Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future-A systematic review

Alabi, Rasheed Omobolaji

2021-05

Alabi , R O , Youssef , O , Pirinen , M , Elmusrati , M , Mäkitie , A , Leivo , I & Almangush , A 2021 , ' Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future-A systematic review ' , Artificial Intelligence in Medicine , vol. 115 , 102060 . https://doi.org/10.1016/j.artmed.2021.102060

http://hdl.handle.net/10138/342581
https://doi.org/10.1016/j.artmed.2021.102060

Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future - A systematic review

Rasheed Omobolaji Alabi M.Sc [a], Omar Youssef MD, PhD [b,c], 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.**

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

Background: Oral cancer can show heterogenous patterns of behavior. For proper and effective management of oral cancer, early diagnosis and accurate prediction of 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 systematic review of diagnostic and prognostic application of machine learning in oral squamous cell carcinoma (OSCC) and also highlights some of the limitations and concerns of clinicians towards the implementation of machine learning-based models for daily clinical practice. **Data sources:** We searched OvidMedline, PubMed, Scopus, Web of Science, and Institute of Electrical and Electronics Engineers (IEEE) databases from inception until February 2020 for articles that used machine learning for diagnostic or prognostic purposes of OSCC. **Eligibility criteria:** Only original studies that examined the application of machine learning models for prognostic and/or diagnostic purposes were considered. **Data extraction:** Independent extraction of articles was done by two researchers (A.R. & O.Y) using predefine study selection criteria. We used the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) in the searching and screening processes. We also used Prediction model Risk of Bias Assessment Tool (PROBAST) for assessing the risk of bias (ROB) and quality of included studies. **Results**: A total of 41 studies were published to have used machine learning to aid in the diagnosis/or prognosis of OSCC. The majority of these studies used the 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 can be grouped as either the challenges inherent to the science of machine learning or relating to the clinical implementations. **Conclusion:** Machine learning models have been reported to show promising performances for diagnostic and prognostic analyses in studies of oral cancer. These models should be developed to further enhance explainability, interpretability, and externally validated for generalizability in order to be safely integrated into daily clinical practices. Also, regulatory frameworks for the adoption of these models in clinical practices are necessary.

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 is 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.

However, this 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 [3]. 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 a machine learning approach is more accurate in prognostication than the traditional statistical analyses [3–7].

The 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 the 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 clinicians in making informed decisions, reducing diagnostics errors, improving, and promoting the overall patient health.

This study, therefore, aims to systematically review the published studies that applied machine learning to aid in the diagnosis and prediction of the prognosis of oral squamous cell carcinoma (OSCC). This gives an overview of the current status of machine learning-based models in OSCC. Additionally, this study examines the concerns towards the actual implementation of machine learning-based models in clinical settings of OSCC. These concerns were considered from the limitations, shortcomings, and clinicians' concerns in the published studies regarding the application of machine learning for OSCC prognosis. In addition, the required approaches needed to translate these potentially transformative models into daily clinical practice were explored. 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')]. The potentially relevant articles were exported to RefWorks reference manager software and duplicate were removed.

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. Furthermore, the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) was followed in the searching and screening processes (Figure 1) [35]. We used the corresponding PRISMA checklist (Supplementary 1) to ensure that essential aspects of a systematic review were considered.

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, and expert views were all excluded. Similarly, articles in languages other than English were excluded. Studies that examined machine learning applications for normal oral mucosa, oral lesions (without cancer), or dental caries, oral mucosa, DNA and RNA microarray genes, proteomics, fluorescence spectroscopy, and genetic programming 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 (R.A., & 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.

2.4. Parameters extracted from the included studies. The extracted information from each study included author (s) name, year of publication, country, site of mouth cancer, number of study participants, machine learning algorithms examined in the study, the 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. Overall, the reported accuracy in each of the included studies serves as the technical performance (summary measure) of the developed machine learning model described in that study. 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.

2.5. Quality assessment of the included studies. We used the Prediction model Risk of Bias Assessment Tool (PROBAST) for evaluating and assessing the risk of bias (ROB) and quality of included studies (Table 2). To further ensure that the included studies meet the required standard, 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 3. 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 4.

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 [15–30, 37–60]. The main 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. In addition, some of the published 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 [16–30,37,40–46,48,49,52–61], 4 studies focused on oral tongue squamous cell carcinoma [3,15,50,51], while 2 studies considered other sites in addition to oral cavity [38,47]. Furthermore, 19 studies examined the use of machine learning applications in the prognostic analysis, 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). 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 [3,15,17–21,24,25,27,28,37,45,48,49,57–59]. 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 (e.g. combination of risk habits, or histopathologic, demographics, clinicopathologic, 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 [3,15,17,18,26,39,42,45,46,50,51,57,58] 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]. The concerns to the successful deployment of artificial intelligent-based model into daily clinical practice can be broadly divided into those that are inherent to the science of machine learning (sometimes generalized as the black box concern) and clinician concerns relating to the implementations of machine learning models in healthcare.

The concerns that are intrinsic to the science of machine learning include 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) [25,62], result and model interpretability (what aspect of the data or the input features led to the prediction) [25,63,64], the amount and quality of the data used in the training [25,30], unintended fitting of cofounders as input variables [25,30], and generalizability of the model (the predictive model can be used outside the data on which it was trained initially) [3,15,25].

