, respectively.

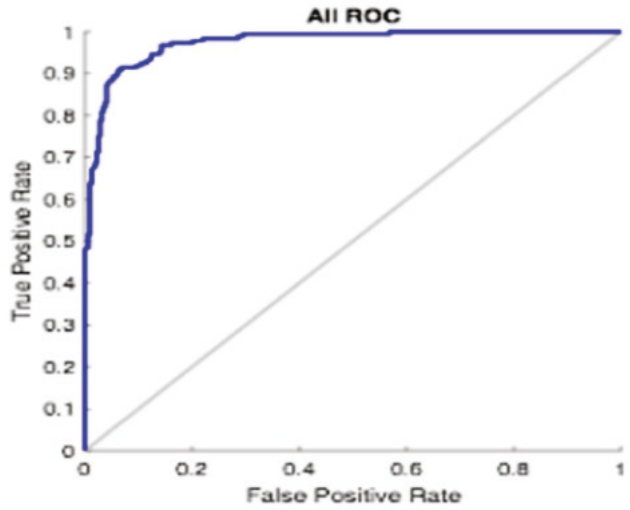

Figure 28. The area under characteristics curve of the trained neural network

Similarly, the overall receiving operating characteristic curve (AUC of ROC), which is also the C-statistics (C-index) value was 97.3%. Comparing the trained neural network with traditional methods such as the logistic regression model, the neural network outperformed the logistic regression. While the neural network gave an overall accuracy of 92.7%, the logistic regression model which gave an accuracy of 86.5%.

5.1 Comparison of machine learning algorithms to predict locoregional recurrences

Four widely used algorithms were compared to predict locoregional recurrences in early stage oral tongue cancer. These algorithms were support vector machine, naive Bayes, decision forest, and boosted decision. The basic four thresholds from the training process are given in Figure 29. These thresholds are true positive (TP), false positives (FP), true negatives (TN), and false negatives (FN).

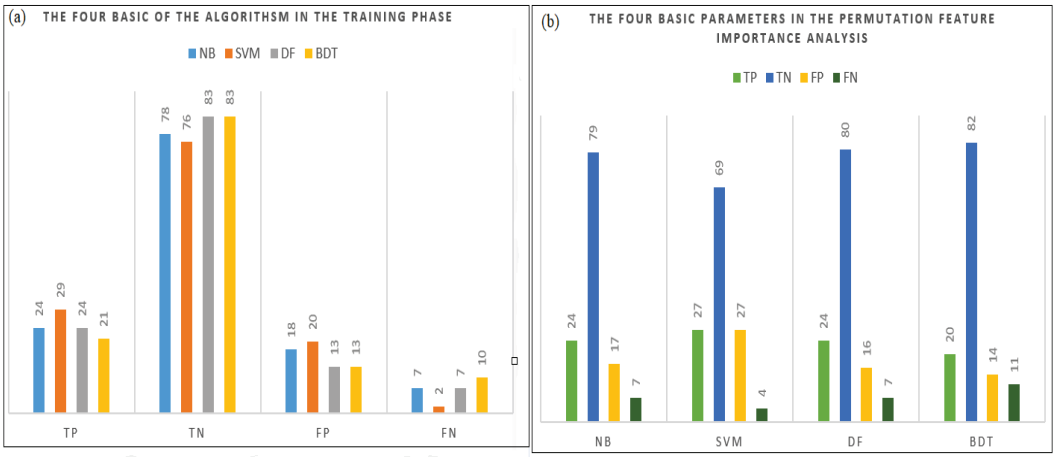

Figure 29. The four basic thresholds after the training phase of the compared algorithms.

Based on these thresholds, the overall accuracy of the compared algorithms after the training process were: support vector machine (82.7%), naive Bayes (80.0%), decision forest (84.0%), and boosted decision tree (82.0%). Therefore, the decision forest outperformed the other compared algorithms. In addition, naïve Bayes and decision forest showed the best area under the receiving operating characteristic curve of 0.89 each.

5.2 External validation algorithms to predict locoregional recurrences

The machine learning algorithms were validated externally with new cases (59 external validation cases) that were not used in the training. The artificial neural network gave an accuracy of 81.4% (study I). Additionally, the overall accuracy of the compared algorithms with the external validation cohorts were (study II): support vector machine (68%), naive Bayes (70%), decision forest (78%), and boosted decision tree (81%). Similarly, the algorithm that showed the best accuracy with external cohorts outperformed the traditional method of depth of invasion (DOI) which gave an accuracy of 63% (study II).

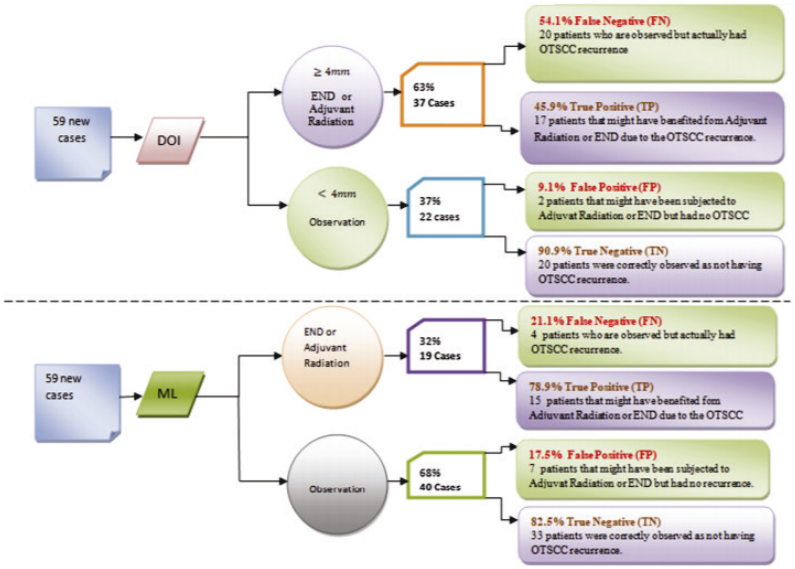

Figure 30. Comparison of depth of invasion model with machine learning model

5.3 Feature importance of the parameters to predict locoregional recurrences

The artificial neural network model identified tumor budding and depth of invasion as the most important prognosticators to predict locoregional recurrence (study I). Similarly, the boosted decision tree identified tumor budding, depth of invasion in addition to age, perineural and worst pattern of invasions as important parameters for the model to predict locoregional recurrences in early-stage oral tongue cancer (study II).

5.4 A web based tool to predict locoregional recurrences

In studies I & II, the artificial neural network outperformed all the compared algorithms in terms of accuracy when externally validated with new cohorts. The trained neural network model was integrated as a web based tool using Microsoft Azure cloud service. The web-based tool is freely available. Users of the website can enter the prognostic factors to generate a personalized estimation of locoregional recurrence for the patient. The web-based tool is available: https://predictrecurrence.azurewebsites.net/Default.aspx

5.5 Comparison of machine learning algorithm with a nomogram (study III)

The machine learning algorithm (boosted decision tree) performed better than the nomogram in predicting overall survival in patients with tongue cancer. When these two approaches were compared using the external validation cases mentioned in this thesis, the machine learning-based algorithm showed an accuracy of 88.7% while the nomogram (with surgical treatment) showed 66.0%, and the nomogram (with radiotherapy) produced an accuracy of 60.4%.

5.6 Ethical challenges of machine learning model in cancer management (study IV)

Based on the systematic review, privacy and confidentiality of patients' data, bias in the model, peer disagreement, responsibility gap, client-patient relationship, and patients' autonomy were the ethical challenges identified to the use of a machine learning model in daily clinical practices.

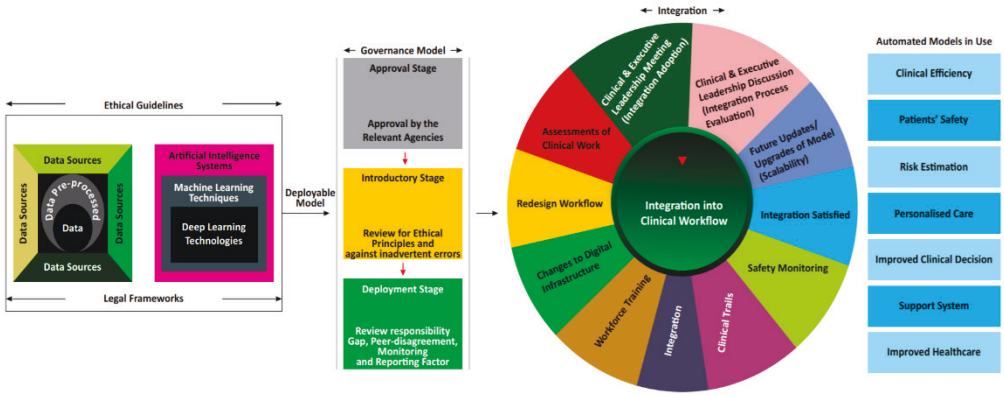

Figure 31. Proposed framework for smooth integration of machine learning

Furthermore, to enhance hitch-free integration of a machine learning model into daily clinical practice of oral cancer management, a framework for smooth integration has been proposed as shown in Figure 31.

5.7 Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future (study V)

The artificial neural network and support vector machine were the most widely used machine learning algorithms in prognostication studies in oral cancer. The accuracy of the machine learning algorithms examined in the reviewed paper ranged from 63.4% to 100.0%. The clinical concerns of these machine learning models include an explainable model (model interpretability), how the model produced the result (result interpretability), concern about rendering the oncologists less important in the management of oral cancer patients, and privacy and confidentiality concerns. For the future, it is important that the machine learning model to be used in daily clinical diagnosis should provide quality explanations. The quality of explanations may be measured using the system causability scale (SCS) (Holzinger et al., 2020).

6 DISCUSSION

Early-stage oral tongue cancer (cT1-2N0M0) is characterized by a high risk of recurrences (locoregional), occult nodal metastases, and cancer-related mortality. Traditionally, the American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) staging system has been shown to be an objective and accurate tool for predicting the prognosis for an entire population of cancer patients. Thus, TNM offers an important approach for planning effective strategies for oral tongue cancer patients (American Joint Committee on Cancer, 2002; Li et al., 2017). Despite the promising prospects shown by the TNM staging scheme, it has been criticized as ineffective for individualized prediction of outcome. Also, it fails to consider other tumor- and patient-related risk factors (S. G. Patel & Lydiatt, 2008; Sobin, 2003). Consequently, this leads to inappropriate management of oral cancer. To this end, a tool that considers these factors together to accurately predict patients' outcomes would be pertinent (Li et al., 2017).

Of note, clinicopathologic markers are used by clinicians to make decisions regarding cancer patients. Interestingly, numerous prognostic markers have been published in various studies for oral tongue cancer prognosis. Unfortunately, most of these studies have shown certain methodological concerns such as a mixture of tumor sites (including all oral cavity subsites), single institution consideration, and cumbersome protocol (immunostaining consideration).

In this study, we have attempted to overcome such obstacles by introducing a subfield of artificial intelligence, machine learning techniques that combines the readily available prognosticators (i.e. that could be easily included in the pathology reports) for the prediction of locoregional recurrences and survival assessments of tongue cancer patients.

The artificial neural network and boosted decision tree machine learning algorithms showed promising performance in the stratification of patients into low- and high-risk recurrence and low- and high-chance of survival of tongue cancer, respectively. The neural network works in a way similar to that of the human brain by analyzing the dataset used in the training and recognizing patterns that could be used to make meaningful inference from a new set of data. In addition, the neural network is able to build a nonlinear statistical model to examine biological systems. Therefore, in this study, a feedforward neural network type was used where enough neurons were used in the hidden layer. Also, the performance of the network was computed using cross-entropy. The cross-entropy ensured that outputs that are inaccurate were heavily penalized with little or no penalty for the accurate classifications. Hence, the trained neural network showed

good classification ("low-risk" or "high-risk" of recurrence) performance. The fact that the neural network was found to be posited for effective stratification of patients was corroborated by the studies of Faradmal et al., and Kazemnejad where ANN was found to show improved predictive performance over the traditional methods of logistic regression and log-logistic regression (Faradmal et al., 2014; Kazemnejad et al., 2010).

The importance of accurate stratification of patients as either "low-risk" or "high-risk" of recurrence lies in the fact that effective and informed decision of multimodality treatment can be taken for those cases at high risk although they are diagnosed at an early stage (Alabi et al., 2019). Additionally, high-risk cases may benefit from elective neck dissection (END) and postoperative oncological therapy. Also, an enhanced post-treatment follow-up program can be effectively tailored to the high-risk patients (Alabi et al., 2019).

Besides the neural network, the boosted decision tree algorithm showed an encouraging performance in the prediction of locoregional recurrences in oral tongue cancer patients. The performance of the boosted decision tree algorithm can be attributed to its ability as an ensemble method whereby it is able to create a fleet of algorithms with relatively similar bias and then combining their outputs to minimize variance (Alabi et al., 2019). Therefore, the boosted decision tree algorithm is poised to offer accurate prediction of outcome (low or high risk of recurrence) in early-stage oral tongue cancer. A similar result was obtained in other published studies where the decision tree outperformed other compared algorithms in the prognostication of other cancers (de Melo et al., 2018; Sumbaly et al., 2014; Tseng et al., 2015; B. Zhang et al., 2017).

6.1 Prognostic significance of the examined parameters

Of note, one of the major challenges in the effective treatment of patients with early oral tongue squamous carcinoma is finding the appropriate parameters that can stratify the patients into risk groups or offer an accurate prognosis. With this possibility, the incidence of treatment failure in patients with oral tongue cancer can be minimized (Safi et al., 2017). To this end, the permutation feature importance (PFI) evaluated the precise contribution of each parameter to the overall predictive ability of the machine learning algorithms. For the boosted decision tree to predict locoregional recurrence, tumor budding, depth of invasion, age at diagnosis, perineural invasion, and worst pattern of invasion were identified as important parameters (Alabi et al., 2019). Interestingly, training a machine learning algorithm with these identified parameters (tumor budding, depth of

invasion, age at diagnosis, perineural invasion, and worst pattern of invasion) produced a machine learning model with predictive performance as a model that includes other additional parameters such as grade, lymphocytic host response, stage of cancer, and gender (Alabi et al., 2019) (study II).

Therefore, it is important to ensure that the parameters to be used in developing the machine learning-based model are independent of each other, which prevents a collinearity problem of input parameters. In the training of the machine learning model used in our analysis, the input parameters were dissimilar (Figure 32).

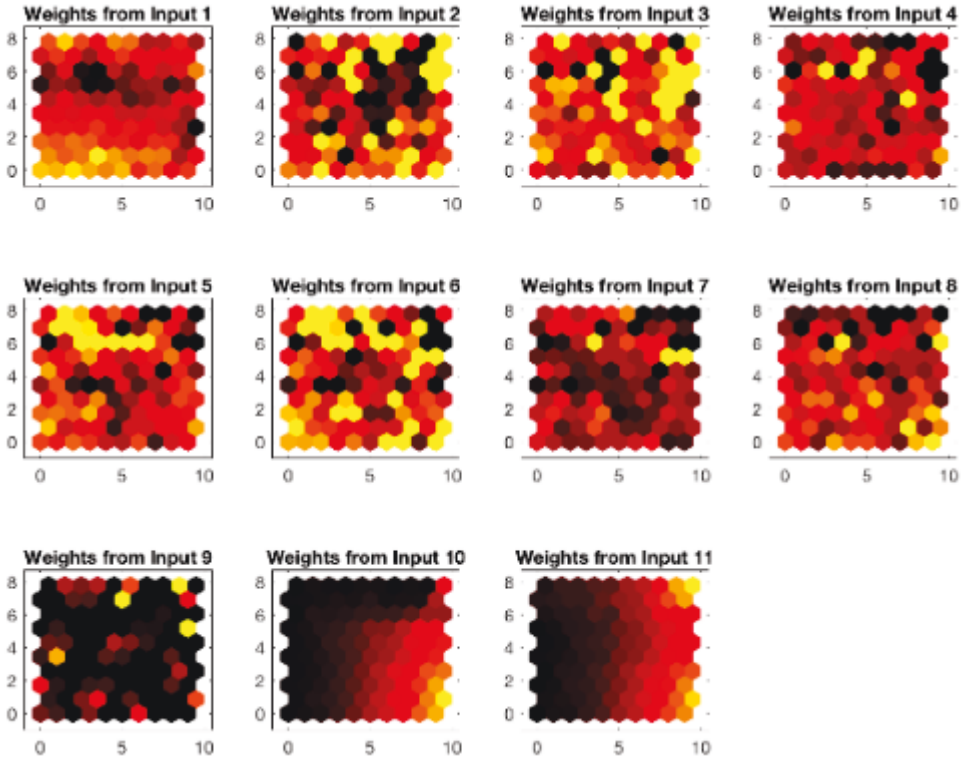

Figure 32. The heatmap of the input variables. (Input 1 = Age [Input 1], Gender [Input 2], Stage [Input 3], Grade [Input 4], Tumor Budding [Input 5], Depth [Input 6], Worst Pattern of Invasion [Input 7], Lymphocytic Host Response [Input 8], Perineural Invasion [Input 9], Disease free months [Input 10], Follow-up time [Input 11]).

Independent input parameters ensure that each input parameter has a unique effect on the target (output variable). For example, with independent variables, it makes it possible to perform a preliminary investigation on the ability of these variables to clearly and distinctly stratify the target variable into clusters as shown in Figure 33.

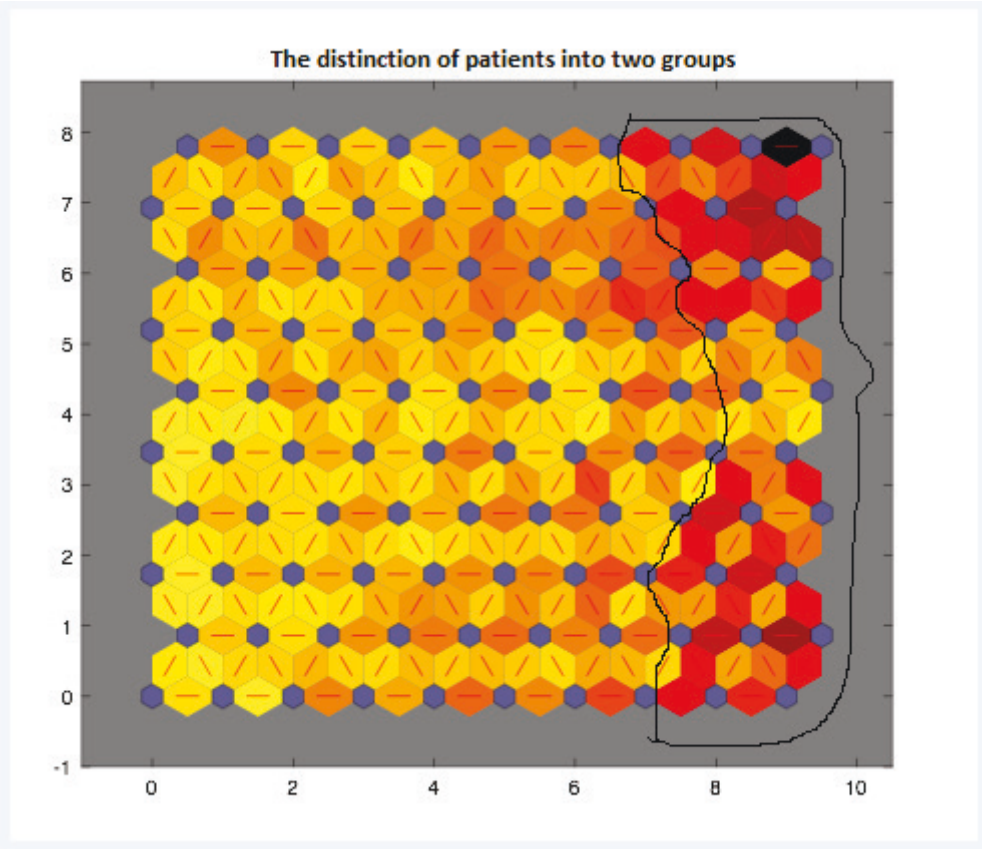

Figure 33. The weight distance matrix of input variables to form cluster.

As shown in Figure 33, a segment of the dark red band from the lower-right region to the upper right region partitions the potential risk groups associated with recurrence of oral tongue cancer.

Therefore, it is advisable to use prognostic markers that have been reported useful in the prognostication of that type of cancer. In the case of oral tongue cancer, Almangush et al., reported the prognostic importance of tumor budding, depth of invasion and worst pattern of invasion (A. Almangush et al., 2015; Alhadi Almangush et al., 2015, 2018). Similarly, perineural invasion was touted as a marker that showed promising prognostic significance (Yang et al., 2018). These findings were corroborated by the study of Arora et al. and others which underlined the prognostic significance of the selected parameters used in this thesis (A. Almangush et al., 2015; Arora et al., 2017; Ganly et al., 2015; Tai et al., 2012; Yang et al., 2018). Hence, to ensure that the machine learning-based model shows accurate predictive capabilities, it is important to ensure that the input parameters are devoid of collinearity issues, that is, they are independent, and also these input variables have been reported to have a significant influence in the prognostication of such cancers.

6.2 Comparison of a machine learning model with a nomogram

Traditionally, the experience of clinicians plays a significant role in estimating the risk. However, experience is also subjective and therefore represents a great risk of bias and risk estimation errors (Elstein, 1999; Kudo, 2019; Vlaev & Chater, 2006). The nomogram has been suggested to improve diagnosis and it has been used extensively in head and neck cancer (Ali et al., 2014; Cho et al., 2015; Ganly et al., 2015; Gross et al., 2008; Montero et al., 2014). Admittedly, nomograms had been shown to minimize risk estimates in cancer patients (Shariat et al., 2008). However, the emergence of technologies such as machine learning techniques have been shown to provide improved risk estimation for patients (Alabi, Elmusrati, Sawazaki-Calone, et al., 2019; Alabi, Elmusrati, Sawazaki-Calone, et al., 2019; Bur et al., 2019; Karadaghy et al., 2019) (study III). Of note, the improved predictive accuracy exhibited by this model is important in the proper management of cancer for personalized medicine (Cruz & Wishart, 2007).

The machine learning model outperformed the nomogram because it was able to identify and understand the hard-to-discern relationships between the input variables. Despite the improved risk estimation of the machine learning model, a concern relating to model and result interpretability exist. That is, how explainable is the model? In contrast, the nomogram provides a transparent, non-computer dependent, graphical, and appealing approach to estimating the risk of patients. These important features of the nomogram are also worthy of consideration in clinical decision making despite the lower predictive accuracy compared to the machine learning techniques. This is because these important characteristics offered by the nomogram address the concern that the results from machine learning models are not easily interpretable.

Based on the improved risk estimation offered by the machine learning model and also the transparency provided by the nomogram, our study proposed a hybrid approach known as a Nomogram-Machine learning model (NomoML) for the effective and accurate risk estimation of patients. This hybrid approach is intended to provide both transparency and improved risk estimation in cancer management. The transparency feature is important to ensure that the shared decision-making between the patient and clinician can be strengthened. Similarly, improved risk estimation gives confident to the patients regarding the recommended treatment approach. Therefore, this hybrid approach (NomoML) is posited to offer individualized assessment and proper recommendation of the most appropriate adjuvant treatment for tongue cancer patients.

6.3 Web-based tool towards personalized medicine

The artificial neural network performed better than all other algorithms explored in our studies (studies I, II, and III). With the web-based tool developed in this study (https://predictrecurrence.azurewebsites.net/Default.aspx), medical treatment tailored to the needs of an individual can be fulfilled. The methodological approach fulfilled the vision of personalized medicine because there was an integration of retrospective patients' with evidence-based prognostic markers for risk estimation of patients using machine learning techniques, a subfield of artificial intelligence. Our web-based prognostic tool seeks to offer personalized medicine to the patients through an accurate risk estimate that enhances effective treatment planning and informed clinical decision making.

6.4 Ethical concerns of machine learning models in medicine

Despite the promising performance of machine learning models in effective cancer management, certain ethical concerns are raised across different quarters (ethicists, clinicians, patients, agencies, human right activists, and government). Some of these ethical concerns include data privacy and confidentiality (Arambula & Bur, 2020; Ma et al., 2019; Nabi, 2018), peer disagreement (contradictory risk estimation between the clinician and the model), reduced patients – clinicians' relationship, and possible lack of shared decision making between the patients and the clinicians on the type of treatment plan.

Regarding data privacy and confidentiality, various hospitals have embraced the digitalization of health data, especially, the electronic health records (EHR). These records contain the details of the patients, their ailments, clinicopathological parameters, genetic information, treatments, and outcomes. Therefore, the patients consent should be adequately sought (Geis et al., 2019; Powles & Hodson, 2017). This is because the development of machine learning models involves significant usage of the patient's data. Therefore, informed consent of the patients regarding the potential usage of their data seeks to prevent illegal use of their data and privacy breaches (Bali et al., 2019; Balthazar et al., 2018; Char et al., 2018; Nabi, 2018; Powles & Hodson, 2017; Yuste et al., 2017). Hospital managements should have a standardized data use agreement mechanism (Kohli & Geis, 2018). Additionally, a modern and secure scheme to prevent data privacy violations can be introduced (Geis et al., 2019; Y. Liu et al., 2017; Ma et al., 2019; Vayena et al., 2018; G. Wang et al., 2015; X. Zhang et al., 2018). Of note, data privacy and

confidentiality usually lead to an important issue regarding data ownership. However, this issue (data ownership) is beyond the scope of this dissertation.

Another important ethical concern is the trustworthiness of the machine learning model (Figure 34). This concern begins with how transparent is the machine learning model? In addition, possible error/malfunctioning of the model such as data imbalance in the training should be clearly mentioned (England & Cheng, 2019; Park & Kressel, 2018; Vayena et al., 2018; Zou & Schiebinger, 2018) to give transparency to the model and consequently, the results from these models (Geis et al., 2019; S. H. Park et al., 2019).

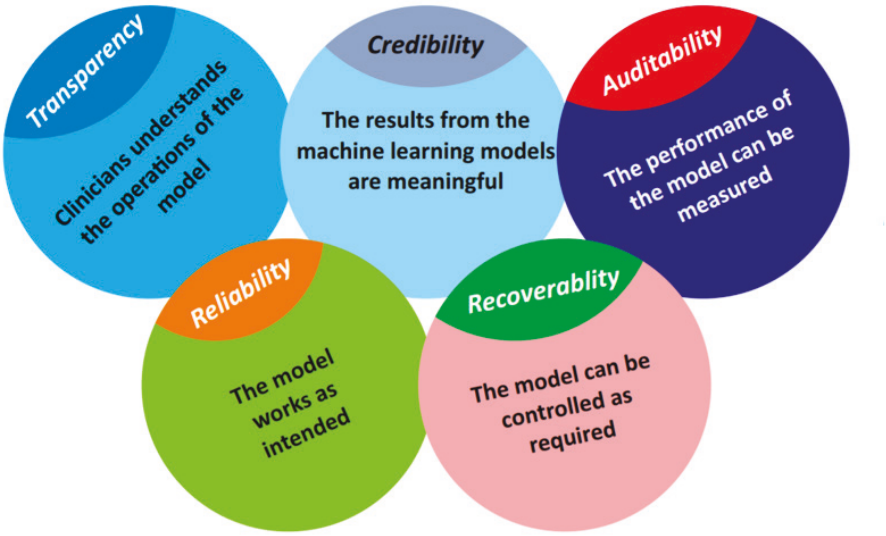

Figure 34. The trustworthiness principles expected from a machine learning model

Therefore, the trustworthiness principles expected from a machine learning model include transparency, credibility, auditability, reliability, and recovery (Figure 34). Essentially, it is important to follow the guidelines related to transparent reporting (Bossuyt et al., 2015; Collins et al., 2015; England & Cheng, 2019). Bearing in mind the concern of transparency and a trustworthy machine learning model, any machine learning model to be used in daily clinical practice for prognostication should uphold the fundamental pillars of medical ethics (autonomy, beneficence, nonmaleficence and justice) (Arambula & Bur, 2020) and the ethical principles of transparency, credibility, auditability, reliability and recoverability (Keskinbora, 2019) (Figure 34).

Peer disagreement is another ethical concern (Christensen, 2007; Kelly, 2010). That is, what happens when the machine learning model and the clinician have contrasting opinions regarding the estimated risk for a patient (Frances & Matheson, 2018)? It is impossible to have a dialogical engagement with the model, as proposed in the argumentative theory of reasoning (Mercier & Sperber, 2017). Should the clinician follow the risk estimate given by the ML model (Christensen, 2007) or adhere to his/her self-convinced estimates (Enoch, 2010)? Therefore, there is a challenge regarding the decision to be made in this scenario. To address this concern, an ethical guidelines and legal frameworks to guide the usage of machine learning models in daily clinical practice become imperative (Figure 35).

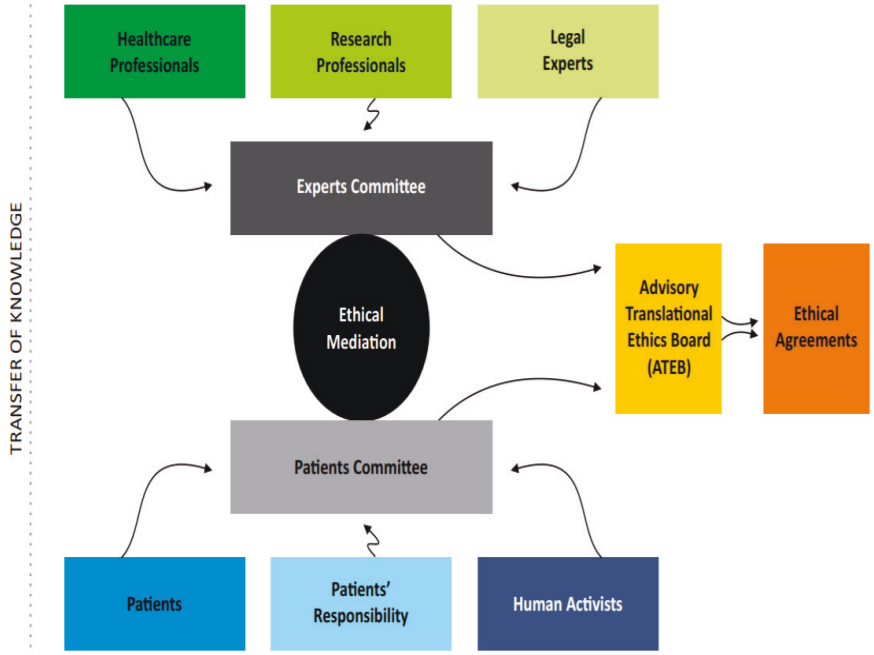

Figure 35. Ethical and legal frameworks for ethical agreements.

The ethical guidelines are expected to detail what to do in such types of situations and address the responsibility gap, i.e., who is to be held responsible if the estimate from the model is wrong or when the model malfunctions. Also, the hospital and medical professional guidelines should be considered in the development of the ethical guidelines for the application of the machine learning model in daily clinical practices.

Moreover, the patients' autonomy and participation in shared decision making should not be violated with the use of a machine learning based model (Grote &

Berens, 2020). Usually, the discussion in terms of the risk estimates and treatment plan is a two-way conversation (face – to – face) between the clinician and the patient. Considering the possible introduction of machine learning models in daily clinical practices, the need to change from a two-way conversation to a triangular-type of conversation (clinician – model – patient) appears inevitable (Figure 36). That is, in a two-way conversation, the clinicians can take into consideration other important factors regarding the patient to suggest a treatment plan that could minimise the suffering of the patient. However, with a machine learning-based tool, the major consideration for the model is usually to maximize the life span and overall survival of the patient, thereby making this model paternalistic in nature. In this case, the suffering and other important conditions regarding the patients are not considered by the machine learning model in the determination of treatment plan. This generally triggers the ethical concern of a shared decision making between the clinician and the patient (McDougall, 2019). Therefore, it is pertinent to establish relevant standards to determine which information from the machine learning model is important to be explained to the patient during the shared decision making process. Additionally, the patients should be duly informed regarding the use of a machine learning-based model in the decision making (Grote & Berens, 2020; McDougall, 2019; Mittelstadt & Floridi, 2016).

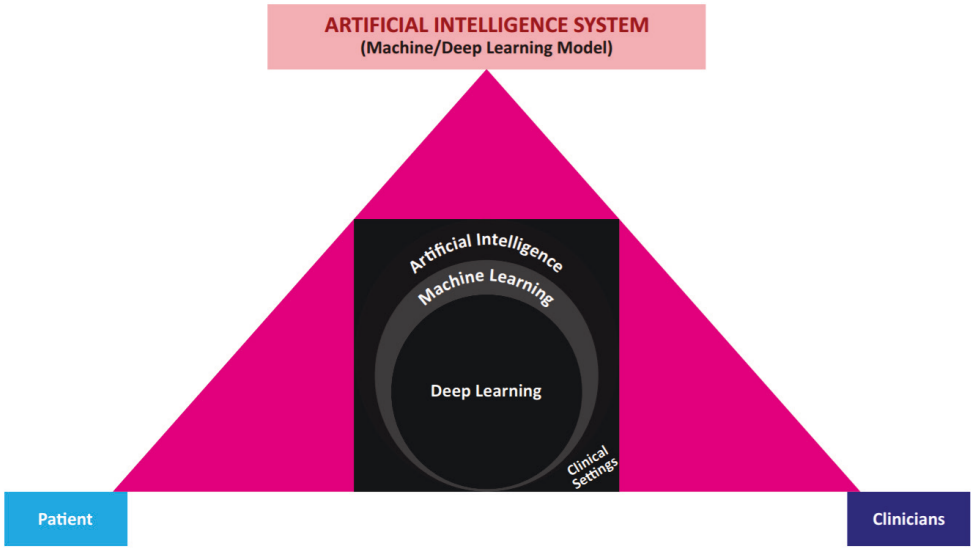

Figure 36. Shared decision making between patients and clinicians.

The possible change of paradigm to a patients–models–clinicians (three-way diagnostic procedure) raises the concern of humanness of clinicians and their role of cognitive empathy, trust, responsibility and confidentiality among clinicians (Boers et al., 2020). Will the clinicians still have empathy for the patients? Will the

role of the clinicians regarding confidentiality and responsibility be compromised? In addition, self-medication and self-management become serious concerns when the models are made available to the public. Therefore, the fundamental aspects of patient care may be affected (Boers et al., 2020). To mitigate these concerns, the ethical guidelines should be developed to guarantee patients' empathy and maintain patient – clinician relationships. This is because such relationships have been reported to positively affect how the patients respond to their conditions, treatments, and overall quality of care (Kelley et al., 2014). In addition, the integration of the model for daily clinical practices should be done in such a way that restricts patients' access.

Therefore, it is important to set up a dedicated ethical research agenda (Boers et al., 2020) aimed at developing structured, standardized, and internationally acceptable ethical guidelines for the application of a machine learning model in daily clinical practice of cancer management (Arambula & Bur, 2020; Gruson et al., 2019; Johnson, 2019, 2019). These ethical concerns are essential to achieve trustworthy AI. The ethical guidelines should ensure compliance to ethical norms and principles, uphold fundamental human rights, offer acceptable moral values and entitlement (European Commission, 2019). These ethical norms and principles include respect for patients autonomy, encourages shared decision making, prevent harm to the patient, treat the patient fairly, and offer explainability in terms of the model itself and the result produced therein (European Commission, 2019). Moreover, other fundamental ethical principles such as honesty, transparency, non-malevolence, truthfulness, and benevolence should be upheld by the model (Keskinbora, 2019). Apart from these principles, other criticisms regarding the application of ML-based models in actual clinical practice should be considered in the ethical guidelines (Figure 37).

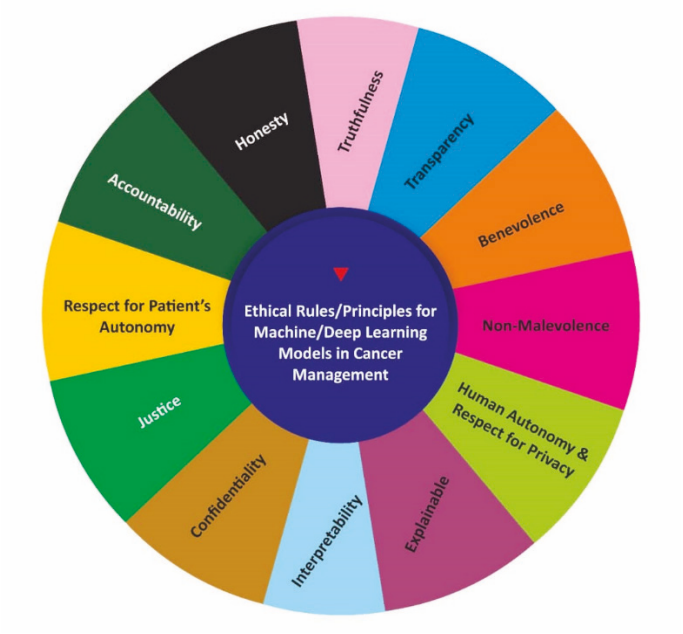

Figure 37. The features of a trustworthy machine learning model.

Corresponding laws both locally (nationally) and internationally should be enacted by the respective governments to ensure legally binding principles (e.g., the European General Data Protection Regulations) (Flaumenhaft & Ben-Assuli, 2018; Vayena et al., 2018) and jurisdictional mechanisms for enforcement (Robles Carrillo, 2020). Without a doubt, the application of machine learning is poised to revolutionize and improve prognosis in cancer. Hence, an ethical and legal framework should be taken into consideration from the data collection point to the development and actual integration of these models into daily clinical practice. Some of the important ethical questions that should be answered are presented in Table 4.

Table 4. Ethical concerns of machine learning models in cancer prognostication

Ethical concerns	**Meaning**	**The structural aspect of the ethical concerns** Ethical and moral concerns	
Privacy and confidentiality of patients' data	Approval of patients' consent and the concerned	*Concern 1:* Will the ML model developer use the extracted patients' information	*Concern VII*: How can the developer seek informed consent from the

	authority to use patients' data	from the hospital registry without their consent?	patient, hospital authority and national agency?
Bias in the data used to develop the model	Data may tend towards a particular race, geographical location, sexual orientation and so on	*Concern II:* Will the developed ML model be biased due to the imbalance in the data?	*Concern VIII*: How can the developer handle the possible data imbalance in the developed ML model?
Peer disagreement	Contradictory diagnostic or prognostic opinion between the model and the clinician	*Concern III:* Will the clinician follow his/her own diagnostic decision in cases in which the ML model gives a contrary opinion?	*Concern IX:* How can the discrepancy between conflicting diagnostic opinions be balanced? Is there an ethical guideline or standard that guides the use of a ML model in cancer management?
Responsibility gap	Assignment of responsibility when the ML models gave a wrong prediction	*Concern IV:* Will the clinician be held responsible when the ML model gives a wrong prediction?	*Concern X:* How should the clinicians interpret the hospital guidelines on the use of ML models? What does medical ethics stipulate? What are ethical guidelines or standards that guide the use of ML models in cancer management?
Clinician–patient relationship	Fiduciary interaction between the physicians and patients may change	*Concern V:* Will the patient feel comfortable and confident about the diagnostic decision made by a machine/computer?	*Concern X1:* How will the clinician explain to the patient that the ML model is capable of making an accurate

			decision and justify the decision made by the model? Will this continue to uphold clinician–patient relationship?
Patients' autonomy	Ability of the patient to determine the best treatment and take part in a shared decision-making process	*Concern VI:* Will the patient be allowed to choose the treatment approach that suits him/her when the model gives a different treatment plan?	*Concern XII:* How can the clinician take into consideration the treatment plan that best considers the daily activities of the patient?

Considering the benefits of machine learning models in proper cancer management, a dedicated, decisive and proactive role is expected from the government, clinical experts, patients' representatives, data scientists, ML experts and legal and human rights activists in defining these ethical guidelines. This is important for the machine learning models to achieve the touted benefits of providing supports to clinicians in making informed decisions, optimize health systems, and improve the quality of patient care.

6.5 Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for the future

Through proper stratification of patients into groups (high- and low risk) using machine learning applications, clinical practice can be guided to ensure precision and personalized medicine. Proper counselling regarding realistic expectations such as specific treatment and follow-up (post-operative adjuvant treatment) can be offered to the patients. For instance, high-risk patients might benefit from adjuvant oncological therapy after surgery.

The support vector machine and artificial neural network are the two widely used machine learning models for head and neck cancer prognostication (Patil et al., 2019) (Sharma & Om, 2014). The reason for the use of the support vector machine is because it is an empirical risk minimizer algorithm and avoids the danger of being trapped in local minima (Levitin, 2007). This makes it resistant to the challenge of overfitting. Thus, it is posited to discern the complex relationships

between the input and output parameters. Similarly, the success recorded in the application of the neural network for risk estimation in cancer gave rise to modification to contain multiple hidden layers known as deep learning neural networks. It has found application in complex problems such as image analysis (LeCun et al., 2015; Michie et al., 1994), especially in oral cancer prognostication (Ariji et al., 2019; Aubreville et al., 2017; Chan et al., 2019; Das et al., 2018; Jeyaraj & Samuel Nadar, 2019; Shams & Htike, 2017; Uthoff et al., 2018; Xu et al., 2019; M. Yu et al., 2019).

Considering the prospects for the future, machine learning models should be explainable (both in terms of the model and results) and avoid black-box criticism (Bur et al., 2019; Castelvecchi, 2016; M. K. Yu et al., 2018) (Figure 38).

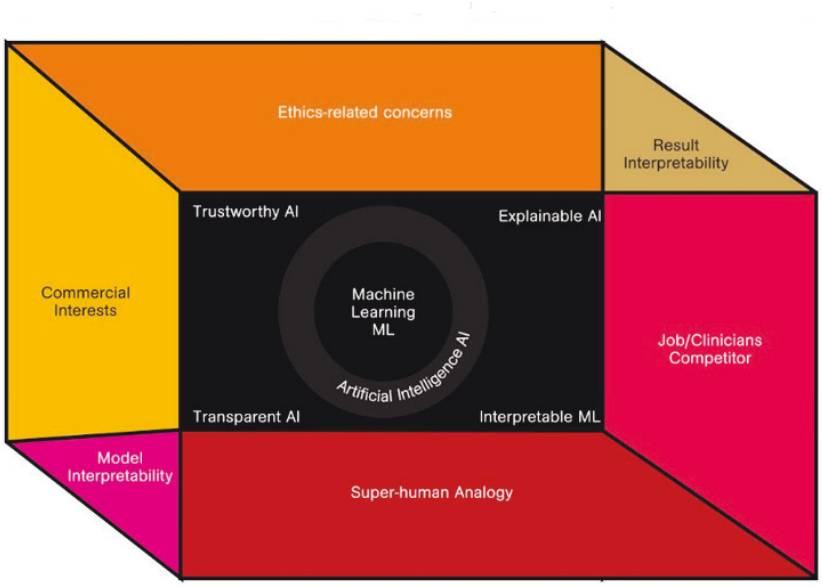

Figure 38. Summary of the black-box of a typical machine learning model.

Interestingly, as a result of the development trend of machine learning algorithms from direct algorithms to ensemble algorithms and to deep learning, the black box concern and explainable AI (model and results explainability) become more noticeable. To address this concern, it is important to evaluate the quality of the explanations provided by this model. An important approach to accurately evaluate the quality of explanations by the machine learning models is to use the system causability scale (SCS) (Holzinger et al., 2020). The results from the SCS

measures how explainable the model is, and also the level of transparency provided by such model. Thus, both explainability and transparency (Altmann et al., 2010; Bur et al., 2019; Karadaghy et al., 2019; M. K. Yu et al., 2018) (Figure 4) are important characteristics of a potential machine learning model for cancer prognostication. They offer the clinicians the opportunities to trust, understand, and be able to explain how the model arrived at a particular decision. Several terms including explainable AI, transparent ML, interpretable ML, and trustworthy AI have been employed to describe this phenomenon (Bernease Herman, 2017; European Union, 2019; Holzinger et al., 2020; Zachary, 2017).

Furthermore, the misconception regarding the super-human nature of the developed model should be put into proper perspective. These notions have led to mixed reactions in terms of the acceptance of these models in actual daily clinical practices. The clinical community appeared to be divided. The first group appeared to appreciate these models and opined that they help revolutionize clinical oncology. On the other hand, there are trepidation and concerns that these models could replace the need for professional experience-based consideration in the near future (Grace et al., 2018). Therefore, it is important to correct the notion that machine learning models are magical diagnostic tools. Rather, several factors such as the amount and quality of data and experience of the machine learning experts – input parameters selected and training approach used play significant roles in the performance of the machine learning model (Bur et al., 2019; Karadaghy et al., 2019; Shah et al., 2018).

The generalizability of the model is another important concern for the future. Of note, a generalized model means that the inherent bias in the dataset has been accounted for in the development of the model (Heinrichs & Eickhoff, 2019). Therefore, a model trained with a limited amount of data may not offer reasonable generalizability when exposed to external data (not used in the training of the model) (Alabi et al., 2019; Aubreville et al., 2017; Chang et al., 2013, 2014; Exarchos et al., 2012; Shams & Htike, 2017; Sharma & Om, 2015; Tseng et al., 2015; C.-Y. Wang et al., 2003; X. Wang et al., 2020). Thus, the health data economy should be improved to ensure that the patients' health data can be shared relatively easily. With improved data infrastructure of healthcare organizations', a machine learning model can be developed using integrated data (data fusion) to enhance generalizability. However, in the quest to ensure vibrant health data economy, privacy and ethical usage of data should be significantly considered (Bur et al., 2019; Karadaghy et al., 2019). Also, an effective health data economy address the concern for reduction in revenue for healthcare organizations or rendering the clinicians less important (Bur et al., 2019).

Admittedly, machine learning has a huge potential in the management of oral cancer. Therefore, resolving these issues related to the concerns – ethical and methodological – highlights important steps towards the implementation of this approach in daily clinical practice. These potentials include informed clinical decision-making, improved quality of care, precise diagnosis, effective treatment planning, and accurate prognostication of oral cancer.

7 CONCLUSION

In this multicenter international study, we have examined and combined evidence-based clinicopathologic parameters that have been suggested to play significant roles in the prognostication of tongue cancers for risk estimation of patients in order to achieve precision and personalized medicine. Specifically:

A. We applied machine learning techniques that considered the shortcomings of the Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) staging to estimate and predict tongue cancer patients' outcomes such as locoregional recurrences and overall survival. We found that the artificial neural network and boosted decision tree showed promising performance in the risk estimation of patients' outcomes.

B. We evaluated the prognostic significance of the input parameters using machine learning techniques. Our models showed that tumor budding and depth of invasion are promising parameters in the prognostication of oral tongue cancers.

C. We developed a machine learning-based web prognostic tool that was targeted at providing personalized medicine for oral tongue cancer patients. Our web-based tool stratifies oral cancer patients into a low- or high-risk locoregional recurrence.

D. We compared the performance of a machine-learning based model to a model based on depth of invasion. We found that the depth of invasion alone is not enough for risk estimation of early oral tongue cancer patients.

E. Additionally, we compared the performance of machine learning techniques to nomograms in the prognostication of outcomes (overall survival) for tongue cancer patients. The machine learning outperformed the nomogram, while the nomogram showed transparency and clarity (explainability) in the risk estimation of patients. Thus, we proposed a hybrid approach that combines improved performance (machine learning model) and transparency and explainability (nomogram). This hybrid was termed a nomogram – machine learning model (NomoML).

F. We highlighted some ethical challenges that can affect the implementation of machine learning models for daily clinical practice.

G. We examined the current status of ML applications in oral cancer in addition to clinical concerns and offer prospects for the future.

References

Adam, S. (2012). *Is coursera the begining of the end for traditional higher education?* Higher Education.

Agnihotri, R., & Gaur, S. (2014). Implications of tobacco smoking on the oral health of older adults: Smoking and geriatric oral health. *Geriatrics & Gerontology International*, *14*(3), 526–540. https://doi.org/10.1111/ggi.12285

Ahmad, L., & Eshlaghy, A. (2013). Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Journal of Health & Medical Informatics*, *04*(02). https://doi.org/10.4172/2157-7420.1000124

Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., Mäkitie, A. A., Salo, T., Almangush, A., & Leivo, I. (2019). Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *International Journal of Medical Informatics*, 104068. https://doi.org/10.1016/j.ijmedinf.2019.104068

Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., Mäkitie, A. A., Salo, T., Leivo, I., & Almangush, A. (2019). Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. *Virchows Archiv*, *475*(4), 489–497. https://doi.org/10.1007/s00428-019-02642-5

Al-Amad, S. H., Awad, M. A., & Nimri, O. (2014). Oral cancer in young Jordanians: potential association with frequency of narghile smoking. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, *118*(5), 560–565. https://doi.org/10.1016/j.0000.2014.08.002

Ali, S., Palmer, F. L., Yu, C., DiLorenzo, M., Shah, J. P., Kattan, M. W., Patel, S. G., & Ganly, I. (2014). Postoperative nomograms predictive of survival after surgical management of malignant tumors of the major salivary glands. *Annals of Surgical Oncology*, *21*(2), 637–642. https://doi.org/10.1245/s10434-013-3321-y

Almangush, A., Coletta, R. D., Bello, I. O., Bitu, C., Keski-Säntti, H., Mäkinen, L. K., Kauppila, J. H., Pukkila, M., Hagström, J., Laranne, J., Tommola, S., Soini, Y., Kosma, V.-M., Koivunen, P., Kowalski, L. P., Nieminen, P., Grénman, R., Leivo, I., & Salo, T. (2015). A simple novel prognostic model for early stage oral tongue cancer. *International Journal of Oral and Maxillofacial Surgery*, *44*(2), 143–150. https://doi.org/10.1016/j.ijom.2014.10.004

Almangush, Alhadi. (2015). Histopathological predictors of early stage oral tongue cancer. *University of Helsinki, Academic Dissertation*. Department of Pathology, Haartman Institute.

Almangush, Alhadi, Bello, I. O., Coletta, R. D., Mäkitie, A. A., Mäkinen, L. K., Kauppila, J. H., Pukkila, M., Hagström, J., Laranne, J., Soini, Y., Kosma, V.-M., Koivunen, P., Kelner, N., Kowalski, L. P., Grénman, R., Leivo, I., Läärä, E., & Salo, T. (2015). For early-stage oral tongue cancer, depth of invasion and worst pattern of invasion are the strongest pathological predictors for locoregional recurrence

and mortality. *Virchows Archiv*, *467*(1), 39–46. https://doi.org/10.1007/s00428-015-1758-z

Almangush, Alhadi, Pirinen, M., Heikkinen, I., Mäkitie, A. A., Salo, T., & Leivo, I. (2018). Tumour budding in oral squamous cell carcinoma: a meta-analysis. *British Journal of Cancer*, *118*(4), 577–586. https://doi.org/10.1038/bjc.2017.425

Alpaydin, E. (2014). *Introduction to machine learning* (Third edition). The MIT Press.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

American Cancer Society. (2018). Cancer A-Z: What are oral cavity and oropharyngeal cancers? *Cancer*. https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/about/what-is-oral-cavity-cancer.html

American Cancer Society. (2020). Key statistics for oral cavity and oropharyngeal cancers. *A-Z Cancer*. https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/about/key-statistics.html

American Joint Committee on Cancer. (2002). *AJCC Cancer Staging Manual* (F. L. Greene, D. L. Page, I. D. Fleming, A. G. Fritz, C. M. Balch, D. G. Haller, & M. Morrow (Eds.)). Springer New York. https://doi.org/10.1007/978-1-4757-3656-4

Amit, M., Yen, T.-C., Liao, C.-T., Chaturvedi, P., Agarwal, J. P., Kowalski, L. P., Ebrahimi, A., Clark, J. R., Kreppel, M., Zöller, J., Fridman, E., Bolzoni, V. A., Shah, J. P., Binenbaum, Y., Patel, S. G., Gil, Z., & The International Consortium for Outcome Research (ICOR) in Head and Neck Cancer. (2013). Improvement in survival of patients with oral cavity squamous cell carcinoma: An international collaborative study: OCSCC Postoperative Survival Trends. *Cancer*, *119*(24), 4242–4248. https://doi.org/10.1002/cncr.28357

Arambula, A. M., & Bur, A. M. (2020). Ethical Considerations in the Advent of Artificial Intelligence in Otolaryngology. *Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, *162*(1), 38–39. https://doi.org/10.1177/0194599819889686

Ariji, Y., Fukuda, M., Kise, Y., Nozawa, M., Yanashita, Y., Fujita, H., Katsumata, A., & Ariji, E. (2019). Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, *127*(5), 458–463. https://doi.org/10.1016/j.0000.2018.10.002

Arora, A., Husain, N., Bansal, A., Neyaz, A., Jaiswal, R., Jain, K., Chaturvedi, A., Anand, N., Malhotra, K., & Shukla, S. (2017). Development of a New Outcome Prediction Model in Early-stage Squamous Cell Carcinoma of the Oral Cavity Based on Histopathologic Parameters With Multivariate Analysis: The Aditi-Nuzhat Lymph-node Prediction Score (ANLPS) System. *The American Journal of*

Surgical Pathology, *41*(7), 950–960. https://doi.org/10.1097/PAS.0000000000000843

Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., Bohr, C., Neumann, H., Stelzle, F., & Maier, A. (2017). Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-12320-8

Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E., & Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer*, *116*(14), 3310–3321. https://doi.org/10.1002/cncr.25081

Balachandran, V. P., Gonen, M., Smith, J. J., & DeMatteo, R. P. (2015). Nomograms in oncology: more than meets the eye. *The Lancet. Oncology*, *16*(4), e173-180. https://doi.org/10.1016/S1470-2045(14)71116-7

Bali, J., Garg, R., & Bali, R. T. (2019). Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian Journal of Ophthalmology*, *67*(1), 3–6. https://doi.org/10.4103/ijo.IJO_1292_18

Balthazar, P., Harri, P., Prater, A., & Safdar, N. M. (2018). Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. *Journal of the American College of Radiology*, *15*(3), 580–586. https://doi.org/10.1016/j.jacr.2017.11.035

Bartholomai, J. A., & Frieboes, H. B. (2018). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 632–637. https://doi.org/10.1109/ISSPIT.2018.8642753

Bauer, E., & Kohavi, R. (2003). An empirical comparison of voting classification algorithms: Bagging, Boosting and Variants. *Journal of Machine Learning*, *36*, 105–139.

Bektas-Kayhan, K., Karagoz, G., Kesimli, M. C., Karadeniz, A. N., Meral, R., Altun, M., & Unur, M. (2014). Carcinoma of the Tongue: A Case-control Study on Etiologic Factors and Dental Trauma. *Asian Pacific Journal of Cancer Prevention*, *15*(5), 2225–2229. https://doi.org/10.7314/APJCP.2014.15.5.2225

Bernease Herman. (2017). *The Promise and Peril of Human Evaluation for Model Interpretability*. Presented at NIPS 2017 Symposium on Interpretable Machine Learning, Interpretable ML Symposium, Cornell University, USA. http://interpretable.ml/

Biglarian, A., Hajizadeh, E., Kazemnejad, A., & Zali, M. (2011). Application of artificial neural network in predicting the survival rate of gastric cancer patients. *Iranian Journal of Public Health*, *40*(2), 80–86.

Biglarian, Akbar, Hajizadeh, E., Kazemnejad, A., & Zayeri, F. (2010). Determining of prognostic factors in gastric cancer patients using artificial neural networks. *Asian Pacific Journal of Cancer Prevention: APJCP*, *11*(2), 533–536.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bishop, C. M. (2014). Embracing uncertainty - probabilistic inference. *Microsoft: Machine Learning Documentation*. https://docs.microsoft.com/en-us/archive/blogs/machinelearning/embracing-uncertainty-probabilistic-inference

Bishop, C. M., & Nabney, I. . (2008). *Pattern Recognition and Machine Learning: A Matlab Companion*. Springer.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-106

Boers, S. N., Jongsma, K. R., Lucivero, F., Aardoom, J., Büchner, F. L., de Vries, M., Honkoop, P., Houwink, E. J. F., Kasteleyn, M. J., Meijer, E., Pinnock, H., Teichert, M., van der Boog, P., van Luenen, S., van der Kleij, R. M. J. J., & Chavannes, N. H. (2020). SERIES: eHealth in primary care. Part 2: Exploring the ethical implications of its application in primary care practice. *European Journal of General Practice*, *26*(1), 26–32. https://doi.org/10.1080/13814788.2019.1678958

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, h5527. https://doi.org/10.1136/bmj.h5527

Bottaci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P. W., Macintyre, I. M., Duthie, G. S., & Monson, J. R. (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *The Lancet*, *350*(9076), 469–472. https://doi.org/10.1016/S0140-6736(96)11196-X

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (1998). Arcing classifiers (with distinction). *The Annals of Statistics*, *26*, 801–849.

Brinkmann, O., Kastratovic, D. A., Dimitrijevic, M. V., Konstantinovic, V. S., Jelovac, D. B., Antic, J., Nesic, V. S., Markovic, S. Z., Martinovic, Z. R., Akin, D., Spielmann, N., Zhou, H., & Wong, D. T. (2011). Oral squamous cell carcinoma detection by salivary biomarkers in a Serbian population. *Oral Oncology*, *47*(1), 51–55. https://doi.org/10.1016/j.oraloncology.2010.10.009

Bur, A. M., Holcomb, A., Goodwin, S., Woodroof, J., Karadaghy, O., Shnayder, Y., Kakarala, K., Brant, J., & Shew, M. (2019). Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. *Oral Oncology*, *92*, 20–25. https://doi.org/10.1016/j.oraloncology.2019.03.011

Casparis, S., Borm, J. M., Tektas, S., Kamarachev, J., Locher, M. C., Damerau, G., Grätz, K. W., & Stadlinger, B. (2015). Oral lichen planus (OLP), oral lichenoid lesions (OLL), oral dysplasia, and oral cancer: retrospective analysis of clinicopathological data from 2002–2011. *Oral and Maxillofacial Surgery*, *19*(2), 149–156. https://doi.org/10.1007/s10006-014-0469-y

Castelvecchi, D. (2016). Can we open the black box of AI. *Nature*, *538*, 20–23.

Chan, C.-H., Huang, T.-T., Chen, C.-Y., Lee, C.-C., Chan, M.-Y., & Chung, P.-C. (2019). Texture-Map-Based Branch-Collaborative Network for Oral Cancer Detection. *IEEE Transactions on Biomedical Circuits and Systems*, *13*(4), 766–780. https://doi.org/10.1109/TBCAS.2019.2918244

Chang, S.-W. (2013). *The application of artificial intelligent techniques in oral cancer prognosis based on clinicopathologic and genomic markers*. Faculty of Computer Science & Information Technology, University of Malaya.

Chang, S.-W., Abdul-Kareem, S., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-170

Chang, S.-W., Sameem, A., Amir Feisal Merican, A. M., & Rosnah Binti, Z. (2014). A Hybrid Prognostic Model for Oral Cancer based on Clinicopathologic and Genomic Markers. *Sains Malaysiana*, *43*(4), 567–573.

Chao, C.-M., Yu, Y.-W., Cheng, B.-W., & Kuo, Y.-L. (2014). Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree. *Journal of Medical Systems*, *38*(10). https://doi.org/10.1007/s10916-014-0106-1

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Chen, Y.-C., Ke, W.-C., & Chiu, H.-W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in Biology and Medicine*, *48*, 1–7. https://doi.org/10.1016/j.compbiomed.2014.02.006

Cho, J.-K., Lee, G.-J., Yi, K.-I., Cho, K.-S., Choi, N., Kim, J. S., Kim, H., Oh, D., Choi, S.-K., Jung, S.-H., Jeong, H.-S., & Ahn, Y. C. (2015). Development and external validation of nomograms predictive of response to radiation therapy and overall survival in nasopharyngeal cancer patients. *European Journal of Cancer (Oxford, England: 1990)*, *51*(10), 1303–1311. https://doi.org/10.1016/j.ejca.2015.04.003

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *Philosophical Review*, *116*(2), 187–217. https://doi.org/10.1215/00318108-2006-035

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, *350*(jan07 4), g7594–g7594. https://doi.org/10.1136/bmj.g7594

Criminisi, A., & Shotton, J. (Eds.). (2013). *Decision forests for computer vision and medical image analysis*. Springer.

Cruz, J. A., & Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, *2*, 59–77.

Das, D. K., Bose, S., Maiti, A. K., Mitra, B., Mukherjee, G., & Dutta, P. K. (2018). Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis. *Tissue and Cell*, *53*, 111–119. https://doi.org/10.1016/j.tice.2018.06.004

de Melo, N. B., Bernardino, Í. de M., de Melo, D. P., Gomes, D. Q. C., & Bento, P. M. (2018). Head and neck cancer, quality of life, and determinant factors: a novel approach using decision tree analysis. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, *126*(6), 486–493. https://doi.org/10.1016/j.0000.2018.07.055

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Elmusrati, M. (2020). *Lecture notes: Machine learning course*. Universityof Vaasa (Moodle).

Elstein, A. S. (1999). Heuristics and biases: selected errors in clinical reasoning. *Academic Medicine: Journal of the Association of American Medical Colleges*, *74*(7), 791–794. https://doi.org/10.1097/00001888-199907000-00012

England, J. R., & Cheng, P. M. (2019). Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. *American Journal of Roentgenology*, *212*(3), 513–519. https://doi.org/10.2214/AJR.18.20490

Enoch, D. (2010). Not Just a Truthometer: Taking Oneself Seriously (but not Too Seriously) in Cases of Peer Disagreement. *Mind*, *119*(476), 953–997. https://doi.org/10.1093/mind/fzq070

European Commission. (2019). *High-level expert group on artificial intelligence. In: Ethics Guidelines for Tustrworthy AI*. European Union.

European Union. (2019). *High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI*. Brussel.

Exarchos, K. P., Goletsis, Y., & Fotiadis, D. I. (2012). Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, *16*(6), 1127–1134. https://doi.org/10.1109/TITB.2011.2165076

Expert System. (2020). Machine Learning. *Definition of Machine Learning*. https://expertsystem.com/machine-learning-definition/

Faradmal, J., Soltanian, A. R., Roshanaei, G., Khodabakhshi, R., & Kasaeian, A. (2014). Comparison of the Performance of Log-logistic Regression and Artificial Neural Networks for Predicting Breast Cancer Relapse. *Asian Pacific Journal of Cancer Prevention*, *15*(14), 5883–5888. https://doi.org/10.7314/APJCP.2014.15.14.5883

Flaumenhaft, Y., & Ben-Assuli, O. (2018). Personal health records, global policy and regulation review. *Health Policy*, *122*(8), 815–826. https://doi.org/10.1016/j.healthpol.2018.05.002

Frances, B., & Matheson, J. (2018). *Disagreement. In: Zalta EN,ed. The Stanford encyclopedia of philosophy. Spring*. 8. https://plato.stanford.edu/archives/spr2018/entries/ disagreement/

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, *121*(2), 256–285. https://doi.org/10.1006/inco.1995.1136

Freund, Y., & Schapire, R. E. (1996). Experiments With a New Boosting Algorithm. In: Proceedings of the Thirteenth International Conference on Machine Leaning Theory. *Morgan Kaufmann Publishers Inc*, 148–156.

Gandhi, R. (2018). Introduction to machine learning algorithms. *Towards Data Science*. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Ganly, I., Amit, M., Kou, L., Palmer, F. L., Migliacci, J., Katabi, N., Yu, C., Kattan, M. W., Binenbaum, Y., Sharma, K., Naomi, R., Abib, A., Miles, B., Yang, X., Lei, D., Bjoerndal, K., Godballe, C., Mücke, T., Wolff, K.-D., … Patel, S. G. (2015). Nomograms for predicting survival and recurrence in patients with adenoid cystic carcinoma. An international collaborative study. *European Journal of Cancer (Oxford, England: 1990)*, *51*(18), 2768–2776. https://doi.org/10.1016/j.ejca.2015.09.004

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Borondy Kitts, A., Birch, J., Shields, W. F., van den Hoven van Genderen, R., Kotter, E., Wawira Gichoya, J., Cook, T. S., Morgan, M. B., Tang, A., Safdar, N. M., & Kohli, M. (2019). Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *Radiology*, *293*(2), 436–440. https://doi.org/10.1148/radiol.2019191586

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., & Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, *22*(14), e184–e190. https://doi.org/10.1093/bioinformatics/btl230

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *ArXiv:1705.08807 [Cs]*. http://arxiv.org/abs/1705.08807

Grimes, D. A. (2008). The nomogram epidemic: resurgence of a medical relic. *Annals of Internal Medicine*, *149*(4), 273–275. https://doi.org/10.7326/0003-4819-149-4-200808190-00010

Gross, N. D., Patel, S. G., Carvalho, A. L., Chu, P.-Y., Kowalski, L. P., Boyle, J. O., Shah, J. P., & Kattan, M. W. (2008). Nomogram for deciding adjuvant treatment after surgery for oral cavity squamous cell carcinoma. *Head & Neck*, *30*(10), 1352–1360. https://doi.org/10.1002/hed.20879

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Gruson, D., Helleputte, T., Rousseau, P., & Gruson, D. (2019). Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation. *Clinical Biochemistry*, *69*, 1–7. https://doi.org/10.1016/j.clinbiochem.2019.04.013

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Heinrichs, B., & Eickhoff, S. B. (2019). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*. https://doi.org/10.1002/hbm.24886

Hillbertz, N. S., Hirsch, J.-M., Jalouli, J., Jalouli, M. M., & Sand, L. (2012). Viral and molecular aspects of oral cancer. *Anticancer Research*, *32*(10), 4201–4212.

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *KI - Künstliche Intelligenz*, *34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z

Hubert Low, T.-H., Gao, K., Elliott, M., & Clark, J. R. (2015). Tumor classification for early oral cancer: Re-evaluate the current TNM classification: TNM Classification for Early Oral Tumor. *Head & Neck*, *37*(2), 223–228. https://doi.org/10.1002/hed.23581

Jaiswal, A. (2018). *Integrative Bioinformatics of Functional and Genomic Profiles for Cancer Systems Medicine. University of Helsinki*, *PhD Dissertation*, 82.

Jalouli, J., Jalouli, M. M., Sapkota, D., Ibrahim, S. O., Larsson, P.-A., & Sand, L. (2012). Human papilloma virus, herpes simplex virus and epstein barr virus in oral squamous cell carcinoma from eight different countries. *Anticancer Research*, *32*(2), 571–580.

Jeong, H. ., Obaidat, S. ., Yen, N. ., & Park, J. . (2013). *Advanced in computer science and its applications: CSA 2013* (1st edition). Springer.

Jeyaraj, P. R., & Samuel Nadar, E. R. (2019). Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm.

Journal of Cancer Research and Clinical Oncology, *145*(4), 829–837. https://doi.org/10.1007/s00432-018-02834-7

Johnson, S. L. J. (2019). AI, Machine Learning, and Ethics in Health Care. *Journal of Legal Medicine*, *39*(4), 427–441. https://doi.org/10.1080/01947648.2019.1690604

Karadaghy, O. A., Shew, M., New, J., & Bur, A. M. (2019). Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma. *JAMA Otolaryngology–Head & Neck Surgery*, *145*(12), 1115. https://doi.org/10.1001/jamaoto.2019.0981

Kazemnejad, A., Batvandi, Z., & Faradmal, J. (2010). Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes. *Eastern Mediterranean Health Journal = La Revue De Sante De La Mediterranee Orientale = Al-Majallah Al-Sihhiyah Li-Sharq Al-Mutawassit*, *16*(6), 615–620.

Kelley, J. M., Kraft-Todd, G., Schapira, L., Kossowsky, J., & Riess, H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. *PloS One*, *9*(4), e94207. https://doi.org/10.1371/journal.pone.0094207

Kelly T. (2010). *Peer disagreement and higher order evidence. In: Goldman Al, Whitcomb D, eds. Social Epistemology: essential readings. Oxford University Press.* Oxford University Press.

Keogan, M. T., Lo, J. Y., Freed, K. S., Raptopoulos, V., Blake, S., Kamel, I. R., Weisinger, K., Rosen, M. P., & Nelson, R. C. (2002). Outcome Analysis of Patients with Acute Pancreatitis by Using an Artificial Neural Network. *Academic Radiology*, *9*(4), 410–419. https://doi.org/10.1016/S1076-6332(03)80186-1

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of Clinical Neuroscience*, *64*, 277–282. https://doi.org/10.1016/j.jocn.2019.03.001

Kim, J., & Shin, H. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association: JAMIA*, *20*(4), 613–618. https://doi.org/10.1136/amiajnl-2012-001570

Kim, W., Kim, K. S., Lee, J. E., Noh, D.-Y., Kim, S.-W., Jung, Y. S., Park, M. Y., & Park, R. W. (2012). Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. *Journal of Breast Cancer*, *15*(2), 230. https://doi.org/10.4048/jbc.2012.15.2.230

Kohli, M., & Geis, R. (2018). Ethics, Artificial Intelligence, and Radiology. *Journal of the American College of Radiology*, *15*(9), 1317–1319. https://doi.org/10.1016/j.jacr.2018.05.020

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction.

Computational and Structural Biotechnology Journal, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In J. S. Gero & F. Sudweeks (Eds.), *Artificial Intelligence in Design '96* (pp. 151–170). Springer Netherlands. https://doi.org/10.1007/978-94-009-0279-4_9

Kudo, Y. (2019). Predicting cancer outcome: Artificial intelligence vs. pathologists. *Oral Diseases*, *25*(3), 643–645. https://doi.org/10.1111/odi.12954

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Levitin, G. (2007). *Computational Intelligence in Reliability Engineering Evolutionary Techniques in Reliability Analysis and Optimization*. Springer Berlin Heidelberg. http://link.springer.com/book/10.1007/978-3-540-37368-1

Li, Y., Zhao, Z., Liu, X., Ju, J., Chai, J., Ni, Q., Ma, C., Gao, T., & Sun, M. (2017). Nomograms to estimate long-term overall survival and tongue cancer-specific survival of patients with tongue squamous cell carcinoma. *Cancer Medicine*, *6*(5), 1002–1013. https://doi.org/10.1002/cam4.1021

Lisboa, P. J. G. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*, *15*(1), 11–39. https://doi.org/10.1016/S0893-6080(01)00111-3

Listl, S., Jansen, L., Stenzinger, A., Freier, K., Emrich, K., Holleczek, B., Katalinic, A., Gondos, A., Brenner, H., & the GEKID Cancer Survival Working Group. (2013). Survival of Patients with Oral Cavity Cancer in Germany. *PLoS ONE*, *8*(1), e53415. https://doi.org/10.1371/journal.pone.0053415

Liu, J., Geng, Q., Liu, Z., Chen, S., Guo, J., Kong, P., Chen, Y., Li, W., Zhou, Z., Sun, X., Zhan, Y., & Xu, D. (2016). Development and external validation of a prognostic nomogram for gastric cancer using the national cancer registry. *Oncotarget*, *7*(24), 35853–35864. https://doi.org/10.18632/oncotarget.8221

Liu, Y., Li, Y., Fu, Y., Liu, T., Liu, X., Zhang, X., Fu, J., Guan, X., Chen, T., Chen, X., & Sun, Z. (2017). Quantitative prediction of oral cancer risk in patients with oral leukoplakia. *Oncotarget*, *8*(28). https://doi.org/10.18632/oncotarget.17550

Lu, C., Lewis, J. S., Dupont, W. D., Plummer, W. D., Janowczyk, A., & Madabhushi, A. (2017). An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. *Modern Pathology*, *30*(12), 1655–1665. https://doi.org/10.1038/modpathol.2017.98

Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques.

International Journal of Medical Informatics, *108*, 1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013

Ma, H., Guo, X., Ping, Y., Wang, B., Yang, Y., Zhang, Z., & Zhou, J. (2019). PPCD: Privacy-preserving clinical decision with cloud support. *PLOS ONE*, *14*(5), e0217349. https://doi.org/10.1371/journal.pone.0217349

Maclin, P. S., Dempsey, J., Brooks, J., & Rand, J. (1991). Using neural networks to diagnose cancer. *Journal of Medical Systems*, *15*(1), 11–19. https://doi.org/10.1007/BF00993877

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Manoharan, S., Nagaraja, V., & Eslick, G. D. (2014). Ill-fitting dentures and oral cancer: A meta-analysis. *Oral Oncology*, *50*(11), 1058–1061. https://doi.org/10.1016/j.oraloncology.2014.08.002

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

Meir, R., & Rätsch, G. (2003). *An introduction to boosting and leveraging*.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Meurman, J. H. (2010). Infectious and dietary risk factors of oral cancer. *Oral Oncology*, *46*(6), 411–413. https://doi.org/10.1016/j.oraloncology.2010.03.003

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (Eds.). (1994). *Machine learning, neural and statistical classification*. Ellis Horwood.

Microsoft Azure Machine Learning Studio. (2018). *Azure Machine Learning Studio: In Documentation*.

Mishra, D. (2019). Regression: An explanation of regression metrics and what can go wrong. *Towards Data Science*. https://towardsdatascience.com/regression-an-explanation-of-regression-metrics-and-what-can-go-wrong-a39a9793d914

Mitchell, T. . (2006). *The Discipline of Machine Learning*. Carnegie Mellon University.

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, *22*(2), 303–341. https://doi.org/10.1007/s11948-015-9652-2

Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, *24*(1), 31–42. https://doi.org/10.3233/THC-151071

Montero, P. H., Yu, C., Palmer, F. L., Patel, P. D., Ganly, I., Shah, J. P., Shaha, A. R., Boyle, J. O., Kraus, D. H., Singh, B., Wong, R. J., Morris, L. G., Kattan, M. W.,

& Patel, S. G. (2014). Nomograms for preoperative prediction of prognosis in patients with oral cavity squamous cell carcinoma. *Cancer*, *120*(2), 214–221. https://doi.org/10.1002/cncr.28407

Mroueh, R., Haapaniemi, A., Grénman, R., Laranne, J., Pukkila, M., Almangush, A., Salo, T., & Mäkitie, A. (2017). Improved outcomes with oral tongue squamous cell carcinoma in Finland: Oral tongue carcinoma in Finland. *Head & Neck*, *39*(7), 1306–1312. https://doi.org/10.1002/hed.24744

Nabi, J. (2018). How Bioethics Can Shape Artificial Intelligence and Machine Learning. *The Hastings Center Report*, *48*(5), 10–13. https://doi.org/10.1002/hast.895

National Cancer Institute. (2017). *Head and Neck Cancers*. https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet

Ng, J. H., Iyer, N. G., Tan, M.-H., & Edgren, G. (2017). Changing epidemiology of oral squamous cell carcinoma of the tongue: A global study: Changing epidemiology of tongue cancer. *Head & Neck*, *39*(2), 297–304. https://doi.org/10.1002/hed.24589

Nocedal, J., & Wright, S. . (1999). *Numerical Oprimization*. Springer.

Oji, C., & Chukwuneke, F. (2012). Poor oral Hygiene may be the Sole Cause of Oral Cancer. *Journal of Maxillofacial and Oral Surgery*, *11*(4), 379–383. https://doi.org/10.1007/s12663-012-0359-5

Park, K., Ali, A., Kim, D., An, Y., Kim, M., & Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, *26*(9), 2194–2205. https://doi.org/10.1016/j.engappai.2013.06.013

Park, S. H., Kim, Y.-H., Lee, J. Y., Yoo, S., & Kim, C. J. (2019). Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review. *Science Editing*, *6*(2), 91–98. https://doi.org/10.6087/kcse.164

Park, S. H., & Kressel, H. Y. (2018). Connecting Technological Innovation in Artificial Intelligence to Real-world Medical Practice through Rigorous Clinical Validation: What Peer-reviewed Medical Journals Could Do. *Journal of Korean Medical Science*, *33*(22). https://doi.org/10.3346/jkms.2018.33.e152

Patel, J., & Goyal, R. (2007). Applications of Artificial Neural Networks in Medical Science. *Current Clinical Pharmacology*, *2*(3), 217–226. https://doi.org/10.2174/157488407781668811

Patel, S. G., & Lydiatt, W. M. (2008). Staging of head and neck cancers: is it time to change the balance between the ideal and the practical? *Journal of Surgical Oncology*, *97*(8), 653–657. https://doi.org/10.1002/jso.21021

Patil, S., Habib Awan, K., Arakeri, G., Jayampath Seneviratne, C., Muddur, N., Malik, S., Ferrari, M., Rahimi, S., & Brennan, P. A. (2019). Machine learning and

its potential applications to the genomic study of head and neck cancer—A systematic review. *Journal of Oral Pathology & Medicine*, *48*(9), 773–779. https://doi.org/10.1111/jop.12854

Peng, K. A., Chu, A. C., Lai, C., Grogan, T., Elashoff, D., Abemayor, E., & St. John, M. A. (2014). Is there a role for neck dissection in T1 oral tongue squamous cell carcinoma? The UCLA experience. *American Journal of Otolaryngology*, *35*(6), 741–746. https://doi.org/10.1016/j.amjoto.2014.06.019

Piazza, C., Taglietti, V., Paderno, A., & Nicolai, P. (2014). End-to-end versus end-to-side venous microanastomoses in head and neck reconstruction. *European Archives of Oto-Rhino-Laryngology*, *271*(1), 157–162. https://doi.org/10.1007/s00405-013-2496-y

Platt, J., Christianini, N., & Shawe-Taylor, J. (1999). Large margin DAGs for multiclass classification. *Advances in Neural Information Procession Systems (NIPS)*, *12*, 547–553.

Po Wing Yuen, A., Lam, K. Y., Lam, L. K., Ho, C. M., Wong, A., Chow, T. L., Yuen, W. F., & Wei, W. I. (2002). Prognostic factors of clinically stage I and II oral tongue carcinoma?A comparative study of stage, thickness, shape, growth pattern, invasive front malignancy grading, martinez-gimeno score, and pathologic features. *Head & Neck*, *24*(6), 513–520. https://doi.org/10.1002/hed.10094

Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and Technology*, *7*(4), 351–367. https://doi.org/10.1007/s12553-017-0179-1

Puri, M., Pathak, Y., Sutariya, V. B., Tipparaju, S., & Moreno, W. (Eds.). (2016). *Artificial neural network for drug design, delivery, and disposition*. Elsevier/AP, Academic Press is an imprint of Elsevier.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *32*, 172–181.

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, 101937. https://doi.org/10.1016/j.telpol.2020.101937

Rosado, P., Lequerica-Fernández, P., Villallaín, L., Peña, I., Sanchez-Lasheras, F., & de Vicente, J. C. (2013). Survival model in oral squamous cell carcinoma based on clinicopathological parameters, molecular markers and support vector machines. *Expert Systems with Applications*, *40*(12), 4770–4776. https://doi.org/10.1016/j.eswa.2013.02.032

Roser, M., & Ritchie, H. (2019, November). Cancer. *Our World in Data*. https://ourworldindata.org/cancer

Rusthoven, K., Ballonoff, A., Raben, D., & Chen, C. (2008). Poor prognosis in patients with stage I and II oral tongue squamous cell carcinoma. *Cancer*, *112*(2), 345–351. https://doi.org/10.1002/cncr.23183

Safi, A.-F., Kauke, M., Grandoch, A., Nickenig, H.-J., Zöller, J. E., & Kreppel, M. (2017). Analysis of clinicopathological risk factors for locoregional recurrence of oral squamous cell carcinoma – Retrospective analysis of 517 patients. *Journal of Cranio-Maxillofacial Surgery*, *45*(10), 1749–1753. https://doi.org/10.1016/j.jcms.2017.07.012

Sarode, S. C., Sarode, G. S., & Karmarkar, S. (2012). Early detection of oral cancer: Detector lies within. *Oral Oncology*, *48*(3), 193–194. https://doi.org/10.1016/j.oraloncology.2011.11.018

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Schapire, R. E., & Freund, Y. (2012). *Boosting: foundations and algorithms*. MIT Press.

Scully, C. (2011). Oral cancer aetiopathogenesis; past, present and future aspects. *Medicina Oral Patología Oral y Cirugia Bucal*, e306–e311. https://doi.org/10.4317/medoral.16.e306

SEER, P. (2012). *Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data 1973–2009, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.*

Selaru, F. M., Xu, Y., Yin, J., Zou, T., Liu, T. C., Mori, Y., Abraham, J. M., Sato, F., Wang, S., Twigg, C., Olaru, A., Shustova, V., Leytin, A., Hytiroglou, P., Shibata, D., Harpaz, N., & Meltzer, S. J. (2002). Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology*, *122*(3), 606–613. https://doi.org/10.1053/gast.2002.31904

Shah, N. D., Steyerberg, E. W., & Kent, D. M. (2018). Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA*, *320*(1), 27. https://doi.org/10.1001/jama.2018.5602

Shams, W., & Htike, Z. (2017). Oral cancer prediction using gene expression profilling and machine learning. *International Journal of Applied Engineering Research*, *12*(15), 4893–4898.

Shariat, S. F., Karakiewicz, P. I., Suardi, N., & Kattan, M. W. (2008). Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *14*(14), 4400–4407. https://doi.org/10.1158/1078-0432.CCR-07-4713

Sharma, N., & Om, H. (2013). Data mining models for predicting oral cancer survivability. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *2*(4), 285–295. https://doi.org/10.1007/s13721-013-0045-7

Sharma, N., & Om, H. (2014). Using MLP and SVM for predicting survival rate of oral cancer patients. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *3*(1). https://doi.org/10.1007/s13721-014-0058-x

Sharma, N., & Om, H. (2015). Usage of Probabilistic and General Regression Neural Network for Early Detection and Prevention of Oral Cancer. *The Scientific World Journal*, *2015*, 1–11. https://doi.org/10.1155/2015/234191

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

Siegel, R., Ma, J., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, *64*(1), 9–29. https://doi.org/10.3322/caac.21208

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, *1*, 958–963. https://doi.org/10.1109/ICDAR.2003.1227801

Simes, R. J. (1985). Treatment selection for cancer patients: Application of statistical decision theory to the treatment of advanced ovarian cancer. *Journal of Chronic Diseases*, *38*(2), 171–186. https://doi.org/10.1016/0021-9681(85)90090-6

Sobin, L. H. (2003). TNM: Evolution and relation to other prognostic factors. *Seminars in Surgical Oncology*, *21*(1), 3–7. https://doi.org/10.1002/ssu.10014

Søland, T. M., & Brusevold, I. J. (2013). Prognostic molecular markers in cancer - quo vadis? *Histopathology*, *63*(3), 297–308. https://doi.org/10.1111/his.12184

Spelt, L., Nilsson, J., Andersson, R., & Andersson, B. (2013). Artificial neural networks – A method for prediction of survival following liver resection for colorectal cancer metastases. *European Journal of Surgical Oncology (EJSO)*, *39*(6), 648–654. https://doi.org/10.1016/j.ejso.2013.02.024

Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of Breast Cancer using Decision Tree DatanMining Technique. *International Journal of Computer Applications*, *98*(10), 16–24.

Sun, L., Feng, J., Ma, L., Liu, W., & Zhou, Z. (2013). CD133 expression in oral lichen planus correlated with the risk for progression to oral squamous cell carcinoma. *Annals of Diagnostic Pathology*, *17*(6), 486–489. https://doi.org/10.1016/j.anndiagpath.2013.06.004

Sun, W., Jiang, Y.-Z., Liu, Y.-R., Ma, D., & Shao, Z.-M. (2016). Nomograms to estimate long-term overall survival and breast cancer-specific survival of patients with luminal breast cancer. *Oncotarget*, *7*(15), 20496–20506. https://doi.org/10.18632/oncotarget.7975

Sun, Y., Goodison, S., Li, J., Liu, L., & Farmerie, W. (2007). Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, *23*(1), 30–37. https://doi.org/10.1093/bioinformatics/btl543

Sunasra, M. (2017). Performance metrics for classfication problems in machine learning. *Medium*. https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Swaminathan, S. (2018). Logistic regression - detailed overview [Medium]. *Towards Data Science*. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Tai, S.-K., Li, W.-Y., Chu, P.-Y., Chang, S.-Y., Tsai, T.-L., Wang, Y.-F., & Huang, J.-L. (2012). Risks and clinical implications of perineural invasion in T1-2 oral tongue squamous cell carcinoma. *Head & Neck*, *34*(7), 994–1001. https://doi.org/10.1002/hed.21846

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (1st ed). Pearson Addison Wesley.

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*, *7*(3), 293–299. https://doi.org/10.1016/j.cegh.2018.10.003

Tseng, W.-T., Chiang, W.-F., Liu, S.-Y., Roan, J., & Lin, C.-N. (2015). The Application of Data Mining Techniques to Oral Cancer Prognosis. *Journal of Medical Systems*, *39*(5). https://doi.org/10.1007/s10916-015-0241-3

Urbanowicz, R. J., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2013). Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach. *Journal of the American Medical Informatics Association*, *20*(4), 603–612. https://doi.org/10.1136/amiajnl-2012-001574

Uthoff, R. D., Song, B., Sunny, S., Patrick, S., Suresh, A., Kolur, T., Keerthi, G., Spires, O., Anbarani, A., Wilder-Smith, P., Kuriakose, M. A., Birur, P., & Liang, R. (2018). Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities. *PLOS ONE*, *13*(12), e0207493. https://doi.org/10.1371/journal.pone.0207493

van der Waal, I., de Bree, R., Brakenhoff, R., & Coebergh, J.-W. (2011). Early diagnosis in primary oral cancer: is it possible? *Medicina Oral, Patologia Oral Y Cirugia Bucal*, *16*(3), e300-305. https://doi.org/10.4317/medoral.16.e300

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Vázquez-Mahía, I., Seoane, J., Varela-Centelles, P., Tomás, I., García, A. Á., & López Cedrún, J. L. (2012). Predictors for Tumor Recurrence After Primary Definitive Surgery for Oral Cancer. *Journal of Oral and Maxillofacial Surgery*, *70*(7), 1724–1732. https://doi.org/10.1016/j.joms.2011.06.228

Vlaev, I., & Chater, N. (2006). Game relativity: how context influences strategic decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(1), 131–149. https://doi.org/10.1037/0278-7393.32.1.131

Wang, C.-Y., Tsai, T., Chen, H.-M., Chen, C.-T., & Chiang, C.-P. (2003). PLS-ANN based classification model for oral submucous fibrosis and oral carcinogenesis.