The clinical concerns include the explainability of the machine learning models. That is, the models should be convenient and easy to use in such a way that the clinicians could explain the performance metrics and how the model arrived at the prognostication [25,63,64]. Other concerns of the clinicians include how will these potentially transformative technologies change the patient-clinicians' relationships [25]. Additionally, super-human analogy (the assumption that the diagnosis or prognosis from the machine learning algorithm is close to perfect or better than the performance of the clinicians) [63] and job-competitor (concerns that the adoption of machine learning model would replace the pathologists) are also some of the challenges.

3.4. Quality assessment of the studies included in the review

According to the PROBAST assessment, most (90.2%) of the included studies showed an overall low risk of bias while 92.7% of the included studies also exhibited low concern regarding applicability (Table 2). In another measure of the quality of the studies included in this study which was scaled from satisfactory to excellent, most of the studies were generally good (Table 4). Although some of the studies did not properly follow the guidelines provided by Luo et al. (Table 3).

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 [65].

Interestingly, the performance metrics reported in the included studies suggest high performance of machine learning models in oral cancer. 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 [66]. This finding was corroborated by another study that examined machine learning and its potential applications to genomic studies of the head and neck [67]. 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 [3]. Similarly, the 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 the 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, the 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 [67]. In another study, the support vector machine was concluded to be the most favorable algorithm for predicting the survival rate of oral cancer [45]. The support vector machine is frequently used because it is an empirical risk minimizer algorithm [68]. 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 an artificial neural network (ANN) [16]. Consequently, the neural network was also one of the most widely used algorithms. The success recorded from the use of neural networks 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 [69,70]. 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 [3,15–17,20,23,25,30,38,71,72]. 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 science of machine learning (sometimes

generalized as the black box concern) and clinician concerns relating to the implementations of machine learning models in healthcare viewpoints.

The first and most frequent issue for the clinical implementation is the black-box concern [25,62,73] (Figure 3). It comes in from two distinct yet interacting perspectives, namely the result and model interpretability concerns [63]. 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 [25,63]. 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 [25,30,62,64].

Clinicians should be able to understand and 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 [74–76]. Holzinger et al proposed a system causability scale (SCS) to measure the quality of explanations offered by the machine learning models [77,78]. Notably, recent research emphasized the need for explainability and re-traceability on demands for models that can significantly affect users [79]. Similarly, these models should be reported using best practice reporting guidelines such as the Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [80,81] or its extension that is peculiar to machine learning (TRIPOD-ML) [82].

Intertwined with the issue of result and model interpretability concerns is the fact that most of these models were developed using retrospective data (historically labelled data).

However, the true performance of the machine learning models may be achieved with prospective data. Therefore, for the future, it is important that machine learning models are developed or validated with prospective data. Also, clinicians are expected to be aware of the performance of these models in metrics that gives better comprehension to the clinicians. The decision curve analysis, which seeks to present the net benefit of these models, may offer the clinicians the picture of the actual performance of these models in a relatable manner [83]. Furthermore, randomized controlled trials may be used to evaluate the true performance of these models as the higher technical accuracy reported for these models does not necessarily correspond to better patient prognostication [84].

Many of the current challenges in translating machine learning models for use in daily clinical practice 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. However, 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 [25,30,85].

Furthermore, a fundamental component to achieving safe and effective deployment of machine learning models in clinical practices is for the models to achieve reliable generalizability. That is, the 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 [3,15,25,29,38]. Thus, for the machine learning model to create sustainable benefits in medical diagnosis, the data infrastructure of healthcare organizations needs to be improved so that machine learning models are developed using heterogeneous and aggregated data from multiple sources (big data) [86]. In addition, the model produced should be externally validated to avoid biases and to enhance the generalizability of the model [3,15,25,87,88]. This will ensure that relevant

variations of the model in real clinical settings are adequately captured [88]. Of note, the practice to externally validate the developed model is rare as few of the included studies in this systematic review performed external validation [3,15,25,52].

Considering the concerns inherent to the science of machine learning, the limited amount of data used in the machine learning analyses represents a major concern [3,17,19,23,28,38,43,44,46,55]. Of note, data represents an essential backbone for any machine learning model. Therefore, the nature of the data in terms of quality and quantity plays a significant role in the performance of the model [25,30,85]. The concern of the limited amount of data can be addressed by the aggregation of data (data fusion technique). Unfortunately, such data is not readily available for machine learning analysis. The data is usually stored in different locations and formats ranging from electronic health records (EHR), pathology systems, medical imaging archives, insurance data, and electronic prescribing tools [89]. In fact, these medical data are characterized as being messy, voluminous, and complex [90]. This makes it challenging for data fusion and aggregation [89]. Therefore, it is advisable to pre-process (carefully labelled and curated) the data prior to the attempt to aggregate the data [90]. The Fast Healthcare Interoperability Resources (FHIR) has been suggested to offer an approach for better unification of data formats [91].

The limited amount of data used for the training of machine learning models can also give rise to algorithmic bias. This concern is closely related to the generalizability of the developed model [89]. Retrospective data that are usually used to train machine learning models have been reported to have significant biases towards under-represented groups that have been affected by factors such as gender, race, and socioeconomic background [92,93]. Examples of biased algorithms have been reported in the mortality prediction model [94] and the dermoscopic melanoma recognition model [95,96]. The problem of biases in algorithms can be addressed by improving the nature (quality and quantity) of the training data using big

data [87,88]. Also, the performance of the models should be evaluated within population subgroups such as gender, age, ethnicity, socioeconomic background, location, and other under-represented factors in the data.