Lasers in Surgery and Medicine, *32*(4), 318–326. https://doi.org/10.1002/lsm.10153

Wang, G., Lu, R., & Huang, C. (2015). PSLP: Privacy-preserving single-layer perceptron learning for e-Healthcare. *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)*, 1–5. https://doi.org/10.1109/ICICS.2015.7459925

Wang, X., Yang, J., Wei, C., Zhou, G., Wu, L., Gao, Q., He, X., Shi, J., Mei, Y., Liu, Y., Shi, X., Wu, F., Luo, J., Guo, Y., Zhou, Q., Yin, J., Hu, T., Lin, M., Liang, Z., & Zhou, H. (2020). A personalized computational model predicts cancer risk level of oral potentially malignant disorders and its web application for promotion of non-invasive screening. *Journal of Oral Pathology & Medicine*. https://doi.org/10.1111/jop.12983

Weber, M. E., Yiannias, J. A., Hougeir, F. G., Kyle, A., Noble, B. N., Landry, A. M., & Hinni, M. L. (2012). Intraoral Metal Contact Allergy as a Possible Risk Factor for Oral Squamous Cell Carcinoma. *Annals of Otology, Rhinology & Laryngology*, *121*(6), 389–394. https://doi.org/10.1177/000348941212100605

Weinberg, R. A. (2014). *The biology of cancer* (Second edition). Garland Science, Taylor & Francis Group.

Witten, I. H., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier. https://doi.org/10.1016/C2009-0-19715-5

World Cancer Research Fund. (2018). Global cancer data by country. *American Institute for Cancer Research*. https://www.wcrf.org/dietandcancer/cancer-trends/data-cancer-frequency-country

World Health Organization. (2018, September 12). *Cancer: Fact Sheet*. https://www.who.int/en/news-room/fact-sheets/detail/cancer

Xu, S., Liu, Y., Hu, W., Zhang, C., Liu, C., Zong, Y., Chen, S., Lu, Y., Yang, L., Ng, E. Y. K., Wang, Y., & Wang, Y. (2019). An Early Diagnosis of Oral Cancer based on Three-Dimensional Convolutional Neural Networks. *IEEE Access*, *7*, 158603–158611. https://doi.org/10.1109/ACCESS.2019.2950286

Yanamoto, S., Yamada, S., Takahashi, H., Kawasaki, G., Ikeda, H., Shiraishi, T., Fujita, S., Ikeda, T., Asahina, I., & Umeda, M. (2013). Predictors of locoregional recurrence in T1-2N0 tongue cancer patients. *Pathology Oncology Research: POR*, *19*(4), 795–803. https://doi.org/10.1007/s12253-013-9646-9

Yang, X., Tian, X., Wu, K., Liu, W., Li, S., Zhang, Z., & Zhang, C. (2018). Prognostic impact of perineural invasion in early stage oral tongue squamous cell carcinoma: Results from a prospective randomized trial. *Surgical Oncology*, *27*(2), 123–128. https://doi.org/10.1016/j.suronc.2018.02.005

Yao, Q.-W., Zhou, D.-S., Peng, H.-J., Ji, P., & Liu, D.-S. (2014). Association of periodontal disease with oral cancer: a meta-analysis. *Tumour Biology: The*

Journal of the International Society for Oncodevelopmental Biology and Medicine, *35*(7), 7073–7077. https://doi.org/10.1007/s13277-014-1951-8

Yu, M. K., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J., & Ideker, T. (2018). Visible Machine Learning for Biomedicine. *Cell*, *173*(7), 1562–1565. https://doi.org/10.1016/j.cell.2018.05.056

Yu, M., Yan, H., Xia, J., Zhu, L., Zhang, T., Zhu, Z., Lou, X., Sun, G., & Dong, M. (2019). Deep convolutional neural networks for tongue squamous cell carcinoma classification using Raman spectroscopy. *Photodiagnosis and Photodynamic Therapy*, *26*, 430–435. https://doi.org/10.1016/j.pdpdt.2019.05.008

Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., … Wolpaw, J. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, *551*(7679), 159–163. https://doi.org/10.1038/551159a

Zachary, L. (2017). *The Doctor Just Won't Accept That!* Presented at NIPS 2017 Symposium on Interpretable Machine Learning, Interpretable ML Symposium, Cornell University, USA. http://interpretable.ml/

Zhang, B., He, X., Ouyang, F., Gu, D., Dong, Y., Zhang, L., Mo, X., Huang, W., Tian, J., & Zhang, S. (2017). Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Letters*, *403*, 21–27. https://doi.org/10.1016/j.canlet.2017.06.004

Zhang, X., Chen, X., Wang, J., Zhan, Z., & Li, J. (2018). Verifiable privacy-preserving single-layer perceptron training scheme in cloud computing. *Soft Computing*, *22*(23), 7719–7732. https://doi.org/10.1007/s00500-018-3233-7

Zheng, M.-H., Shi, K.-Q., Lin, X.-F., Xiao, D.-D., Chen, L.-L., Liu, W.-Y., Fan, Y.-C., & Chen, Y.-P. (2013). A model to predict 3-month mortality risk of acute-on-chronic hepatitis B liver failure using artificial neural network. *Journal of Viral Hepatitis*, *20*(4), 248–255. https://doi.org/10.1111/j.1365-2893.2012.01647.x

Zheng, Y., Xia, P., Zheng, H.-C., Takahashi, H., Masuda, S., & Takano, Y. (2010). The screening of viral risk factors in tongue and pharyngolaryngeal squamous carcinoma. *Anticancer Research*, *30*(4), 1233–1238.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Virchows Archiv (2019) 475:489–497
https://doi.org/10.1007/s00428-019-02642-5

ORIGINAL ARTICLE

Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool

Rasheed Omobolaji Alabi[1] · Mohammed Elmusrati[1] · Iris Sawazaki-Calone[2] · Luiz Paulo Kowalski[3] · Caj Haglund[4,5] · Ricardo D. Coletta[6] · Antti A. Mäkitie[7,8,9] · Tuula Salo[10,11,12] · Ilmo Leivo[13] · Alhadi Almangush[8,10,13,14]

Received: 2 April 2019 / Revised: 26 July 2019 / Accepted: 31 July 2019 / Published online: 17 August 2019

Abstract
Estimation of risk of recurrence in early-stage oral tongue squamous cell carcinoma (OTSCC) remains a challenge in the field of head and neck oncology. We examined the use of artificial neural networks (ANNs) to predict recurrences in early-stage OTSCC. A Web-based tool available for public use was also developed. A feedforward neural network was trained for prediction of locoregional recurrences in early OTSCC. The trained network was used to evaluate several prognostic parameters (age, gender, T stage, WHO histologic grade, depth of invasion, tumor budding, worst pattern of invasion, perineural invasion, and lymphocytic host response). Our neural network model identified tumor budding and depth of invasion as the most important prognosticators to predict locoregional recurrence. The accuracy of the neural network was 92.7%, which was higher than that of the logistic regression model (86.5%). Our online tool provided 88.2% accuracy, 71.2% sensitivity, and 98.9% specificity. In conclusion, ANN seems to offer a unique decision-making support predicting recurrences and thus adding value for the management of early OTSCC. To the best of our knowledge, this is the first study that applied ANN for prediction of recurrence in early OTSCC and provided a Web-based tool.

Keywords Oral tongue cancer · Artificial neural network · Machine learning · Locoregional recurrence · Prediction

Ilmo Leivo and Alhadi Almangush contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00428-019-02642-5) contains supplementary material, which is available to authorized users.

✉ Alhadi Almangush
alhadi.almangush@helsinki.fi

[1] Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland

[2] Oral Pathology and Oral Medicine, Dentistry School, Western Parana State University, Cascavel, PR, Brazil

[3] Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, SP, Brazil

[4] Research Programs Unit, Translational Cancer Biology, University of Helsinki, Helsinki, Finland

[5] Department of Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

[6] Department of Oral Diagnosis, School of Dentistry, University of Campinas, Piracicaba, São Paulo, Brazil

[7] Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland

[8] Research Programme in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[9] Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden

[10] Department of Pathology, University of Helsinki, Helsinki, Finland

[11] Department of Oral and Maxillofacial Diseases, University of Helsinki, Helsinki, Finland

[12] Cancer and Translational Medicine Research Unit, Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland

[13] Institute of Biomedicine, Pathology, University of Turku, Turku, Finland

[14] Faculty of Dentistry, University of Misurata, Misurata, Libya

Introduction

Oral tongue squamous cell carcinoma (OTSCC) typically displays aggressive behavior even at an early stage [1, 2]. Inaccurate assessment of OTSCC behavior may lead to improper management either as ineffective treatment or as unnecessary overtreatment. Therefore, identifying patients with low-risk or high-risk OTSCC can influence management decision-making and guide the selection of treatment approach. Several prognostic markers have been suggested to improve the prognostication of OTSCC [3, 4]. The advantages of evaluating some histopathologic prognostic markers in the examination of routine hematoxylin and eosin (HE)-stained slides include their low cost and time-saving aspects (as there is no need for additional staining) as well as the fact that these markers are potentially ready to be included in the routine pathology reports. Such advantages have motivated researchers to study various histopathologic features, and the recent evidence has confirmed the prognostic value of certain markers including, for example, tumor budding [5], depth of invasion [6], worst pattern of invasion [7], and perineural invasion [8]. It is necessary to mention that the previous studies on these markers have used traditional tools for data analysis, which have not produced any simple approach to utilize them as multiple prognostic factors should be applied to aid decision making.

The use of machine learning, a branch of artificial intelligence, in medical applications has increased widely in recent years; this has been driven by the rapidly accumulating volume of medical data. Similarly, artificial neural networks (ANNs) are an integral part and a subfield of machine learning. An ANN is an innovative hardware/software model that functions in a way inspired by the human brain [9–12]. In addition, ANN seems effective since the complex relationship between input and output can be accurately modeled with a relatively simple computer programming code. Structurally, ANN comprises input, hidden, and output layers (Fig. 1).

ANNs have an effective learning ability, and they can learn the relationship within a dataset. This effective learning characteristic has made ANN a good choice for predictive inferences that can be used to provide support for clinical decision-making. Many recent studies have applied ANNs for the prognostication of different cancers [9–12]. ANNs are adapted statistical models that analyze data for the prediction of outcomes in medical applications [13, 14], such as in colorectal cancer [15] and acute pancreatitis [16]. Spelt and collaborators applied ANN to predict survival in colorectal cancer, revealing that ANN produces better C-index than Cox regression [17].

The use of ANNs specifically for early OTSCC has not been previously studied. Thus, this study examined the use of ANN in prognostication of early OTSCC. We examined the use of ANNs to estimate the risk of locoregional recurrence in early-stage OTSCC. The neural network toolbox of MATLAB (R2018b version) was used to create, train, and simulate ANN for pattern recognition and classification [18]. Furthermore, the Microsoft Azure machine learning studio (Azure, 2018) was used to develop a Web-based prognostic estimator that can provide a prediction for each individual case in daily practice.

Material and methods

Patients

The clinicopathologic characteristics of 311 patients with cT1-2cN0cM0 OTSCC treated between 1979 and 2009 at the University Hospitals of Helsinki, Oulu, Turku, Tampere, and Kuopio (all in Finland) and at the A.C. Camargo Cancer Center in Sao Paulo, Brazil, were collected. The histopathologic parameters are briefly summarized in Table 1. The use of patient samples and data inquiry in this study were approved by the

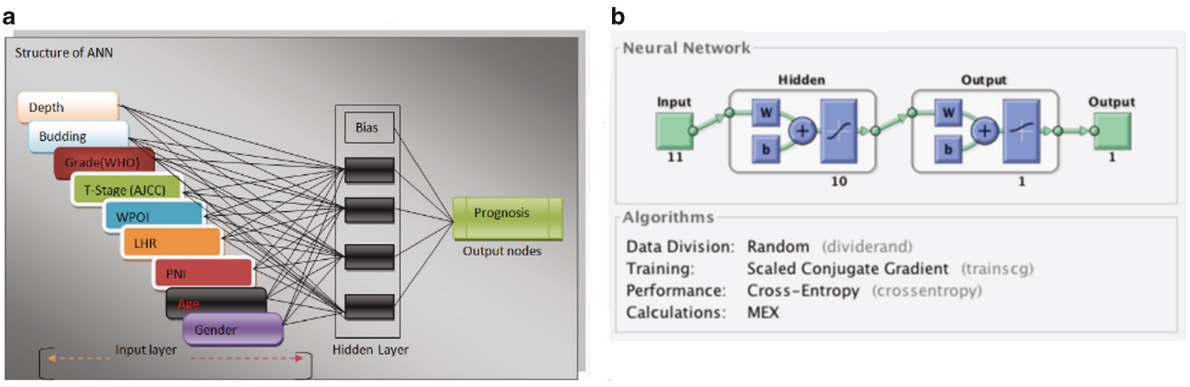

Fig. 1 Structure of ANN with prognostic factors for training the network (WPOI worst pattern of invasion, LHR lymphocytic host response, PNI perineural invasion)

Table 1 Summary of histopathologic parameters included for neural network analysis and development of the Web-based tool

Variable	Categories	Definition	Total	Recurrence
WHO grade				
	Grade I	Well-differentiated tumor	105	28
	Grade II	Moderately differentiated tumor	131	38
	Grade III	Poorly differentiated tumor	75	23
Tumor budding*				
	None	There is no tumor budding	114	26
	Low	Tumor has less than five buds	102	24
	High	Tumor has five buds or more at the invasive front	95	39
Depth of invasion				
	Superficial	Tumor less than 4 mm in depth	116	26
	Deep	Tumor with 4 mm in depth or deeper	195	63
Worst pattern of invasion (WPOI)				
	Type 1; Type 2; Type 3**	Pushing border; finger-like growth; large tumor islands	78	15
	Type 4	Small tumor islands (≤ 15 cancer cells)	190	61
	Type 5	Tumor satellites	43	13
Lymphocytic host response (LHR)				
	Type 1	Strong	53	16
	Type 2	Intermediate	116	35
	Type 3	Weak	142	38
Perineural invasion (PNI)				
	Absent	PNI was not observed	269	73
	Present	PNI was observed	42	16

*Tumor budding defined as a single cancer cell or cancer cluster of four cancer cells or less

**Types 1, 2, and 3 of worst pattern of invasion were considered in one risk group

Finnish National Supervisory Authority for Welfare and Health (VALVIRA), and by the Ethics Committee in Research of the Piracicaba Dental School, University of Campinas, São Paulo, Brazil.

Prognostic parameters

Clinicopathologic variables including age, gender, T stage (AJCC 7th), and WHO grade were included as classic prognostic factors. All histopathologic parameters were evaluated on postoperative surgical specimens stained with routine hematoxylin and eosin. The histopathologic parameters include the WHO histological grade, tumor budding, depth of invasion, worst pattern of invasion (WPOI), lymphocytic host response (LHR), and perineural invasion (PNI). We selected these prognostic factors based on our recent reports on the significance of tumor budding [19, 20], depth of invasion, and worst pattern of invasion in early OTSCC [7]. Of note, a recent study on a large cohort of OSCC [21] underlined the prognostic significance of all the prognostic factors that we used to construct the ANN.

ANN for prediction of locoregional recurrence

The dataset of 311 cases was loaded into the MATLAB workspace (The MathWorks, Inc., USA). An example of a feedforward neural network used was the basic feedforward network also known as multi-layer perceptron (MLP), with sigmoid hidden and softmax output activation function [22]. It is a two-layer network where the training of the network is based on the definition of a suitable error function, which is optimized with respect to the weights and biases in the network [22].

Prediction of locoregional recurrence

A supervised learning method was used in this study. After loading the dataset into the MATLAB workspace, prognostic factors including age, gender, stage, WHO histologic grade, tumor budding, tumor depth, WPOI, LHR, and PNI were set as inputs for the neural network, and locoregional recurrence was considered as the output. The neural network representation of the inputs, hidden neurons, and the outputs of the training process is shown in Fig. 1.

The dataset is usually divided into 70% training, 15% validation, and 15% testing sets [18, 23, 24]. In some instances, the validation and testing sets can be combined and considered to be testing sets only. This was the case with the Azure machine learning studio [25]. The prediction of locoregional recurrence was thought to be a classification task which is a form of pattern recognition. Therefore, the network was trained using *patternnet* function. It creates a standard solution neural network that classifies inputs into a target. The final process involves training the configured network for prediction [26, 27]. Follow-up time and disease-free time were included in the training of the network. The network was trained using scaled conjugate gradient backpropagation and the performance of the network was computed using cross-entropy as

shown in Fig. 2. The overall performance of this trained network was measured in terms of accuracy and area under receiving characteristic curve. Additionally, we compared the performance of this ANN model with logistic regression model in terms of accuracy.

Analyses of the importance of prognostic parameters

To examine the importance of each of the prognostic factors, each factor was removed from the inputs and the network was re-trained. The performance error was observed. This process was repeated for all the inputs shown in Fig. 1. Furthermore, it is important to gain insight into the input variables, recognize the pattern between them, and their level of correlation. Therefore, clustering offers one unique approach to achieve these. It is another excellent application of neural network, though mostly used in unsupervised learning. In this study, clustering was performed with a self-organizing map (SOM). The SOM is the most commonly used type of neural network for clustering. It has a competitive layer with neurons arranged in a grid form and hexagonal topology. The SOM network is trained with the input variables with each of them being connected to each of the neurons using the weight vector. The input data have been visualized in 2D using heatmap. Heatmaps visualize data through setting variations in coloring. The heatmap (weight planes/component planes) showing different input variables is shown in Supplementary Fig. 1. Additionally, the clustering of patients into two groups of either high- or low-risk recurrence is given in Supplementary Fig. 2.

Implementation of the Web-based prognostic tool

The process of Web deployment using the Azure machine learning Web application templates (Microsoft Corporation, USA) involves two phases. The first phase is to develop a predictive model using a machine learning studio. In the second phase, the predictive model was then accessed and thus set up as a Web service directly from Azure machine learning studio. The Web tool for recurrence prediction can be freely accessed on the Microsoft Azure cloud service [25]. Users of this Web site can enter prognostic factors to generate a personalized estimation of locoregional recurrence for the patient. The Web page is https://predictrecurrence.azurewebsites.net/Default.aspx. We tested the accuracy of our Web-based prognostic tool using 59 new cases of early OSCC treated between 1998 to 2008 at the UOPECCAN Cancer Hospital (Cascavel, Parana, Brazil). These cases were included in our previous study [28], but they were not included in the training of the ANN and were not included in the development of our Web page.

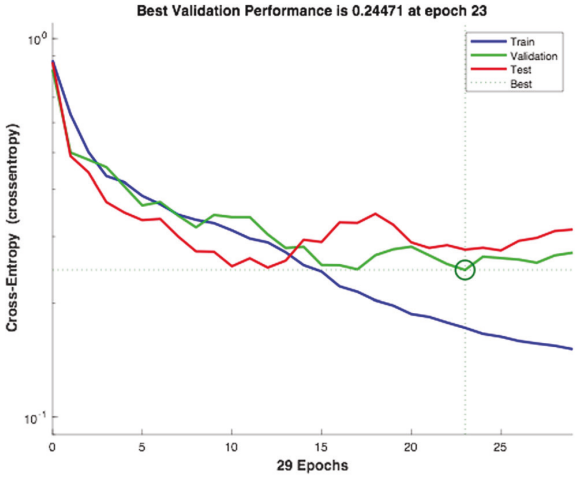

Fig. 2 The network training performance measure using cross-entropy

Results

The clinicopathologic characteristics of these patients have been previously reported [19]. This cohort consists of 165 men and 146 women. The distribution of tumors according to their diameter showed that 124 cases were staged T1 and 187 were T2. The number of patients with disease recurrences was 89 (28.6%). All cases were clinically N0 and M0. Similarly, the new cohort of 59 cases (46 men, 13 women) differs from the first one of 311 cases used in the training. The distribution according to tumor diameter showed that 22 patients had stage T1 and 37 stage T2. In terms of the distribution according to tumor budding, 14 patients showed no budding, 19 patients had less than five buds, and 26 patients had five buds or more. The mean age at diagnosis within this cohort was 56.2 (range 31–84). The number of patients with a disease recurrence was 19 (32.2%) in this cohort that was used to test the Web-based tool.

Our ANN model recognized tumor budding and depth of invasion as the most important histopathological prognostic parameters for the network to effectively predict locoregional recurrence. The heatmap presented (Supplementary Fig. 1) showed that the prognostic significance of input variables were independent. Also, the SOM network appeared to have clustered the patient into two distinct groups of high- and low-risk recurrence (Supplementary Fig. 2). In terms of accuracy of the network, the ANN yielded an overall accuracy of 92.7%. The accuracy of the ANN was higher than that given by the logistic regression model which gave an accuracy of 86.5%. The receiving operating characteristic curve of the network is given in Fig. 3. The error histogram of the training, validation, and testing phases is shown in Fig. 4a.

An overall accuracy of 88.2% was obtained with the Web prognostication tool. This was actually the overall proportion of properly classified instances between the outputs and the targets. Other metrics from the evaluation model included

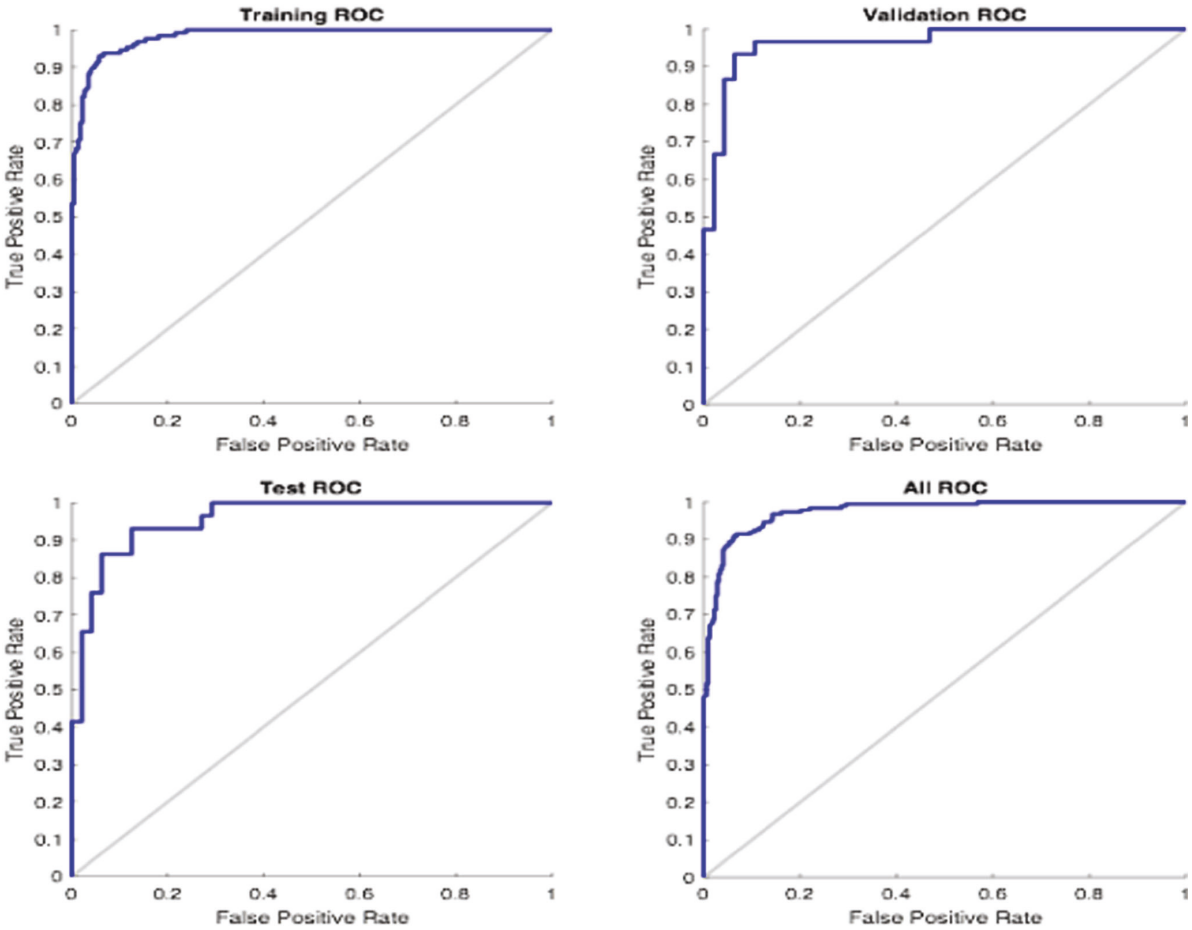

Fig. 3 The ROC curve of the trained network in MATLAB

recall, precision, and area under receiving operating characteristics curve (AUC). Recall, which is also known as sensitivity, was 71.2% and specificity was 98.9%. The positive and negative predictive values were 97.7% and 84.5%, respectively. The C-statistics (C-index) value was 97.3%. It is necessary to mention that C-index equals to the area under the receiving operating characteristics (ROC) curve shown in Fig. 4b. The performance measures for both MATLAB and Azure Web services are summarized in Table 2.

Testing/validation of the Web site with new cases

The Web site was tested with a new cohort of cases. Of the 59 cases tested, 48 cases were predicted correctly while 11 cases gave incorrect predictions when compared with the actual status of locoregional recurrence recorded by the hospital. For this new cohort of cases, an 81.4% overall accuracy was achieved using this Web-based tool. A sensitivity value of 78.9% was recorded indicating more than two thirds of the cases under consideration. With this high value of sensitivity, false-negative cases would be greatly reduced and would lead to reduction in classification (prediction) error. The overall performance metrics of the tested cohorts using our Web-based prognostic tool is presented in Table 3.

Furthermore, a specificity value of 82.5% was achieved with these test cases. In other words, the data from 33/40 patients with "low-risk" truly gave a prediction of "low-risk" in the Web-based tool. A positive predictive value of 68.2% was observed, pointing out the likelihood of a high-risk test result in individuals who actually developed a recurrence. Conversely, a negative predictive value of 89.2% was observed. The latter value indicates the probability of a low-risk result in the Web-based tool in individuals who are cases of a true low risk for recurrence. Finally, from the abovementioned performance information, a positive likelihood ratio (LR^+) of 4.5 and a negative likelihood ratio (LR^-) of 0.25 were computed. A LR^+ indicates how much more likely it is for the Web-based tool to predict a high risk for

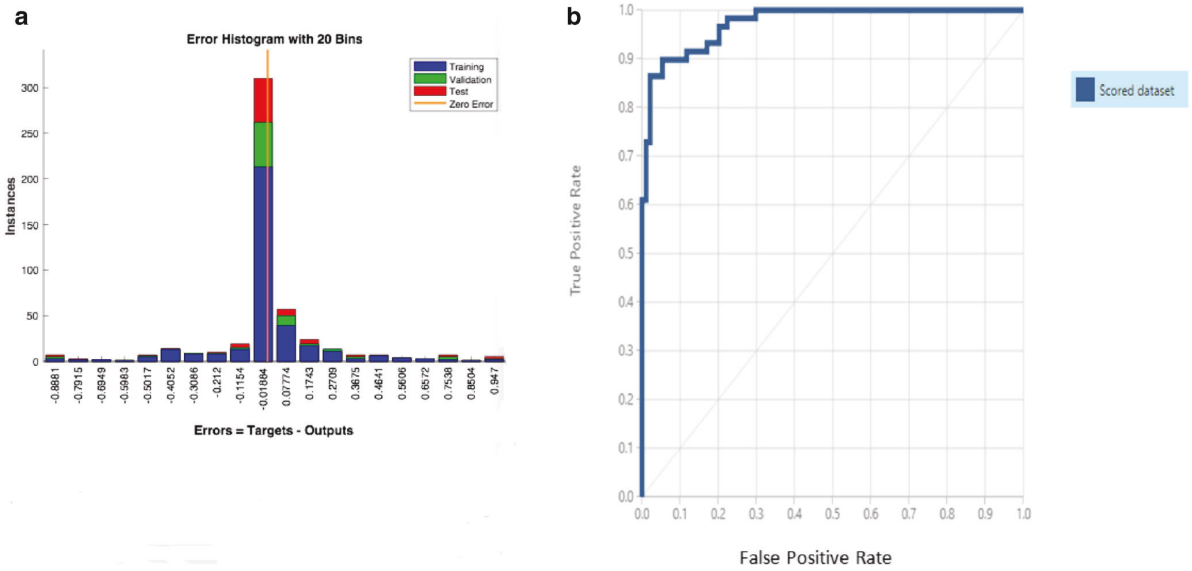

Fig. 4 **a** Error histogram showing the difference between the targets and outputs. **b** An indicative receiver operating characteristics (ROC) curve from Azure for the Web deployment

recurrence compared to a low risk for recurrence. Similarly, the LR^- value indicates how much less likely it is for the Web-based tool to predict a low risk for recurrence.

Discussion

In this study, we explored the use of ANN to predict locoregional recurrences in early-stage oral tongue cancer and we reported a better performance for the ANN compared with logistic regression. We also examined the odd ratios of each of these prognostic parameters. In addition, we developed a Web-based tool that provides the prediction as "low-risk" or "high-risk" of recurrence. The histopathologic parameters used in this study (and involved in our Web-based tool) were selected based on findings in our previous study [19] and our recent meta-analysis of many studies reporting the importance of tumor budding [20]. Depth of invasion and worst pattern of invasion have also shown promising prognostic significance in recent research by our group [7] and others [29, 30]. Perineural invasion was a valuable marker in other recent studies [8, 31]. In early OSCC, Arora et al. [21] have recently introduced a prognostic model including all the histopathologic parameters that were included in our current study. Of note, all histopathologic parameters included in this study can be evaluated using routine hematoxylin and eosin staining, and some of those parameters are routinely included in pathology reports. Moreover, multivariate analyses of many studies have underlined the prognostic significance of the selected parameters [8, 19, 21, 29, 31].

This is the first study that used an ANN and provided a Web tool for the prediction of recurrence in early-stage OTSCC. A neural network seems to have the potential to offer a better approach to data analyses and pattern recognition within data. It can build a nonlinear statistical model to examine biological systems. There is no need to identify key prognostic markers or to form a hypothesis in analyses using ANN. Interestingly, the input variables were shown to be independent of each other (Supplementary Fig. 1) as the connection patterns of these inputs are dissimilar; hence, each of these variables represents a different concept on the target variable. Thus, the issue of collinearity in machine learning is prevented

Table 2 The overall performance measures of the network

Software	Performance measures for the training of the network				
MATLAB	0.24471		92.7%		
	Network performance error		Accuracy		
Azure Machine Learning (ML) Studio	88.2%	71.2%	98.9%	0.824	97.3%
	Accuracy	Sensitivity	Specificity	F1 Score	C-index/AUC
	97.7%		84.5%		
	Positive predictive value (PPV)		Negative predictive value (NPV)		

Table 3 The performance of the Web-based tool on the newly tested cases

	Patients with OTSCC			
Web-based tool for the prediction of OTSCC recurrences	High-risk OTSCC recurrences	Low-risk OTSCC recurrences	Total	Other performance metrics
	15	7	22	68.2%
	True positive	False positive	Total_Test-positive	Positive predictive value (PPV)
	4	33	37	89.2%
	False negative	True negative	Total_Test-negative	Negative predictive value (NPV)
	19	40	59	4.5/0.35
	Total_High-risk OTSCC recurrences	Total_Low-risk OTSCC recurrences	Total_Test-cases	Positive/negative likelihood ratios
	78.9%	82.5%		
	Sensitivity	Specificity		

(no two input variables have the same effects on the target variable). Therefore, a band of dark segments from the lower-right region to the upper-right region demonstrate the potential group associated with recurrence of OTSCC (Supplementary Fig. 2).

While age of patient, tumor budding, depth, worst pattern of invasion, and perineural invasion showed significant association in terms of odd ratios to the recurrence of OTSCC, other parameters such as gender, clinical stage, histopathological grade, lymphocytic host response, and follow-up time showed low odd ratios but were included in the neural network as confounders to report independence of the significant markers and to improve the performance of ANN. Our study assessed histopathologic parameters based on postoperative surgical specimens. Therefore, patients that were recognized as "high-risk cases" (according to our Web-based tool) might benefit from postoperative adjuvant treatment (e.g., radiotherapy). Of note, recent research has showed that some histopathologic parameters (e.g., PNI, depth of invasion, and tumor budding) that were included in our study can be evaluated preoperatively either using magnetic resonance imaging [32] or satisfactory diagnostic biopsies [20]. All these are additional parameters to tumor grade, which is routinely reported for preoperative biopsies. Thus, further research should consider examining our Web-based tool in a large cohort with preoperative assessment of these histopathologic parameters. Such approach has the potential of being of great importance for treatment planning.

In the Microsoft Azure machine learning studio, a two-class neural network algorithm was used to develop the Web-based prognostic tool. It was able to produce reasonably well true positive and false negative values in recurrence prediction and had a high precision value of 97.7%. This value is also known as the positive predictive value, which explains the performance of our Web-based tool. The true positive and false positive rates can be inspected in the receiving operating characteristics (ROC) plot, also known as a precision/recall plot, and the corresponding area under the ROC curve (Fig. 4b). In our study, the area under the characteristic curve was 97.3% with a curve that tends towards the upper left corner (Fig. 4b) and far from the diagonal. This suggests a good performance of the model. The values of the likelihood ratios (LR^+ 4.5/LR^- 0.25) implied that our Web-based tool could effectively predict the cases associated with or without a recurrence of OTSCC.

This study also showed that the performance accuracy of ANN was higher than the logistic regression model. Other studies have compared ANN with traditional statistical models. For example, the study by Faradmal et al. demonstrated that the ability of prediction with ANN was higher than with the log-logistic regression model in predicting breast cancer relapse [33]. Similarly, Kazemnejad et al. compared ANN with binary logistic regression based on their performance in differentiating between disease-free patients and patients with impaired glucose tolerance or diabetes mellitus diagnosed by fasting plasma glucose [34].

In this study, the feedforward neural network produced a better performance and predicted the recurrences reasonably well by using enough neurons in the hidden layer. Computing the performance of the network using cross-entropy ensures a trained network that heavily penalizes outputs that are extremely inaccurate, with only little penalty for fairly correct classifications. Thereby, it proved to be a network with good classification capabilities. Hence, our findings indicated that ANN is an effective approach for predicting recurrences in early OTSCC. The Web-based tool provides the prediction as "low-risk" or "high-risk" of recurrence. Thus, the decision of multimodality treatment can be taken for those cases at high risk although they are diagnosed at early stage.

It is important to mention that our Web-based tool was trained with a limited number of cases. Therefore, it is possible that it will miss some predictions. In addition, the values within the follow-up time in months and disease-free time columns are not sufficiently diverse. This means that prediction from the Web-based tool for extremely high values of follow-up time could not be relied upon. Accordingly, feedback from users of this Web-based tool would be greatly appreciated. It is also hoped that this tool would be re-trained at certain intervals for better prediction based on the anticipated feedback for better prediction capacity. In addition, our current neural network did not include some parameters such as margin status and pTNM stage due to unavailability of such information

for several cases in our multicenter cohort of six institutions. Thus, we were not able to include these two parameters during the construction of our neural network, and we advise including such parameters in the further development of the neural network of early OTSCC.

In conclusion, the use of ANN is an efficient means to predict recurrence in early OTSCC. The combination of markers that were presented in our Web-based tool were able to predict recurrence successfully. With our Web-based tool, patients could be identified as high or low-risk individuals, which makes it easier to assess their prognoses. Those high-risk cases were identified with aggressive histopathologic characteristics (e.g., high intensity of tumor budding and deep invasion). Thus, such cases might benefit from elective neck dissection and postoperative oncological therapy in addition to an individualized enhanced posttreatment follow-up program. To further develop this Web-based tool, a multicenter setting should be applied to add more data to improve its effectiveness.

Acknowledgments Moral support, encouragement, and confidence received from Dr. Reno Paulo Kunz from CEONC- Oncology Center of Cascavel are acknowledged.

Author contributions Institutional coordinators: Salo T, Coletta RD, Kowalski LP, Leivo I, Mäkitie AA, Haglund C. Study concepts and study design: Alabi RO, Elmusrati M, Almangush A, Coletta RD, Salo T, Leivo I. Data acquisition and quality control of data: Mäkinen L, Sawazaki-Calone I, Kowalski LP, Leivo I. Data analysis and interpretation: Alabi RO, Elmusrati M, Almangush A, Sawazaki-Calone I, Mäkitie AA, Salo T, Leivo I. Manuscript preparation: Alabi RO, Elmusrati M, Almangush A, Mäkitie AA, Coletta RD. Manuscript review: Mäkitie AA, Leivo I, Salo T, Kowalski LP, Sawazaki-Calone I. Manuscript editing: Salo T, Leivo I, Mäkitie AA, Haglund C. All authors approved the final manuscript for submission.

Funding Information Open access funding provided by University of Turku (UTU) including Turku University Central Hospital. This work was supported by the Finnish Dental Society, the Rauha Ahokas Foundation, the K. Albin Johanssons Foundation, Turku University Hospital Fund, Helsinki University Hospital Research Fund, the Finnish Cancer Society, Finska Läkaresällskapet, the Maritza and Reino Salonen Foundation, and the UOPECCAN Center of Study and Research.