One of the most widely used sources of data for machine learning analysis is the hospital database such as the electronic health record (EHR). Unfortunately, this hospital environment is characterized by changes in clinical and operational practices over time, thereby, causing a shift in the patient populations and characteristics [97]. Therefore, earlier developed models should be retrained periodically [98]. This can be achieved by simple recalibration or full retraining of the model [98]. This approach offers an important step to addressing biases and further enhances the generalizability of the model [90].

Therefore, it is important to aggregate the available dataset siloed at different locations mentioned above. These aggregated data can be preprocessed (cleaned, re-organized, and stored) to form big data. In oncology, one of the insightful ways to achieve big data is to ensure that the size of the data is big enough (volume) with multiple parameters such as socio-economic, risk factors, clinical, radiology, pathological, treatment data, and complications [99]. Additionally, the data should be preprocessed and accessed at a relatively fast speed (velocity). Furthermore, the data should contain varieties (variety) of data types such as discrete, continuous, binary, descriptive, structured, and unstructured data. Also, it is important that data is highly variable –parameters contained in the data are well defined and include minimum parameters that can make the data useful. It is important that the data being collected is valuable [99]. All these are coined under a general term of 5 Vs of big data (volume, velocity, variety, variability, and value) [99].

These big data can be used to develop machine learning models that offer insightful prognostication which could assist clinicians in making informed decisions [90]. Also, with big data, complex patterns can be derived from population-level rather than from the small

number of samples [90]. Thus, poised to address algorithmic bias and generalizability of the resulting model [90]. With the increasing number of patient registries and health databases, phenotypic and genotypic data are now linked to research data to have robust big data for machine learning analysis. Thereby, producing a model that is capable of prognostic analytics [90]. If these models are successfully validated and implemented, they could be of significant assistance for clinicians in making informed decisions.

Connected to the concerns relating to the science of machine learning is that cofounders may be unintentionally fitted as part of the input variables to training the models. In some cases, these inputs may not be reliable in the clinical setting. To address this concern, machine learning analyses have been suggested to include principal component analysis, feature selection, or feature importance analysis in order to reduce the incidence of fitting confounders during model training. As shown in this systematic review, some of included studies performed either feature selection or feature importance analysis to reduce the incidence of unintentional fitting of confounders [3,15,23,29,44,46]. Although, this process may not be needed in deep learning analysis.

In the quest to successfully translate these potentially transformative models from research into daily clinical practices, the privacy of patient information and ethical use of the data should also be considered [25,30]. Therefore, to address the concern of privacy and illegal exploitation of patients' data, informed consent of the patients is necessary regarding the usage of patients' data [100–103]. Other ethical (sociocultural) concerns include the balance between the benefits to potential harm concern, defining who will be responsible if the model fails [25,30], and commercial related interests (integration of machine learning-based model may actually reduce the revenue of the health systems and consequently of the clinicians) [25]. Other ethical related issues relating to the deployment of the machine learning models in daily clinical practices have been recently summarized by Alabi et al. [104]. Most importantly,

considering the impressive array of studies that had examined the application of machine learning in oral cancer prognostication as presented in this study, proactive ethical, regulatory, governance, and legal frameworks are necessary to ensure that machine learning models progress safely to daily clinical practices [90,105].

Our systematic review has several limitations. The main limitation of this systematic review is that most of the included studies did not evaluate the challenges of the integration of machine learning models into daily clinical practices. Thus, possible solutions could not be inferred from the included studies. In addition, the qualities of the included studies varied.

In conclusion, our systematic review reveals the potential usefulness 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 faster implementation of this approach in clinical practice. This would further enhance informed clinical decision-making and offer a better diagnosis and prognostication of oral cancer. Future work to improve explainability and interpretability of the machine learning models and using clinically applicable performance metrics would be necessary to translate these models for use in daily clinical practice. The developers of machine learning models should be conversant with the data to be used in the training process and with unintended algorithmic bias, and they should ensure that the developed models are externally validated to enhance generalization. The development of insightful regulatory frameworks is essential for the safe integration of these models into daily clinical practices.

Authors Contribution

Study concepts and study design: Alabi RO, Elmusrati M, Almangush A, Leivo I. **Studies extraction:** Alabi RO, Omar Y. **Acquisition and quality control of included studies:** Alabi RO, Omar Y, Almangush A. **Data analysis and interpretation:** Alabi RO, Elmusrati M, Almangush A, Mäkitie AA, Pirinen M, Leivo I. **Manuscript preparation:** Alabi RO, Omar Y, Almangush A, Mäkitie AA, Pirinen M. **Manuscript review:** Mäkitie AA, Leivo I, Elmusrati M, Pirinen M. **Manuscript editing:** Almangush, Alabi RO, Omar Y. All authors approved the final manuscript for submission.

Summary points

What was already known on the topic:

- There are several published studies on the application of machine learning techniques to analyze oral squamous cell carcinoma (OSCC).
- 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 analyze oral squamous cell carcinoma (OSCC).
- It examines the concerns and limitations to the actual implementation of machine learning-based models in clinical settings. This study also discusses possible solutions to these concerns.
- Support vector machine and artificial neural network are the most widely used algorithms for oral cancer prognostication.
- Addressing the limitations as suggested in this study may ensure that the models are useful for effective oral cancer management.

Acknowledgement/Funding

The School of Technology and Innovations, University of Vaasa Scholarship Fund. Turku University Hospital Fund, 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

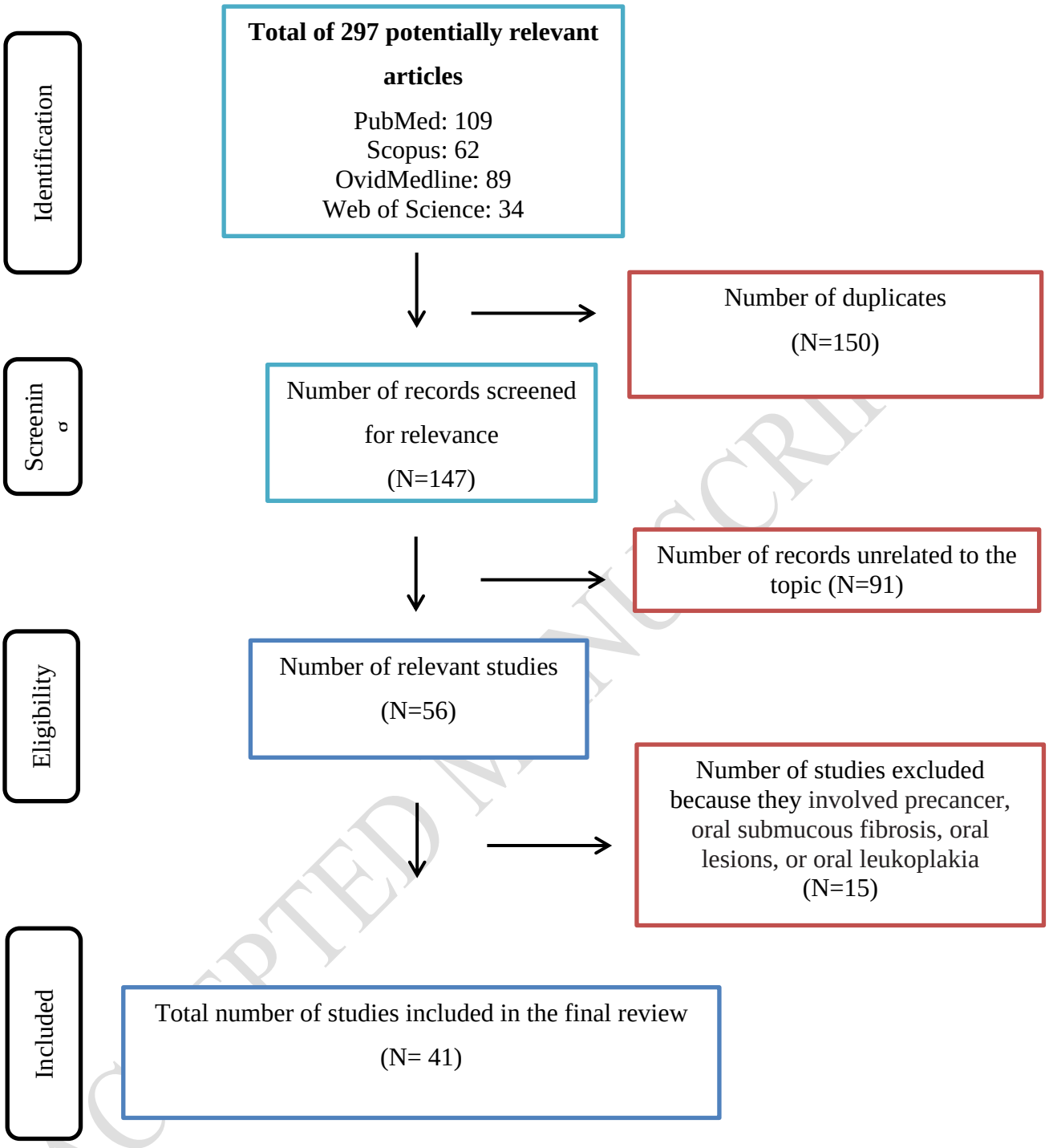

Figure 1. The flow diagram highlighting the search strategy and the search results**.**