Compliance with ethical standards The institutional review boards of the Helsinki, Turku, Tampere, Oulu and Kuopio University Hospitals approved the study. The Brazilian Human Research Ethics Committee and the Finnish National Supervisory Authority for Welfare and Health (VALVIRA) also approved this study.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Rusthoven K, Ballonoff A, Raben D, Chen C (2008) Poor prognosis in patients with stage I and II oral tongue squamous cell carcinoma. Cancer 112:345–351. https://doi.org/10.1002/cncr.23183
2. Bello IO, Soini Y, Salo T (2010) Prognostic evaluation of oral tongue cancer: means, markers and perspectives (I). Oral Oncol 46:630–635. https://doi.org/10.1016/j.oraloncology.2010.06.006
3. Bello IO, Soini Y, Salo T (2010) Prognostic evaluation of oral tongue cancer: means, markers and perspectives (II). Oral Oncol 46:636–643. https://doi.org/10.1016/j.oraloncology.2010.06.008
4. Almangush A, Heikkinen I, Makitie AA, Coletta RD, Laara E, Leivo I, Salo T (2017) Prognostic biomarkers for oral tongue squamous cell carcinoma: a systematic review and meta-analysis. Br J Cancer 117:856–866. https://doi.org/10.1038/bjc.2017.244
5. Yamakawa N, Kirita T, Umeda M, Yanamoto S, Ota Y, Otsuru M, Okura M, Kurita H, Yamada SI, Hasegawa T, Aikawa T, Komori T, Ueda M (2018) Japan Oral oncology G (2019) Tumor budding and adjacent tissue at the invasive front correlate with delayed neck metastasis in clinical early-stage tongue squamous cell carcinoma. J Surg Oncol 119:370–378. https://doi.org/10.1002/jso.25334
6. Tam S, Amit M, Zafereo M, Bell D, Weber RS (2019) Depth of invasion as a predictor of nodal disease and survival in patients with oral tongue squamous cell carcinoma. Head Neck 41:177–184. https://doi.org/10.1002/hed.25506
7. Almangush A, Bello IO, Coletta RD, Makitie AA, Makinen LK, Kauppila JH, Pukkila M, Hagstrom J, Laranne J, Soini Y, Kosma VM, Koivunen P, Kelner N, Kowalski LP, Grenman R, Leivo I, Laara E, Salo T (2015) For early-stage oral tongue cancer, depth of invasion and worst pattern of invasion are the strongest pathological predictors for locoregional recurrence and mortality. Virchows Arch 467:39–46. https://doi.org/10.1007/s00428-015-1758-z
8. Yang X, Tian X, Wu K, Liu W, Li S, Zhang Z, Zhang C (2018) Prognostic impact of perineural invasion in early stage oral tongue squamous cell carcinoma: results from a prospective randomized trial. Surg Oncol 27:123–128. https://doi.org/10.1016/j.suronc.2018.02.005
9. Zheng MH, Shi KQ, Lin XF, Xiao DD, Chen LL, Liu WY, Fan YC, Chen YP (2013) A model to predict 3-month mortality risk of acute-on-chronic hepatitis B liver failure using artificial neural network. J Viral Hepat 20:248–255. https://doi.org/10.1111/j.1365-2893.2012.01647.x
10. Biglarian A, Hajizadeh E, Kazemnejad A, Zayeri F (2010) Determining of prognostic factors in gastric cancer patients using artificial neural networks. Asian Pac J Cancer Prev 11:533–536
11. Biglarian A, Hajizadeh E, Kazemnejad A, Zali M (2011) Application of artificial neural network in predicting the survival rate of gastric cancer patients. Iran J Public Health 40:80–86
12. Amiri Z, Mohammad K, Mahmoudi M, Zeraati H, Fotouhi A (2008) Assessment of gastric cancer survival: using an artificial hierarchical neural network. Pak J Biol Sci 11:1076–1084
13. Lisboa PJ (2002) A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 15:11–39
14. Patel JL, Goyal RK (2007) Applications of artificial neural networks in medical science. Curr Clin Pharmacol 2:217–226
15. Selaru FM, Xu Y, Yin J, Zou T, Liu TC, Mori Y, Abraham JM, Sato F, Wang S, Twigg C, Olaru A, Shustova V, Leytin A, Hytiroglou P, Shibata D, Harpaz N, Meltzer SJ (2002) Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. Gastroenterology 122:606–613
16. Keogan MT, Lo JY, Freed KS, Raptopoulos V, Blake S, Kamel IR, Weisinger K, Rosen MP, Nelson RC (2002) Outcome analysis of patients with acute pancreatitis by using an artificial neural network. Acad Radiol 9:410–419
17. Spelt L, Nilsson J, Andersson R, Andersson B (2013) Artificial neural networks–a method for prediction of survival following liver

resection for colorectal cancer metastases. Eur J Surg Oncol 39: 648–654. https://doi.org/10.1016/j.ejso.2013.02.024
18. Yashwant Pathak MP, Tipparaju S, Sutariya VK, Moreno W (2015) Artificial neural network for drug design, delivery and disposition. Academic Press
19. Almangush A, Coletta RD, Bello IO, Bitu C, Keski-Santti H, Makinen LK, Kauppila JH, Pukkila M, Hagstrom J, Laranne J, Tommola S, Soini Y, Kosma VM, Koivunen P, Kowalski LP, Nieminen P, Grenman R, Leivo I, Salo T (2015) A simple novel prognostic model for early stage oral tongue cancer. Int J Oral Maxillofac Surg 44:143–150. https://doi.org/10.1016/j.ijom.2014.10.004
20. Almangush A, Pirinen M, Heikkinen I, Makitie AA, Salo T, Leivo I (2018) Tumour budding in oral squamous cell carcinoma: a meta-analysis. Br J Cancer 118:577–586. https://doi.org/10.1038/bjc.2017.425
21. Arora A, Husain N, Bansal A, Neyaz A, Jaiswal R, Jain K, Chaturvedi A, Anand N, Malhotra K, Shukla S (2017) Development of a new outcome prediction model in early-stage squamous cell carcinoma of the oral cavity based on histopathologic parameters with multivariate analysis: the Aditi-Nuzhat Lymph-node Prediction Score (ANLPS) system. Am J Surg Pathol 41:950–960. https://doi.org/10.1097/PAS.0000000000000843
22. Bishop C (2006) Pattern recognition and machine learning. Springer, New York
23. Jeong H-YOM, Yen NY, James J-H (2013) Advances in computer science and its application. Springer, New York
24. Chen GLF, Shojafar (2016) Fuzzy system and data mining: proceedings of FSDM 2015. IOS Press, Amsterdam
25. Studio AM (2018) Azure machine learning documentation. In Docs.Microsoft. Redmond,Washington: Microsoft Corporation
26. T.M M (2006) The discipline of machine learning: Carnegie Mellon University. Pittsburg. Carnegie Mellon University, School of Computer Science, Machine Learning Department, Pennsylvania, United States
27. WIaF E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, San Francisco
28. Sawazaki-Calone I, Rangel A, Bueno AG, Morais CF, Nagai HM, Kunz RP, Souza RL, Rutkauskis L, Salo T, Almangush A, Coletta RD (2015) The prognostic value of histopathological grading systems in oral squamous cell carcinomas. Oral Dis 21:755–761. https://doi.org/10.1111/odi.12343
29. Ganly I, Patel S, Shah J (2012) Early stage squamous cell cancer of the oral tongue–clinicopathologic features affecting outcome. Cancer 118:101–111. https://doi.org/10.1002/cncr.26229
30. Sinha N, Rigby MH, McNeil ML, Taylor SM, Trites JR, Hart RD, Bullock MJ (2018) The histologic risk model is a useful and inexpensive tool to assess risk of recurrence and death in stage I or II squamous cell carcinoma of tongue and floor of mouth. Mod Pathol 31:772–779. https://doi.org/10.1038/modpathol.2017.183
31. Tai SK, Li WY, Chu PY, Chang SY, Tsai TL, Wang YF, Huang JL (2012) Risks and clinical implications of perineural invasion in T1-2 oral tongue squamous cell carcinoma. Head Neck 34:994–1001. https://doi.org/10.1002/hed.21846
32. Chatzistefanou I, Lubek J, Markou K, Ord RA (2017) The role of perineural invasion in treatment decisions for oral cancer patients: a review of the literature. J Craniomaxillofac Surg 45:821–825. https://doi.org/10.1016/j.jcms.2017.02.022
33. Faradmal J, Soltanian AR, Roshanaei G, Khodabakhshi R, Kasaeian A (2014) Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse. Asian Pac J Cancer Prev 15:5883–5888
34. Kazemnejad A, Batvandi Z, Faradmal J (2010) Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes. East Mediterr Health J 16:615–620

International Journal of Medical Informatics 136 (2020) 104068

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer

Rasheed Omobolaji Alabi[a,*], Mohammed Elmusrati[a], Iris Sawazaki-Calone[b], Luiz Paulo Kowalski[c], Caj Haglund[d], Ricardo D. Coletta[e], Antti A. Mäkitie[f,g,h], Tuula Salo[i,j,k], Alhadi Almangush[l,m,n,1], Ilmo Leivo[m,1]

[a] *Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland*
[b] *Oral Pathology and Oral Medicine, Dentistry School, Western Parana State University, Cascavel, PR, Brazil*
[c] *Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo, SP, Brazil*
[d] *Research Programs Unit, Translational Cancer Biology, University of Helsinki, Helsinki, Finland*
[e] *Department of Oral Diagnosis, School of Dentistry, State University of Campinas, Piracicaba, São Paulo, Brazil*
[f] *Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[g] *Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland*
[h] *Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden*
[i] *Department of Pathology, University of Helsinki, Helsinki, Finland*
[j] *Department of Oral and Maxillofacial Diseases, University of Helsinki, Helsinki, Finland*
[k] *Cancer and Translational Medicine Research Unit, Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland*
[l] *Department of Pathology, University of Helsinki, Helsinki, Finland*
[m] *Institute of Biomedicine, Pathology, University of Turku, Turku, Finland*
[n] *Faculty of Dentistry, University of Misurata, Misurata, Libya*

ARTICLE INFO

Keywords:
Artificial intelligence
Oral tongue cancer
Machine learning
Prediction

ABSTRACT

Background: The proper estimate of the risk of recurrences in early-stage oral tongue squamous cell carcinoma (OTSCC) is mandatory for individual treatment-decision making. However, this remains a challenge even for experienced multidisciplinary centers.
Objectives: We compared the performance of four machine learning (ML) algorithms for predicting the risk of locoregional recurrences in patients with OTSCC. These algorithms were Support Vector Machine (SVM), Naive Bayes (NB), Boosted Decision Tree (BDT), and Decision Forest (DF).
Materials and methods: The study cohort comprised 311 cases from the five University Hospitals in Finland and A.C. Camargo Cancer Center, São Paulo, Brazil. For comparison of the algorithms, we used the harmonic mean of precision and recall called F1 score, specificity, and accuracy values. These algorithms and their corresponding permutation feature importance (PFI) with the input parameters were externally tested on 59 new cases. Furthermore, we compared the performance of the algorithm that showed the highest prediction accuracy with the prognostic significance of depth of invasion (DOI).
Results: The results showed that the average specificity of all the algorithms was 71% · The SVM showed an accuracy of 68% and F1 score of 0.63, NB an accuracy of 70% and F1 score of 0.64, BDT an accuracy of 81% and F1 score of 0.78, and DF an accuracy of 78% and F1 score of 0.70. Additionally, these algorithms outperformed the DOI-based approach, which gave an accuracy of 63%. With PFI-analysis, there was no significant difference in the overall accuracies of three of the algorithms; PFI-BDT accuracy increased to 83.1%, PFI-DF increased to 80%, PFI-SVM decreased to 64.4%, while PFI-NB accuracy increased significantly to 81.4%.
Conclusions: Our findings show that the best classification accuracy was achieved with the boosted decision tree algorithm. Additionally, these algorithms outperformed the DOI-based approach. Furthermore, with few parameters identified in the PFI analysis, ML technique still showed the ability to predict locoregional recurrence. The application of boosted decision tree machine learning algorithm can stratify OTSCC patients and thus aid in their individual treatment planning.

* Corresponding author.
E-mail address: rasheed.alabi@student.uwasa.fi (R.O. Alabi).
[1] The last two authors have equal contributions.

https://doi.org/10.1016/j.ijmedinf.2019.104068
Received 28 May 2019; Received in revised form 10 December 2019; Accepted 26 December 2019

1. Introduction

Oral tongue squamous cell carcinoma (OTSCC) refers to squamous cell carcinoma that arises from the anterior two thirds of the tongue (also known as mobile tongue). It is usually reported as part of oral squamous cell carcinoma (OSCC), which includes all anatomical subsites of the oral cavity. A recent international study including 22 registries reported 89,212 incident cases of OTSCC and an increasing annual incidence [1], which has been confirmed by others [2]. The primary treatment of choice for OTSCC is surgical excision. However, even early-stage tumors may express a pattern of aggressive behavior [3,4]. Thus, OTSCC with aggressive behavior and those with advanced stage require multimodality treatment including neck dissection and adjuvant (chemo)radiotherapy. Therefore, it is important to precisely estimate the clinical behavior and outcome of OTSCC. Predicting the risk of recurrences is one of the important assessments for the clinician during treatment planning. More importantly, early diagnosis and predicting the risk of recurrences form a milestone in the management of OTSCC as the recent analysis of Finnish cases reported that about 67% of OTSCC cases were diagnosed at an early stage (I-II) [5]. With accurate and timely recurrence prediction, high-risk cases of OTSCC can be identified and multimodality treatment applied accordingly. In a large cohort of early OTSCC, about one fourth of cases (27.8%) developed a recurrence, and all of them might have benefitted from early prediction and corresponding treatment planning [6].

Many recent studies have examined the use of machine learning (ML) techniques for prognostication of different cancers [7,8]. Interestingly, predicting patient outcome by ML techniques has shown better accuracy than Cox regression [9]. This is why the use of ML has been in active research focus during recent years. For instance, ML techniques have been used to predict the outcome of various cancer types [10–12] and a web-based tool based on artificial neural network to predict outcome in cancer has been reported [13].

In this study, we examined four different ML algorithms, namely, support vector machine (SVM), naive Bayes (NB), boosted decision tree (BDT), and decision forest (DF) in terms of their performances to predict locoregional recurrence in OTSCC patients. Also, the predictive performance of a permutation feature importance (PFI) of these algorithms was evaluated. Many researchers have used this approach for comparing ML techniques for survival prediction in different malignancies like breast and lung cancers [14–17]. Tapak et al. examined six ML algorithms and two traditional methods for the prediction of breast cancer survival and metastasis [15]. In our study, we aimed to identify the best algorithm that would effectively classify patients as either low-risk or high-risk OTSCC recurrence. The algorithm with the overall best classification performance was further compared to a recently reported risk model based on the depth of invasion (DOI) [18]. This comparison was a result of the fact that DOI of 4 mm or deeper has been considered to be a factor that accurately predicts locoregional recurrence [6]. Moreover, the recent American Joint Committee on Cancer (AJCC) 8th edition incorporated depth of invasion (DOI) into T-stage [19]. Similarly, the study by Almangush et al. suggested that DOI is one of the strongest pathological predictors for locoregional recurrence [6]. This suggestion is in agreement with reports by others [20,21].

We hypothesize that the application of the above-mentioned supervised learning classifiers may be used in the prediction of OTSCC locoregional recurrences and will thereby add value for the management of OTSCC.

2. Material and methods

2.1. Patients

We used data from a study cohort comprising patients treated at the five Finnish University Hospitals of Helsinki, Oulu, Turku, Tampere, and Kuopio and at the A.C. Camargo Cancer Center, Sao Paulo, Brazil. This is a multicenter study from six institutions and data were provided for many cases as locoregional recurrences without specification. The clinicopathologic characteristics of this cohort have been previously reported and summarized [22]. The primary treatment for all cases was surgical excision. In addition, some cases received neck dissection and/or adjuvant radiotherapy. The parameters included were age, gender, T-

Table 1
The parameters contained in the dataset and their respective descriptors.

Number	Parameters	Description	Type
1	Age	Age at the time of diagnosis.	Discrete
2	Gender	The sexual orientation of the patient	Categorical 1 = Male; 2 = Female
3	T-stage	T stage describing tumor size	Categorical 1 = T1; 2 = T2.
4	WHO Grade	Histopathologic grading according to World Health Organization (WHO) criteria	Categorical 1 = Grade I; 2 = Grade II; 3 = Grade III
5	Tumor budding	Tumor budding is defined as the presence of single cells or small clusters of cancer cells detached from the main tumor mass	Categorical 0 = No budding; 1 < 5 buds; 2 for ≥ 5 buds.
6	Tumor depth	This is the measure of tumor depth of invasion. It was measured in millimetres (mm)	Categorical 1 for < 4 mm, 2 for ≥ 4 mm
7	WPOI	Worst pattern of invasion	Categorical Value of 0 for WPOI type 1 to 3; Value of 1 for WPOI type 4; Value of 3 for WPOI type 5.
8	LHR	Lymphocytic host response	Categorical Value of 0 for LHR type 1; Value of 1 for LHR type 2; Value of 3 for LHR type 3.
9	PNI	Perineural invasion	Categorical 0 = Absent; 1 = Present
10	Treatment	This indicates the type of treatment offered for the patient. It could either be surgery alone or adjuvant (chemo)radiotherapy in addition to the surgery	Categorical 0 = Surgery alone 1 = Surgery + Adjuvant (chemo)radiotherapy
11	Neck treatment	This variable indicates whether neck dissection was performed or not	Categorical 0 = No neck dissection 1 = Neck dissection performed.
12	Recurrence[a]	The occurrence of disease after treatment	Categorical 0 = Low-Risk; 1 = High-Risk

[a] Recurrence was considered as the output/target label.

stage, WHO grade, tumor budding, depth of invasion, worst pattern of invasion (WPOI), lymphocytic host response (LHR), and perineural invasion (PNI) as shown in Table 1. Several studies have confirmed the prognostic importance of these variables [6,13,22–25]. Neck dissection and adjuvant radiotherapy were also included in the machine learning algorithms due to the impact of variation in the treatment modality that might influence the risk of recurrence. The use of patient samples and data inquiry were approved by the Hospital Research Ethics Committees of all individual hospitals, by the Finnish National Supervisory Authority for Welfare and Health (VALVIRA) and by the Brazilian Human Research Ethics Committee.

2.2. The classification algorithms examined

The algorithms considered in this study are basic and have been commonly used in other cancer studies [14–18].

2.2.1. Support vector machine (SVM)

Support vector machine (SVM) is an elegant and powerful ML technique extensively used for both classification and regression problems [26]. This is due to its ability to classify non-linearly separable patterns by projecting the original features into a higher dimensional space (hyperplane) [27].

2.2.2. Naive Bayes (NB)

Naive Bayes (NB) is known as Bayes point machine in the Azure ML studio and it is based on the generally-known Bayes theorem [26,27]. The algorithm operates by learning and estimating the prior probability of belonging to each class using the training data [27,28].

2.2.3. Boosted Decision Tree (BDT)

Boosted Decision Tree (BDT) with gradient boosting machine was the subtype of BDT used in this study. It is an ensemble learning method where the second tree corrects the errors of the first tree, the third tree corrects the errors in the second trees, the fourth tree corrects the errors in the third trees, etc. Predictions are based on the entire ensemble of trees [27,28].

2.2.4. Decision Forest (DF)

Decision Forest (DF) relies on the combination of multiple related models to get better results and a more generalized model. Therefore, it works by using a bootstrapped sample of data to build each tree where only a proportion of the variable set is considered for each tree. Each tree in the decision forest outputs a frequency histogram of labels that is non-normalized. These frequency histograms were aggregated in the process that sums these histograms and then normalizes the results to get the probabilities for each label [27].

2.2.5. Permutation Feature Importance (PFI)

Permutation Feature Importance (PFI) is a model-agnostic ranker feature ranker that computes the scores for each of the variables contained in a dataset. It basically examines the contribution of each feature to the overall predictive performance of the algorithm [27].

2.3. Evaluation of the performance of the algorithms

The performance metrics were aimed to evaluate how the algorithms performed [29–31]. Most of these metrics have been previously used in other studies [15,32]. However, in addition to accuracy, only two (F1 score and specificity) of these statistical measures that are medically more relevant in the clinic, were discussed in the current study.

3. The training-validation phase for the algorithms in Microsoft Azure for prediction of recurrence

Microsoft Azure Machine Learning Studio (Azure ML 2019) was used in this study [27]. The data was preprocessed to handle missing values. The input parameters were age, gender, stage, grade, tumor budding, depth of invasion (DOI), worst pattern of invasion (WPOI), lymphocytic host response (LHR), perineural invasion (PNI) and treatment given, while the target output was locoregional recurrence. Disease-free survival (DFS) time of the cases ranged from 1 to 267 months. Specifically, the DFS in cases with recurrence varied between 1 and 120 months. Firstly, a potential class imbalance with respect to the number of patients who experienced a tumor recurrence in the target class (locoregional recurrence) was handled by up-sampling in order to balance the classes used in the training. Synthetic minority oversampling technique (SMOTE) [33] offers a better way to handle imbalance than simply duplicating existing cases. The dataset and the corresponding samples are therefore more general [33]. The dataset was divided into two sets of training and validation. Due to the relatively limited amount of data, a 5-fold cross validation was used with 50% training and 50% validation {50:50} percentage splitting sets [15]. Each of the algorithms of interest was then configured as shown in Fig. 1 [27,28]. After training, the algorithms were evaluated for the various quality metrics (Table 3.

Furthermore, these algorithms were further tested with new cases (Section 3.1). The result obtained from this approach was considered as the gold standard in this study as it gives an account of how the algorithm is expected to predict in reality. Also, it addresses any concerns about the generalizability of the trained models. In addition, the contribution of each of the input variables on the predictive ability of each model was examined using permutation feature importance (PFI) analysis. Their contributions were given in the form of PFI-performance scores. To avoid bias in the algorithm, disease-free survival and treatment were removed from the PFI analysis that was aimed to examine the predictive ability of each variables. The input features with positive scores were selected. Also, only one of the inputs was selected when two or more inputs give the same negative score. The variables with least scores were not selected. These selected variables were used to train the algorithms. The given accuracies in the PFI analysis were compared with the accuracies obtained without PFI. Similarly, the PFI-based algorithms were tested with new cases.

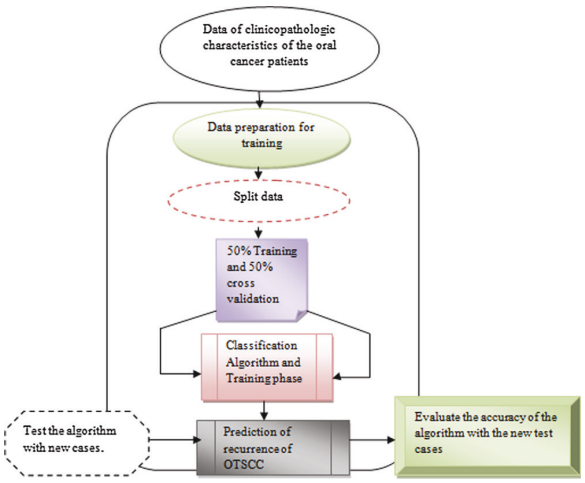

Fig. 1. The training process in azure machine learning studio.

3.1. Testing performance of the model with new cases

In this phase, the trained algorithms were tested with 59 new cohort cases that were not included in the training or in the validation sets (Fig. 1). These new independent data were obtained from a cancer center in Brazil. The results are presented in Table 4. The PFI-based models were also tested with these new cases as presented in Table 5.

3.2. Comparison with the depth of invasion (DOI)

The algorithm that showed the highest overall accuracy when tested with these new external cases (Section 3.1) was also compared with the depth of invasion (DOI) based model as shown in Fig. 3.

4. Results

4.1. Data description

The study cohort included 311 patients with cT1-T2cN0M0 OTSCC; 165 men and 146 women, resulting in a male-to-female range of 1.1:1. Out of these 311 cases, 57 cases had missing details about any post-operative treatment information. Therefore, these cases were excluded and the machine learning training was performed with 254 cases. These cases included 141 men and 113 women with the mean age at diagnosis was 61.51 (SD ± 14.81: range 10–95) and the median age was 62.0 years. The distribution according to tumor diameter showed that 100 patients had stage T1 and 154 stage T2. The histopathologic parameters are briefly summarized in Table 2. In terms of the treatment, 157 patients had surgery alone while 97 had adjuvant (chemo)radiotherapy (92 radiotherapy and 5 chemoradiotherapy). Similarly, 185 had neck dissection while 69 had no neck dissection performed. Thus, from the 185 patients who had neck dissection, 43% were exposed to adjuvant radiotherapy while 57% had only surgery as single-modality treatment. Similarly, out of the 69 cases who had no dissection performed, 25% were exposed to adjuvant radiotherapy while 75% had only surgery.

The number of patients with disease recurrences was 68 (26.8%). While the disease-free survival (DFS) time ranged from 1 to 267 months, the DFS time for cases with a locoregional recurrence was between 1 to 120 months. Overall, 89.6% of the recurrences occurred in the first 2 years, while 10.45% recurrence was recorded after 2 years. The mean follow-up time was 75 months (SD ± 64.6; range 1–258 months) and the median was 60 months. Similarly, for the 59 new OSCC cases used for external testing, DFS time varied between 1 to 146 months. Also, 74% had a recurrence in the first year, 16% after the first and before end of second year, and 10% of the patients recurred after the second year. The mean age in this external validation cohort was 56.2 years (range, 31–84 years). All these new cases had neck dissection, where 34 cases had surgery alone while 25 had adjuvant (chemo) radiotherapy (22 radiotherapy and 3 chemoradiotherapy). The DOI model performance in terms of accuracy in the training set was 47.2% and the overall accuracy in the new cohorts used for external validation was 63%.

4.2. Performance metrics for the algorithms

The distribution of true and false positives, true and false negatives, and other performance metrics for the algorithms in the training phase are given in Fig. 2a and Table 3, respectively. During the training phase, decision forest showed the highest accuracy while naive Bayes and decision forest showed the best area under receiving operating characteristic (AUC of ROC). When these algorithms were tested on the 59 new external cases from the cancer center in Brazil, the average specificity of all the algorithms was 71%. The tested algorithms i.e. support vector machine, naive Bayes, decision forest, and boosted decision tree gave an overall accuracy of 68%, 70%, 78% and 81%, respectively. The details of the performance of parameters with this new cohorts are given in Table 4. Considering the harmonic mean of precision and recall, that is, F1 score, the support vector machine, naive Bayes, decision forest, and boosted decision tree gave 0.63, 0.64, 0.70 and 0.78, respectively. Therefore, the best overall classification performance to predict recurrence was achieved with the boosted decision tree algorithm. Comparison of the boosted decision tree algorithm and the DOI model is shown in Fig. 3; the DOI model showed an accuracy of 63% where 54.1% of the patients would be observed, thereby not subjected to adjuvant therapy or elective neck dissection (END). The boosted decision tree on the other hand showed 81% overall accuracy where 21.1% of the patients would have been observed and not subjected to END. Similarly, about half (49.5%) of the patients were correctly identified as having OTSCC recurrence using the DOI model. Boosted decision tree machine learning technique correctly identified 78.9% as having OTSCC recurrence as shown in Fig. 3. Thus, each of these algorithms performed significantly better than the DOI-based model.

The results of the permutation feature importance (PFI) analyses are given in Table 5. The PFI scores were calculated for each feature independently. A zero score is returned when there is no difference in the performance metrics before and after PFI of that feature. Similarly, a negative score is returned when a random PFI of that feature produced a higher accuracy and lower error (better performance metrics) compared to the performance before PFI was applied. Moreover, a higher importance score (positive) gives an indication of the contribution of that feature to the predictive ability of the model. The PFI of boosted decision tree (PFI-BDT) showed the highest accuracy (83.1%). Also, it was observed that the accuracy of BDT increased from 81.0% to 83.1% and DF increased from 78% to 80%, while SVM showed a reduction in accuracy from 68% to 64.4% in the PFI analysis. Interestingly, the accuracy of NB increased significantly from 70.0% to 81.4% in the permutation feature importance fitting. The ranking of the scores of the

Table 2
Summary of histopathologic parameters included for the machine learning training.

Variable	Category (Definition)	Number
WHO grade		
	Grade I (Well-differentiated tumor)	78
	Grade II (Moderately-differentiated tumor)	103
	Grade III (Poorly-differentiated tumor)	73
Tumor budding		
	None (There is no tumor budding)	93
	Low (Tumor has less than five buds)	85
	High (Tumor has five buds or more at the invasive front)	76
Depth of invasion		
	Superficial (Tumor < 4 mm in depth)	96
	Deep (Tumor ≥ 4 mm in depth)	158
Worst pattern of invasion (WPOI)		
	Type 1 (Pushing border)	64
	Type 2 (Finger-like growth)	
	Type 3 (Large tumor islands)	
	Type 4(Small tumor islands of ≤ 15 cancer cells)	158
	Type 5 (Tumor satellites)	32
Lymphocytic host response (LHR)		
	Type 1 (Strong)	36
	Type 2 (Intermediate)	88
	Type 3 (Weak)	130
Perineural invasion (PNI)		
	Absent (PNI was not observed)	223
	Present (PNI was observed)	31

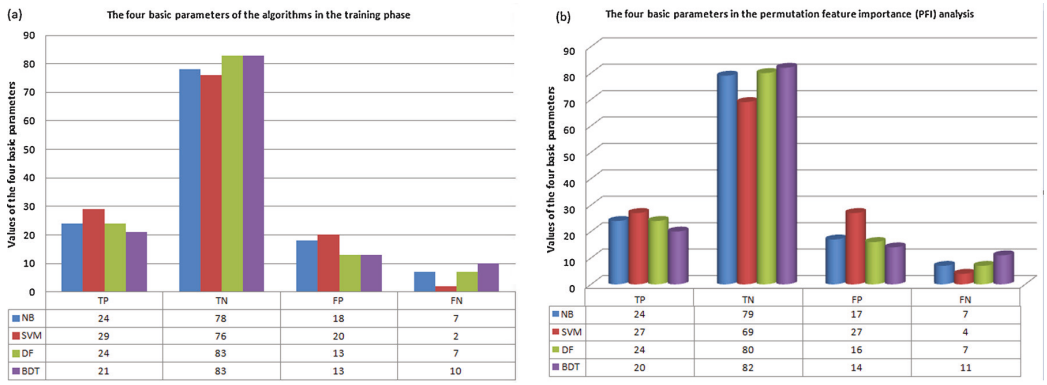

Fig. 2. The classification results of the four basic parameters for each algorithm in the training and also for PFI analysis.
(TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, BDT: Boosted Decision Tree, SVM: Support Vector Machine, NB: Naive Bayes, and DF: Decision Forest).

features is as shown in Table 5.

5. Discussion

The present study compared the performance of ML algorithms to stratify patients with OTSCC into low or high-recurrence risk group. In this regard, four ML algorithms, namely, boosted decision tree, naive Bayes, support vector machine, and decision forest were examined. We found that the performance of these techniques was higher than that of depth of invasion (DOI) based approach. Our multicenter cohort of cases is one of the largest published series. Majority of the previous publications including hundreds of cases have mixed early-stage cases with those with advanced stage, and/or have mixed different subsites of the oral cavity (e.g. oral tongue with floor of mouth and retromolar region). Therefore, heterogeneity of such series makes it difficult to identify robust prognostic markers. The advantage of our homogenous cohort (only early stage and only oral tongue) allows for reaching definitive conclusions that can be considered to be applied in daily practice.

Although significant progress has been made in early diagnostics, treatment strategies and prevention of OTSCC in recent years, the prognosis of OTSCC is poor due to aggressive local invasion and metastasis, leading to recurrence. The mortality rates in cases with recurrence has been reported to be very high [34]. When recurrence is diagnosed earlier, the mortality rates have been reported to decrease [35,36]. The reported rates of recurrence in oral squamous cell carcinoma range from 6.9% to 37.4% of patients [37,38]. This is in accordance with the 26.8% locoregional recurrence rate within the dataset used in this study. Improved prediction of locoregional recurrences in early-stage OTSCC can lead to an adjusted, patient-oriented follow-up program. For example, based on prediction of the patient as a high-risk case, a customized surveillance could be organized instead of the general follow-up program.

Abundant studies exist that have considered DOI as a strong histologic feature that correlates with locoregional recurrence. The machine learning algorithms examined in this study, however, outperformed the power of prediction of locoregional recurrence based on DOI. However, it will offer a better approach with significant accuracy in stratifying the patients as carrying a high- or low-risk for recurrence. Therefore, it seems obvious, that the described challenge in the treatment-decision making would be successfully addressed by the machine learning model due to increased specificity, F1 score and overall accuracies of the ML

Table 4
The performance of the algorithms with external cases.

Parameter	SVM	NB	BDT	DF
True Positive (TP)	16	16	15	15
False Positive (FP)	16	15	07	09
True Negative (TN)	24	25	33	31
False Negative (FN)	03	03	04	04
Sensitivity	0.84	0.84	0.79	0.79
Specificity	0.60	0.63	0.83	0.78
Precision (PPV)	0.50	0.52	0.76	0.63
NPV	0.89	0.89	0.89	0.89
LR^+	2.10	2.27	4.65	3.59
LR^-	0.27	0.25	0.25	0.27
F1 Score	0.63	0.64	0.78	0.70
Accuracy	**68%**	**70%**	**81%**	**78%**

Table 3
The overall performance metrics of the classifiers in the training phase.

50% Training and 50% Testing Cross Validation Scheme									
Algorithm	Sensitivity	Specificity	Precision	NPV	LR^+	LR^-	F1 Score	AUC	Accuracy %
NB	0.67	0.81	0.77	0.92	3.53	0.41	0.66	0.89	80.0
SVM	0.94	0.79	0.59	0.97	4.48	0.08	0.73	0.88	82.7
DF	0.77	0.86	0.65	0.92	5.50	0.27	0.71	0.89	84.0
BDT	0.68	0.87	0.62	0.89	5.23	0.37	0.65	0.82	82.0
PFI-NB	0.77	0.83	0.59	0.92	4.53	0.28	0.67	0.89	81.0
PFI-SVM	0.87	0.72	0.50	0.95	3.11	0.18	0.64	0.87	76.0
PFI-DF	0.77	0.83	0.60	0.92	4.53	0.28	0.68	0.85	82.0
PFI-BDT	0.65	0.85	0.59	0.88	4.33	0.41	0.62	0.84	80.0

BDT = Boosted Decision Tree, SVM = Support Vector Machine, BPM = Bayes Point Machine, DF = Decision Forest, Precision (PPV = Predictive positive value), NPV = Negative predictive value, LR^+ = Positive likelihood ratio and LR^- = Negative likelihood ratio, Sensitivity (recall), Area under receiving operating characteristics curve (AUC), and CDE = Custom Designed Ensemble.

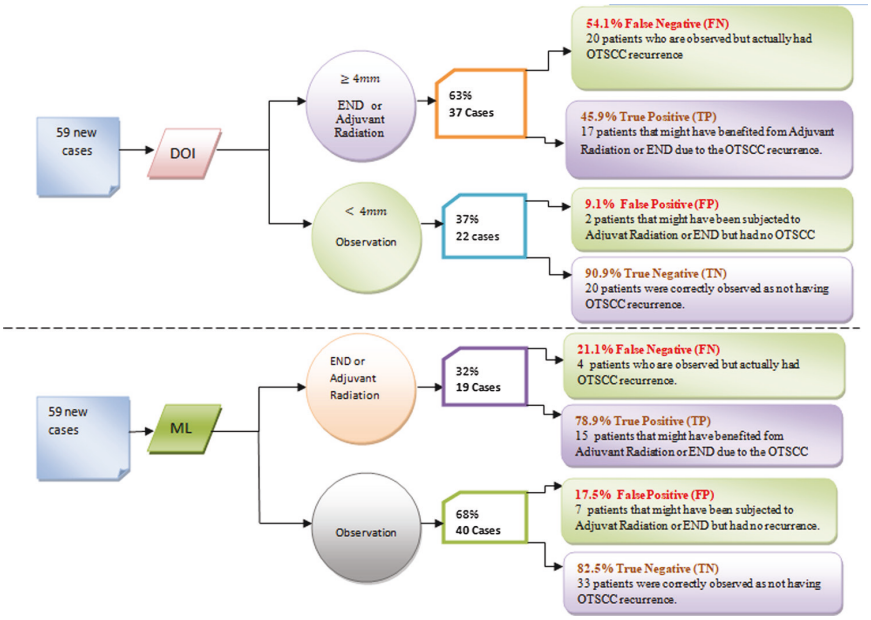

Fig. 3. The comparison of the boosted decision tree algorithm to the depth of Invasion model [18].

Table 5
Permutation Feature Importance (PFI) of the algorithms.

PFI-DF		PFI-BDT		PFI-SVM		PFI-NB	
Features	Scores	Features	Scores	Features	Scores	Features	Scores
PNI	0.0078	Age	0.0315	Gender	0.0079	Age	0.0079
Depth	0.0000	Depth	0.0236	Stage	0.0079	Gender	0.0079
Tumor Budding	0.0158[a]	WPOI	0.0236	Tumor Budding	0.0079	Stage	0.0079
Stage	0.0315[a]	PNI	0.0079	Depth	0.0079	Depth	0.0079
LHR	0.0315[a]	Tumor Budding	0.0000	LHR	0.0079	Grade	0.0000
Gender	0.0394[a]	LHR	0.0079[a]	PNI	0.0079	Tumor Budding	0.0079[a]
Grade	0.0394[a]	Stage	0.0158[a]	Age	0.0000	LHR	0.0079[a]
WPOI	0.0472[a]	Grade	0.0158[a]	Grade	0.0000	PNI	0.0079[a]
Age	0.0551[a]	Gender	0.0236[a]	WPOI	0.0000	WPOI	0.0236[a]
Accuracy (External Testing)	80.0%	**Accuracy (External Testing)**	83.1%	**Accuracy (External Testing)**	64.4%	**Accuracy (External Testing)**	81.4%

[a] Negative score. DF: Decision Forest, BDT: Boosted Decision Forest, SVM: Support Vector Machine, NB: Naive Bayes. WPOI: Worst Pattern of Invasion, PNI: Perineural Invasion, LHR: Lymphocytic host response.

algorithms. Thus, this study has potentially high impact to clinicians in the management of early OTSCC.

With regards to the performance metrics examined, F1 score used as the benchmark to choose the best algorithm as it finds the optimal blend between two other performance metrics (precision and recall). As shown in Table 4, the F1 score for the boosted decision tree algorithm showed to be very good at stratifying the patients as having either low-risk or high-risk of recurrence of OTSCC. This justifies why boosted decision tree was compared to the DOI as shown in Fig. 3 [18]. It is important to note that the support vector machine showed promising evaluation performance metrics in the training phase. This is due to the fact that it is an empirical risk minimizer algorithm. Hence, it is not usually prone to overfitting related issue as it avoids the danger of getting trapped into local minima [39]. However, the ensemble algorithms performed better than the support vector machine because they were able to create a fleet of algorithms with relatively similar bias and subsequently combining their outputs to reduce variance.

Furthermore, a major challenge in the treatment of patients with early OTSCC is in finding the right parameters that predict prognosis and help to properly identify patients at high risk of locoregional recurrences. This would carry the potential to minimize the incidence treatment failure of patients with OTSCC [35]. With the PFI-analyses, the exact contribution of each parameter to the predictive ability of the machine learning algorithms was known. Interestingly, there was no significant difference in the overall accuracies achieved in the ensemble methods (decision forest and boosted decision tree) with reduced parameters identified in the PFI analyses compared to the algorithms without PFI. Therefore, the cost and resources associated with getting numerous parameters can be properly managed. Also, the time taken to properly prepare an individualized treatment plan for the patients can be improved. This is because a few but important features that are needed for the ML algorithms were identified in the PFI analysis while producing the same range of prediction accuracies. Thus, predicting recurrence with such accuracy as shown in this study would be crucial to the clinicians in terms of management decisions.

Numerous studies have compared the performance of various machine learning classifiers to predict an outcome of interest in cancer. Tapak et al. compared various machine learning classifiers in series of 550 breast cancer patients, and found that the support vector machine predicted survival better than other classifiers [15]. Similarly, the study

by Tseng et al. compared decision tree ML technique with a traditional statistical model such as logistic regression in series of 673 oral cancer patients and the decision tree was found to perform better [40]. De Melo et al. used decision tree to evaluate the quality of life among patients with head and neck cancer [41]. Similarly, Sumbaly et al. used the decision tree in the diagnosis of breast cancer [42]. The decision forest also produced the highest prognostic performance when compared with other machine algorithms by Zhang et al. for the radiomics-based prediction of failure in advanced nasopharyngeal carcinoma [43].

In conclusion, this study investigated four different ML algorithms and found that the boosted decision tree algorithm showed the best overall performance accuracy. Due to the sensitive nature of the application of machine learning in medicine, it is important for these algorithms to produce very high accuracies. In this study, the ensemble algorithms such as the boosted decision tree and the decision forest algorithms performed better than non-ensemble algorithms such as support vector machine, naive Bayes and a method based on depth of invasion. Therefore, the ensemble machine algorithms should be considered in medical applications. Presently, it is challenging for clinicians to assess the outcomes of clinical early-stage oral cancer. For the clinicians, knowledge of potential locoregional prediction to stratify the patients into low-risk or high-risk groups using machine learning applications can help to guide clinical practice. Patients can be counseled accordingly with realistic expectations and clinicians can be guided in making informed decisions. Furthermore, this contributes to the individual data regarding patient and tumor-related factors and thereby helps the clinician in planning the optimal patient-specific treatment and follow-up (post-operative adjuvant treatment). For instance, high-risk patients might benefit from adjuvant oncological therapy after surgery. Future research should consider including other prognostic parameters as inputs for the selected algorithms. In terms of the limitation of this study, we are limited by the number of available cases as this was a retrospective study of five teaching hospitals in Finland and one in Brazil. Also, the external data used to test the performance of the algorithms were relatively limited. Therefore, with larger external data, the performance of the algorithms could be improved.

Authors contribution

Institutional Coordinators: Salo T, Coletta RD, Kowalski LP, Leivo I, Mäkitie AA, Haglund C. **Study concepts and study design:** Alabi RO, Elmusrati M, Almangush A, Coletta RD, Salo T, Leivo I. **Data acquisition and quality control of data:** Sawazaki-Calone I, Kowalski LP, Leivo I. **Data analysis and interpretation:** Alabi RO, Elmusrati M, Almangush A, Sawazaki-Calone I, Mäkitie AA, Salo T, Leivo I. **Manuscript preparation:** Alabi RO, Elmusrati M, Almangush A, Mäkitie AA, Coletta RD. **Manuscript review:** Mäkitie AA, Leivo I, Salo T, Kowalski LP, Sawazaki-Calone I. **Manuscript editing:** Salo T, Leivo I, Mäkitie AA, Haglund C. All authors approved the final manuscript for submission.

Summary Points

What was already known on the topic

- There are few published studies on the comparison of machine learning techniques to predict locoregional recurrence of oral tongue squamous cell carcinoma (OTSCC).
- Accuracy value is the most considered performance metrics to choosing the machine learning technique for prediction.

What knowledge this study adds

- To the best of our knowledge, this is the first study that analyzed more than three machine learning techniques to predict risk of locoregional recurrence in oral tongue squamous cell carcinoma (OTSCC) as low-risk or high-risk.
- It is important to consider other performance metrics such as specificity and F1 score (weighted average of precision and recall) in medical applications.
- The permutation importance feature (PFI) algorithm to extract important features does not correspond to better overall prediction and does not necessarily perform better than the ensemble algorithms.
- The application of these supervised learning techniques to stratify the patients as having low-risk or high-risk for the recurrence of OTSCC may be useful for effective cancer management.

;1;

Declaration of Competing Interest

The authors declare no conflicts of interest.

Acknowledgments

We would like to include the funding as follow: The School of Technology and Innovations, University of Vaasa Scholarship Fund. Turku University Hospital Research Fund, Helsinki University Hospital Research Fund, and the Finnish Cancer Society.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2019.104068.

References

[1] J.H. Ng, N.G. Iyer, M.-H. Tan, G. Edgren, Changing epidemiology of oral squamous cell carcinoma of the tongue: a global study: changing epidemiology of tongue cancer, Head Neck 39 (2017) 297–304, https://doi.org/10.1002/hed.24589.

[2] J.E. Tota, W.F. Anderson, C. Coffey, J. Califano, W. Cozen, R.L. Ferris, M.St. John, E.E.W. Cohen, A.K. Chaturvedi, Rising incidence of oral tongue cancer among white men and women in the United States, 1973–2012, Oral Oncol. 67 (2017) 146–152, https://doi.org/10.1016/j.oraloncology.2017.02.019.

[3] K. Rusthoven, A. Ballonoff, D. Raben, C. Chen, Poor prognosis in patients with stage I and II oral tongue squamous cell carcinoma, Cancer 112 (2008) 345–351, https://doi.org/10.1002/cncr.23183.

[4] I.O. Bello, Y. Soini, T. Salo, Prognostic evaluation of oral tongue cancer: means, markers and perspectives (I), Oral Oncol. 46 (2010) 630–635, https://doi.org/10.1016/j.oraloncology.2010.06.006.

[5] R. Mroueh, A. Haapaniemi, R. Grénman, J. Laranne, M. Pukkila, A. Almangush, T. Salo, A. Mäkitie, Improved outcomes with oral tongue squamous cell carcinoma in Finland: oral tongue carcinoma in Finland, Head Neck 39 (2017) 1306–1312, https://doi.org/10.1002/hed.24744.

[6] A. Almangush, I.O. Bello, R.D. Coletta, A.A. Mäkitie, L.K. Mäkinen, J.H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, Y. Soini, V.-M. Kosma, P. Koivunen, N. Kelner, L.P. Kowalski, R. Grénman, I. Leivo, E. Läärä, T. Salo, For early-stage oral tongue cancer, depth of invasion and worst pattern of invasion are the strongest pathological predictors for locoregional recurrence and mortality, Virchows Arch. 467 (2015) 39–46, https://doi.org/10.1007/s00428-015-1758-z.

[7] S. Anand, K. Rajesh, Analysis of SEER dataset for breast Cancer diagnosis using C4.5 classification algorithm, Int. J. Adv. Res. Comput. Commun. Eng. 1 (2012) 72–77.

[8] B. Zheng, S.W. Yoon, S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, Expert Syst. Appl. 41 (2014) 1476–1482, https://doi.org/10.1016/j.eswa.2013.08.044.

[9] L. Zhu, W. Luo, M. Su, H. Wei, J. Wei, X. Zhang, C. Zou, Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients, Biomed. Rep. 1 (2013) 757–760, https://doi.org/10.3892/br.2013.140.

[10] D. Delen, N. Patil, Knowledge extraction from prostate cancer data, Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) (2006), https://doi.org/10.1109/HICSS.2006.240 92b-92b.

[11] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2005) 113–127, https://doi.org/10.1016/j.artmed.2004.07.002.

[12] D. Delen, Analysis of Cancer Data: A Data Mining Approach, Expert Systems, (2009), https://doi.org/10.1111/j.1468-0394.2008.00480.x (Accessed 5 September 2019).

[13] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R.D. Coletta, A.A. Mäkitie, T. Salo, I. Leivo, A. Almangush, Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool, Virchows Arch. (2019) 489–497, https://doi.org/10.