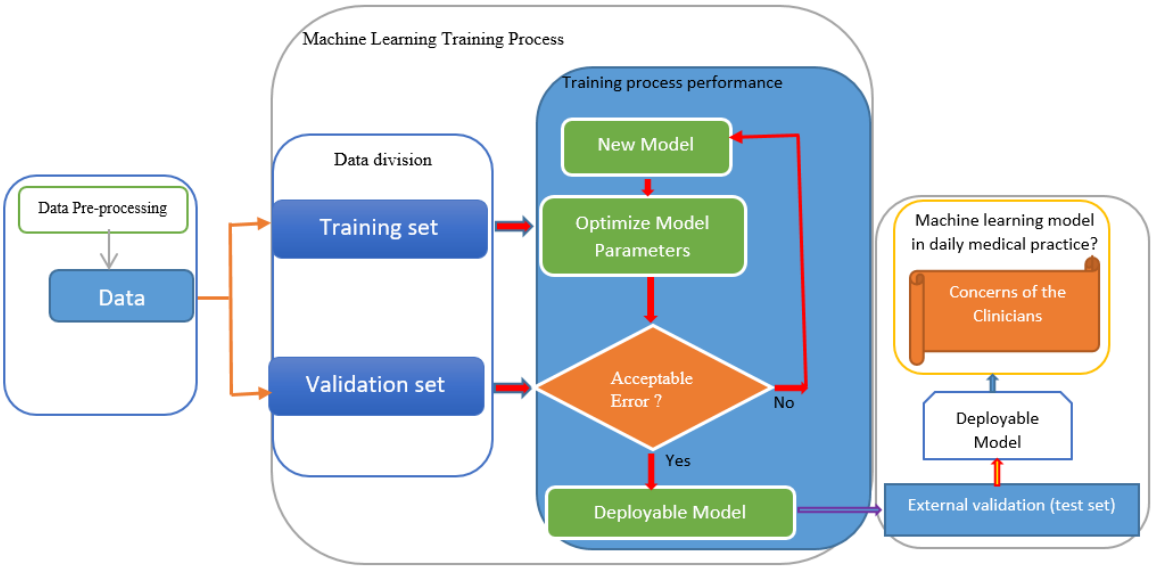

Figure 2. Machine learning training scheme showing the concern to actual implementation.

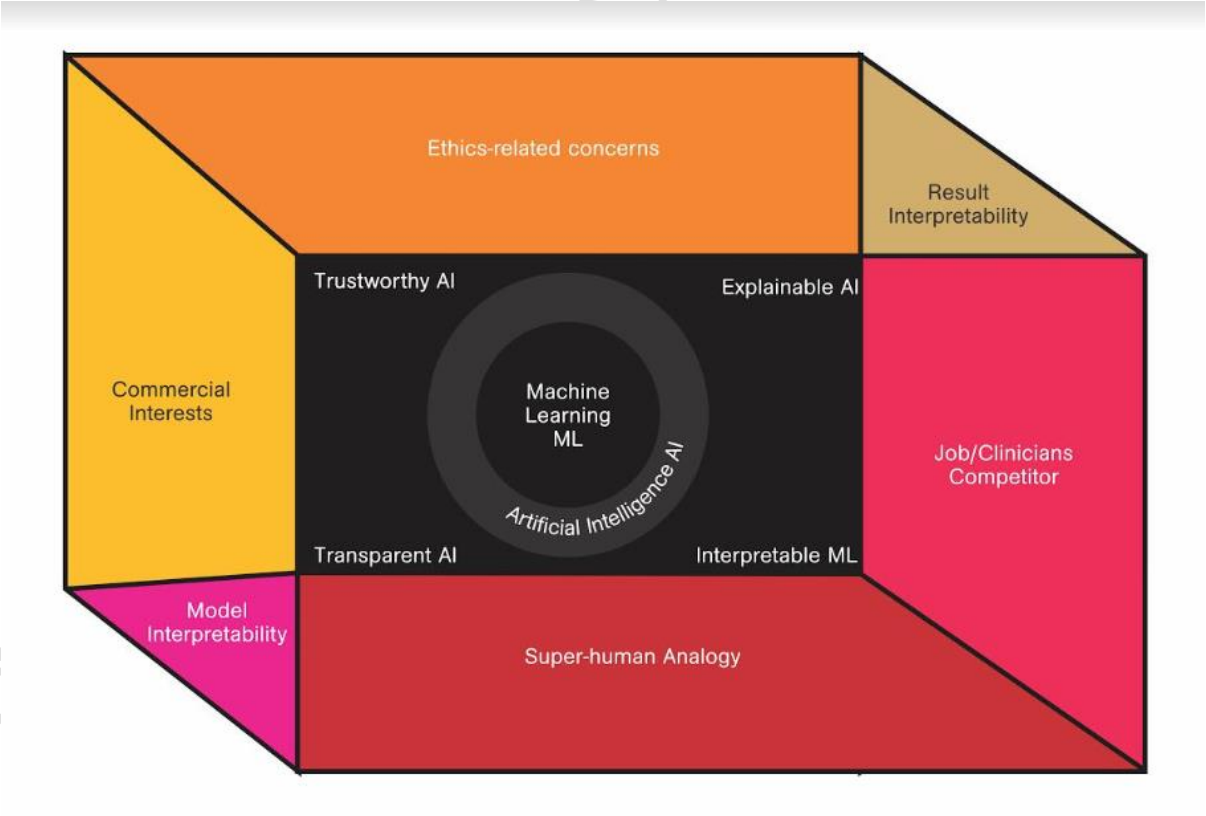

Figure 3. The black-box concern of the machine learning models in oral cancer management

Table 1. Extracts of the main findings from the included studies

Authors, year (country)	Site	No of Cases [date type]	Machine Learning Methods	Use of Machine Learning in Oral cancer	Study Aim	Results	Performance metric (s)	Conclusion
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	The neural network may be valuable in the identification of patients with a high risk of oral cancer.
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	The hybrid technique proposed in this study significantly improved the identification efficiency.
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%	The algorithm showed significant accuracy for the prediction of lymph node metastasis.
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	Sensitivity: 0.91	The Bayesian framework addressed some of the

						by the Bayesian framework of RVM was comparable to the traditional SVM.	Specificity: 0.96	concern other traditional algorithms while producing comparable performance.
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	The examined algorithm distinguished between normal, premalignant, and malignant oral tissues.
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%	Predicting lymph node status before surgery may help to decide whether additional preoperative treatment or surgical lymph node dissection is needed.
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%	The prediction of potential relapse may offer decision support avenue for the clinicians.
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	The effective prediction of survival in oral cancer gives an overall better management of oral cancer.

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%	The presented hybrid method offers superior prognosis. Also, the selected features suggests the potential of becoming a significant milestone in oral cancer studies.
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 (clinicopathologic and genomic) than group 1 (clinicopathologic markers only)	Accuracy: 93.8%	The study identified important markers and produced model that can support effective clinical decisions.
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%	The support vector machine may be the most favorable model for predicting survival in oral cancer patients.
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%	The survival rate is influence by factors such as treatment and poor cell differentiation. Patients with stage IV with certain characteristics have low survival rate.

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%	The developed variants of neural network performed better than the widely used classifiers.
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%	Two models of neural network predicted chances of survival of oral cancer patients.
Shams & Htike, 2017 (Malaysia)	Oral cavity	86	Support Vector Machine (SVM), Deep Neural Network (DNN), Regularized Least Squares (RLS) & Multi-layer perceptron (MLP)	Prognostic (Gene expression data).	To predict the risks of oral cancer in oral premalignant lesion (OPL) patients.	The DNN technique performed better than others.	Sensitivity:0.98 Specificity: 0.94 Accuracy: 96%	ML technique with gene expression profiling predicted the possibility of oral cancer development in OPL patients.
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	This approach provides an automatic diagnosis using deep learning.
Lu *et al.*, 2017 (China & USA)	Oral cavity	115	Linear Discriminant Analysis (LDA), Quadratic Discriminant	Prognostic (Clinicopathologic + image analysis).	To predict the disease-specific survival.	The study properly associa	AUC: 0.72	Nuclear morphology can risk stratify patients for

			Analysis (QDA), Support Vector Machine (SVM), Random Forest (RF)			ted local nuclear morphologic heterogeneity with long term outcomes.		disease-specific survival.
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	The approach offered early detection and diagnosis, minimize disease progression and reduce death rate.
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%.	The early-detection of oral cancer helps to reduce the death rate associated with oral cancer.
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.	The boosting probabilistic versions of SVM improved the performance in the oral cancer discrimination tasks.
Cheng *et al.*, 2018 (Taiwan)	Oral cavity	1,429	K-Nearest Neighbor (KNN), K-shortest paths (K-STAR), Randomizable	Diagnostic (Clinicopathological data)	To predict recurrence	Important risk factors for	Specificity: 0.75	The application of this model is poised to reduce the

			Filtered Classifier (RFC), & Random Tree (RT)			recurrence were identified. Also, KSTAR algorithm showed the best performance	Accuracy: 77.0%	incidence of recurrence.
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%	Clinically relevant regions from oral mucosa image were distinguished
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%	The approach produced automatic screening of biopsy images
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%	Raman spectroscopy combined with deep learning has a great potential for the intraoperative evaluation of the margin resection of oral tongue squamous cell carcinoma.

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%	Raman spectral characterization and DCNN classification of normal and tongue tumor tissue.
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	A model for the detection of cancer of the oral cavity developed.
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.	Sensitivity: 0.917 Specificity: 0.576 AUC: 0.84	The machine learning approach improves prediction of pathologic nodal metastasis
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 from healthy individuals. The PCA-LDA outperformed other examined techniques.	Specificity: 89% Accuracy: 95%	The ANN was used to detect subtle changes in the conformations of proteins, lipids, and nucleic acids. Thus, this non-invasive method was able to make distinction between oral cancer and healthy individuals.

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%	The machine learning (ML) predicted locoregional recurrence and also performed better than depth of invasion (DOI) based model
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%	The proposed hybrid method provided an efficient method to classify oral cancer.
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%	ML method provided effective technique to classify oral cancer into stages
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	The machine learning model was able to classify the patients as either high-risks or low-risks. Thereby providing precise, cost effective and personalized treatments.
Alabi *et al.*, 2019 (Finland & Brazil)	Oral tongue squamous cell carcino	311	Artificial Neural Network (ANN)	Prognostic (Clinicopathological data)	Prediction of locoregional recurrences	The accuracy of the neural networ	Sensitivity: 0.71 Specificity: 0.98	The machine learning approach offers a unique decision-making for

	ma (OTSCC)					k was significantly higher.	Accuracy: 88.2%	predicting locoregional recurrences.
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%	Machine learning approach produced a model to predict survival of OSCC patients.
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%	The tele-cytology approach showed to be an effective method for accurate and remote diagnosis.
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 %	The application of regression-based partitioned CNN improves diagnosis. Thereby, improving early detection and cancer treatments.
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%.	Although, the performance of the CNN is no different from the pathologists, it can be a useful method for diagnostic support.
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 perfor	Accuracy: 75.4%	The examined variant showed promising results in stratifying between

						mance than the 2DCNN in differentiating between benign and malignant.		benign and malignant oral cancer.
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%	A radiomic machine learning (ML) techniques was able to predict tumor grade and nodal status in oral cancer patients
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 assessment.	Accuracy: 99.3%	The approach facilitates effective screening of PMOL
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, immunohistochemical)	To predict occult lymph node metastases (OLNM)	The examined algorithm offered a clinical management strategies to identify patients that would benefit from neck dissection	Sensitivity: 0.8 Specificity: 0.9 Accuracy: 90%	The developed model could significantly improve the management of patients with early-stage OSCC

Abbreviations: OCSCC: Oral Cavity Squamous-Cell Carcinoma, AUC: Area Under Receiving Operating Characteristics (ROC) curve, MAUC: Mean Area Under Receiving Operating Characteristics (ROC) curve. * Other sites were considered in the study as well. + Where more than one algorithm was considered, the algorithm with the best performance metrics was reported in the above table (Table 2). Similarly, when the performance metrics were reported differently for training and validation, only the validation performance metrics was considered.