1007/s00428-019-02642-5.
[14] C.-M. Chao, Y.-W. Yu, B.-W. Cheng, Y.-L. Kuo, Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree, J. Med. Syst. 38 (2014) 106, https://doi.org/10.1007/s10916-014-0106-1.
[15] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, J. Poorolajal, Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, Clin. Epidemiol. Glob. Health (2018) 293–299, https://doi.org/10.1016/j.cegh.2018.10.003.
[16] M. Montazeri, M. Montazeri, M. Montazeri, A. Beigzadeh, Machine learning models in breast cancer survival prediction, THC 24 (2016) 31–42, https://doi.org/10.3233/THC-151071.
[17] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R.N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inform. 108 (2017) 1–8, https://doi.org/10.1016/j.ijmedinf.2017.09.013.
[18] A.M. Bur, A. Holcomb, S. Goodwin, J. Woodroof, O. Karadaghy, Y. Shnayder, K. Kakarala, J. Brant, M. Shew, Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma, Oral Oncol. 92 (2019) 20–25, https://doi.org/10.1016/j.oraloncology.2019.03.011.
[19] W.M. Lydiatt, S.G. Patel, B. O'Sullivan, M.S. Brandwein, J.A. Ridge, J.C. Migliacci, A.M. Loomis, J.P. Shah, Head and neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual: Head and Neck Cancers-Major 8th Edition Changes, CA Cancer J. Clin. 67 (2017) 122–137, https://doi.org/10.3322/caac.21389.
[20] M.J. Lin, A. Guiney, C.E. Iseli, M. Buchanan, T.A. Iseli, Prophylactic neck dissection in early oral tongue squamous cell carcinoma 2.1 to 4.0 mm depth, Otolaryngol. Head Neck Surg. 144 (2011) 542–548, https://doi.org/10.1177/0194599810394988.
[21] P. O-charoenrat, G. Pillai, S. Patel, C. Fisher, D. Archer, S. Eccles, P. Rhys-Evans, Tumour thickness predicts cervical nodal metastases and survival in early oral tongue cancer, Oral Oncol. 39 (2003) 386–390.
[22] A. Almangush, R.D. Coletta, I.O. Bello, C. Bitu, H. Keski-Säntti, L.K. Mäkinen, J.H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, S. Tommola, Y. Soini, V.-M. Kosma, P. Koivunen, L.P. Kowalski, P. Nieminen, R. Grénman, I. Leivo, T. Salo, A simple novel prognostic model for early stage oral tongue cancer, Int. J. Oral Maxillofac. Surg. 44 (2015) 143–150, https://doi.org/10.1016/j.ijom.2014.10.004.
[23] N. Yamakawa, T. Kirita, M. Umeda, S. Yanamoto, Y. Ota, M. Otsuru, M. Okura, H. Kurita, S. Yamada, T. Hasegawa, T. Aikawa, T. Komori, M. Ueda, Japan Oral Oncology Group (JOOG), Tumor budding and adjacent tissue at the invasive front correlate with delayed neck metastasis in clinical early-stage tongue squamous cell carcinoma, J. Surg. Oncol. (2018) jso.25334, https://doi.org/10.1002/jso.25334.
[24] X. Yang, X. Tian, K. Wu, W. Liu, S. Li, Z. Zhang, C. Zhang, Prognostic impact of perineural invasion in early stage oral tongue squamous cell carcinoma: results from a prospective randomized trial, Surg. Oncol. 27 (2018) 123–128, https://doi.org/10.1016/j.suronc.2018.02.005.
[25] A. Arora, N. Husain, A. Bansal, A. Neyaz, R. Jaiswal, K. Jain, A. Chaturvedi, N. Anand, K. Malhotra, S. Shukla, Development of a new outcome prediction model in early-stage squamous cell carcinoma of the oral cavity based on histopathologic parameters with multivariate analysis: the Aditi-Nuzhat Lymph-node Prediction Score (ANLPS) system, Am. J. Surg. Pathol. 41 (2017) 950–960, https://doi.org/10.1097/PAS.0000000000000843.
[26] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2011, https://doi.org/10.1016/C2009-0-19715-5.
[27] Microsoft Azure Machine Learning Studio, Azure Machine Learning Studio: In Documentation, (2018).
[28] R. Barga, V. Fontama, W.-H. Tok, Predictive Analytics with Microsoft Azure Machine Learning, second edition, Apress, Berkeley, CA, 2015.
[29] P.A. Flach, The geometry of ROC space: understanding machine learning metrics through ROC isometrics, Icml, (2003).
[30] J. Fürnkranz, P.A. Flach, ROC 'n' rule learning—towards a better understanding of covering algorithms, Mach. Learn. 58 (2005) 39–77, https://doi.org/10.1007/s10994-005-5011-x.
[31] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.
[32] A.K. Dwivedi, Analysis of computational intelligence techniques for diabetes mellitus prediction, Neural Comput. Appl. 30 (2018) 3837–3845.
[33] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinf. 14 (2013) 106, https://doi.org/10.1186/1471-2105-14-106.
[34] J. Berdugo, L.D.R. Thompson, B. Purgina, C.D. Sturgis, M. Tuluc, R. Seethala, S.I. Chiosea, Measuring depth of invasion in early squamous cell carcinoma of the oral tongue: positive deep margin, extratumoral perineural invasion, and other challenges, Head Neck Pathol. 13 (2019) 154–161, https://doi.org/10.1007/s12105-018-0925-3.
[35] A.-F. Safi, M. Kauke, A. Grandoch, H.-J. Nickenig, J.E. Zöller, M. Kreppel, Analysis of clinicopathological risk factors for locoregional recurrence of oral squamous cell carcinoma – retrospective analysis of 517 patients, J. Cranio-Maxillofacial Surg. 45 (2017) 1749–1753, https://doi.org/10.1016/j.jcms.2017.07.012.
[36] I. Vázquez-Mahía, J. Seoane, P. Varela-Centelles, I. Tomás, A.Á. García, J.L. López Cedrún, Predictors for tumor recurrence after primary definitive surgery for oral cancer, J. Oral Maxillofac. Surg. 70 (2012) 1724–1732, https://doi.org/10.1016/j.joms.2011.06.228.
[37] M.A. Ermer, K. Kirsch, G. Bittermann, T. Fretwurst, K. Vach, M.C. Metzger, Recurrence rate and shift in histopathological differentiation of oral squamous cell carcinoma – a long-term retrospective study over a period of 13.5 years, J. Cranio-Maxillofacial Surg. 43 (2015) 1309–1313, https://doi.org/10.1016/j.jcms.2015.05.011.
[38] D.R. Camisasca, M.A.N.C. Silami, J. Honorato, F.L. Dias, P.A.S. de Faria, S. de Q.C. Lourenço, Oral squamous cell carcinoma: clinicopathological features in patients with and without recurrence, ORL. 73 (2011) 170–176, https://doi.org/10.1159/000328340.
[39] G. Levitin, Computational Intelligence in Reliability Engineering, Springer, Berlin, 2007 (Accessed 6 September 2019), https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=186979.
[40] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, C.-N. Lin, The application of data mining techniques to oral Cancer prognosis, J. Med. Syst. 39 (2015) 59, https://doi.org/10.1007/s10916-015-0241-3.
[41] N.B. de Melo, Í. de M. Bernardino, D.P. de Melo, D.Q.C. Gomes, P.M. Bento, Head and neck cancer, quality of life, and determinant factors: a novel approach using decision tree analysis, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 126 (2018) 486–493, https://doi.org/10.1016/j.oooo.2018.07.055.
[42] R. Sumbaly, N. Vishnusri, S. Jeyalatha, Diagnosis of breast Cancer using decision tree data mining technique, Int. J. Comput. Appl. 98 (2014) 16–24.
[43] B. Zhang, X. He, F. Ouyang, D. Gu, Y. Dong, L. Zhang, X. Mo, W. Huang, J. Tian, S. Zhang, Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma, Cancer Lett. 403 (2017) 21–27, https://doi.org/10.1016/j.canlet.2017.06.004.

International Journal of Medical Informatics 145 (2021) 104313

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer

Rasheed Omobolaji Alabi [a,*], Antti A. Mäkitie [b,c,d], Matti Pirinen [e,f,g], Mohammed Elmusrati [a], Ilmo Leivo [h], Alhadi Almangush [b,h,i,j]

[a] *Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland*
[b] *Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland*
[c] *Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland*
[d] *Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden*
[e] *Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland*
[f] *Department of Public Health, University of Helsinki, Helsinki, Finland*
[g] *Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*
[h] *University of Turku, Institute of Biomedicine, Pathology, Turku, Finland*
[i] *Department of Pathology, University of Helsinki, Helsinki, Finland*
[j] *Faculty of Dentistry, University of Misurata, Misurata, Libya*

ARTICLE INFO

Keywords:
Machine learning
Nomogram
tongue cancer
Predict
overall survival

ABSTRACT

Background: The prediction of overall survival in tongue cancer is important for planning of personalized care and patient counselling.
Objectives: This study compares the performance of a nomogram with a machine learning model to predict overall survival in tongue cancer. The nomogram and machine learning model were built using a large data set from the Surveillance, Epidemiology, and End Results (SEER) program database. The comparison is necessary to provide the clinicians with a comprehensive, practical, and most accurate assistive system to predict overall survival of this patient population.
Methods: The data set used included the records of 7596 tongue cancer patients. The considered machine learning algorithms were logistic regression, support vector machine, Bayes point machine, boosted decision tree, decision forest, and decision jungle. These algorithms were mainly evaluated in terms of the areas under the receiver-operating characteristic (ROC) curve (AUC) and accuracy values. The performance of the algorithm that produced the best result was compared with a nomogram to predict overall survival in tongue cancer patients.
Results: The boosted decision-tree algorithm outperformed other algorithms. When compared with a nomogram using external validation data, the boosted decision tree produced an accuracy of 88.7% while the nomogram showed an accuracy of 60.4%. In addition, it was found that age of patient, T stage, radiotherapy, and the surgical resection were the most prominent features with significant influence on the machine learning model's performance to predict overall survival.
Conclusion: The machine learning model provides more personalized and reliable prognostic information of tongue cancer than the nomogram. However, the level of transparency offered by the nomogram in estimating patients' outcomes seems more confident and strengthened the principle of shared decision making between the patient and clinician. Therefore, a combination of a nomogram – machine learning (NomoML) predictive model may help to improve care, provides information to patients, and facilitates the clinicians in making tongue cancer management-related decisions.

* Corresponding author.
E-mail address: rasheed.alabi@student.uwasa.fi (R.O. Alabi).

https://doi.org/10.1016/j.ijmedinf.2020.104313
Received 30 June 2020; Received in revised form 4 October 2020; Accepted 20 October 2020
Available online 24 October 2020

1. Introduction

The most common epithelial neoplasm affecting the oral cavity is the oral cavity squamous cell carcinoma (OCSCC) [1]. Tongue squamous cell carcinoma (TSCC) accounts for the most common cases in oral cavity cancers [2–6]. It is characterized by an aggressive clinical behavior [7] such as rapid local invasion and early lymph node metastasis [8]. This aggressive behavior leads to a high rate of recurrence and mortality [9]. Despite the advancement in cancer diagnostic and management approaches in recent years, the 5-year relative overall survival (OS) for patients treated with curative intent was 61% in a recent study [10] and 63% in another report [11].

The prediction of tongue cancer survival outcomes is of utmost interest to both clinicians and patients. This is because determining cancer outcomes may crucially contribute to personalized treatment planning and even to avoiding unnecessary therapies [12]. Also, it provides a useful insight into effective management decision-making and may guide the selection of a protocol treatment approach. Of note, predicting TSCC survival is challenging due to different patient-related factors, tumor characteristics, and available treatment modalities. The American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) staging system has shown to be an objective and accurate tool for predicting the prognosis for an entire population of cancer patients. Thus, it was widely used for planning of treatment strategies for TSCC patients [8,13]. However, for an individual patient, it is ineffective for predicting outcome due to its inability to consider other tumor- and patient-related risk factors [14,15]. To this end, a tool that considers these factors together to accurately predict patients' outcomes would be pertinent [8].

Nomogram is defined as a pictorial representation of a complex mathematical formula that uses certain variables such as demographics, clinical, or treatment variables to graphically depict a statistical prognostic model [16,17]. This graphical representation of the prognostic model can be used for the prognostication of clinical events such as recurrence, disease-specific survival or overall survival for a given patient [17]. Nomograms have been used in predicting survival in breast cancer [18], gastric cancer [19], and head and neck cancer [8,20,21]. Similarly, machine learning techniques have been touted for effective prediction of outcomes. These include for instance, predicting locoregional recurrence [22,23], occult node metastasis [24,25], and survival rates [26–29].

In this study, we aim to compare the performance of a nomogram with machine learning techniques in predicting the overall survival of tongue cancer patients. The survival time in months of tongue cancer patients was considered as the time from the beginning of treatment until the last follow-up or death [30]. The examined machine learning algorithms were logistic regression, support vector machine (SVM), naive Bayes (NB), neural network, boosted decision tree, decision forest, and decision jungle algorithms. This comparison is pertinent as it is aimed at providing the clinicians with a comprehensive, practical, and most accurate assistive system to predict overall survival for patients. Additionally, this system will assist the clinicians to provide a more personalized and precise therapeutic decision. This study is based on multi-population data obtained from the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH).

2. Material and Methods

2.1. National Cancer Institute Database

The study data were obtained from the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH). It is one of the largest cancer database that is available publicly [31]. It gives non-identifiable information on cancer statistics of the United States population. These important characteristics of this database make it a database of choice, thus facilitating large-scale outcome analysis research. The ethical permission to use the SEER database was approved with the user identification numbers of 10455-Nov 2018 and 11522-Nov2019, respectively.

2.2. Selection of patient attributes

The SEER database was chosen as it was considered as a high-quality database of different cancer patients [8]. The SEER program of the National Cancer Institute was searched for Nov 2015 submission [1973 – 2013] (Fig. 1). These years (1973 – 2013) under consideration were also selected because the nomogram to be used for comparison was built using the same date range.

The inclusion criteria included that the patients were diagnosed with histologically confirmed (positive) tongue cancer. Additionally, the patient must have known basic data such as gender and age at diagnosis. Therefore, the included known clinical and pathologic characteristics were age at diagnosis, race, marital status, grade, TNM status according to AJCC 7th edition, treatment (surgery, and radiotherapy) [8]. The survival period (in months) and overall survival status of the patients were also extracted. All patients whose diagnostic information were unknown were excluded. A total of 7649 cases were found eligible to be included in this study (Table 1). The data extraction process is shown in Fig. 1. The explanation of the included variables and categorization is shown in Table 2.

2.3. Separate external validation cases

Out of 7649 cases, we reserved the last 53. These cases were not used in the machine learning training and testing phase. It was reserved to externally validate the model that showed the best performance metrics in terms of the accuracy. The external validation cases, who had been labelled as dead, had died within 5 years from the first treatment. Similarly, the individuals who were labelled as alive were alive at least 5 years from the first treatment. It is important to externally validate the model to address the possible concern about the generalizability of the model.

2.4. Nomogram

We used the nomogram constructed in another previously published study for evaluating the 5- and 8-year overall survival in tongue squamous cell carcinoma (Figs. 2,3) [8]. It was chosen because it considered overall survival as a distinct event in its construction. Additionally, it was well-validated (internal and external validation) and calibrated [8].

2.5. Machine learning training process

Microsoft Azure Machine Learning Studio (Azure ML 2019) was used in this study [32]. The input parameters were age at diagnosis, race, sex, marital status at diagnosis, tumor grade, AJCC TNM staging system, survival time, treatment (radiotherapy, surgical resection). The output variable was overall survival (alive or dead). The survival months ranged from 1 to 47 months. Firstly, the extracted data were checked to ensure that these were properly preprocessed. In addition, all the variables were converted to numeric to reduce possible spelling errors and omissions in each variable (Table 2). Also, potential class imbalance in the target variable was handled by up-sampling using synthetic minority oversampling technique (SMOTE) [33]. This approach offers a reasonable approach to handling potential imbalance than simply duplicating existing cases.

The data set was divided into two sets of training and validation using a 5-fold cross-validation method in the ratio 80% training and 20% validation {80:20} percentage splitting sets [34] (Fig. 4). Using cross-validation, hyperparameters were fine-tuned to maximize the area

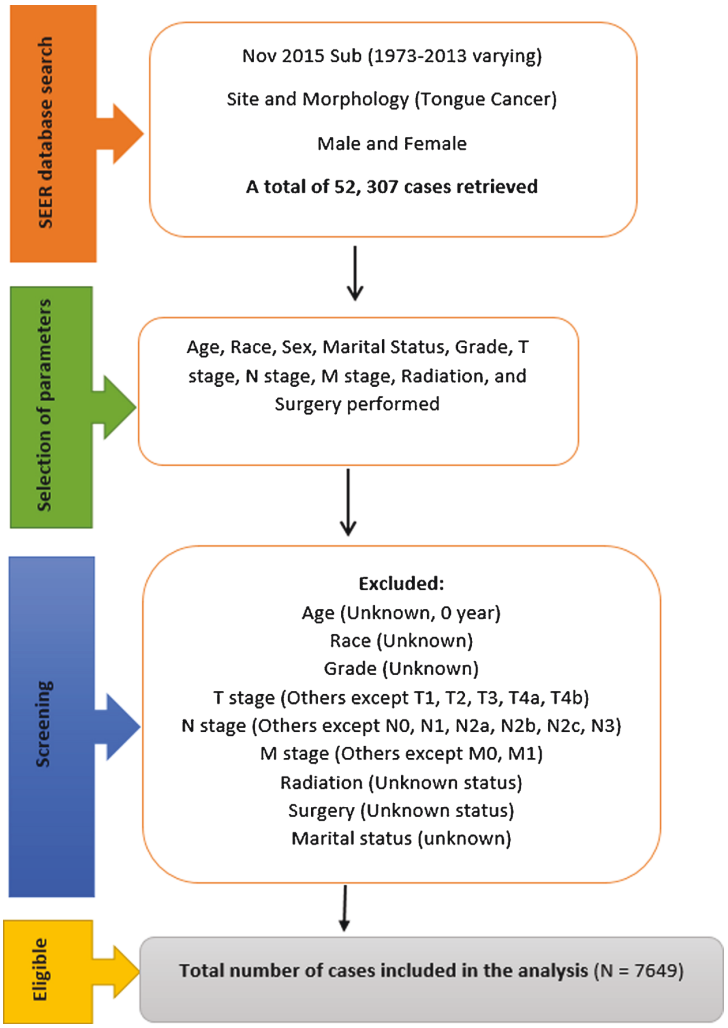

Fig. 1. Flowchart for data extraction for the Surveillance, Epidemiology, and End Results data selection.

under receiving operating characteristics curve (AUC) or concordant partial AUC especially when imbalanced dataset was used in the training [35] for each of the examined algorithms. Each of the algorithms of interest was then configured and used for the whole training data set [36,37]. After training, the algorithms were evaluated for the performance metrics of interest (Fig. 4, Table 3). The performance of classification algorithms was compared mainly in terms of accuracy, AUC (internal validation), F1 score, and likelihood ratios as represented in Table 3. The algorithm that showed the best AUC values was used for external validation and comparison with the nomogram. The data used for external validation were not used during the training phase. The result obtained for validating this model externally was considered as the true performance of the algorithm (Table 4). Also, it addressed possible concerns relating to the generalization of the algorithm.

2.6. Comparison of the performance of the machine learning with a nomogram

The nomogram and the machine learning algorithm that showed the best accuracy were compared using the external validation data (Section 2.3). The machine learning algorithm was compared with a nomogram built with surgical treatment (Fig. 2). The result of this comparison is presented in Table 4. Likewise, the algorithm was compared with a nomogram built with radiotherapy (Fig. 3). The result of this comparison is given in Table 4. The overall performance of these two predictive tools in terms of accuracy, sensitivity, specificity, and F1 score is shown in Table 4. The comparison was necessary to ensure that the predictive tool used in medicine is convenient, accurate, and explainable (enables clinicians to understand why the algorithm produced certain result). This was corroborated by the study of Holzinger et al., where human and machine explanations were compared using system causability scale (SCS) to allow for explainable AI [38].

3. Results

3.1. Data Description

The study cohort included 7596 patients with tongue squamous cell carcinoma; 5322 male and 2274 female in a male-to-female ratio of 2.3:1. The mean age at diagnosis was 62.3 (SD ± 12.7: range 12-102) and the median age was 62.0 years. In terms of the ethnicity, 6597 (86.8%) were from the white origin, 516 (6.8%) were black, and 483 (6.4%) were from other origins including American Indian/AK Native, Asian/Pacific Islander. Considering marital status, 4430 (58.3%) were

Table 1
Baseline demographic and tumor characteristics of patients in SEER database.

Variables	Overall survival, N = 7596 Training and testing cohort	Overall survival, N = 53 External validation cohort
Age at diagnosis (years)		
1 – 18	5 (0.1%)	0 (0.0%)
19 – 44	515 (6.8%)	2 (3.8%)
45 – 54	1412 (18.6%)	14 (26.4%)
55 – 64	2497 (32.9%)	18 (34.0%)
65 – 74	1877 (24.7%)	13 (24.5%)
75+	1290 (16.9%)	6 (11.3%)
Ethnic origin		
White	6597 (86.8%)	44 (83.0%)
Black	516 (6.8%)	9 (17.0%)
Other*	483 (6.4%)	
Sex		
Male	5322 (70.0%)	38 (71.7%)
Female	2274 (30.0%)	15 (28.3%)
Marital status		
Married	4430 (58.0%)	29 (54.7%)
Unmarried	3166 (42.0%)	24 (15.3%)
Grade		
Grade I	1215 (16.0%)	8 (15.1%)
Grade II	3768 (49.6%)	29 (54.7%)
Grade III	2543 (33.5%)	15 (28.3%)
Grade IV	70 (0.9%)	1 (1.9%)
T stage (2010+)		
T1	2942 (38.7%)	19 (35.8%)
T2	2492 (32.8%)	16 (30.2%)
T3	1159 (15.3%)	6 (11.3%)
T4a	920 (12.1%)	11 (20.8%)
T4b	83 (1.1%)	1 (1.9%)
N Stage (2010+)		
N0	3485 (45.9%)	14 (26.4%)
N1	1220 (16.1%)	10 (18.9%)
N2a	327 (4.3%)	5 (9.4%)
N2b	1498 (19.7%)	14 (26.4%)
N2c	880 (11.6%)	9 (17.0%)
N3	186 (2.4%)	1 (1.9%)
M stage (2010+)		
M0	7425 (97.7%)	50 (94.3%)
M1	171 (2.3%)	3 (5.7%)
Surgery performed		
Yes	4654 (61.3%)	22 (41.5%)
None	2942 (38.7%)	31 (58.5%)
Radiotherapy		
Yes	4489 (59.1%)	37 (69.8%)
None	3107 (40.9%)	16 (30.2%)
Overall survival status		
Alive	5743 (75.6%)	47 (88.7%)
Dead	1853 (24.4%)	6 (11.3%)

* Other including American Indian (native), Asian/Pacific Islander.

married while 3166 (41.7%) were considered unmarried (single, divorced, widowed, and separated) at the time of diagnosis. For 1215 (16.0%) out of the 7596 patients the grade was well-differentiated, for 3768 (49.6%) moderately differentiated, 2543 (33.5%) poorly differentiated, and 70 (0.9%) undifferentiated.

Similarly, the distribution according to the AJCC TNM staging scheme, the tumor diameter showed that 2942 (38.7%) patients had stage T1, 2492 (32.8%) stage T2, 1159 (15.3%) stage T3, 920 (12.1%) stage T4a, and 83 (1.1%) stage T4b. Also, 3485 (45.9%) had N0, 1220 (16.1%) had N1, 327 (4.3%) N2a, 1498 (19.7%) N2b, 880 (11.6%) N2c, 186 (2.4%) N3; 7425 M0, and 171 M1. The histopathologic characteristics are briefly summarized in Table 2. In terms of the treatment, 4654 (61.3%) had surgery while 2942 (38.7) had no surgery. Adjuvant radiotherapy was administered to 4489 (59.1%) patients. The follow-up time ranged from 0 to 47 months (Mean 18.3; SD ± 13.3). The number of patients who were alive at last follow-up was 5743 (75.6%).

For the testing series (n = 53) that was not included in the construction of the machine learning model, the mean age at diagnosis was 61.1 years (SD ± 11.1; range from 31 to 85 years with 44 (83.0%) male and 9 (17.0%) female. Additionally, 38 (71.7%) were married and 15 (28.3%) unmarried. For histological grade, 29 (54.7%) were well-differentiated, 10 (18.9%) moderately-differentiated, 9 (17.0%) poorly-differentiated, and 5 (9.4%) undifferentiated grade. In terms of treatment, 37 (69.8%) patients had radiotherapy while 22 (41.5%) had surgery performed. The mean follow-up time was 4.2 months (SD ± 5.2; range 0-23 months) and 47 (88.6%) patients were alive at the end of follow-up. The detailed characteristics of the external validation data are given in Table 1.

Table 2
Selected SEER attributes, description, and categorization used in machine learning training.

Attribute	Description	Categorization for machine learning training	Type
Age	Age at time of diagnosis	No categorization	Discrete
Race/ Ethnicity	This describes the ethnicity of the patient	0 = White; 1 = Black; 2= Others (American Indian /AK Native, Asian pacific	Numeric
Sex	Biological sex	0 = Male; 1 = Female	Numeric
Marital Status	The marital status of the patient at diagnosis of TSCC.	0 = Married; 1 = Single (never married, Unmarried or domestic partner); 2 = Divorced (separated); 3 = Widowed; 4 = Separated.	Numeric
Grade	The differentiation of the cancer cell.	1 = Grade 1 (Well differentiated), 2 = Grade 2 (Moderately differentiated), 3 = Grade 3 (poorly differentiated), 4 =(Undifferentiated)	Numeric
Derived AJCC T, 7th edition (2010+) stage	T1: The tumor is ≤ 2 cm or less in greatest dimension. T2: The tumor is > 2 cm & ≤ 4 cm. T3: Tumor is > 4 cm. T4a: Moderately advanced local disease. T4b: Significantly advanced local disease.	T1 = 1; T2 = 2; T3 = 3; T4a = 4; T4b = 5	Numeric
Derived AJCC N, 7th edition (2010+) stage	N0; No regional lymph node metastasis N1: Regional lymph node metastasis (single node). N2a: Cancer has spread to a single lymph node. N2b: The present of multiple lymph nodes. N2c: There are lymph nodes in the neck either on the opposite side as the main cancer or on both sides. N3: There is spread to one or more neck lymph nodes	N0 = 0; N1 = 1; N2a = 2; N2b = 3; N2c = 4; N3 = 5;	Numeric
Derived AJCC M, 7th edition (2010+) stage	M0; No distant metastasis M1; Distant metastasis	M0 = 0; M1 = 1	Numeric
Radiation	Indication of whether patient has received radiation	0 = None, 1 = exposed to radiation	Numeric
Surgical resection	This describes if surgery was performed	0 = No surgery performed; 1 = Surgery not performed	Numeric
Overall survival	The time from the beginning of treatment until the last follow-up time or death	0 = Alive; 1 = Dead	Discrete

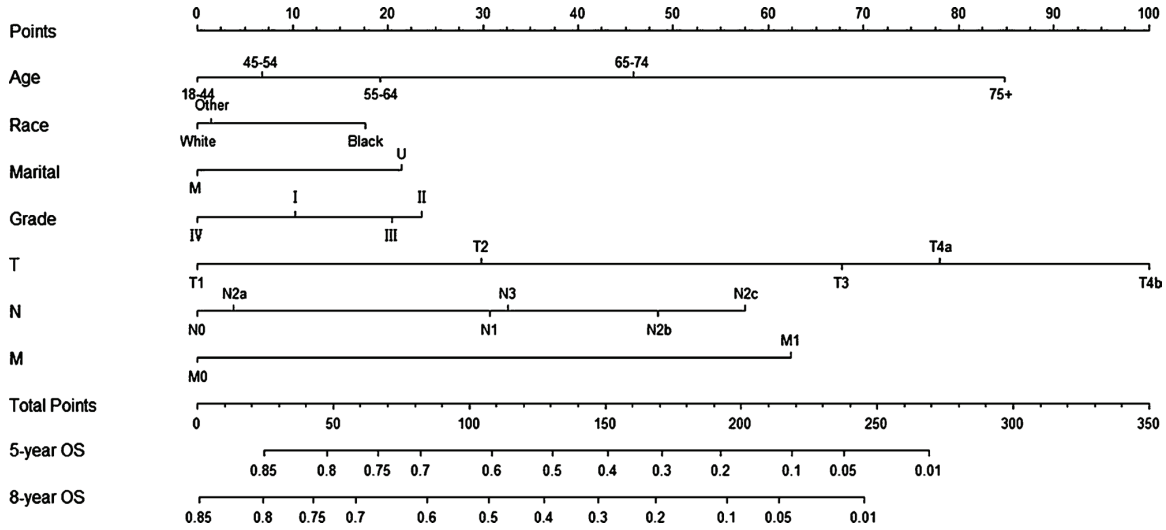

Fig. 2. Nomogram for predicting 5- and 8-year overall survival with surgical treatment.

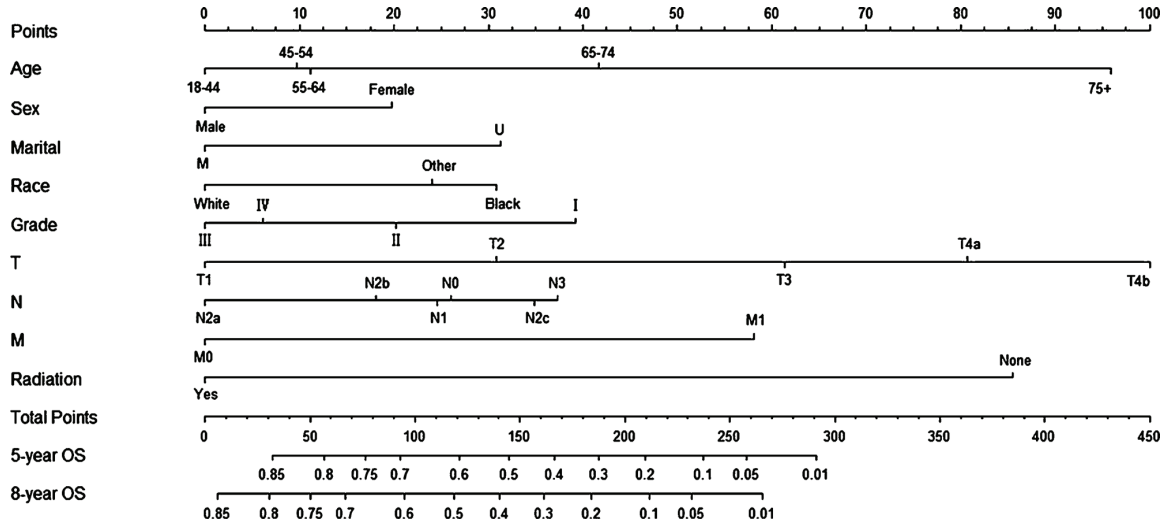

Fig. 3. Nomogram for predicting 5- and 8-year overall survival with radiation treatment.

3.2. Performance metrics for the algorithms

The performance metrics of the algorithms are presented in Table 3. The average specificity of the examined algorithms was 0.89. Similarly, the average sensitivity was 0.76. In terms of the accuracy and the area under receiving operating characteristics curve (AUC), the boosted decision tree outperformed all other algorithms.

3.3. Evaluating the input variables for importance

The permutation feature importance of the input variables showed that the AJCC T stage, radiotherapy, age and surgical status are the most prominent features that had significant influence on the model's performance to predict the overall survival in tongue cancer patients.

3.4. Comparison of the performance of the nomogram with machine learning algorithm

The nomogram (with surgical treatment) showed 66.0% accuracy, likewise, the nomogram (with radiotherapy) produced an accuracy of 60.4% when tested with the external validation data. The machine learning algorithm (boosted decision tree) showed an accuracy of 88.7% when tested with external validation data. All the examined methods showed 100% sensitivity (Table 4). Considering the specificity and F1 score, the nomogram (with surgical treatment) showed 0.62 and 0.40, machine learning model gave 0.87 and 0.66, and nomogram (with radiation) produced 0.55 and 0.36 (Table 4).

4. Discussion

In this study, a nomogram and several machine learning algorithms were utilized and compared in the prediction of overall survival in patients with tongue cancer. These machine learning algorithms used were

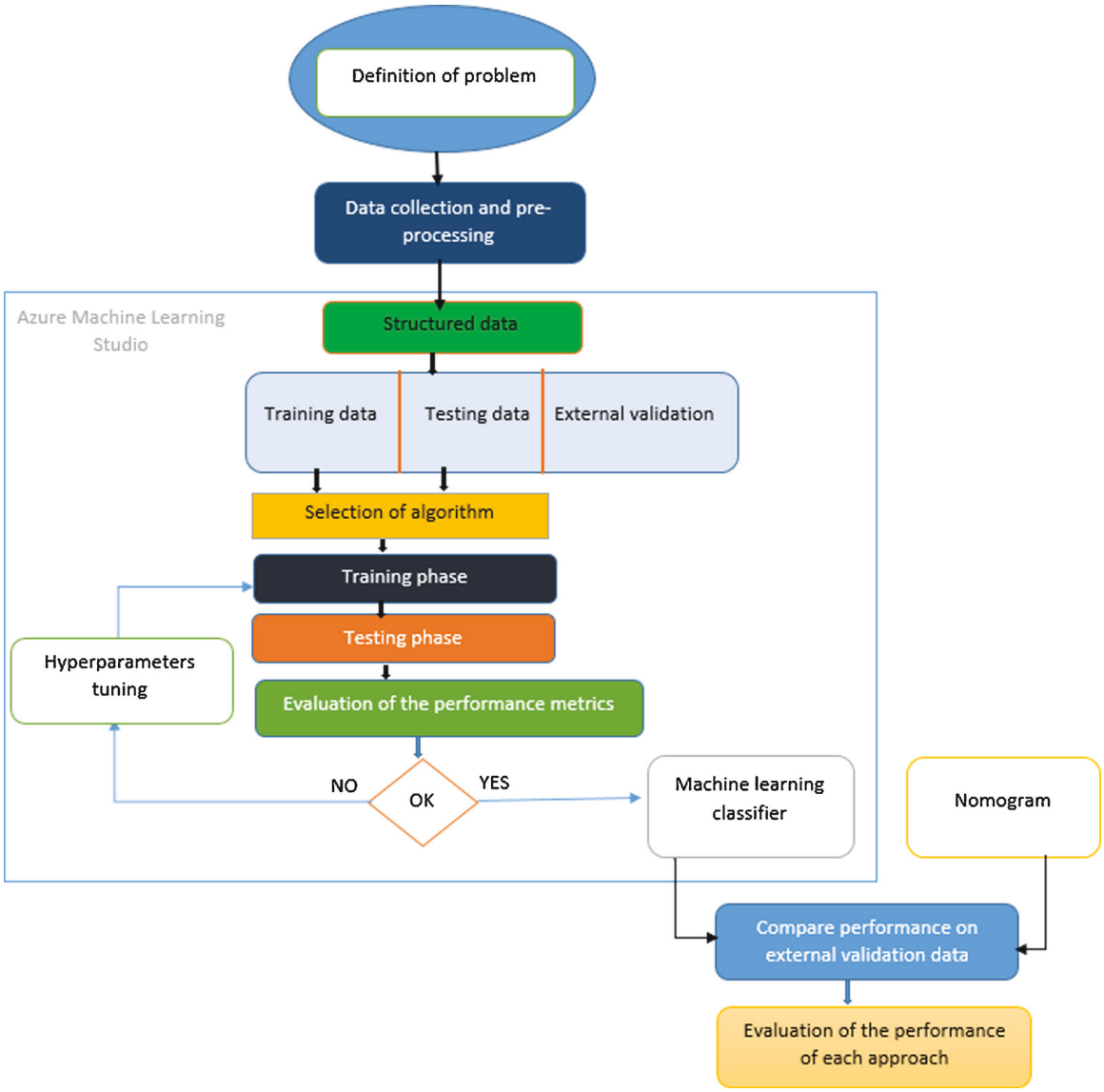

Fig. 4. The flowchart for the machine learning process and comparison with nomogram.

Table 3
The performance metrics of the cross-validated machine learning algorithms on the training data.

Algorithms	Accuracy (%)	AUC	Precision	F1 score	Recall
Logistic Regression	69.6	0.76	0.71	0.69	0.66
Naive Bayes	69.6	0.76	0.71	0.67	0.67
Support Vector Machine	69.5	0.76	0.71	0.68	0.66
Neural Network	73.1	0.83	0.78	0.70	0.63
Boosted Decision Tree	83.1	0.90	0.82	0.83	0.85
Decision Forest	81.5	0.89	0.81	0.82	0.83
Decision Jungle	79.6	0.88	0.80	0.79	0.79

Area Under Receiving Operating Characteristic (ROC) Curve (AUC); Recall = Sensitivity.

Table 4
The performance metrics of the comparison between the nomogram and machine learning model.

Parameters	Nomogram (with surgical treatment)	Machine learning model	Nomogram (with radiotherapy)
True positive	6	6	6
False positive	18	6	21
True negative	29	41	26
False negative	0	0	0
Sensitivity	1	1	1
Specificity	0.62	0.87	0.55
F1 score	0.40	0.66	0.36
Accuracy (%)	**66.0**	**88.7**	**60.4**

logistic regression, support vector machine, naive Bayes, neural network (NN), boosted decision tree, decision forest, and decision jungle. The algorithm that produced the best accuracy was compared with a nomogram. This comparison was based on a separate cohort that was not used in the training or testing phase. This comparison is necessary to ensure that the best model is selected for the specific management of patients with tongue cancer.

Several studies have examined the significance of machine learning (shallow and deep learning) techniques for oral cancer prognostication. For example, Tseng et al. developed a machine learning model for survival risk stratification of patients with advanced OSCC using clinicopathologic and genetic data [39]. This approach was corroborated by Karadaghy et al., where social, demographic, and clinicopathologic features were used to develop a machine learning-based model to predict 5-year overall survival of OSCC patients [29]. These studies concluded that a machine learning-based approach augments the clinicians' ability to properly estimate the survival risk of OSCC patients. Thus, effective and efficient treatment plan – intensifying or deintensifying the regimen, can be mapped out to improve the quality of care and survival of OSCC patients [39].

Besides survival risk estimate with machine learning techniques, Alabi et al., and Bur et al., have published promising results regarding predicting clinical outcome of a progressive disease such as locoregional tumor recurrence and/or distant metastases [22–24,40]. This technique has also been found to be better than such conventional methods as tumor depth of invasion (DOI), neutrophil-to-lymphocyte ratio (NLR), or tumor budding in predicting clinical outcomes [24,41]. Also, deep learning techniques have shown to be a promising noninvasive approach in early diagnosis [42], assessment of cervical lymph node metastasis [43], and discriminating between well-differentiated and poorly differentiated OSCC [44].

In this study, the boosted decision tree outperformed other algorithms and the nomogram. It uses the gradient boosting approach to create an ensemble of classification trees needed to stratify the patients in terms of their overall survival of tongue cancer. Each of the tree created is dependent on the prior trees. The algorithm learns by fitting the errors of the tree that proceeded it. Consequently, the second tree fits the errors of the first tree, the third tree fits the errors in the second trees, the sequence of error fitting continues in that order until the final tree. Predictions are therefore based on the entire ensemble of trees [23,36,37]. With a reasonable amount of data used in the study, the boosted decision tree algorithm was able to minimize errors due to the large coverage of the relationship between the data to improve the accuracy.

Tumor stage (T stage), radiotherapy, age of patient, and surgical resection were the input variables that had significant importance on the machine learning model's ability to predict overall survival in tongue cancer patients. For the stage of the disease at diagnosis, it has been reported that it is strongly correlated with prognosis [45]. The survival of patients with stage I (T1N0) of the disease exceeds 80% while stage III-IV (T3-T4) reduces below 40% [46,47]. Interestingly, most of the oral cancer patients are usually found to be at stage III or IV at the time of diagnosis [48,49]. This further corroborated the importance of T stage on the predictive model. Similarly, the age of the patient at the time of diagnosis was found to play an important role in the model's predictive ability. This result was emphasized in other studies that reported that the survival of oral cancer patients steadily decreases with age of the patient [30,50]. As the cohort contained largely early-stage tongue cancer, it is no surprise that the treatment options had a significant impact on the predictive performance of the model. This is because the treatment of choice for early-stage tongue cancer can either be surgery, radiotherapy or combination of both [45].

Traditionally, the clinicians' judgments have formed the foundation for estimating the risk of patients, counseling and decision making. Therefore, the experience of clinicians plays a significant role in accurate risk estimation and decision making. This approach poses a great risk of bias and the predicted outcomes of the patients may be highly subjective [12,51,52]. The nomogram has been used to predict survival in various head and neck cancers [20,21,53–55]. Its performance was reported to provide superior disease-related risk estimations for patients [56]. Likewise, machine learning models have shown encouraging risk estimation for patients [22–24,29]. Therefore, the introduction of nomogram and machine learning models have been touted to providing the clinicians with a decision-making assistive tool that gives more accurate predictions of patients' outcome. When these two approaches were compared as presented in this study, the machine learning model outperformed the nomogram in predicting the overall survival in tongue cancer patients. To the best of our knowledge, this is the first study that compares the performance of a nomogram with machine learning for tongue cancer.

The machine learning model showed that it was able to identify and understand the hard-to-discern relationships between the input variables. The predictive accuracy exhibited by this model is particularly well-suited to medical applications for personalized and predictive medicine [57]. The boosted decision-tree algorithm was able to build formidable overall survival classification trees in a step-wise manner, where the error in each step is measured and corrected in the next step to produce a model with improved predictive accuracy. Despite the better predictive performance of the machine learning model over the nomogram, the fact that the nomogram offers an appealing, transparent means of estimating the risk of patients without the use of the internet or computer is worthy of consideration for clinical decision-making.

Of note, the transparency offered by the nomogram addresses the concerns that the results from machine learning models are not easily interpretable. With this level of transparency in calculating the patients' outcomes, it is obvious that the patient will be more confident in the recommended treatment approach. More importantly, the principle of shared decision-making between the patient and clinician can be strengthened. Therefore, a combination of nomogram – machine learning (NomoML) approach may offer a more transparent approach for individualized assessment and add to the planning of the most appropriate adjuvant treatment for tongue cancer patients. The level of transparency in calculating the patients' outcomes in addition to the significant accuracy offered by the machine learning model is poised to give confidence to the patient for the recommended treatment approach.

In addition to the accuracy and transparency that the proposed NomoML seeks to offer, it is also poised to allow for explainable artificial intelligence (AI). However, as this proposed tool seeks to combine human and automatic (autonomous) machine learning approaches, the need to examine the causability (property of a person that measures the quality of explanations [58]) and explainability (property of a system that measures why an algorithm/system came up with certain result [38,58]) of this tool becomes imperative. An example of a viable tool to measure the quality of these explanations is the systematic causability scale (SCS) proposed by Holzinger et al. [38]. This tool (SCS) combines causability and explainability to reach the level of explainable medicine [59]. Therefore, for future study, it would be important to examine our proposed diagnostic tool for SCS evaluation. Undoubtedly, concerns about human-AI relationship and the extent to which AI-based model can or should support clinical decisions is growing. However, it is important to properly understand causability and explainability prior to addressing the former concerns [59].

In this study, there were certain limitations to be considered. Both the nomogram and machine learning were developed using retrospective cohorts; it remains important to validate these with a prospective cohort for the comparison to be a representation of the performance of these tools. Also, there may be a possibility of bias in the data set as a significant number of the patients were alive at the end of follow-up. In addition, information about some variables such as perineural invasion that have been reported to have significant influence on the overall survival are not available. Therefore, it would be worthwhile to further calibrate the nomogram to include these variables and to compare it with deep machine learning technologies.