Table 2. Tabular presentation of PROBAST results.

Study	ROB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Speight et al., 1995	+	+	+	?	+	+	+	-	+
Wang et ., 2003	+	+	+	+	+	+	+	+	+
Kawazu et al., 203	+	+	+	+	+	+	+	+	+
Majumder et al., 2005	+	+	+	+	+	+	+	+	+
Nayak et al., 2006	+	+	+	+	+	+	+	+	+
Kim & Cha, 2011	+	+	+	+	+	+	+	+	+
Exarchos et al., 2012	+	+	+	+	+	+	+	+	+
Sharma and Om, 2013	+	+	+	+	+	+	+	+	+
Chang et al., 2013	+	+	+	+	+	+	+	+	+
Chang et al., 2014	+	+	+	+	+	+	+	+	+
Sharma & Om, 2014	+	+	+	+	+	+	+	+	+
Tseng et al., 2015	+	+	+	+	+	+	+	+	+
Sharma and Om, 2015	+	+	+	+	+	+	+	+	+
Sharma & Om, 2015	+	?	+	+	+	?	+	-	-
Shams & Htike 2017	+	?	+	+	+	?	+	-	-

Aubreville et al., 2017	+	+	+	+	+	+	+	+	+
Lu et al., 2017	+	+	+	+	+	+	+	+	+
Uthoff et al., 2018	+	+	+	+	+	+	+	+	+
Al-Ma'aitah & AlZubi, 2018	+	+	+	+	+	+	+	+	+
Turki & Wei, 2018	+	+	+	+	+	+	+	+	+
Cheng et al., 2018	+	?	+	+	+	?	+	-	-
Das et al., 2018	+	+	+	+	+	+	+	+	+
Nawandhar et al., 2019	+	+	+	+	+	+	+	+	+
Yan et al., 2019	+	+	+	+	+	+	+	+	+
Yu et al., 2019	+	+	+	+	+	+	+	+	+
Chan et al., 2019	+	+	+	+	+	+	+	+	+
Bur et al., 2019	+	+	+	+	+	+	+	+	+
Zlotogorski-Hurvitz et al., 2019	+	+	+	+	+	+	+	+	+
Alabi et al., 2019	+	+	+	+	+	+	+	+	+
Lalithamani et al., 2019	+	+	+	+	+	+	+	+	+
Lavanya & Chandra, 2019	+	+	+	+	+	+	+	+	+
Wang et al., 2019	+	+	+	+	+	+	+	+	+
Alabi et al., 2019	+	+	+	+	+	+	+	+	+
Karadaghy et al., 2019	+	+	+	+	+	+	+	+	+
Sunny et al., 2019	+	+	+	+	+	+	+	+	+

Jeyaraj & Samuel Nadar, 2019	+	+	+	+		+	+	+	+	+
Ariji et al., 2019	+	+	+	+		+	+	+	+	+
Xu et al., 2019	+	+	+	+		+	+	+	+	+
Romeo et al., 2020	+	+	+	+		+	+	+	+	+
McRae et al., 2020	+	+	+	+		+	+	+	+	+
Mermod et al., 2020	+	+	+	+		+	+	+	+	+

PROBAST = Prediction model Risk Of Bias Assessment Tool; ROB = Risk of Bias.
+ Indicates Low ROB/Low concern regarding applicability.
– Indicates High ROB/high concern regarding applicability.
? Indicates unclear ROB/unclear concern regarding applicability.

Table 3. 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 4. Quality scores of the included studies based on the guidelines provided Luo et al., 2016 [36, guidelines modified]

Studies	Title	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%
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%
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												90.0%
Lavanya & Chandra, 2019												90.0%
Wang et al., 2019												100.0%
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%

Yes No

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] 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.

[4] 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.

[5] 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.

[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] T. Kawazu, K. Araki, K. Yoshiura, E. Nakayama, S. Kanda, Application of neural networks to the prediction of lymph node metastasis in oral cancer, Oral Radiol. 19 (2003) 35–40. https://doi.org/10.1007/BF02493239.
[62] 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.
[63] 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.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] 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).
[69] D. Michie, D.J. Spiegelhalter, C.C. Taylor, eds., Machine learning, neural and statistical classification, Ellis Horwood, New York, 1994.
[70] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.
[71] 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.
[72] 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.
[73] D. Castelvecchi, Can we open the black box of AI., Nature. 538 (2016) 20–23.
[74] Bernease Herman, The Promise and Peril of Human Evaluation for Model Interpretability, (2017). http://interpretable.ml/ (accessed January 15, 2020).
[75] European Union, High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI, (2019).
[76] L. Zachary, The Doctor Just Won't Accept That!, (2017). http://interpretable.ml/ (accessed January 15, 2020).
[77] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Min. Knowl. Discov. 9 (2019). https://doi.org/10.1002/widm.1312.
[78] A. Holzinger, A. Carrington, H. Müller, Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations, KI - Künstl. Intell. 34 (2020) 193–198. https://doi.org/10.1007/s13218-020-00636-z.
[79] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain?, ArXiv171209923 Cs Stat. (2017). http://arxiv.org/abs/1712.09923 (accessed January 6, 2021).
[80] K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration, Ann. Intern. Med. 162 (2015) W1. https://doi.org/10.7326/M14-0698.
[81] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, members of the TRIPOD group, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement, Eur. Urol. 67 (2015) 1142–1151. https://doi.org/10.1016/j.eururo.2014.11.025.
[82] G.S. Collins, K.G.M. Moons, Reporting of artificial intelligence prediction models, The Lancet. 393 (2019) 1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6.
[83] A.J. Vickers, A.M. Cronin, E.B. Elkin, M. Gonen, Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers, BMC Med. Inform. Decis. Mak. 8 (2008). https://doi.org/10.1186/1472-6947-8-53.
[84] P. Brocklehurst, D. Field, K. Greene, E. Juszczak, R. Keith, S. Kenyon, L. Linsell, C. Mabey, M. Newburn, R. Plachcinski, M. Quigley, E. Schroeder, P. Steer, Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial, The Lancet. 389 (2017) 1719–1729. https://doi.org/10.1016/S0140-6736(17)30568-8.
[85] 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.
[86] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N.G. Campeau, V.K. Venugopal, V. Mahajan, P. Rao, P. Warier, Deep learning algorithms for detection of critical findings in