In conclusion, concerted efforts should be made towards the construction of a more accurate nomogram and machine learning models. This includes use of larger data sets, inclusion of novel biomarkers, improved data collection, calibration, and validation methods. With an improved NomoML, accurate estimation of the likelihood of recurrences, tongue-specific and overall survival of tongue cancer can be greatly improved and management-related decisions for tongue cancer patients can be enhanced.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Authors Contribution

Study concepts and study design: Alabi RO, Makitie AA, Pirinen M, Almangush, A. Date extraction: Alabi RO, Makitie A, Almangush A. Data quality: Almangush A, Elmusrati M. Data analysis and interpretation: Alabi RO, Elmusrati M, Almangush A, Mäkitie AA, Pirinen M, Leivo I. Manuscript preparation: Alabi RO, Almangush A, Mäkitie AA, Pirinen M. Manuscript review: Mäkitie AA, Leivo I, Elmusrati M, Pirinen M. Manuscript editing: Almangush, A. Makitie AA, Pirinen M, Institution Head: Elmusrati M, Leivo I, All authors approved the final manuscript for submission.

Acknowledgment

The School of Technology and Innovations, University of Vaasa Scholarship Fund. The Helsinki University Hospital Research Fund. A special appreciation and citation to the National Cancer Institute for permission to acesss Nov 2019 sub (1975- 2017) SEER*Stat Database released in April 2020.

References

[1] M.L. Wallace, B.W. Neville, Squamous cell carcinoma of the gingiva with an atypical appearance, J Periodontol 67 (1996) 1245–1248, https://doi.org/10.1902/jop.1996.67.11.1245.
[2] V.C. Rodrigues, S.M. Moss, H. Tuomainen, Oral cancer in the UK: to screen or not to screen, Oral Oncol 34 (1998) 454–465, https://doi.org/10.1016/s1368-8375(98)00052-9.
[3] S.R. Moore, N.W. Johnson, A.M. Pierce, D.F. Wilson, The epidemiology of mouth cancer: a review of global incidence, Oral Dis 6 (2000) 65–74, https://doi.org/10.1111/j.1601-0825.2000.tb00104.x.
[4] S. Marur, A.A. Forastiere, Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment, Mayo Clin Proc 91 (2016) 386–396, https://doi.org/10.1016/j.mayocp.2015.12.017.
[5] F.R. Pires, A.B. Ramos, J.B.C. Oliveira de, A.S. Tavares, P.S.R. Luz da, T.C.R. B. Santos dos, Oral squamous cell carcinoma: clinicopathological features from 346 cases from a single oral pathology service during an 8-year period, J Appl Oral Sci 21 (2013) 460–467, https://doi.org/10.1590/1679-775720130317.
[6] R. Li, W.M. Koch, C. Fakhry, C.G. Gourin, Distinct epidemiologic characteristics of oral tongue cancer patients, Otolaryngol Head Neck Surg 148 (2013) 792–796, https://doi.org/10.1177/0194599813477992.
[7] I.O. Bello, Y. Soini, T. Salo, Prognostic evaluation of oral tongue cancer: means, markers and perspectives (I), Oral Oncol 46 (2010) 630–635, https://doi.org/10.1016/j.oraloncology.2010.06.006.
[8] Y. Li, Z. Zhao, X. Liu, J. Ju, J. Chai, Q. Ni, et al., Nomograms to estimate long-term overall survival and tongue cancer-specific survival of patients with tongue squamous cell carcinoma, Cancer Medicine 6 (2017) 1002–1013, https://doi.org/10.1002/cam4.1021.
[9] A. Almangush, I.O. Bello, R.D. Coletta, A.A. Mäkitie, L.K. Mäkinen, J.H. Kauppila, et al., For early-stage oral tongue cancer, depth of invasion and worst pattern of invasion are the strongest pathological predictors for locoregional recurrence and mortality, Virchows Archiv 467 (2015) 39–46, https://doi.org/10.1007/s00428-015-1758-z.
[10] R. Mroueh, A. Haapaniemi, R. Grénman, J. Laranne, M. Pukkila, A. Almangush, et al., Improved outcomes with oral tongue squamous cell carcinoma in Finland: Oral tongue carcinoma in Finland, Head & Neck 39 (2017) 1306–1312, https://doi.org/10.1002/hed.24744.
[11] B.A.C. van Dijk, M.T. Brands, S.M.E. Geurts, M.A.W. Merkx, J.L.N. Roodenburg, Trends in oral cavity cancer incidence, mortality, survival and treatment in the Netherlands, Int J Cancer 139 (2016) 574–583, https://doi.org/10.1002/ijc.30107.
[12] Y. Kudo, Predicting cancer outcome: Artificial intelligence vs. pathologists, Oral Diseases 25 (2019) 643–645, https://doi.org/10.1111/odi.12954.
[13] American Joint Committee on Cancer, AJCC Cancer Staging Manual, Springer New York, New York, NY, 2002, https://doi.org/10.1007/978-1-4757-3656-4.
[14] L.H. Sobin, TNM: Evolution and relation to other prognostic factors, Seminars in Surgical Oncology 21 (2003) 3–7, https://doi.org/10.1002/ssu.10014.
[15] S.G. Patel, W.M. Lydiatt, Staging of head and neck cancers: is it time to change the balance between the ideal and the practical? J Surg Oncol 97 (2008) 653–657, https://doi.org/10.1002/jso.21021.
[16] D.A. Grimes, The nomogram epidemic: resurgence of a medical relic, Ann Intern Med 149 (2008) 273–275, https://doi.org/10.7326/0003-4819-149-4-200808190-00010.
[17] V.P. Balachandran, M. Gonen, J.J. Smith, R.P. DeMatteo, Nomograms in oncology: more than meets the eye, Lancet Oncol 16 (2015) e173–180, https://doi.org/10.1016/S1470-2045(14)71116-7.
[18] W. Sun, Y.-Z. Jiang, Y.-R. Liu, D. Ma, Z.-M. Shao, Nomograms to estimate long-term overall survival and breast cancer-specific survival of patients with luminal breast cancer, Oncotarget 7 (2016) 20496–20506, https://doi.org/10.18632/oncotarget.7975.
[19] J. Liu, Q. Geng, Z. Liu, S. Chen, J. Guo, P. Kong, et al., Development and external validation of a prognostic nomogram for gastric cancer using the national cancer registry, Oncotarget 7 (2016) 35853–35864, https://doi.org/10.18632/oncotarget.8221.
[20] N.D. Gross, S.G. Patel, A.L. Carvalho, P.-Y. Chu, L.P. Kowalski, J.O. Boyle, et al., Nomogram for deciding adjuvant treatment after surgery for oral cavity squamous cell carcinoma, Head Neck 30 (2008) 1352–1360, https://doi.org/10.1002/hed.20879.
[21] P.H. Montero, C. Yu, F.L. Palmer, P.D. Patel, I. Ganly, J.P. Shah, et al., Nomograms for preoperative prediction of prognosis in patients with oral cavity squamous cell carcinoma, Cancer 120 (2014) 214–221, https://doi.org/10.1002/cncr.28407.
[22] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R. D. Coletta, et al., Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool, Virchows Archiv 475 (2019) 489–497, https://doi.org/10.1007/s00428-019-02642-5.
[23] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R. D. Coletta, et al., Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer, International Journal of Medical Informatics (2019) 104068, https://doi.org/10.1016/j.ijmedinf.2019.104068.
[24] A.M. Bur, A. Holcomb, S. Goodwin, J. Woodroof, O. Karadaghy, Y. Shnayder, et al., Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma, Oral Oncology 92 (2019) 20–25, https://doi.org/10.1016/j.oraloncology.2019.03.011.
[25] M. Mermod, E. Jourdan, R. Gupta, M. Bongiovanni, G. Tolstonog, C. Simon, et al., Development and validation of a multivariable prediction model for the identification of occult lymph node metastasis in oral squamous cell carcinoma, Head & Neck, 2020, https://doi.org/10.1002/hed.26105.
[26] N. Sharma, H. Om, Data mining models for predicting oral cancer survivability, Network Modeling Analysis in Health Informatics and Bioinformatics 2 (2013) 285–295, https://doi.org/10.1007/s13721-013-0045-7.
[27] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, C.-N. Lin, The Application of Data Mining Techniques to Oral Cancer Prognosis, Journal of Medical Systems 39 (2015), https://doi.org/10.1007/s10916-015-0241-3.
[28] C. Lu, J.S. Lewis, W.D. Dupont, W.D. Plummer, A. Janowczyk, A. Madabhushi, An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival, Modern Pathology 30 (2017) 1655–1665, https://doi.org/10.1038/modpathol.2017.98.
[29] O.A. Karadaghy, M. Shew, J. New, A.M. Bur, Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma, JAMA Otolaryngology–Head & Neck Surgery 145 (2019) 1115, https://doi.org/10.1001/jamaoto.2019.0981.
[30] S.-W. Chen, Q. Zhang, Z.-M. Guo, W.-K. Chen, W.-W. Liu, Y.-F. Chen, et al., Trends in clinical features and survival of oral cavity cancer: fifty years of experience with 3,362 consecutive cases from a single institution, Cancer Management and Research 10 (2018) 4523–4535, https://doi.org/10.2147/CMAR.S171251.
[31] P. SEER, Surveillance, Epidemiology, and End Results (SEER) Program, Research Data 1973–2009, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, 2012 released April 2012, based on the November 2011 submission, www.seer.cancer.gov.
[32] Microsoft Azure Machine Learning Studio, Azure Machine Learning Studio. Documentation, 2018.
[33] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinformatics 14 (2013) 106, https://doi.org/10.1186/1471-2105-14-106.
[34] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, J. Poorolajal, Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, Clinical Epidemiology and Global Health (2018), https://doi.org/10.1016/j.cegh.2018.10.003.
[35] A.M. Carrington, P.W. Fieguth, H. Qazi, A. Holzinger, H.H. Chen, F. Mayr, et al., A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Medical Informatics and Decision Making 20 (2020), https://doi.org/10.1186/s12911-019-1014-6.
[36] Microsoft Azure Machine Learning Studio, Azure Machine Learning Studio. Documentation, 2018.
[37] R. Barga, V. Fontama, W.-H. Tok, Predictive analytics with Microsoft Azure Machine Learning, Second edition, Apress, Berkeley, CA, 2015.

[38] A. Holzinger, A. Carrington, H. Müller, Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations, KI - Künstliche Intelligenz 34 (2020) 193–198, https://doi.org/10.1007/s13218-020-00636-z.
[39] Y.-J. Tseng, H.-Y. Wang, T.-W. Lin, J.-J. Lu, C.-H. Hsieh, C.-T. Liao, Development of a Machine Learning Model for Survival Risk Stratification of Patients With Advanced Oral Cancer, JAMA Network Open 3 (2020) e2011768, https://doi.org/10.1001/jamanetworkopen.2020.11768.
[40] C.S. Chu, N.P. Lee, J. Adeoye, P. Thomson, S. Choi, Machine learning and treatment outcome prediction for oral cancer, Journal of Oral Pathology & Medicine (2020), https://doi.org/10.1111/jop.13089.
[41] J. Shan, R. Jiang, X. Chen, Y. Zhong, W. Zhang, L. Xie, et al., Machine Learning Predicts Lymph Node Metastasis in Early-Stage Oral Tongue Squamous Cell Carcinoma, Journal of Oral and Maxillofacial Surgery (2020), https://doi.org/10.1016/j.joms.2020.06.015.
[42] P.R. Jeyaraj, E.R. Samuel Nadar, Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm, Journal of Cancer Research and Clinical Oncology 145 (2019) 829–837, https://doi.org/10.1007/s00432-018-02834-7.
[43] Y. Ariji, M. Fukuda, Y. Kise, M. Nozawa, Y. Yanashita, H. Fujita, et al., Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology 127 (2019) 458–463, https://doi.org/10.1016/j.oooo.2018.10.002.
[44] J. Ren, M. Qi, Y. Yuan, S. Duan, X. Tao, Machine Learning–Based MRI Texture Analysis to Predict the Histologic Grade of Oral Squamous Cell Carcinoma, American Journal of Roentgenology (2020) 1–7, https://doi.org/10.2214/AJR.19.22593.
[45] R. Arrangoiz, F. Cordera, D. Caba, E. Moreno, E. Luque de Leon, M. Munoz, Oral Tongue Cancer: Literature Review and Current Management, Cancer Reports and Reviews 2 (2018), https://doi.org/10.15761/CRR.1000153.
[46] R. Siegel, J. Ma, Z. Zou, A. Jemal, Cancer statistics, 2014, CA Cancer J Clin 64 (2014) 9–29, https://doi.org/10.3322/caac.21208.
[47] L.B. Harrison, R.B. Sessions, M.S. Kies (Eds.), Head and neck cancer: a multidisciplinary approach, Fourth edition, Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA, 2014.
[48] S.B. Edge, in: American Joint Committee on Cancer, American Cancer Society (Eds.), AJCC cancer staging handbook: from the AJCC cancer staging manual, 7th ed., Springer, New York, 2010.
[49] P.E. Wallner, G.E. Hanks, S. Kramer, C.J. McLean, Patterns of Care Study. Analysis of outcome survey data-anterior two-thirds of tongue and floor of mouth, Am J Clin Oncol 9 (1986) 50–57, https://doi.org/10.1097/00000421-198602000-00013.
[50] S. Listl, L. Jansen, A. Stenzinger, K. Freier, K. Emrich, B. Holleczek, et al., Survival of Patients with Oral Cavity Cancer in Germany, PLoS ONE 8 (2013) e53415, https://doi.org/10.1371/journal.pone.0053415.
[51] A.S. Elstein, Heuristics and biases: selected errors in clinical reasoning, Acad Med 74 (1999) 791–794, https://doi.org/10.1097/00001888-199907000-00012.
[52] I. Vlaev, N. Chater, Game relativity: how context influences strategic decision making, J Exp Psychol Learn Mem Cogn 32 (2006) 131–149, https://doi.org/10.1037/0278-7393.32.1.131.
[53] I. Ganly, M. Amit, L. Kou, F.L. Palmer, J. Migliacci, N. Katabi, et al., Nomograms for predicting survival and recurrence in patients with adenoid cystic carcinoma. An international collaborative study, Eur J Cancer 51 (2015) 2768–2776, https://doi.org/10.1016/j.ejca.2015.09.004.
[54] S. Ali, F.L. Palmer, C. Yu, M. DiLorenzo, J.P. Shah, M.W. Kattan, et al., Postoperative nomograms predictive of survival after surgical management of malignant tumors of the major salivary glands, Ann Surg Oncol 21 (2014) 637–642, https://doi.org/10.1245/s10434-013-3321-y.
[55] J.-K. Cho, G.-J. Lee, K.-I. Yi, K.-S. Cho, N. Choi, J.S. Kim, et al., Development and external validation of nomograms predictive of response to radiation therapy and overall survival in nasopharyngeal cancer patients, Eur J Cancer 51 (2015) 1303–1311, https://doi.org/10.1016/j.ejca.2015.04.003.
[56] S.F. Shariat, P.I. Karakiewicz, N. Suardi, M.W. Kattan, Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature, Clin Cancer Res 14 (2008) 4400–4407, https://doi.org/10.1158/1078-0432.CCR-07-4713.
[57] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Inform 2 (2007) 59–77.
[58] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019), https://doi.org/10.1002/widm.1312.
[59] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019), https://doi.org/10.1002/widm.1312.

Machine learning for prognosis of oral cancer: What are the ethical challenges?

Long paper

Rasheed Omobolaji Alabi[1] [0000-0001-7655-5924], Tero Vartiainen[2] [0000-0003-3843-8561], Mohammed Elmusrati[1] [0000-0001-9304-6590]

[1] Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland
[2] Department of Computer Science, School of Technology and Innovations, University of Vaasa, Vaasa, Finland
rasheed.alabi@student.uwasa.fi

Abstract.
Background: Machine learning models have shown high performance, particularly in the diagnosis and prognosis of oral cancer. However, in actual everyday clinical practice, the diagnosis and prognosis using these models remain limited. This is due to the fact that these models have raised several ethical and morally laden dilemmas. **Purpose:** This study aims to provide a systematic state-of-the-art review of the ethical and social implications of machine learning models in oral cancer management. **Methods:** We searched the OvidMedline, PubMed, Scopus, Web of Science and Institute of Electrical and Electronics Engineers databases for articles examining the ethical issues of machine learning or artificial intelligence in medicine, healthcare or care providers. The Preferred Reporting Items for Systematic Review and Meta-Analysis was used in the searching and screening processes. **Findings:** A total of 33 studies examined the ethical challenges of machine learning models or artificial intelligence in medicine, healthcare or diagnostic analytics. Some ethical concerns were data privacy and confidentiality, peer disagreement (contradictory diagnostic or prognostic opinion between the model and the clinician), patient's liberty to decide the type of treatment to follow may be violated, patients–clinicians' relationship may change and the need for ethical and legal frameworks. **Conclusion:** Government, ethicists, clinicians, legal experts, patients' representatives, data scientists and machine learning experts need to be involved in the development of internationally standardised and structured ethical review guidelines for the machine learning model to be beneficial in daily clinical practice.

Keywords: Ethics, machine learning, oral tongue cancer, systematic review

1 Introduction

Cancer is the second leading cause of death, with an estimated 9.6 million deaths worldwide in 2018 (Bray et al., 2018). From this estimation, oral cancer accounts for 354,864 new cases and 177,384 deaths (Bray et al., 2018), making it one of the most common cancers and thus a source of significant health concern. Notably, oral squamous cell carcinoma is the most frequent of all cases of oral cancer (Ng et al., 2017). It represents about 90% of all the reported cases of oral cancer (Le Campion et al., 2017; Neville et al., 2009). Oral tongue cancer has been reported to have a worse prognosis than squamous cell carcinoma arising from other subsites of the oral cavity (Rusthoven et al., 2008). Therefore, an accurate tool for the effective prognostication of oral cancer is necessary.

Artificial intelligence (AI), or its subfield machine learning (ML), holds great promise in effective oral cancer diagnosis and prognosis (Amato et al., 2013), clinical decision making (Bennett & Hauser, 2013; Esteva et al., 2019; Topol, 2019) and personalised medicine (Dilsizian & Siegel, 2014) because of the improved availability of large datasets (big data), increased computational power and advances in ML training algorithms. In the era of unprecedented technological advancements, AI or ML is recognised as one of the most important application areas. It is currently positioned at the apex of the hype curve and is touted to facilitate improved diagnostics, prognostics, workflow and treatment planning and monitoring of oral cancer patients.

Several studies have been published emphasising the importance of ML techniques in prediction outcomes, such as recurrence (Alabi, Elmusrati, Sawazaki-Calone, et al., 2019; Alabi, Elmusrati, Sawazaki-Calone, et al., 2019), occult node metastasis (Bur et al., 2019) or five-year overall survival in oral cancer patients (Karadaghy et al., 2019). Despite the reported high accuracy in the application of ML techniques in head and neck cancer studies, there is also some trepidation among clinicians regarding its uncertain effect on the demand and training of the current and future workforce. Some clinicians have considered the introduction of ML to daily routine medical practice as a transformative improvement in the ability to diagnose the disease early enough and more accurately, and others have expressed concerns about the assessment of and consensus on possible ethical pitfalls. Interestingly, this is usually the case with most disruptive technologies.

The adoption of AI technology in actual daily medical practice has been argued to threaten patients' preference, safety and privacy (Michael, 2019). Considering the progress made by AI technology and ML-based models in cancer management, the current policy and ethical guidelines are lagging (Michael, 2019). Although there are some efforts to engage in these ethical discussions (Luxton, 2014, 2016; Peek et al., 2015), the medical community needs to be informed about the complexities surrounding the application of AI technology and ML-based models in actual clinical practice (Michael, 2019).

Studies have examined the ethical challenges in the implementation of AI, or its subfield ML, in healthcare or medicine. As this approach seems general, few published works have focused on the ethical challenges in AI or ML in oral cancer. Therefore, our study aims to systematically review the research on the ethics of AI in medicine. This study mainly focuses on ML models. These ethical dilemmas are

adapted to when these ML models are used in oral cancer management. To this end, this systematic review addresses the following research questions (RQ):

RQ. What are the ethical challenges in the integration of the ML model into the daily clinical practice of oral cancer management?

RQ. What are the generic approaches to addressing these ethical challenges?

This paper is organised as follows. Section 2 describes the methodology. Section 3 examines the results obtained from the systematic review. Section 4 discusses the results and the implications for daily clinical practices.

2 Materials and methods

2.1 Search protocol

In this study, we systematically retrieved all studies that examined ethics in ML or AI. The systematic search included the databases of OvidMedline, PubMed, Scopus, Institute of Electrical and Electronics Engineers, Web of Science and Cochrane Library from their inception until 17 March 2020. The search approach was developed by combining the following search keywords: [('machine learning OR artificial intelligence') AND ('ethics')]. The retrieved hits were further analysed for possible duplicates and irrelevant studies. To further minimise the omission of any study, the reference lists of all eligible articles were manually searched to ensure that all the relevant studies were duly included. In addition, the Preferred Reporting Items for Systematic Review and Meta-Analysis was used in the searching and screening processes (Figure 1).

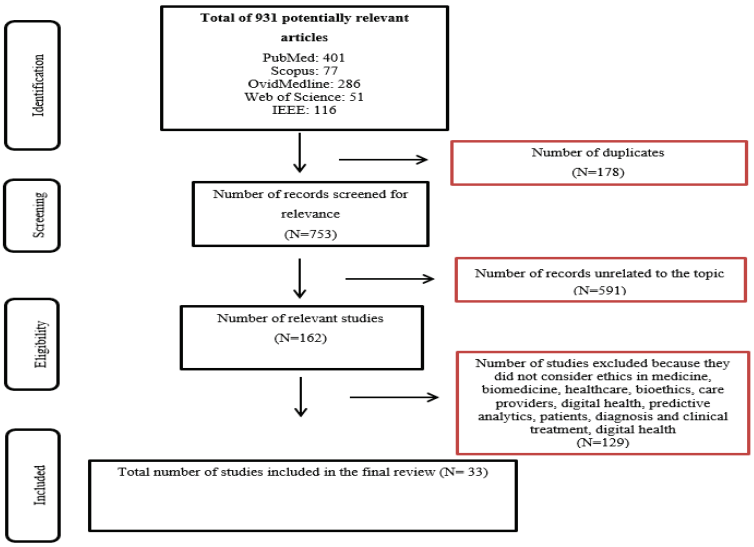

Figure 1. The number of articles included that examined the ethical concerns of ML models in medicine.

2.2 Inclusion and exclusion criteria

All original articles that considered the ethics of ML or AI in medicine or healthcare were included in this study. The eligible studies must have evaluated the ethical considerations or concerns of ML or AI in medicine. Studies that examined privacy issues, ethics of data practice and stewardship were also deemed eligible. Owing to the nature of the research questions in this study, perspectives, editorials and reviews were included. However, studies on animals, abstracts and conference papers were omitted. Articles in languages other than English were also excluded (Figure 1).

2.3 Screening

A data extraction sheet was used to minimise errors due to the omission of eligible studies.

2.4 Data extraction

The extracted parameters from each study included the author's/authors' name, year of publication, country of authors, title of studies and summary of the ethical issues mentioned in the study (Supplementary Table 1). Other important parameters, such as how to address such ethical challenges, were noted and discussed collectively in the discussion section.

3 Results

3.1 Results of the search strategy

The flow chart (Figure 1) describes the study selection process. A total of 931 hits were retrieved. Among them, 178 studies were found to be duplicate studies, and 591 were found to be irrelevant to the research questions in this review. Additionally, 129 studies did not consider ethics in medicine, biomedicine, healthcare, predictive analytics, digital health or patients. Thus, they were all excluded. Overall, 33 studies were found eligible for this systematic review (Figure 1, Supplementary Table 1). The findings of these studies indicated the ethical consideration of AI or ML in medicine. They were examined on how they relate to the implementation of ML models in oral cancer management. The ethical concerns discussed in these studies were privacy and confidentiality of patients' data, bias in the data used to develop the model, peer disagreement (Grote & Berens, 2020), responsibility or accountability gap (Grote & Berens, 2020; Jaremko et al., 2019; Kwiatkowski, 2018), fiduciary relationship between physicians and patients may change (Char et al., 2018; Nabi, 2018; Reddy et al., 2020) and patients' autonomy may be violated (Arambula & Bur, 2020; Boers et al., 2020; Grote & Berens, 2020; Johnson, 2019). These ethical concerns, brief definitions and corresponding structural aspects (what and how to address these concerns) are presented in Table 1.

Table 1. Ethical concerns of ML models in oral cancer prognostication.

Ethical concerns	**Meaning**	**The structural aspect of the ethical concerns** Ethical and moral concerns	
Privacy and confidentiality of patients' data	Approval of patients' consent and the concerned authority to use patients' data	*Concern 1:* Will the ML model developer use the extracted patients' information from the hospital registry without their consent?	*Concern VII*: How can the developer seek informed consent from the patient, hospital authority and national agency?
Bias in the data used to develop the model	Data may tend towards a particular race, geographical location, sexual orientation and so on	*Concern II:* Will the developed ML model be biased due to the imbalance in the data?	*Concern VIII*: How can the developer handle the possible data imbalance in the developed ML model?
Peer disagreement	Contradictory diagnostic or prognostic opinion between the model and the clinician	*Concern III:* Will the clinician follow his/her own diagnostic decision in cases in which the ML model gives a contrary opinion?	*Concern IX:* How can I find balance between conflicting diagnostic opinions? Is there an ethical guideline or standard that guides the use of a ML model in cancer management?
Responsibility gap	Assignment of responsibility when the ML models gave a wrong prediction	*Concern IV:* Will the clinician be held responsible when the ML model gives a wrong prediction?	*Concern X:* How should the clinicians interpret the hospital guidelines on the use of ML models? What does medical ethics stipulate? What are ethical guidelines or standards that guide the use of ML models in cancer management?
Clinician–patient relationship	Fiduciary interaction between the physicians and patients may change	*Concern V:* Will the patient feel comfortable and confident about the diagnostic decision made by a machine/computer?	*Concern X1:* How will I explain to the patient that the ML model is capable of making an accurate decision? How can I further justify the decision made by the model? How can I uphold clinician–patient relationship?
Patients' autonomy	Ability of the patient to determine the best treatment and take part in a shared decision-making process	*Concern VI:* Will the patient be allowed to choose the treatment approach that suits him/her when the model gives a different treatment plan?	*Concern XII:* How can the clinician take into consideration the treatment plan that best considers the daily activities of the patient?

The title of each concern (Table 1) addresses the core ethical challenge: in the case of ethical and moral concerns, 'Will the clinician, ML developer or the corresponding model perform the unethical action?' and in the case of morally acceptable actions, 'How can the clinician, ML developer or the corresponding model resolve the ethical concerns'? From these findings, it is important for the ML model to be trustworthy before it can be considered in actual medical practice. To ensure the trustworthiness of the model, the five trustworthiness principles of transparency, credibility, auditability, reliability and recoverability should be incorporated (Figure 2) (Keskinbora, 2019;

Rossi, 2016). Moreover, an ethics board has been proposed to discuss ethics in ML models from the perspective of experts and patients (Mamzer et al., 2017) (Figure 3).

3.2 Characteristics of the study

In terms of language, all the studies included were conducted in English. Out of the 33 included studies, 16 (48.5%) emphasised the privacy and confidentiality of patients' data (Bali et al., 2019a; Balthazar et al., 2018; Boers et al., 2020; Geis et al., 2019; Grote & Berens, 2020; Jaremko et al., 2019; Kluge, 1999; Kohli & Geis, 2018; Ma et al., 2019; Nabi, 2018; Nebeker et al., 2019; Reddy et al., 2020; Seddon, 1996; Sethi & Theodos, 2009; Vayena et al., 2018; Yuste et al., 2017), 13 (39.4%) examined the significance of informed consent, data protection, access, usability, sharing and regulatory schemes or rules prior to the use of patients' data (Balthazar et al., 2018; Gruson et al., 2019; Jaremko et al., 2019; Kluge, 1999; Kohli & Geis, 2018, 2018; Ma et al., 2019; Nabi, 2018; Nebeker et al., 2019; Reddy et al., 2020; Sethi & Theodos, 2009; Vayena et al., 2018; Yuste et al., 2017), 12 (36.4%) discussed the possibility bias in the data used for ML applications (Boers et al., 2020; Cahan et al., 2019; Char et al., 2018; Geis et al., 2019; Grote & Berens, 2020; Gruson et al., 2019; Kohli & Geis, 2018; Nabi, 2018; Reddy et al., 2020; Vayena et al., 2018; Wiens et al., 2019; Yuste et al., 2017), 4 (12.1%) suggested that the integration of ML models in clinical settings could assist clinicians to make informed decisions (Berner, 2002; Boers et al., 2020; Grote & Berens, 2020; Kwiatkowski, 2018) and 13 (39.4%) reported the need for ethical principles, guidelines and legal frameworks before ML models could be integrated into medical practice (Arambula & Bur, 2020; Cahan et al., 2019; Char et al., 2018; Gruson et al., 2019; Jian, 2019; Johnson, 2019; Keskinbora, 2019; Mamzer et al., 2017; Morley & Floridi, 2020; Nebeker et al., 2019; Rajkomar et al., 2018; Reddy et al., 2020; Robles Carrillo, 2020).

4 Discussion

This systematic review examined the ethical challenges in ML models in clinical practice. These challenges were examined on how they relate to the integration of ML models in oral cancer management. These ethical challenges carry significant implications in terms of integrating the ML model for daily routine in oral cancer management. The following highlights these ethical challenges and suggests a generic approach to addressing them.

Data privacy and confidentiality: the patient's consent should be sought

The first of these ethical concerns is healthcare data privacy (Arambula & Bur, 2020; Nabi, 2018). Developing ML models involves the substantial usage of healthcare data of patients. Therefore, it raises privacy and patient confidentiality concerns (Ma et al., 2019; Nabi, 2018). To arrest this concern, the patients, or their respective subjects, need to be informed about the collection and usage of their data (Geis et al., 2019; Powles & Hodson, 2017) to ensure informed consent and avoid illegal proprietary exploitation of the data and data privacy breaches (Bali et al., 2019b; Balthazar et al., 2018; Char et al., 2018; Nabi, 2018; Powles & Hodson, 2017; Yuste et al., 2017).

Nevertheless, it is important that data use agreements should be reviewed and approved by the appropriate quarters (Kohli & Geis, 2018). Moreover, a scheme (i.e., privacy-preserving clinical decision with cloud support) that preserves the privacy of the patient in terms of their data can be introduced (Geis et al., 2019; Liu et al., 2017; Ma et al., 2019; Vayena et al., 2018; Wang et al., 2015; Zhang et al., 2018). However, the discussion about the ownership of the data is beyond the scope of this study.

Trustworthy AI: the model should be trustworthy

It is important for the model to work as expected. Therefore, the model should have minimal errors in the training phase. Any form of error/malfunctioning of the model should be mentioned and defined (England & Cheng, 2019; Park & Kressel, 2018; Vayena et al., 2018; Zou & Schiebinger, 2018) to give transparency to the model and consequently, the results from these models (Geis et al., 2019; Park et al., 2019). Therefore, a possible imbalance in the data should be considered when developing the model to ensure the trustworthiness of the model. To address this challenge, related guidelines can be followed for transparent reporting (Bossuyt et al., 2015; Collins et al., 2015; England & Cheng, 2019). With these guidelines, the ML model deployed will be trustworthy and uphold the fundamental pillars of medical ethics (autonomy, beneficence, nonmaleficence and justice) (Arambula & Bur, 2020) and the ethical principles of transparency, credibility, audibility, reliability and recoverability (Keskinbora, 2019) (Figure 2).

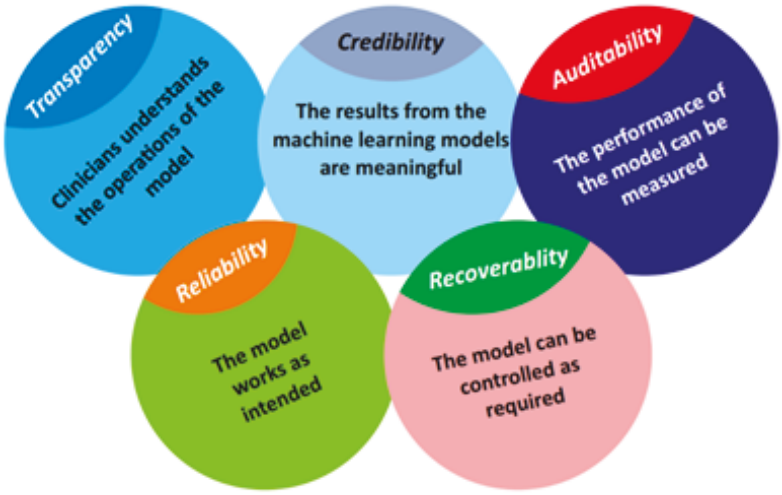

Figure 2. The trustworthiness principles expected from a ML model

In this way, an inherently biased model is avoided (Arambula & Bur, 2020; Collins & Moon, 2018; Reddy et al., 2020; Wiens et al., 2019). Trustworthiness should not only concern the properties of the ML or AI inherent model but also the socio-technical

systems involving the ML or AI applications (European Commission, 2019), that is, the expected trustworthiness of all actors and processes that constitute the socio-technical context in the application of AI for the prognostication of oral tongue cancer. Thus, for trustworthiness in AI, the essential components of trust in design, development, law compliance, ethics and robustness must be present (European Commission, 2019). In addition, the key requirements for a trustworthy AI include human regulatory agency, technical robustness and safety, privacy and data governance, transparency, non-discrimination and fairness, environmental friendliness and compliance, and accountability (European Commission, 2019).

Peer disagreement: the model and clinician should act to protect the patient from harm

As the ML model is viewed as an expert system/model, peer disagreement and its possible resolution guidelines are another important ethical issue (Christensen, 2007; Kelly T, 2010). What happens when the model and the clinicians disagree on the output of a proposition (diagnosis or prognosis) (Frances & Matheson, 2018)? It is impossible to have a dialogical engagement with the model, as proposed by Mercier and Sperber in the argumentative theory of reasoning (Mercier & Sperber, 2017). Should the clinician follow the proposition of the ML model (Christensen, 2007) or adhere to her own proposition (Enoch, 2010)? Therefore, there is a standoff in terms of the possible decision to make by the clinician. In this case, ethical guidelines and legal frameworks become imperative (Figure 3).

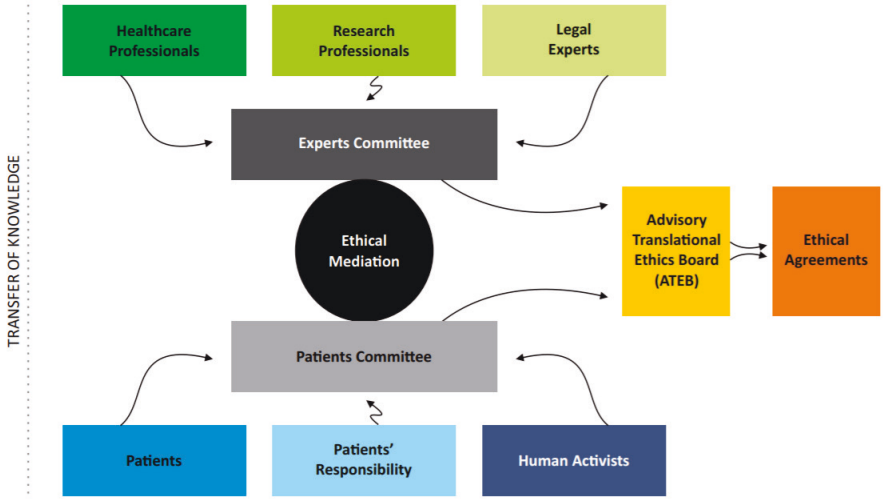

Figure 3. Ethical and legal frameworks for ethical agreements.

The ethical guidelines in this case ensure that clinicians make a decision to protect the safety and improve the overall health condition of the patient. The hospital and ethical guidelines should also address the possible errors that may arise from using the model (responsibility gap).

Patients' autonomy: shared decision making

The ethical question of patients' autonomy also comes to fore (Grote & Berens, 2020). For example, an ML model that predicts the type of treatment for an oral cancer patient should eschew the preferred treatment that could minimise the suffering of the patient. Instead, it should maximise the lifespan and overall survival of the patient, thereby making this model paternalistic in nature. This raises the ethical question of a shared decision making between the clinician and the patient to ensure that the autonomy and dignity of the patient are not violated (McDougall, 2019). Therefore, it is important to establish relevant standards to determine which information from the ML model is essential to be explained to the patient to be regarded as informed consent so that the patient can make an informed decision (Grote & Berens, 2020; McDougall, 2019; Mittelstadt & Floridi, 2016).

Humanness: Empathy and trust from the clinician–patient relationship

Another ethical concern is the 'humanness' of clinicians and the role of cognitive empathy, trust, responsibility and confidentiality among clinicians (Boers et al., 2020). This seems to be a source of concern, as the integration of ML models in oral cancer management may lead to a paradigm shift from the current face-to-face or direct interaction between patients and clinicians (two-way diagnostic procedure) to a triangular relationship of patients–models–clinicians (three-way diagnostic procedure). This concern becomes pronounced especially when the models are publicly available, as the patients may engage in self-medication and self-management. Thus, the fundamental aspects of patients' care may be undermined (Boers et al., 2020). To mitigate this, these models should be integrated in such a way that restricts patients' access. In this way, the patient–clinician relationship can still be maintained, as this type of relationship has been reported to influence how patients respond to their illnesses and treatments (Kelley et al., 2014).

Ethics is one of the essential components to achieve a trustworthy AI. It is important to have a model that ensures compliance to ethical norms and principles, including fundamental human rights, moral entitlements and acceptable moral values (European Commission, 2019). As mentioned previously, some of these principles include respect for human autonomy, prevention of harm, fairness and explicability (European Commission, 2019). To this end, we tend to agree with the suggestion of setting up a dedicated ethical research agenda (Boers et al., 2020). This ethical research agenda is expected to form the required premise for the development of internationally standardised and structured ethical review guidelines (Arambula & Bur, 2020; Gruson et al., 2019; Johnson, 2019, 2019). These guidelines should emphasise the fundamental ethical rules of honesty, truthfulness, transparency, benevolence, non-malevolence and respect for autonomy (Keskinbora, 2019) and address other criticisms surrounding the application of ML-based models in actual clinical practice (Figure 4).

Figure 4. The fundamental ethical principles expected from the clinician and the ML model.

Aside from these ethical guidelines, corresponding laws (internal framework and international sphere) should be enacted by the government to ensure the legal (e.g., the European General Data Protection Regulations) (Flaumenhaft & Ben-Assuli, 2018; Vayena et al., 2018) and jurisdictional mechanisms for their enforcement (Robles Carrillo, 2020).

In conclusion, the development of ML models should take the ethical and legal framework into consideration from the data collection to the ML process and to the integration into clinical practice. A strong and proactive role is expected from the government, clinical experts, patients' representatives, data scientists, ML experts and legal and human rights activists in defining these ethical guidelines. Through this, ML models can achieve the touted benefits of optimising health systems and decision support for professionals and improve the overall health of patients. As oral tongue cancer was considered in this study, the ethical concerns mentioned and the proposed solution are peculiar to other cancer types.

Supplementary

Supplementary Table 1. Included studies and the main ethical points discussed (below)

Name of Author	Country	Year Published	Title	Ethical points/Summary
Seddon A.M	United Kingdom	1996	Predicting our health: ethical implications of neural networks and outcome potential predictions	▪ Privacy concern
Kluge E.H	Canada	1999	Medical narratives and patient analogs: ethical implications of electronic patient records	▪ Privacy ▪ Accessibility of the data
Bernie E.S	United States	2002	Ethical and legal issues in the use of clinical decision support systems	▪ Informed decision ▪ Transparency
Sethi & Theodos	United States	2009	Translational bioinformatics and healthcare informatics: computational and ethical challenges	▪ Sensitive nature of genetic data ▪ Privacy and confidentiality
Mamzer et al.	France	2017	Partnering with patients in translational oncology research: ethical approach	▪ To establish a long-term partnership integrating patient's expectations ▪ Expert and Patient ▪ Cancer research and personalised medicine (CARPEM) develops translational research of precision medicine for cancer
Yuste et al.	United States	2017	Four ethical priorities for neuroethologies and AI	▪ Privacy and consent ▪ Agency and identity ▪ Augmentation ▪ Biases
Balthazar et al.	United States	2017	Protecting patients' interest in the era of big data, artificial intelligence and predictive analytics	▪ Privacy ▪ Confidentiality ▪ Data ownership ▪ Informed consent ▪ Epistemology ▪ Inequalities
Kwiatkowski W	Poland	2018	Medicine and technology. Remarks on the notion of responsibility in	▪ The notion of responsibility

			technology-assisted healthcare	
Rajkomar et al.	United States	2018	Ensuring fairness in machine learning to advance health equity	▪ The principle of distributed justice ▪ Health equity ▪ Four medical ethics principles
Char et al.	United States	2018	Implementing machine learning in healthcare– Addressing ethical challenges	▪ Biases ▪ Fiduciary relationship between physicians and patients ▪ Ethical guidelines ▪ Proposition of policy enactment, programming approaches, task force or a combination of these strategies
Nabi Junaid	United States	2018	How bioethics can shape artificial intelligence and machine learning	▪ Biases ▪ Privacy of patients ▪ Informed consent ▪ Fairness, trust, equity and confidentiality ▪ Patient–clinician relationship might change
Vayena et al.	Switzerland & United States	2018	Machine learning in medicine: Addressing ethical challenges	▪ Data protection ▪ Privacy preservation ▪ Biased dataset ▪ Fairness and transparency
Kohli & Geis	United States	2018	Ethics, artificial Intelligence and radiology	▪ Informed consent ▪ Privacy ▪ Objectivity ▪ Data protection ▪ Ownership ▪ Bias ▪ Data use agreement ▪ Review of agreement ▪ Safety, transparency and value alignment
Geis et al.	Canada	2019	Ethics of artificial intelligence in radiology	▪ Bias ▪ Informed consent ▪ Privacy ▪ Data protection ▪ Ownership ▪ Objectivity and transparency

Cahan et al.	United States	2019	Putting the data before the algorithm in big data addressing personalised healthcare	▪ Biases in the data ▪ Handling the confluence between data and algorithm ▪ Generalisability of the model ▪ Introducing the quality standard for the dataset guidelines
Gruson et al.	Belgium	2019	Data science, artificial intelligence and machine learning: opportunities for laboratory medicine and the value of positive regulation	▪ Biases ▪ Patient information and consent ▪ Ethical and legal frameworks ▪ AI human warranty ▪ Regulation of health data according to their level of sensitivity
Jaremko et al.	Canada	2019	Canadian Association of Radiologists white paper on ethical and legal issues related to artificial intelligence in radiology	▪ Data value and ownership ▪ Data privacy ▪ Data sharing rules ▪ Reliability gap (liability)
Nebeker et al.	United States	2019	Building the case for actionable ethics in digital health research supported by artificial intelligence	▪ Privacy ▪ Data management ▪ Risks and benefits ▪ Access and usability ▪ Ethical principles (respect for persons, beneficence, justice)
Wiens at al.	United States and Canada		Do not harm: a roadmap for responsible machine learning for healthcare	▪ Choosing the right problems ▪ Developing a useful solution ▪ Biases ▪ Proper evaluation of the performance of the model ▪ Thoughtful reporting of the models' results ▪ Integration (making it to the market)
Guan Jian			Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance	▪ Role of government in the ethical auditing ▪ Stakeholders' responsibilities in ethical governance system

Nikola et al.		2019	Algorithm-aided prediction of patient preference: an ethics sneak peek	▪ Protection of patients and clinicians (safety, validity, reproducibility, usability, reliability) ▪ Transparency and comprehensibility ▪ Quality control and monitoring of models
Ma et al.	China	2019	PPCD: Privacy-preserving clinical decision with cloud support	• Data usage privacy concern scheme
Bali et al.	India	2019	Artificial intelligence in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required	▪ Data privacy ▪ Confidentiality ▪ Do not harm principle should be upheld
Mazurowski	United Kingdom	2019	Artificial intelligence in radiology: some ethical considerations for radiologists and algorithm developers	• When it is unethical for a radiologist to oppose AI • Conflicts of interests between radiologists and AI developers
Keskinbora	Turkey	2019	Medical ethics considerations on artificial intelligence	▪ Trustworthy AI ▪ Important ethical principles should be embraced
Park et al.	Korea	2019	Ethical challenges in artificial intelligence in medicine from the perspective of science editing and peer review	▪ Transparency in training, testing and validation dataset ▪ Clearly explains the data preparation processes
Johnson Sandra	United Kingdom	2019	AI, machine learning and ethics in healthcare	▪ Ethical guidelines recommendation ▪ Vigilant to potential errors and biases ▪ Medical ethics
Arambula & Bur	United States	2019	Ethical considerations in the advent of artificial intelligence in otolaryngology	▪ Four ethical principles (respect for patient's autonomy, beneficence, nonmaleficence, and justice)

Reddy et al.,	Australia	2019	A governance model for the application of AI in healthcare	▪ AI biases ▪ Privacy ▪ Patient and clinician trust ▪ Regulatory guidelines ▪ Proposed governance for AI in healthcare ▪ Stages for monitoring and evaluating AI-enabled services
Boers et al.,	Netherlands and United Kingdom	2019	SERIES: eHealth in primary care. Part 2: Exploring the ethical implications of its application in primary care practice	▪ Biased and discriminatory algorithm ▪ Patient's autonomy ▪ Shared decision making ▪ Data privacy, trust and confidentiality
Morley & Floridi	United Kingdom	2020	An ethical mindful approach to AI for healthcare	• Internationally standardised and structured ethical review guidelines
Carrillo et al.,	Spain	2020	Artificial intelligence: From ethics to law	▪ Distinguish between legal and ethical aspects ▪ Non-formalistic approach to law ▪ International law is identified as the principal legal framework for the regulation of AI models
Grote & Berens	Germany	2020	On the ethics of algorithmic decision making in healthcare	• Peer disagreement • Patients' autonomy • Shared decision making • Obscuration of accountability • Biased and discriminatory algorithm • Data privacy

References

Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., Mäkitie, A. A., Salo, T., Almangush, A., & Leivo, I. (2019). Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *International Journal of Medical Informatics*, 104068. https://doi.org/10.1016/j.ijmedinf.2019.104068

Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., Mäkitie, A. A., Salo, T., Leivo, I., & Almangush, A. (2019). Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool. *Virchows Archiv*, *475*(4), 489–497. https://doi.org/10.1007/s00428-019-02642-5

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, *11*(2), 47–58. https://doi.org/10.2478/v10136-012-0031-x

Arambula, A. M., & Bur, A. M. (2020). Ethical Considerations in the Advent of Artificial Intelligence in Otolaryngology. *Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, *162*(1), 38–39. https://doi.org/10.1177/0194599819889686

Bali, J., Garg, R., & Bali, R. T. (2019a). Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian Journal of Ophthalmology*, *67*(1), 3–6. https://doi.org/10.4103/ijo.IJO_1292_18

Bali, J., Garg, R., & Bali, R. T. (2019b). Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian Journal of Ophthalmology*, *67*(1), 3–6. https://doi.org/10.4103/ijo.IJO_1292_18

Balthazar, P., Harri, P., Prater, A., & Safdar, N. M. (2018). Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. *Journal of the American College of Radiology*, *15*(3), 580–586. https://doi.org/10.1016/j.jacr.2017.11.035

Bennett, C. C., & Hauser, K. (2013). Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial Intelligence in Medicine*, *57*(1), 9–19. https://doi.org/10.1016/j.artmed.2012.12.003

Berner, E. S. (2002). Ethical and legal issues in the use of clinical decision support systems. *Journal of Healthcare Information Management: JHIM*, *16*(4), 34–37.

Boers, S. N., Jongsma, K. R., Lucivero, F., Aardoom, J., Büchner, F. L., de Vries, M., Honkoop, P., Houwink, E. J. F., Kasteleyn, M. J., Meijer, E., Pinnock, H., Teichert, M., van der Boog, P., van Luenen, S., van der Kleij, R. M. J. J., & Chavannes, N. H. (2020). SERIES: eHealth in primary care. Part 2: Exploring the ethical implications of its application in primary care practice. *European Journal of General Practice*, *26*(1), 26–32. https://doi.org/10.1080/13814788.2019.1678958

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). STARD 2015: an

updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, h5527. https://doi.org/10.1136/bmj.h5527

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Bur, A. M., Holcomb, A., Goodwin, S., Woodroof, J., Karadaghy, O., Shnayder, Y., Kakarala, K., Brant, J., & Shew, M. (2019). Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. *Oral Oncology*, *92*, 20–25. https://doi.org/10.1016/j.oraloncology.2019.03.011

Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *Npj Digital Medicine*, *2*(1). https://doi.org/10.1038/s41746-019-0157-2

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *Philosophical Review*, *116*(2), 187–217. https://doi.org/10.1215/00318108-2006-035

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, *350*(jan07 4), g7594–g7594. https://doi.org/10.1136/bmj.g7594

Collins, & Moon. (2018). Is digital medicine different? *The Lancet*, *392*(10142), 95. https://doi.org/10.1016/S0140-6736(18)31562-9

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current Cardiology Reports*, *16*(1), 441. https://doi.org/10.1007/s11886-013-0441-8

England, J. R., & Cheng, P. M. (2019). Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. *American Journal of Roentgenology*, *212*(3), 513–519. https://doi.org/10.2214/AJR.18.20490

Enoch, D. (2010). Not Just a Truthometer: Taking Oneself Seriously (but not Too Seriously) in Cases of Peer Disagreement. *Mind*, *119*(476), 953–997. https://doi.org/10.1093/mind/fzq070

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Commission. (2019). *High-level expert group on artificial intelligence. In: Ethics Guidelines for Tustrworthy AI.* European Union.

Flaumenhaft, Y., & Ben-Assuli, O. (2018). Personal health records, global policy and regulation review. *Health Policy*, *122*(8), 815–826. https://doi.org/10.1016/j.healthpol.2018.05.002

Frances, B., & Matheson, J. (2018). *Disagreement. In: Zalta EN,ed. The Stanford encyclopedia of philosophy. Spring.* 8. https://plato.stanford.edu/archives/spr2018/entries/disagreement/

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Borondy Kitts, A., Birch, J., Shields, W. F., van den Hoven van Genderen, R., Kotter, E., Wawira Gichoya, J., Cook, T. S., Morgan, M. B., Tang, A., Safdar, N. M., & Kohli, M. (2019). Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. *Radiology*, *293*(2), 436–440. https://doi.org/10.1148/radiol.2019191586

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Gruson, D., Helleputte, T., Rousseau, P., & Gruson, D. (2019). Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation. *Clinical Biochemistry*, *69*, 1–7. https://doi.org/10.1016/j.clinbiochem.2019.04.013

Jaremko, J. L., Azar, M., Bromwich, R., Lum, A., Alicia Cheong, L. H., Gibert, M., Laviolette, F., Gray, B., Reinhold, C., Cicero, M., Chong, J., Shaw, J., Rybicki, F. J., Hurrell, C., Lee, E., Tang, A., & Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. (2019). Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology.

Canadian Association of Radiologists Journal = Journal l'Association Canadienne Des Radiologistes, *70*(2), 107–118. https://doi.org/10.1016/j.carj.2019.03.001

Jian, G. (2019). Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges, and Governance. *Chinese Medical Sciences Journal*, *0*(0), 99. https://doi.org/10.24920/003611

Johnson, S. L. J. (2019). AI, Machine Learning, and Ethics in Health Care. *Journal of Legal Medicine*, *39*(4), 427–441. https://doi.org/10.1080/01947648.2019.1690604

Karadaghy, O. A., Shew, M., New, J., & Bur, A. M. (2019). Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma. *JAMA Otolaryngology–Head & Neck Surgery*, *145*(12), 1115. https://doi.org/10.1001/jamaoto.2019.0981

Kelley, J. M., Kraft-Todd, G., Schapira, L., Kossowsky, J., & Riess, H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. *PloS One*, *9*(4), e94207. https://doi.org/10.1371/journal.pone.0094207

Kelly T. (2010). *Peer disagreement and higher order evidence. In: Goldman Al, Whitcomb D, eds. Social Epistemology: essential readings. Oxford University Press.* Oxford University Press.

Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of Clinical Neuroscience*, *64*, 277–282. https://doi.org/10.1016/j.jocn.2019.03.001

Kluge, E. H. (1999). Medical narratives and patient analogs: the ethical implications of electronic patient records. *Methods of Information in Medicine*, *38*(4–5), 253–259.

Kohli, M., & Geis, R. (2018). Ethics, Artificial Intelligence, and Radiology. *Journal of the American College of Radiology*, *15*(9), 1317–1319. https://doi.org/10.1016/j.jacr.2018.05.020

Kwiatkowski, W. (2018). Medicine and technology. Remarks on the notion of responsibility in the technology-assisted health care. *Medicine, Health Care and Philosophy*, *21*(2), 197–205. https://doi.org/10.1007/s11019-017-9788-8

Le Campion, A. C. O. V., Ribeiro, C. M. B., Luiz, R. R., da Silva Júnior, F. F., Barros, H. C. S., dos Santos, K. de C. B., Ferreira, S. J., Gonçalves, L. S., & Ferreira, S. M. S. (2017). Low Survival Rates of Oral and Oropharyngeal Squamous Cell Carcinoma. *International Journal of Dentistry*, *2017*, 1–7. https://doi.org/10.1155/2017/5815493

Liu, Y., Li, Y., Fu, Y., Liu, T., Liu, X., Zhang, X., Fu, J., Guan, X., Chen, T., Chen, X., & Sun, Z. (2017). Quantitative prediction of oral cancer risk in patients with oral leukoplakia. *Oncotarget*, *8*(28). https://doi.org/10.18632/oncotarget.17550

Luxton, D. D. (2014). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, *62*(1), 1–10. https://doi.org/10.1016/j.artmed.2014.06.004

Luxton, D. D. (2016). *Artificial Intelligence in Behavioral and Mental Health Care*. Elsevier. https://doi.org/10.1016/C2013-0-12824-3

Ma, H., Guo, X., Ping, Y., Wang, B., Yang, Y., Zhang, Z., & Zhou, J. (2019). PPCD: Privacy-preserving clinical decision with cloud support. *PLOS ONE*, *14*(5), e0217349. https://doi.org/10.1371/journal.pone.0217349

Mamzer, M.-F., Duchange, N., Darquy, S., Marvanne, P., Rambaud, C., Marsico, G., Cerisey, C., Scotté, F., Burgun, A., Badoual, C., Laurent-Puig, P., & Hervé, C. (2017). Partnering with patients in translational oncology research: ethical approach. *Journal of Translational Medicine*, *15*(1). https://doi.org/10.1186/s12967-017-1177-9

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Michael, R. (2019). Ethical Dimensions of Using Artificial Intelligence in Health Care. *AMA Journal of Ethics*, *21*(2), E121-124. https://doi.org/10.1001/amajethics.2019.121

Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, *22*(2), 303–341. https://doi.org/10.1007/s11948-015-9652-2

Morley, J., & Floridi, L. (2020). An ethically mindful approach to AI for health care. *The Lancet*, *395*(10220), 254–255. https://doi.org/10.1016/S0140-6736(19)32975-7

Nabi, J. (2018). How Bioethics Can Shape Artificial Intelligence and Machine Learning. *The Hastings Center Report*, *48*(5), 10–13. https://doi.org/10.1002/hast.895

Nebeker, C., Torous, J., & Bartlett Ellis, R. J. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Medicine*, *17*(1). https://doi.org/10.1186/s12916-019-1377-7

Neville, B., Damm, D. ., & Allen, C., M. (2009). *Oral and Maxillofacial Pathology* (3rd edition). Elseiver.

Ng, J. H., Iyer, N. G., Tan, M.-H., & Edgren, G. (2017). Changing epidemiology of oral squamous cell carcinoma of the tongue: A global study: Changing epidemiology of tongue cancer. *Head & Neck*, *39*(2), 297–304. https://doi.org/10.1002/hed.24589

Park, S. H., Kim, Y.-H., Lee, J. Y., Yoo, S., & Kim, C. J. (2019). Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review. *Science Editing*, *6*(2), 91–98. https://doi.org/10.6087/kcse.164

Park, S. H., & Kressel, H. Y. (2018). Connecting Technological Innovation in Artificial Intelligence to Real-world Medical Practice through Rigorous Clinical Validation: What Peer-reviewed Medical Journals Could Do. *Journal of Korean Medical Science*, *33*(22). https://doi.org/10.3346/jkms.2018.33.e152

Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial Intelligence in Medicine*, *65*(1), 61–73. https://doi.org/10.1016/j.artmed.2015.07.003

Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and Technology*, *7*(4), 351–367. https://doi.org/10.1007/s12553-017-0179-1

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, *169*(12), 866–872. https://doi.org/10.7326/M18-1990

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, *27*(3), 491–497. https://doi.org/10.1093/jamia/ocz192

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, 101937. https://doi.org/10.1016/j.telpol.2020.101937

Rossi, F. (2016). Artificial Intelligence: Potential Benefits and Ethical Considerations. *Artificial Intelligence*, 8.

Rusthoven, K., Ballonoff, A., Raben, D., & Chen, C. (2008). Poor prognosis in patients with stage I and II oral tongue squamous cell carcinoma. *Cancer*, *112*(2), 345–351. https://doi.org/10.1002/cncr.23183

Seddon, A. M. (1996). Predicting our health: ethical implications of neural networks and outcome potential predictions. *Ethics & Medicine: A Christian Perspective on Issues in Bioethics*, *12*(3), 53–54.

Sethi, P., & Theodos, K. (2009). Translational bioinformatics and healthcare informatics: computational and ethical challenges. *Perspectives in Health Information Management*, *6*, 1h.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Wang, G., Lu, R., & Huang, C. (2015). PSLP: Privacy-preserving single-layer perceptron learning for e-Healthcare. *2015 10th International Conference on Information,*

Communications and Signal Processing (ICICS), 1–5. https://doi.org/10.1109/ICICS.2015.7459925

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, *25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., … Wolpaw, J. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, *551*(7679), 159–163. https://doi.org/10.1038/551159a

Zhang, X., Chen, X., Wang, J., Zhan, Z., & Li, J. (2018). Verifiable privacy-preserving single-layer perceptron training scheme in cloud computing. *Soft Computing*, *22*(23), 7719–7732. https://doi.org/10.1007/s00500-018-3233-7

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future

Rasheed Omobolaji Alabi M.Sc [a], Matti Pirinen [d], Mohammed Elmusrati D.Sc [a], Antti A. Mäkitie MD, PhD [c,e], Ilmo Leivo MD, PhD [f*], Alhadi Almangush DDS, PhD [b,c,f,g*]

[a] Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland.

[b] Department of Pathology, University of Helsinki, Helsinki, Finland.

[c] Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland.

[d] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.
Department of Public Health, University of Helsinki, Helsinki, Finland.
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland.

[e] Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden.

[f] University of Turku, Institute of Biomedicine, Pathology, Turku, Finland.

[g] Faculty of Dentistry, University of Misurata, Misurata, Libya.

***The last two authors have equal contributions.**

Manuscript text count: 2,959

Corresponding Author: Rasheed Omobolaji Alabi
Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland. **E-mail address:** rasheed.alabi@student.uwasa.fi

Disclosure: The authors declare no conflicts of interest.

Abstract

Importance: Oral cancer can show heterogenous patterns of behavior. For proper and effective management of oral cancer, early diagnosis and prognosis are important. To achieve this, artificial intelligence (AI) or its subfield, machine learning, has been touted for its potential to revolutionize cancer management through improved diagnostic precision and prediction of outcomes. Yet, to date, it has made only few contributions to actual medical practice or patient care. **Objectives:** This study provides a state of art the review of diagnostic and prognostic roles of machine learning in oral squamous cell carcinoma (OSCC) and also highlights some of the limitations and concerns of clinicians towards the implementation of these models into daily clinical practice. **Design:** We searched OvidMedline, PubMed, Scopus, Web of Science, and Institute of Electrical and Electronics Engineers (IEEE) databases for articles that used machine learning for diagnostic or prognostic purposes of OSCC. We used the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) in the searching and screening processes. **Main outcomes and measures:** The clinical concerns for the integration of machine learning models for actual daily practice in oral tongue cancer were identified. **Results:** A total of 41 studies were reported to have used machine learning to analyse of OSCC. The majority of these studies used support vector machine (SVM) and artificial neural network (ANN) algorithms as machine learning techniques. Their specificity ranged from 0.57 to 1.00, sensitivity from 0.70 to 1.00, and accuracy from 63.4% to 100.0% in these studies. The main limitations and concerns were a lack of proper understanding of the used machine learning models, inability to interpret which aspect of the data contributes to the result, concern about models possibly rendering the clinicians less important in patient management decisions, and privacy violation. **Conclusion:** The accumulated evidence indicates that machine learning models have a great potential in improving survival of OSCC patients. Therefore, it is important that the concerns of the clinicians are taken into consideration in the development of machine learning models. This would allow for a seamless integration of these models into the daily clinical practice.

KEYWORDS: Machine learning; Oral squamous cell carcinoma; Systematic review; explainable AI

1. Introduction

Oral cancer is an aggressive disease characterized by a low average survival rate [1]. Developments in treatment modalities in the domains of both oncology and surgery have only contributed to a rather limited improvement in outcome. Therefore, accurate diagnosis and prognosis prediction of cancer, especially at an early stage are important in improving survival rate [2]. The availability of different treatment options for oral cancer requires a proper selection of the treatment on a case-by-case basis.

Despite improved effect of the treatment, individualized patient-specific treatments are mostly lacking. Thus, improvements in diagnostic and prognostic accuracy could significantly assist the clinicians in making informed decisions on treatment. To this end, technical advances in statistics and computer software have led to improved prognostication using multi-factor analysis via conventional logistic and Cox regression models. Similarly, the application of machine learning techniques, a subfield of artificial intelligence (AI), plays a major role in the improved prediction of cancer outcomes. Several studies have reported that machine learning approach is more accurate in prognostication than the traditional statistical analyses [3–7].

Machine learning approach was found to be beneficial in the three aspects that are essential to early diagnosis and prognosis. These are an improved accuracy of cancer susceptibility, recurrence, and survival predictions [2], which improve the survival rates through the effective clinical management of patients [8–14]. Over the coming years, the application of the machine learning approach to clinical research continues to increase due to its feasibility and its many advantages. For instance, our group has used machine learning techniques to predict the locoregional recurrence of oral tongue cancer [15]. Similarly, it has been used to detect oral cancer [16–22], and to predict oral cancer recurrence [23,24], occult node metastasis [25,26], and survival rates of oral cancer [27–30]. Additionally, it has been used for the prognostication of other cancers [31–33] and to predict progression of diseases on

the basis of patient records such as from pre-diabetes to type 2 diabetes based on the patients' records [34]. All these applications of machine learning in healthcare are aimed at assisting the doctors in making informed decisions, reducing diagnostics errors, improving and promoting the overall patient health.

Despite numerous studies on the application of machine learning and various intelligent models deployed, the question remains – what are the concerns of clinicians towards the actual implementation of machine learning-based models in clinical settings? These concerns were considered from the limitations, shortcomings, and clinicians' concerns in the published studies regarding the application of machine learning for oral squamous cell carcinoma (OSCC) prognosis. This study, therefore, aims to systematically review the studies on the application of machine learning for diagnosis and prognosis of oral squamous cell carcinoma. OSCC was chosen in this review as it is the most common malignancy of the oral cavity. Also, it constitutes a majority of head and neck squamous cell carcinoma.

2. Methods

2.1. Search protocol. In this study, we systematically retrieved all studies that applied machine learning techniques to oral cancer diagnosis or prognosis. The systematic search included databases of OvidMedline, PubMed, Scopus, Web of Science, and Institute of Electrical and Electronics Engineers (IEEE) from their inception until February 2020. The search approach was developed by combining search keywords: [('oral cancer') AND ('machine learning')]. An additional search was conducted using the search terms: [('oral cancer') AND ('artificial neural network' OR 'ensemble method')]. To minimize the possibility of omission of any study, the reference lists of all the eligible articles were manually searched to ensure that all the relevant studies were duly included. Also, the Preferred

Reporting Items for Systematic Review and Meta-Analysis (PRISMA) was followed in the searching and screening processes (Figure 1) [35].

2.2. Inclusion and exclusion criteria. The eligible studies must have evaluated the diagnostic or prognostic significance of using machine learning algorithms in oral cancer. Invited reviews, review articles, case series, case reports, abstracts, studies on animals, conference papers, editorials, letters to the editors, commentaries, comparative studies, expert views, and general studies on cancer (not specific to oral cancer) were all excluded. Similarly, articles in languages other than English were excluded. Studies that examined machine learning application for normal oral mucosa, oral lesions (without cancer), oral caries, oral mucosa, DNA and RNA microarray genes, proteomics, fluorescence spectroscopy, genetic programming and Fuzzy systems were excluded. The details of the inclusion and exclusion criteria are described in Figure 1.

2.3. Screening. To ensure that all eligible studies were included in this study, a data extraction sheet was used where the studies selected to meet the required criteria for this review. The data extraction process was conducted by two independent reviewers (A.R., & O.Y.). Possible discrepancies were resolved by discussion. A consensus was reached on which studies should be included or excluded after deliberations considering the objectives, and the inclusion and exclusion criteria of the study.

2.3. Data extraction. The extracted parameters from each study included author (s) name, year of publication, country of authors, site of mouth cancer, number of study participants, machine learning algorithms examined in the study, definition of study objective (prognostic or diagnostic), study aim, results, performance metrics (accuracy and/or specificity, or area under receiving operating characteristics (ROC) curve AUC) reported, and conclusion from the study (Table 1). When more than one algorithm was considered in the study, the algorithm with the best performance metrics was extracted, and included in the corresponding column in

Table 1. Similarly, where the results were reported separately for training and validation sets, the reported results for the validation were presented as shown in Table 1. Other important information, such as the limitations of the study and the prognostic significance of the application of the machine learning technique, were noted and summarized in the Discussion section.

2.4. Quality assessment. We used the guidelines for developing and reporting machine learning predictive models to assess the quality of studies that evaluated the application of machine learning in the prognosis of OSCC [36]. We summarized the main guidelines in Table 2. Each point from the guidelines carries a single mark. The threshold was set to be half of the maximum marks. The details of the studies and the final score from these guidelines are given in Table 3.

3. Results

3.1. Results of the database search. The PRISMA flowchart (Figure 1) describes the study selection process. A total of 297 hits were retrieved. After deleting duplicates (N = 150), irrelevant papers (N = 91), and exclusions (N = 15), we found 41 studies eligible to be included in this systematic review as shown in Figure 1 [5, 15–30, 37–60]. The findings of these studies (summarized in Table 1) indicated that the application of machine learning techniques for oral cancer (diagnosis and/or prognosis) could assist the clinicians in making informed decisions regarding diagnostics and prognostic parameters. The results also indicated that these techniques are poised to offer personalized patient care and could improve survival and reduce the death rate associated with oral cancer. In addition, many of these studies mentioned significant limitations for the adoption of such models to actual daily medical practice.

3.2. Characteristics of relevant studies.

All the articles included were published in the English language. Of the 41 included studies, 35 studies considered oral cavity cancer in general [16–30,37,40,41,43,44,46,48,49,52–60], 4 studies focused on oral tongue squamous cell carcinoma [5,15,50,51], while 2 studies considered other sites in addition to oral cavity [38,47]. Furthermore, 19 studies examined the prognostic significance of machine learning applications, 21 studies evaluated the diagnostic significance of machine learning applications, and one study evaluated both (Table 1). Most studies on the application of machine learning techniques in oral cancer were published recently in 2018 and 2019 (N = 24). Over 90% of the data used in the included studies were retrospective in nature. With regards to the origin of relevant articles, 65.8% of the studies were carried out entirely in Asia, 9.6% in Europe, 7.3% in America, and 17.3% of the studies were collaborative efforts from different regions. Furthermore, a total of 4 (9.8%) of the studies used autofluorescence spectral data analysis in addition to the machine learning techniques [38,40,41,52]. Additionally, 18 (43.9%) studies used clinicopathologic or imaging data [5,15,17–21,24,25,27,28,37,45,48,49,57–59]. Also, 2 (4.9%) studies used either clinicopathologic and image [29,56], or clinicopathologic and genomic [43,44], or genomic data only [46,47], or Raman spectral data [50,51]. A single study (2.4%) combined clinical, imaging and genomic data [23]. Similarly, one study (2.4%) used clinical and genomic data [42], while 9 (21.9%) studies used other types of data (combination of risk habits, personal details, and dental attendance, or histopathologic, saliva samples, demographics and histopathologic, pathologic, lesion conditions and histological grade, clinicopathologic and socio-demographic, histologic and brush cytologic parameters, demographics-histopathologic and immunohistochemical).

Most of the included studies considered artificial neural networks (N =12, 29.3%) or support vector machines (N = 14, 34.1%) in their analyses. These two popular algorithms were

followed closely by deep convolutional neural networks (N = 11, 26.8%) [17,19,20,46,48,50–52,57–59]. There was also an increase in the application of deep neural network from the year 2017 onwards. In total, 24 (80%) of the studies had the number of cases less than 500. Similarly, most of the cases used for the analysis were extracted from hospital health records (N = 27, 65.8%). Several metrics were reported in these studies to report the performance of these machine learning algorithms. Of the included studies, 13 (31.7%) reported accuracy as their performance metrics [21–23,28,30,37,43,44,48,49,54,59,60]. Also, 13 (31.7%) used sensitivity, specificity and accuracy [5,15,17,18,26,39,42,45,46,50,51,57,58] while 8 (19.5%) studies employed only sensitivity and specificity [16,20,27,38,40,41,52,55] . Four (7.3%) studies reported only specificity and accuracy [24,25,53,56]. A single study (2.4%) considered sensitivity, specificity, accuracy and area under receiving operating characteristic curve (AUC) [19], while 2 (4.9%) studies used only AUC or its mean (MAUC) [29,47].

A total of 30 studies (73.2%) used a shallow machine learning approach while 11(26.8%) employed a deep machine learning approach. Reported specificity in the reported studies ranged from 0.57 to 1.00 [25,27,41] and sensitivity varied between 0.70 and 1 [16, 27]. Similarly, accuracy ranged from 63.4% to 100%. Notably, only 4 (9.8%) of the included studies reported less than 75% performance accuracy of the machine learning model [18,25,30,45]. Some of the concerns were the black-box concern (inability to interpret how the trained machine learning models make the diagnosis or predictions of the patients on a case-by-case basis) [25,61], result and model interpretability (what aspect of the data or the input features led to the prediction) [25,62,63], the amount and quality of the data used in the training [25,30], super-human analogy (assumption that the diagnosis or prognosis from the machine learning algorithm is close to perfect or better than the performance of the clinicians) [62], generalizability of the model (the predictive model can be used outside the data on which it was trained initially) [5,15,25], job-competitor (concerns that the adoption of machine learning

model would replace the pathologists), commercial interests (integration of machine learning-based model may actually reduce the revenue of the health systems and consequently of the clinicians) [25], and ethical issues (protecting the privacy of the patients information and defining who will be responsible if the model fails) [25,30].

3.4. Quality assessment of the studies included in the review

The quality of the studies included in this study was scaled from satisfactory to excellent. Most of the studies were generally good (Table 3). Although some of the studies did not properly follow the guidelines provided by Luo et al. (Table 2).

4.0 Discussion

The number of studies that focus on the application of machine learning in oral cancer has increased in recent years. In this systematic review, we examined for the first time the studies published on the application of machine learning in oral cancer management. The evaluated studies considered the use of machine learning to analyze clinicopathologic data, genomic data, combination of clinicopathologic and genomic data, image data, and autofluorescence spectral data. These approaches generated models to assist in clinical decision making [64].

Interestingly, the performance metrics reported in the included studies suggest high performance. Thus, the application of machine learning for oral cancer, as well as in other fields of medicine is not merely science fiction, but is becoming a reality [65]. This finding was corroborated by another study that examined machine learning and its potential applications to genomic studies of the head and neck [66]. Of note, sensitivity, specificity and accuracy have been the widely reported performance metrics. This is because accuracy simply considers correct predictions over all the predictions made by the algorithm. Similarly, specificity measures the proportion of patients that did not have oral cancer and were predicted

by the model as non-oral cancer while sensitivity (recall) measures what proportion of patients actually had oral cancer and were identified by the algorithm as having oral cancer.

Using machine learning techniques, a web-based tool has been developed to predict locoregional recurrence [5]. Similarly, machine learning technique was used to automate the diagnosis of oral cancer [49]. Many prognostic factors have been combined together via machine learning techniques for outcome predictions [15,23–30,43,58]. Also, the approach has demonstrated significant accuracy in discriminating between patients with or without oral cancer [16–19,21,22,38,41,47,52,57,59]. In other contexts to enhance effective management of oral cancer, machine learning techniques were used for early-stage detection of precancerous and cancerous lesions [20,40,46,55,60].

Despite the benefits of ensemble machine learning algorithms, support vector machine (SVM) was the most widely used machine learning algorithm for oral cancer diagnosis/prognosis as shown in this systematic review. This was also noted in a study that examined machine learning and its application to genomic data of head and neck cancer [66]. In another study, the support vector machine was concluded to be the most favorable algorithm for predicting survival rate of oral cancer [45]. The support vector machine is frequently used because it is an empirical risk minimizer algorithm. Additionally, it avoids the danger of being trapped in local minima [67]. Thus, it is usually not prone to overfitting, thereby making it capable of producing a good model that can properly capture the complex relationships between the input and output parameters. Of note, the first study that examined the use of artificial intelligence to identify patients at high risks of oral cancer used artificial neural network (ANN) [16]. Consequently, the neural network was also one of the most widely used algorithms. Success recorded from the use of neural network led to its' modification to contain multiple hidden layers. Hence, the name deep neural networks. Deep neural networks are well-positioned to solve most complex problems such as image analysis [68,69]. The application of

deep learning technologies to oral cancer diagnosis and prognosis has increased in recent years [19,20,46,48,51,52,57–59].

All the studies included in this systematic review emphasized that machine learning techniques offer an increased precision approach to clinicians by making informed decisions. This further enhances patient-specific treatments and effective management of hospital resources in a timely, efficient and dynamic manner [5,15–17,20,23,25,30,38,70,71]. Despite these potential benefits, the application of machine learning for medical diagnosis and prognosis has made few contributions to actual medical practice or patient care (Figure 2). Several issues are particularly significant from the clinical and ethical viewpoints.

The first and most frequent issue is the black-box concern [25,61,72] (Figure 3). It comes in from two distinct yet interacting perspectives, namely the result and model interpretability concerns [62]. Result interpretability concern entails an inability of the clinicians to explain which aspect of the dataset used in the training led to the predicted result in a particular case. Similarly, model interpretability reflects the clinicians' ability to understand how the algorithm developed the model [25,62]. As the trend in machine learning techniques moves from direct algorithms, such as support vector machine, to ensemble algorithms, and to deep learning, the black-box concern becomes more pronounced. To address this concern, it is pertinent for the machine learning techniques and the corresponding model to be explainable ("explainable model") and transparent [25,30,61,63] (Figure 4). Clinicians should be able to understand, to trust, to explain and to effectively manage the emerging generation of models to be used for clinical decision making. Several terms have been used to describe this concept. These include explainable AI, transparent ML, interpretable ML, and trustworthy AI [73–75].

The second concerns is the misconceptions of the scope of machine learning in medical diagnosis. The notion that machine learning models are super-human or close to perfect is

erroneous and misleading. This has led to the fear and predictions that these models in the nearest future could replace the need for professional experience-based consideration in diagnostics and prognostication [76]. The experience of the machine learning experts and the quality of the data used in machine learning analyses play a central role in producing a good model. Therefore, it is necessary that the quality of data used for model training should be the best possible and well-structured to produce a high-quality model [25,30,77].

The third concern relates to the limited amount of data used in the machine learning analyses [5,17,19,23,28,38,43,44,46,55]. Therefore, there is concern for generalizability concern of the developed machine learning model. Performance of the model to be applied for external cases outside the data for which the model was trained, is a subject to be highlighted [5,15,25,29,38]. Thus, for the machine learning model to create sustainable benefits in medical diagnosis, the data infrastructure of healthcare organizations' needs to be improved and the model produced should be externally validated to avoid biases and to enhance generalizability of the model. In the quest to improve the healthcare organizations' data infrastructure, also privacy of patient information and ethical use of the data should also be considered [25,30]. Of note, a generalized model does not mean a super-human model [62], which is a concern amongst certain clinicians. Rather, it means that the inherent bias in the dataset has been accounted for in the machine learning process. Therefore, it is important to consider machine learning models as clinical decision support to alleviate the concern for reduction in revenue for healthcare organizations or rendering the clinicians less important [25].

In conclusion, our systematic review reveals the potential of machine learning models in the management of oral cancer. More importantly, resolving the issues related to the concerns highlighted in this systematic review will ensure a faster implementation of this approach in clinical practice. This would further enhance an informed clinical decision-making and offer a better diagnosis, treatment and prognostication of oral cancer.

Authors Contribution

Study concepts and study design: Alabi RO, Elmusrati M, Almangush A, Leivo I. **Studies extraction:** Alabi RO, Omar Y (would be acknowledge in acknowledgment). **Acquisition and quality control of included studies:** Alabi RO, Almangush A. **Data analysis and interpretation:** Alabi RO, Elmusrati M, Almangush A, Mäkitie AA, Pirinen M, Leivo I. **Manuscript preparation:** Alabi RO, Almangush A, Mäkitie AA, Pirinen M. **Manuscript review:** Mäkitie AA, Leivo I, Elmusrati M, Pirinen M. **Manuscript editing:** Almangush, Alabi RO. All authors approved the final manuscript for submission.

Summary points

What was already known on the topic:

- There are published studies on the application of machine learning techniques to analyse oral tongue squamous cell carcinoma (OTSCC).
- The machine model used in actual clinical practice is limited due to certain limitations and concerns.

What knowledge this study adds:

- To the best of our knowledge, this is the first study that systematically review the published studies that examined the application of machine learning techniques to analyse tongue squamous cell carcinoma (OTSCC).
- It examines the concerns and limitations to the actual implementation of machine learning-based models in clinical settings. This study also offers possible solutions to these concerns.
- Support vector machine and artificial neural network are the most widely used algorithms for oral cancer prognostication.

- Addressing these limitations as suggested in this study may ensure that the models are useful for effective oral cancer management.

Acknowledgment

The School of Technology and Innovations, University of Vaasa Scholarship Fund. The Helsinki University Hospital Research Fund.

Figure Legend

Figure 1. The flow diagram highlighting the search strategy and the search results**.**

Figure 2. Machine learning training scheme showing the concern to actual implementation.

Figure 3. The black-box concern of the machine learning models in oral cancer management

Figure 4. An explainable and trustworthy machine learning model.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2020, CA. Cancer J. Clin. 70 (2020) 7–30. https://doi.org/10.3322/caac.21590.
[2] S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges, Cancer Lett. 471 (2020) 61–71. https://doi.org/10.1016/j.canlet.2019.12.007.
[3] L. Zhu, W. Luo, M. Su, H. Wei, J. Wei, X. Zhang, C. Zou, Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients, Biomed. Rep. 1 (2013) 757–760. https://doi.org/10.3892/br.2013.140.
[4] J. Faradmal, A.R. Soltanian, G. Roshanaei, R. Khodabakhshi, A. Kasaeian, Comparison of the Performance of Log-logistic Regression and Artificial Neural Networks for Predicting Breast Cancer Relapse, Asian Pac. J. Cancer Prev. 15 (2014) 5883–5888. https://doi.org/10.7314/APJCP.2014.15.14.5883.
[5] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R.D. Coletta, A.A. Mäkitie, T. Salo, I. Leivo, A. Almangush, Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool, Virchows Arch. 475 (2019) 489–497. https://doi.org/10.1007/s00428-019-02642-5.
[6] C.-W. Chien, Y.-C. Lee, T. Ma, T.-S. Lee, Y.-C. Lin, W. Wang, W.-J. Lee, The application of artificial neural networks and decision tree model in predicting post-operative complication for gastric cancer patients, Hepatogastroenterology. 55 (2008) 1140–1145.
[7] M.R. Gohari, A. Biglarian, E. Bakhshi, M.A. Pourhoseingholi, Use of an artificial neural network to determine prognostic factors in colorectal cancer patients, Asian Pac. J. Cancer Prev. APJCP. 12 (2011) 1469–1472.
[8] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.
[9] H.M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decis. Support Syst. 74 (2015) 150–161. https://doi.org/10.1016/j.dss.2015.04.003.
[10] D. Chen, K. Xing, D. Henson, L. Sheng, A.M. Schwartz, X. Cheng, Developing Prognostic Systems of Cancer Patients by Ensemble Clustering, J. Biomed. Biotechnol. 2009 (2009) 1–7. https://doi.org/10.1155/2009/632786.
[11] C. Denkert, G. von Minckwitz, S. Darb-Esfahani, B. Lederer, B.I. Heppner, K.E. Weber, J. Budczies, J. Huober, F. Klauschen, J. Furlanetto, W.D. Schmitt, J.-U. Blohmer, T. Karn, B.M. Pfitzner, S. Kümmel, K. Engels, A. Schneeweiss, A. Hartmann, A. Noske, P.A. Fasching, C. Jackisch, M. van Mackelenbergh, P. Sinn, C. Schem, C. Hanusch, M. Untch, S. Loibl, Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy, Lancet Oncol. 19 (2018) 40–50. https://doi.org/10.1016/S1470-2045(17)30904-X.
[12] Y. Mintz, R. Brodie, Introduction to artificial intelligence in medicine, Minim. Invasive Ther. Allied Technol. 28 (2019) 73–81. https://doi.org/10.1080/13645706.2019.1575882.
[13] Z. Qian, Y. Li, Y. Wang, L. Li, R. Li, K. Wang, S. Li, K. Tang, C. Zhang, X. Fan, B. Chen, W. Li, Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers, Cancer Lett. 451 (2019) 128–135. https://doi.org/10.1016/j.canlet.2019.02.054.

[14] A. Tan, H. Huang, P. Zhang, S. Li, Network-based cancer precision medicine: A new emerging paradigm, Cancer Lett. 458 (2019) 39–45. https://doi.org/10.1016/j.canlet.2019.05.015.
[15] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R.D. Coletta, A.A. Mäkitie, T. Salo, A. Almangush, I. Leivo, Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer, Int. J. Med. Inf. (2019) 104068. https://doi.org/10.1016/j.ijmedinf.2019.104068.
[16] P.M. Speight, A.E. Elliott, J.A. Jullien, M.C. Downer, J.M. Zakzrewska, The use of artificial intelligence to identify people at risk of oral cancer and precancer, Br. Dent. J. 179 (1995) 382–387. https://doi.org/10.1038/sj.bdj.4808932.
[17] N. Sharma, H. Om, Usage of Probabilistic and General Regression Neural Network for Early Detection and Prevention of Oral Cancer, Sci. World J. 2015 (2015) 1–11. https://doi.org/10.1155/2015/234191.
[18] N. Sharma, H. Om, GMDH polynomial and RBF neural network for oral cancer classification, Netw. Model. Anal. Health Inform. Bioinforma. 4 (2015). https://doi.org/10.1007/s13721-015-0085-2.
[19] M. Aubreville, C. Knipfer, N. Oetter, C. Jaremenko, E. Rodner, J. Denzler, C. Bohr, H. Neumann, F. Stelzle, A. Maier, Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning, Sci. Rep. 7 (2017). https://doi.org/10.1038/s41598-017-12320-8.
[20] R.D. Uthoff, B. Song, S. Sunny, S. Patrick, A. Suresh, T. Kolur, G. Keerthi, O. Spires, A. Anbarani, P. Wilder-Smith, M.A. Kuriakose, P. Birur, R. Liang, Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities, PLOS ONE. 13 (2018) e0207493. https://doi.org/10.1371/journal.pone.0207493.
[21] M. Al-Ma'aitah, A.A. AlZubi, Enhanced Computational Model for Gravitational Search Optimized Echo State Neural Networks Based Oral Cancer Detection, J. Med. Syst. 42 (2018). https://doi.org/10.1007/s10916-018-1052-0.
[22] K. Lalithamani, A. Punitha, Detection of oral cancer using deep neural based adaptive fuzzy system in data mining techniques., Int. J. Recent Technol. Eng. 7 (2019) 397–405.
[23] K.P. Exarchos, Y. Goletsis, D.I. Fotiadis, Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence, IEEE Trans. Inf. Technol. Biomed. 16 (2012) 1127–1134. https://doi.org/10.1109/TITB.2011.2165076.
[24] C.-S. Cheng, P.-W. Shueng, C.-C. Chang, C.-W. Kuo, Adapting an Evidence-based Diagnostic Model for Predicting Recurrence Risk Factors of Oral Cancer, J. Univers. Comput. Sci. 24 (2018) 742–752.
[25] A.M. Bur, A. Holcomb, S. Goodwin, J. Woodroof, O. Karadaghy, Y. Shnayder, K. Kakarala, J. Brant, M. Shew, Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma, Oral Oncol. 92 (2019) 20–25. https://doi.org/10.1016/j.oraloncology.2019.03.011.
[26] M. Mermod, E. Jourdan, R. Gupta, M. Bongiovanni, G. Tolstonog, C. Simon, J. Clark, Y. Monnier, Development and validation of a multivariable prediction model for the identification of occult lymph node metastasis in oral squamous cell carcinoma, Head Neck. (2020). https://doi.org/10.1002/hed.26105.
[27] N. Sharma, H. Om, Data mining models for predicting oral cancer survivability, Netw. Model. Anal. Health Inform. Bioinforma. 2 (2013) 285–295. https://doi.org/10.1007/s13721-013-0045-7.

[28] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, C.-N. Lin, The Application of Data Mining Techniques to Oral Cancer Prognosis, J. Med. Syst. 39 (2015). https://doi.org/10.1007/s10916-015-0241-3.

[29] C. Lu, J.S. Lewis, W.D. Dupont, W.D. Plummer, A. Janowczyk, A. Madabhushi, An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival, Mod. Pathol. 30 (2017) 1655–1665. https://doi.org/10.1038/modpathol.2017.98.