head CT scans: a retrospective study, The Lancet. 392 (2018) 2388–2396. https://doi.org/10.1016/S0140-6736(18)31645-3.
[87] T.P.A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, K.G.M. Moons, A new framework to enhance the interpretation of external validation studies of clinical prediction models, J. Clin. Epidemiol. 68 (2015) 279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018.
[88] D.W. Kim, H.Y. Jang, K.W. Kim, Y. Shin, S.H. Park, Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers, Korean J. Radiol. 20 (2019) 405. https://doi.org/10.3348/kjr.2019.0025.
[89] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019). https://doi.org/10.1186/s12916-019-1426-2.
[90] K.Y. Ngiam, I.W. Khor, Big data and machine learning algorithms for health-care delivery, Lancet Oncol. 20 (2019) e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4.
[91] J.C. Mandel, D.A. Kreda, K.D. Mandl, I.S. Kohane, R.B. Ramoni, SMART on FHIR: a standards-based, interoperable apps platform for electronic health records, J. Am. Med. Inform. Assoc. 23 (2016) 899–908. https://doi.org/10.1093/jamia/ocv189.
[92] K. Crawford, R. Calo, There is blindspot in AI research, Nature. 538 (2016) 311–3.
[93] S. Barocas, A.D. Selbst, Big Data's Disparate Impact, SSRN Electron. J. (2016). https://doi.org/10.2139/ssrn.2477899.
[94] I. Chen Y., F. Johansson D., D. Sontag, Why Is My Classifier Discriminatory? In: 32nd Conference on Neural Information Processing Systems (NeurIPS), (2018). https://proceedings.neurips.cc/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf.
[95] H.A. Haenssle, C. Fink, A. Rosenberger, L. Uhlmann, Reply to the letter to the editor 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists' by H. A. Haenssle et al., Ann. Oncol. 30 (2019) 854–857. https://doi.org/10.1093/annonc/mdz015.
[96] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature. 542 (2017) 115–118. https://doi.org/10.1038/nature21056.
[97] B. Nestor, M.B.A. McDermott, G. Chauhan, T. Naumann, M.C. Hughes, A. Goldenberg, M. Ghassemi, Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation, ArXiv181112583 Cs Stat. (2018). http://arxiv.org/abs/1811.12583 (accessed January 5, 2021).
[98] S.E. Davis, R.A. Greevy, C. Fonnesbeck, T.A. Lasko, C.G. Walsh, M.E. Matheny, A nonparametric updating method to correct clinical prediction model drift, J. Am. Med. Inform. Assoc. 26 (2019) 1448–1457. https://doi.org/10.1093/jamia/ocz127.
[99] S.M. Willems, S. Abeln, K.A. Feenstra, R. de Bree, E.F. van der Poel, R.J. Baatenburg de Jong, J. Heringa, M.W.M. van den Brekel, The potential use of big data in oncology, Oral Oncol. 98 (2019) 8–12. https://doi.org/10.1016/j.oraloncology.2019.09.003.
[100] J.R. Geis, A.P. Brady, C.C. Wu, J. Spencer, E. Ranschaert, J.L. Jaremko, S.G. Langer, A. Borondy Kitts, J. Birch, W.F. Shields, R. van den Hoven van Genderen, E. Kotter, J. Wawira Gichoya, T.S. Cook, M.B. Morgan, A. Tang, N.M. Safdar, M. Kohli, Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement, Radiology. 293 (2019) 436–440. https://doi.org/10.1148/radiol.2019191586.

[101] J. Powles, H. Hodson, Google DeepMind and healthcare in an age of algorithms, Health Technol. 7 (2017) 351–367. https://doi.org/10.1007/s12553-017-0179-1.
[102] J. Bali, R. Garg, R.T. Bali, Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required?, Indian J. Ophthalmol. 67 (2019) 3–6. https://doi.org/10.4103/ijo.IJO_1292_18.
[103] J. Nabi, How Bioethics Can Shape Artificial Intelligence and Machine Learning, Hastings Cent. Rep. 48 (2018) 10–13. https://doi.org/10.1002/hast.895.
[104] R.O. Alabi, V. Tero, E. Mohammed, Machine learning for prognosis of oral cancer: What are the ethical challenges?, CEUR-Workshop Proc. (2020).
[105] Food and Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): FDA, 2019. . https://www.regulations. gov/document?D=FDA-2019-N-1185-0001.