[30] O.A. Karadaghy, M. Shew, J. New, A.M. Bur, Development and Assessment of a Machine Learning Model to Help Predict Survival Among Patients With Oral Squamous Cell Carcinoma, JAMA Otolaryngol. Neck Surg. 145 (2019) 1115. https://doi.org/10.1001/jamaoto.2019.0981.

[31] B. Zheng, S.W. Yoon, S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, Expert Syst. Appl. 41 (2014) 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044.

[32] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R.N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inf. 108 (2017) 1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013.

[33] R. Al-Bahrani, A. Agrawal, A. Choudhary, Colon cancer survival prediction using ensemble data mining on SEER data, in: 2013 IEEE Int. Conf. Big Data, IEEE, Silicon Valley, CA, USA, 2013: pp. 9–16. https://doi.org/10.1109/BigData.2013.6691752.

[34] J.P. Anderson, J.R. Parikh, D.K. Shenfeld, V. Ivanov, C. Marks, B.W. Church, J.M. Laramie, J. Mardekian, B.A. Piper, R.J. Willke, D.A. Rublee, Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records, J. Diabetes Sci. Technol. 10 (2015) 6–18. https://doi.org/10.1177/1932296815620200.

[35] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, PRISMA Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS Med. 6 (2009) e1000097. https://doi.org/10.1371/journal.pmed.1000097.

[36] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, T.B. Ho, S. Venkatesh, M. Berk, Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View, J. Med. Internet Res. 18 (2016) e323. https://doi.org/10.2196/jmir.5870.

[37] V. Romeo, R. Cuocolo, C. Ricciardi, L. Ugga, S. Cocozza, F. Verde, A. Stanzione, V. Napolitano, D. Russo, G. Improta, A. Elefante, S. Staibano, A. Brunetti, Prediction of Tumor Grade and Nodal Status in Oropharyngeal and Oral Cavity Squamous-cell Carcinoma Using a Radiomic Approach, Anticancer Res. 40 (2020) 271–280. https://doi.org/10.21873/anticanres.13949.

[38] C.-Y. Wang, T. Tsai, H.-M. Chen, C.-T. Chen, C.-P. Chiang, PLS-ANN based classification model for oral submucous fibrosis and oral carcinogenesis, Lasers Surg. Med. 32 (2003) 318–326. https://doi.org/10.1002/lsm.10153.

[39] T. Kawazu, K. Araki, S. Kanda, Application of neural networks to the prediction of lymph node metastasis in oral cancer, Int. Congr. Ser. 1230 (2001) 1295–1296. https://doi.org/10.1016/S0531-5131(01)00258-8.

[40] S.K. Majumder, N. Ghosh, P.K. Gupta, Relevance vector machine for optical diagnosis of cancer, Lasers Surg. Med. 36 (2005) 323–333. https://doi.org/10.1002/lsm.20160.

[41] G.S. Nayak, S. Kamath, K.M. Pai, A. Sarkar, S. Ray, J. Kurien, L. D'Almeida, B.R. Krishnanand, C. Santhosh, V.B. Kartha, K.K. Mahato, Principal component analysis and artificial neural network analysis of oral tissue fluorescence spectra: Classification of

normal premalignant and malignant pathological conditions, Biopolymers. 82 (2006) 152–166. https://doi.org/10.1002/bip.20473.
[42] K.-Y. Kim, I.-H. Cha, A novel algorithm for lymph node status prediction of oral cancer before surgery, Oral Oncol. 47 (2011) 1069–1073. https://doi.org/10.1016/j.oraloncology.2011.07.017.
[43] S.-W. Chang, S. Abdul-Kareem, A.F. Merican, R.B. Zain, Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods, BMC Bioinformatics. 14 (2013). https://doi.org/10.1186/1471-2105-14-170.
[44] S.-W. Chang, A. Sameem, A.M. Amir Feisal Merican, Z. Rosnah Binti, A Hybrid Prognostic Model for Oral Cancer based on Clinicopathologic and Genomic Markers, Sains Malays. 43 (2014) 567–573.
[45] N. Sharma, H. Om, Using MLP and SVM for predicting survival rate of oral cancer patients, Netw. Model. Anal. Health Inform. Bioinforma. 3 (2014). https://doi.org/10.1007/s13721-014-0058-x.
[46] W. Shams, Z. Htike, Oral cancer prediction using gene expression profilling and machine learning, Int. J. Appl. Eng. Res. 12 (2017) 4893–4898.
[47] T. Turki, Z. Wei, Boosting support vector machines for cancer discrimination tasks, Comput. Biol. Med. 101 (2018) 236–249. https://doi.org/10.1016/j.compbiomed.2018.08.006.
[48] D.K. Das, S. Bose, A.K. Maiti, B. Mitra, G. Mukherjee, P.K. Dutta, Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis, Tissue Cell. 53 (2018) 111–119. https://doi.org/10.1016/j.tice.2018.06.004.
[49] A. Nawandhar, N. Kumar, V. R, L. Yamujala, Stratified squamous epithelial biopsy image classifier using machine learning and neighborhood feature selection, Biomed. Signal Process. Control. 55 (2020) 101671. https://doi.org/10.1016/j.bspc.2019.101671.
[50] H. Yan, M. Yu, J. Xia, L. Zhu, T. Zhang, Z. Zhu, Tongue squamous cell carcinoma discrimination with Raman spectroscopy and convolutional neural networks, Vib. Spectrosc. 103 (2019) 102938. https://doi.org/10.1016/j.vibspec.2019.102938.
[51] M. Yu, H. Yan, J. Xia, L. Zhu, T. Zhang, Z. Zhu, X. Lou, G. Sun, M. Dong, Deep convolutional neural networks for tongue squamous cell carcinoma classification using Raman spectroscopy, Photodiagnosis Photodyn. Ther. 26 (2019) 430–435. https://doi.org/10.1016/j.pdpdt.2019.05.008.
[52] C.-H. Chan, T.-T. Huang, C.-Y. Chen, C.-C. Lee, M.-Y. Chan, P.-C. Chung, Texture-Map-Based Branch-Collaborative Network for Oral Cancer Detection, IEEE Trans. Biomed. Circuits Syst. 13 (2019) 766–780. https://doi.org/10.1109/TBCAS.2019.2918244.
[53] A. Zlotogorski-Hurvitz, B.Z. Dekel, D. Malonek, R. Yahalom, M. Vered, FTIR-based spectrum of salivary exosomes coupled with computational-aided discriminating analysis in the diagnosis of oral cancer, J. Cancer Res. Clin. Oncol. 145 (2019) 685–694. https://doi.org/10.1007/s00432-018-02827-6.
[54] L. Lavanya, J. Chandra, Oral cancer analysis using machine learnig techniques, Int. J. Eng. Res. Technol. 12 (2019) 596–601.
[55] X. Wang, J. Yang, C. Wei, G. Zhou, L. Wu, Q. Gao, X. He, J. Shi, Y. Mei, Y. Liu, X. Shi, F. Wu, J. Luo, Y. Guo, Q. Zhou, J. Yin, T. Hu, M. Lin, Z. Liang, H. Zhou, A personalized computational model predicts cancer risk level of oral potentially malignant disorders and its web application for promotion of non-invasive screening, J. Oral Pathol. Med. (2020). https://doi.org/10.1111/jop.12983.

[56] S. Sunny, A. Baby, B.L. James, D. Balaji, A. N. V., M.H. Rana, P. Gurpur, A. Skandarajah, M. D'Ambrosio, R.D. Ramanjinappa, S.P. Mohan, N. Raghavan, U. Kandasarma, S. N., S. Raghavan, N. Hedne, F. Koch, D.A. Fletcher, S. Selvam, M. Kollegal, P.B. N., L. Ladic, A. Suresh, H.J. Pandya, M.A. Kuriakose, A smart tele-cytology point-of-care platform for oral cancer screening, PLOS ONE. 14 (2019) e0224885. https://doi.org/10.1371/journal.pone.0224885.

[57] P.R. Jeyaraj, E.R. Samuel Nadar, Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm, J. Cancer Res. Clin. Oncol. 145 (2019) 829–837. https://doi.org/10.1007/s00432-018-02834-7.

[58] Y. Ariji, M. Fukuda, Y. Kise, M. Nozawa, Y. Yanashita, H. Fujita, A. Katsumata, E. Ariji, Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. 127 (2019) 458–463. https://doi.org/10.1016/j.oooo.2018.10.002.

[59] S. Xu, Y. Liu, W. Hu, C. Zhang, C. Liu, Y. Zong, S. Chen, Y. Lu, L. Yang, E.Y.K. Ng, Y. Wang, Y. Wang, An Early Diagnosis of Oral Cancer based on Three-Dimensional Convolutional Neural Networks, IEEE Access. 7 (2019) 158603–158611. https://doi.org/10.1109/ACCESS.2019.2950286.

[60] M.P. McRae, S.S. Modak, G.W. Simmons, D.A. Trochesset, A.R. Kerr, M.H. Thornhill, S.W. Redding, N. Vigneswaran, S.K. Kang, N.J. Christodoulides, C. Murdoch, S.J. Dietl, R. Markham, J.T. McDevitt, Point-of-care oral cytology tool for the screening and assessment of potentially malignant oral lesions, Cancer Cytopathol. (2020). https://doi.org/10.1002/cncy.22236.

[61] M.K. Yu, J. Ma, J. Fisher, J.F. Kreisberg, B.J. Raphael, T. Ideker, Visible Machine Learning for Biomedicine, Cell. 173 (2018) 1562–1565. https://doi.org/10.1016/j.cell.2018.05.056.

[62] B. Heinrichs, S.B. Eickhoff, Your evidence? Machine learning algorithms for medical diagnosis and prediction, Hum. Brain Mapp. (2019). https://doi.org/10.1002/hbm.24886.

[63] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics. 26 (2010) 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

[64] P.A. Keane, E.J. Topol, With an eye to AI and autonomous diagnosis, Npj Digit. Med. 1 (2018). https://doi.org/10.1038/s41746-018-0048-y.

[65] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23 (2001) 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X.

[66] S. Patil, K. Habib Awan, G. Arakeri, C. Jayampath Seneviratne, N. Muddur, S. Malik, M. Ferrari, S. Rahimi, P.A. Brennan, Machine learning and its potential applications to the genomic study of head and neck cancer—A systematic review, J. Oral Pathol. Med. 48 (2019) 773–779. https://doi.org/10.1111/jop.12854.

[67] G. Levitin, Computational Intelligence in Reliability Engineering Evolutionary Techniques in Reliability Analysis and Optimization, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. http://link.springer.com/book/10.1007/978-3-540-37368-1 (accessed February 25, 2020).

[68] D. Michie, D.J. Spiegelhalter, C.C. Taylor, eds., Machine learning, neural and statistical classification, Ellis Horwood, New York, 1994.

[69] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[70] A. Shaban-Nejad, M. Michalowski, D.L. Buckeridge, Health intelligence: how artificial intelligence transforms population and personalized health, Npj Digit. Med. 1 (2018). https://doi.org/10.1038/s41746-018-0058-9.
[71] A.L. Fogel, J.C. Kvedar, Artificial intelligence powers digital medicine, Npj Digit. Med. 1 (2018). https://doi.org/10.1038/s41746-017-0012-2.
[72] D. Castelvecchi, Can we open the black box of AI., Nature. 538 (2016) 20–23.
[73] Bernease Herman, The Promise and Peril of Human Evaluation for Model Interpretability, (2017). http://interpretable.ml/ (accessed January 15, 2020).
[74] European Union, High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI, (2019).
[75] L. Zachary, The Doctor Just Won't Accept That!, (2017). http://interpretable.ml/ (accessed January 15, 2020).
[76] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, When Will AI Exceed Human Performance? Evidence from AI Experts, ArXiv170508807 Cs. (2018). http://arxiv.org/abs/1705.08807 (accessed January 16, 2020).
[77] N.D. Shah, E.W. Steyerberg, D.M. Kent, Big Data and Predictive Analytics: Recalibrating Expectations, JAMA. 320 (2018) 27. https://doi.org/10.1001/jama.2018.5602.

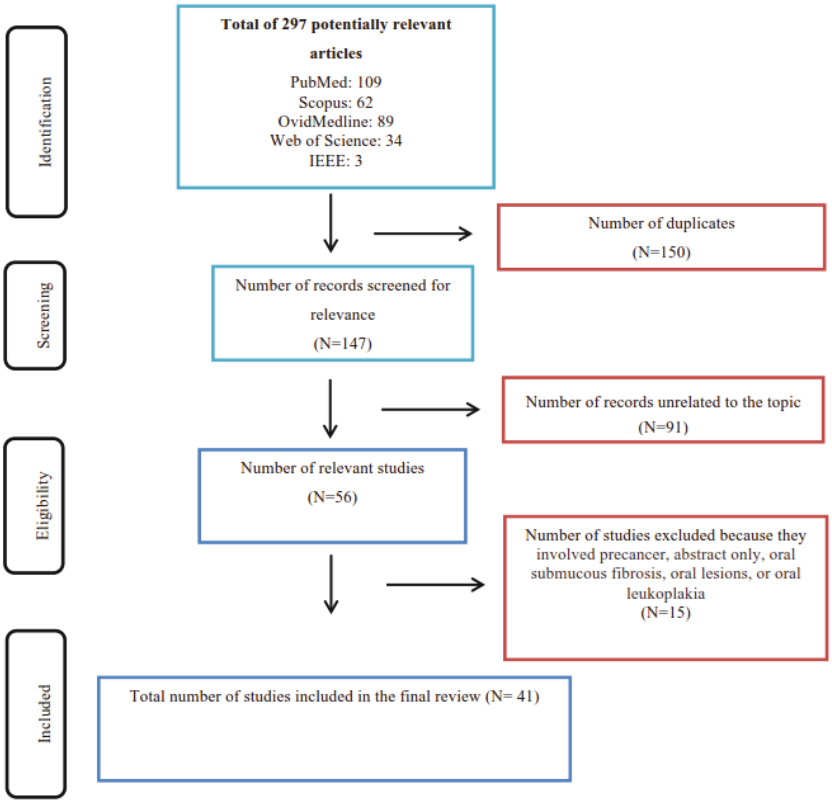

Figure 1. The flow diagram highlighting the search strategy and the search results.

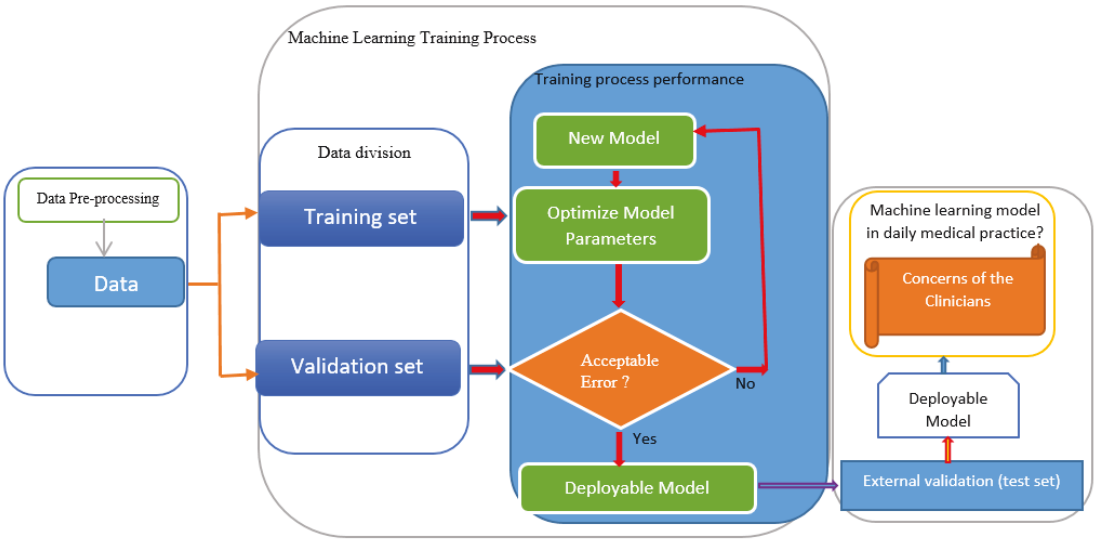

Figure 2. Machine learning training scheme showing the concern to actual implementation.

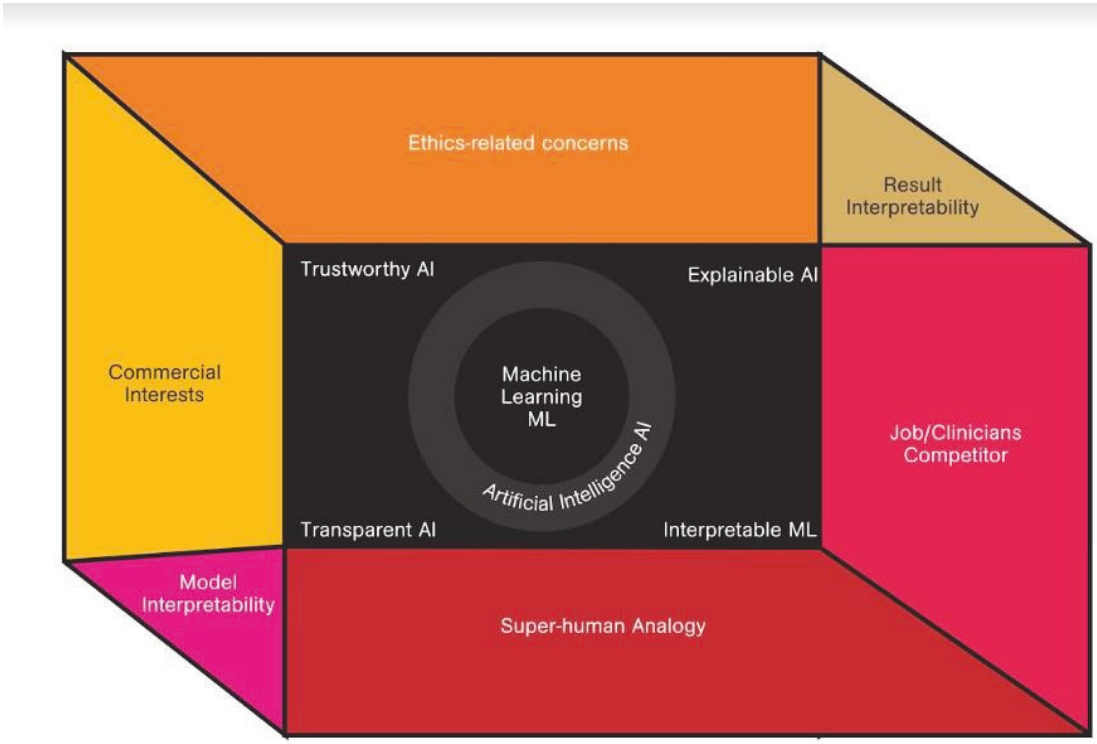

Figure 3. The black-box concern of the machine learning models in oral cancer management

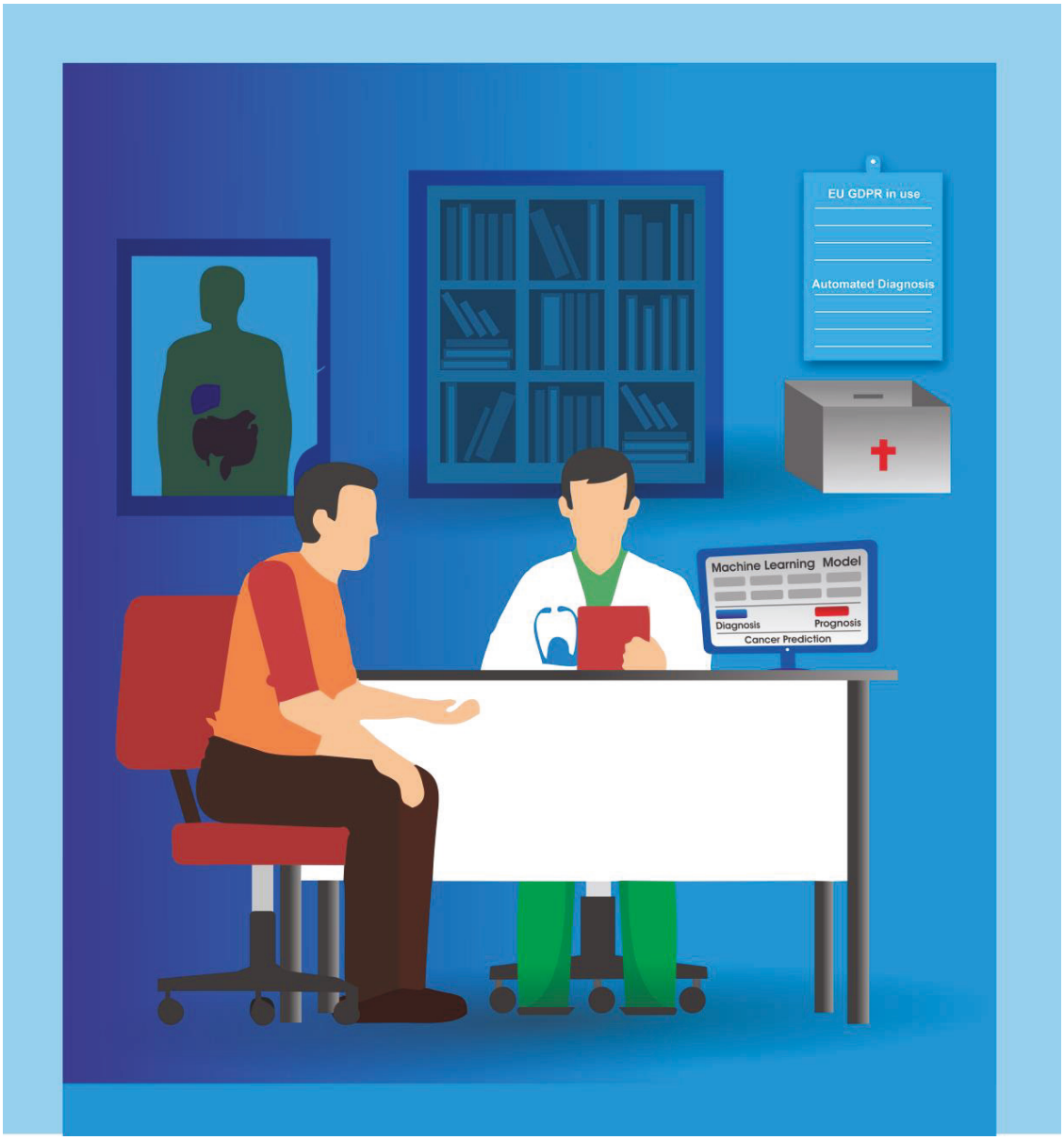

Figure 4. An explainable and trustworthy machine learning model.

Table 1. Extracts of the main findings from the included studies

Authors, year (country of authors)	Site	No of Cases	Machine Learning Methods	Use of Machine Learning in Oral cancer	Study Aim	Results	Performance metric(s)
Speight et al., 1995 (United Kingdom)	Oral cavity	2027	Neural Network	Diagnostic (data of risk habits, personal details, dental attendance).	To predict the likelihood of an individual to having a malignant or potentially malignant oral lesion.	This approach showed promising results compared with the performance of the dentist for the screening exercise.	Sensitivity: 0.80 Specificity: 0.77
Wang et al., 2003 (China)	Oral cavity*	97	Partial Least Squares and Artificial Neural Network (PLS-ANN)	Diagnostic (autofluorescence spectra data analysis).	To differentiate between premalignant and malignant tissues from benign.	The multivariate algorithm differentiated human premalignant and malignant lesions from benign lesions or normal oral mucosa.	Sensitivity: 0.81 Specificity: 0.96
Kawazu et al., 2003 (Japan)	Oral cavity	1,116	Neural Network	Diagnostic (Histopathological)	To predict lymph node metastasis in oral cancer	The prediction performance was comparable to clinical radiologists	Sensitivity: 0.80 Specificity: 0.94 Accuracy: 93.6%
Majumder et al., 2005 (India)	Oral cavity	171	Relevance Vector Machine (RVM) & Support Vector Machine (SVM)	Diagnostic (autofluorescence spectra data analysis)	To diagnose early stage oral cancer	The performance shown by the Bayesian framework of RVM was comparable to the traditional SVM.	Sensitivity: 0.91 Specificity: 0.96

Nayak et al., 2006 (India)	Oral cavity	143	Principal Component Analysis (PCA) & Artificial Neural Network (ANN)	Diagnostic (autoflourescence spectra data analysis).	To classify images into normal, premalignant, and malignant.	The performance of ANN was better than PCA.	Sensitivity: 0.96 Specificity: 1.00
Kim & Cha, 2011 (Korea)	Oral cavity	90	Principal Component Analysis (PCA)	Prognostic (Clinical and genomic)	To predict lymph node status before surgery	The model performed better when the clinical and genomic parameters were combined.	Sensitivity: 0.70 Specificity: 0.88 Accuracy: 84.0%
Exarchos et al., 2012 (Greece)	Oral cavity	41	Bayesian Networks (BN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT) & Random Forest (RF)	Prognostic (Clinical, image and genomic).	To predict oral cancer reoccurrence.	The multiparametric approach presented successfully predicted oral cancer reoccurrence.	Accuracy: 100%
Sharma and Om, 2013 (India)	Oral cavity	1024	Single Tree (ST), Decision Tree Forest (DTF), Tree Boost (TB) model	Prognostic (clinicopathologic)	To predict the survival rate in cancer patients.	The three examined algorithms showed similar results and performances.	Sensitivity: 1.00 Specificity: 1.00
Chang *et al.*, 2013 (Malaysia)	Oral cavity	31	Adaptive Neuro Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression (LR)	Prognostic (Clinicopathologic and genomic)	Oral cancer prognosis using the hybrid of feature selection and several machine learning methods. [*Continuation of previous studies*]	Prognosis is more accurate with the combination of clinicopathologic and genomic markers.	Accuracy: 93.8%
Chang et al., 2014 (Malaysia)	Oral cavity	31	ReliefF-Genetic Algorithm, Feature Selection, Adaptive Neuro Fuzzy Inference System (ANFIS	Prognostic (Clinicopathologic and genomic)	To apply the hybrid of feature selection (Relief-GA) & machine learning technique (ANFIS) in prognosis of oral cancer.	The prognoses was more accurate in group 2 (clinicopa	Accuracy: 93.8%

						thologic and genomic) than group 1 (clinicopathologic markers only)	
Sharma and Om, 2014 (India)	Oral cavity	1024	Support Vector Machine (SVM) & Multi-layer Perceptron (MLP)	Prognostic (Clinicopathologic)	To predict survivability of oral cancer patients.	The performance metrics showed by SVM outperforms the multi-layer perceptron.	Sensitivity: 0.73 Specificity: 0.73 Accuracy: 73.6%
Tseng et al., 2015 (Taiwan)	Oral cavity	673	Decision Tree (DT), Artificial Neural Network (ANN), Logistic Regression (LR), & K-means	Prognostic (Clinicopathologic)	To predict 5-year survival rate and recurrence. Clustering of patients were conducted.	Decision tree and neural network showed superior to traditional method.	Accuracy: 98.4%
Sharma and Om, 2015 (India)	Oral cavity	1025	Probabilistic and General Neural Network (PNN/GRNN), Linear Regression (LR), Decision Tree (DT), Tree Boost (TB), Multi-layer perceptron (MLP), Convolutional Neural Network (CNN)	Diagnostic (Clinicopathologic)	To detect oral cancer.	The model predicted cancer stages and survivability	Sensitivity: 0.92 Specificity: 0.79 Accuracy: 80.0%
Sharma & Om, 2015 (India)	Oral cavity	1025	Group method if data handling (GMDH) polynomial neural network & Radial basis neural network (RBNN)	Diagnostic (Clinicopathologic)	To diagnose new cases of oral cancer.	The two variant of NN showed competitive results in differentiating patients with or without oral cancer.	Sensitivity: 0.77 Specificity: 0.61 Accuracy: 67.8%
Shams & Htike, 2017 (Malaysia)	Oral cavity	86	Support Vector Machine (SVM), Deep Neural Network (DNN),	Prognostic (Gene expression data).	To predict the risks of oral cancer in oral premalignant	The DNN technique performed better	Sensitivity:0.98 Specificity: 0.94 Accuracy: 96%

			Regularized Least Squares (RLS) & Multi-layer perceptron (MLP)		lesion (OPL) patients.	than others.	
Aubreville *et al.*, 2017 (Germany)	Oral cavity	7,894	Deep learning technologies on Confocal Laser Endomicroscopy (CLE) images of oral squamous cell carcinoma (OSCC)	Diagnostic (image analysis)	Detection of oral cancer based on images.	A CNN-based image recognition was successfully applied on confocal laser endomicroscopy images of OSCC.	Sensitivity: 0.86 Specificity: 0.90 Accuracy: 88.3% AUC: 0.96
Lu *et al.*, 2017 (China & USA)	Oral cavity	115	Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Random Forest (RF)	Prognostic (Clinicopathologic + image analysis).	To predict the disease-specific survival.	The study properly associated local nuclear morphologic heterogeneity with long term outcomes.	AUC: 0.72
Uthoff *et al.*, 2018 (USA & India)	Oral cavity	170	Convolutional Neural Network (CNN)	Diagnostic (image analysis)	Early detection of precancerous and cancerous lesions	A low-cost, smartphone-based image system for oral screening was developed	Sensitivity: 0.85 Specificity: 0.88
Al-Ma'aitah & AlZubi, 2018 (Saudi Arabia)	Oral cavity	-	Gravitational Search Optimized Echo State Neural Networks (GSOESNN, Support Vector Machine (SVM), Multi-layer perceptron (MLP), & Neural Network	Diagnostic (image analysis)	Detection of oral cancer	The optimized neural network examined in this study identified oral cancer than other machine learning methods.	Accuracy: 99.2%.

Turki & Wei, 2018 (Saudi Arabia & USA)	Oral cavity*	86	Boosted Support Vector Machine (BSVM)	Prognostic (gene expression data)	Identification of oral cancer	The boosting versions of the examined algorithms outperformed the baseline algorithms.	MAUC: 0.849.
Cheng *et al.*, 2018 (Taiwan)	Oral cavity	1,429	K-Nearest Neighbor (KNN), K-shortest paths (K-STAR), Randomizable Filtered Classifier (RFC), & Random Tree (RT)	Diagnostic (Clinicopathological data)	To predict recurrence	Important risk factors for recurrence were identified. Also, KSTAR algorithm showed the best performance	Specificity: 0.75 Accuracy: 77.0%
Das *et al.*, 2018 (India)	Oral cavity	126	Deep Convolution Neural Network (DCNN)	Diagnostic (image analysis)	Automatic identification of relevant regions for OSCC diagnosis	Keratin pearls region were identified with significant accuracy.	Accuracy: 96.9%
Nawandhar *et al.*, 2019 (India)	Oral cavity	676	Decision Tree (DT), Quadratic Support Vector Machine (QSVM), Cubic SVM (Cu-SVM), Neighborhood Component Analysis (NCA), Random-Subspaces Linear Discriminant Analysis (RS-LDA) & Stratified Squamous Epithelium – Biopsy Image Classifier (SSC–BIC)	Prognostic (Image analysis)	To develop an automatic OSCC image classifier	H&E stained microscopic images were classified as either normal, well, moderately, or poorly differentiated	Accuracy: 95.6%

Yan et al., 2019 (China)	Tongue Squamous Cell Carcinoma (TSCC)	24	Convolutional Neural Networks (CNN)	Diagnostic (Raman Spectroscopy)	To discriminate the border of tongue squamous cell carcinoma from non-tumorous tissue.	The extracted features combined to produce significant accuracy for tongue squamous cell carcinoma discriminations	Sensitivity: 0.99 Specificity: 0.95 Accuracy: 97.2%
Yu et al., 2019 (China)	Oral Tongue Squamous Cell Carcinoma (OTSCC)	36	Deep Convolutional Neural Networks (DCNN), Principle Component Analysis (PCA), Support Vector Machine (SVM), & Linear Discriminant Analysis (LDA)	Diagnostic (Raman spectral data)	To discriminate OTSCC from non-tumorous tissue	DCNN showed better result than the state-of-the-art methods	Sensitivity: 0.99 Specificity: 0.94 Accuracy: 96.9%
Chan *et al.*, 2019 (Taiwan)	Oral cavity	80	Deep Convolutional Neural Networks (DCNN)	Diagnostic (auto-fluorescence data analysis)	To detect oral cancer	The feature extracted by Gabor filter provide more useful information for cancer detection	Sensitivity: 0.93 Specificity: 0.94
Bur *et al.*, 2019 (USA)	Oral cavity	782	Decision Forest (DF), Gradient Boosting (GB)	Prognostic (clinicopathologic)	Predict occult nodal metastasis	The DF and GB performed better at predicting occult nodal metastasis than DOI model.	Specificity: 0.57 Accuracy: 63.4%
Zlotogorski-Hurvitz *et al.*, 2019 (Israel)	Oral cavity	34	Principal Component Analysis – Linear Discriminant Analysis (PCA-LDA), Support Vector Machine (SVM)	Prognostic (saliva samples)	To differentiate between the spectra of oral cancer and healthy individuals.	The mid-infrared (IR) spectra of oral cancer patients was different	Specificity: 89% Accuracy: 95%

						from healthy individuals. The PCA-LDA outperformed other examined techniques.	
Alabi *et al., 2019* (Finland &Brazil)	Oral Tongue Squamous Cell Carcinoma (OTSCC)	254	Support Vector Machine (SVM), Naive Bayes (NB), Boosted Decision Tree (BDT), Decision Forest (DF), & Permutation Feature Importance (PFI)	Prognostic (clinicopathologic)	To predict locoregional recurrence	The BDT produced the highest accuracy. Also, the examined algorithms performed better than the depth of invasion model.	Sensitivity: 0.79 Specificity: 0.83 Accuracy: 81%
Lalithamani *et al.,* 2019 (India)	Oral cavity	-	Deep Neural Based Adaptive Fuzzy System (DNAFS)	Diagnostic (demographics and histopathologic)	To identify oral cancer patients	The novel classifier uses fuzzy logic and DNN for oral cancer identification and detection	Accuracy: 96.3%
Lavanya & Chandra, 2019 (India)	Oral cavity	-	Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Multi-layer perceptron (MLP), Logistic Regression (LR)	Prognostic (Pathological data)	To classify oral cancer into stages	The ML predicted different stages in oral cancer	Accuracy: 90.6%
Wang *et al.,* 2019 (China)	Oral cavity	266	Random Forest (RF)	Prognostic (personal details, smoking & drinking status, lesion conditions, & histological grade)	Predict cancer risk of oral potentially malignant disorders.	The personalized model performed better than the baseline & clinical expert	Sensitivity: 0.82 Specificity: 0.91
Alabi *et al.*, 2019 (Finland & Brazil)	Oral tongue squamous cell	311	Artificial Neural Network (ANN)	Prognostic (Clinicopathological data)	Prediction of locoregional recurrences	The accuracy of the	Sensitivity: 0.71 Specificity: 0.98 Accuracy: 88.2%

	carcinoma (OTSCC)					neural network was significantly higher.	
Karadaghy *et al.*, 2019 (USA)	Oral cavity	33,065	Decision Forest (DF)	Prognostic (Clinicopathological, social and demographic data)	Prediction of 5-year overall survival of OSCC patients	Combining clinicopathological, social and demographics produced better model than TNM-based model.	Accuracy: 71%
Sunny et al., 2019 (India, Germany & America)	Oral cavity	100	Artificial Neural Network (ANN)	Diagnostic (image) & prognostic (clinicopathologic)	To develop a risk stratification model using ANN. Also to enable tele-cytology-based point of care diagnosis (detection of OPML).	The ANN showed higher accuracy.	Specificity: 0.90 Accuracy: 86%
Jeyaraj & Samuel Nadar, 2019 (India)	Oral cavity	100	Convolution Neural Network (CNN)	Diagnostic (image analysis)	To use CNN for the detection of cancerous tumor with benign and cancerous tumor with normal tissue.	The regression-based partitioned CNN performs better than other traditional medical image classification technique examined.	Sensitivity: 0.94 Specificity: 0.91 Accuracy: 91.4 %
Ariji et al., 2019 (Japan)	Oral cavity	45	Convolution Neural Network (CNN)	Diagnostic (image analysis)	To evaluate the performance of CNN for the diagnosis of lymph node metastasis.	The CNN yielded performance that is similar to pathologists.	Sensitivity: 0.75 Specificity: 0.81 Accuracy: 78.2%.
Xu et al., 2019 (China)	Oral cavity	~ 7000	Three-Dimensional Convolutional Neural Networks (3DCNN)	Diagnostic (image analysis)	To differentiate between benign and malignant oral cancers	The 3DCNN variant gave a better performa	Accuracy: 75.4%

						nce than the 2DCNN in differenti ating between benign and malignant .	
Romeo *et al.*, 2020 (Italy)	Oral cavity	40	Naïve Bayes (NB), Bagging of NB, K-Nearest Neighbors (KNN), J48, boosting J48	Prognostic (Image analysis)	Prediction of tumor grade and nodal status in patients with OCSCC & oropharyngeal.	Most accurate subset of features to predict tumor grade and nodal status were identified .	Accuracy: 92.9%
McRae et al., 2020 (USA)	Oral cavity	999	K-Nearest Neighbors (KNN)	Diagnostic (histopathologic and brush cytologic parameters)	To detect potential malignant oral lesions (PMOL).	This approach represent a practical solution for quick PMOL assessme nt.	Accuracy: 99.3%
Mermod et al., 2020 (Switzerland & Australia)	Oral cavity	56 (112 external validation)	Random Forest (RF), linear Support Vector Machine (SVM), LASSO regularized logistic regression, C5.0 decision trees	Prognostic (demographic, histopathologic, immunohistochemic al)	To predict occult lymph node metastases (OLNM)	The examined algorithm offered a clinical managem ent strategies to identify patients that would benefit from neck dissectio n	Sensitivity: 0.8 Specificity: 0.9 Accuracy: 90%

Table 2. Quality measurement guidelines [Adapted from Luo et al., 2016] [36]

Article sections	Parameters	Explanation
Title	▪ Title (Nature of Study)	The study clearly showed that it focused on either diagnostic or prognosis model, or both.
Abstract	• Abstract (Structured summary of the study)	It contains the background, objectives, data sources, performance metrics and conclusion. The data sources and no of data is preferred but can also be optional in the abstract.
Introduction	▪ Rationale ▪ Objectives	Describes the goals of the study. It properly introduced the reader to the study. A brief introduction that reviews the current practice and prediction performance of existing models. Also, identify how the newly proposed model may benefit the clinical practices.
Methods	▪ Describe the available data/describe the setting ▪ Define the problem (diagnostic/prognostic) ▪ Data preparation ▪ Build the model	Describe the data source, size of data sample, year/duration of the available data. The nature of the data (retrospective/prospective), input and target variables definition, cost of prediction errors, performance metrics definition, and the explanation of the success criteria. Data inclusion and exclusion criteria, data processing methods, missing values and how it was handled. Finally, explain how the model was built. (Explaining the nature of data and the external validation are desirable but not mandatory)
Results	• The performance of the model using the external validation dataset	This reports the final model and its performance. It is recommended to compare the performance of the model with other known models, clinical standards or statistical methods. Reporting the confidence intervals is optional but desirable. Similarly, it is highly recommended to validate the model externally. If not possible, internal validation becomes important.
Discussion	▪ Discuss the clinical implications ▪ Discuss the limitations	Discuss the significance of the findings and possible limitations (potential pitfalls) of the study or the model to be specific. Mentioning the financial implications, that is, the amount of money that can be saved using this model is optional.
Conclusion	○ Discuss the overall usage of the model in the clinical arena.	Report the unexpected signs of the model such as collinearity, overfitting, underfitting. Most importantly, evaluates if the objective of the studies was fulfilled.

Table 3. Quality scores of the included studies based on the guidelines provided Luo et al., 2016 [36, guidelines modified]

Studies	Tittle	Abstract	Rationale	Objectives	Setting Description	Problem Definition	Data Preparation	Build Model	Report Performance	Clinical Implications	Limitations	Scores (%)
Speight et al., 1995	●	●	●	●	●	●	●	●	●	●	●	90.0%
Wang et al., 2003	●	●	●	●	●	●	●	●	●	●	●	90.0%
Majumder et al., 2005	●	●	●	●	●	●	●	●	●	●	●	90.0%
Nayak et al., 2006	●	●	●	●	●	●	●	●	●	●	●	100.0%
Exarchos et al., 2012	●	●	●	●	●	●	●	●	●	●	●	90.0%
Sharma & Ohm, 2013	●	●	●	●	●	●	●	●	●	●	●	81.8%
Chang et al., 2013	●	●	●	●	●	●	●	●	●	●	●	81.8%
Chang et al., 2014	●	●	●	●	●	●	●	●	●	●	●	90.0%
Tseng et al., 2015	●	●	●	●	●	●	●	●	●	●	●	100.0%
Sharma & Ohm, 2015	●	●	●	●	●	●	●	●	●	●	●	100.0%
Sharma & Om, 2015	●	●	●	●	●	●	●	●	●	●	●	81.8%
Shams & Htike, 2017	●	●	●	●	●	●	●	●	●	●	●	81.8%
Aubreville et al., 2017	●	●	●	●	●	●	●	●	●	●	●	100.0%
Lu et al., 2017	●	●	●	●	●	●	●	●	●	●	●	90.0%
Uthoff et al., 2018	●	●	●	●	●	●	●	●	●	●	●	81.8%
Al-Ma'aitah & Alzubi, 2018	●	●	●	●	●	●	●	●	●	●	●	81.8%
Turki & Wei, 2018	●	●	●	●	●	●	●	●	●	●	●	81.8%
Cheng et al., 2018	●	●	●	●	●	●	●	●	●	●	●	90.0%
Das et al., 2018	●	●	●	●	●	●	●	●	●	●	●	90.0%
Nawandhar et al., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Yu et al., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Chan et al., 2019	●	●	●	●	●	●	●	●	●	●	●	81.8%
Bur et al., 2019	●	●	●	●	●	●	●	●	●	●	●	100.0%
Zlotogorski-Hurvitz et al., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Alabi et al., 2019	●	●	●	●	●	●	●	●	●	●	●	100.0%
Lalithamani et al., 2019	●	●	●	●	●	●	●	●	●	●	●	100.0%
Lavanya & Chandra, 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Wang et al., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Alabi et al., 2019	●	●	●	●	●	●	●	●	●	●	●	100.0%
Karadaghy et al., 2019	●	●	●	●	●	●	●	●	●	●	●	100.0%
Sunny et al., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Jeyaraj & Samuel Nadar., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Ariji et al., 2019	●	●	●	●	●	●	●	●	●	●	●	100.0%
Xu et al., 2019	●	●	●	●	●	●	●	●	●	●	●	90.0%
Romeo et al., 2020	●	●	●	●	●	●	●	●	●	●	●	100.0%
McRae et al., 2020	●	●	●	●	●	●	●	●	●	●	●	90.0%
Mermod et al., 2020	●	●	●	●	●	●	●	●	●	●	●	100.0%