Advances in Radiation Oncology (2023) 8, 101163

advances
in radiation oncology

www.advancesradonc.org

Scientific Article

Comparison of Machine-Learning and Deep-Learning Methods for the Prediction of Osteoradionecrosis Resulting From Head and Neck Cancer Radiation Therapy

Brandon Reber, BS,[a,*] Lisanne Van Dijk, PhD,[a,b] Brian Anderson, PhD,[a,c] Abdallah Sherif Radwan Mohamed, MD, PhD,[a] Clifton Fuller, MD, PhD,[a] Stephen Lai, MD, PhD,[a] and Kristy Brock, PhD[a]

[a]*Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas;* [b]*University of Groningen, Groningen, Netherlands; and* [c]*University of California, San Diego, San Diego, California*

Received 15 November 2022; accepted 22 December 2022

Abstract

Purpose: Deep-learning (DL) techniques have been successful in disease-prediction tasks and could improve the prediction of mandible osteoradionecrosis (ORN) resulting from head and neck cancer (HNC) radiation therapy. In this study, we retrospectively compared the performance of DL algorithms and traditional machine-learning (ML) techniques to predict mandible ORN binary outcome in an extensive cohort of patients with HNC.

Methods and Materials: Patients who received HNC radiation therapy at the University of Texas MD Anderson Cancer Center from 2005 to 2015 were identified for the ML (n = 1259) and DL (n = 1236) studies. The subjects were followed for ORN development for at least 12 months, with 173 developing ORN and 1086 having no evidence of ORN. The ML models used dose-volume histogram parameters to predict ORN development. These models included logistic regression, random forest, support vector machine, and a random classifier reference. The DL models were based on ResNet, DenseNet, and autoencoder-based architectures. The DL models

Sources of support: Research reported in this publication was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) under award number P30CA016672, the Helen Black Image Guided Fund, resources from the Image Guided Cancer Therapy Research Program at the University of Texas MD Anderson Cancer Center, a generous gift from the Apache Corporation, and support from the Tumor Measurement Initiative through the MD Anderson Strategic Initiative Development Program.

Disclosures: Mr Reber and Dr Brock received support from the NIH/NCI under award number P30CA016672, the Helen Black Image Guided Fund, resources from the Image Guided Cancer Therapy Research Program at the University of Texas MD Anderson Cancer Center, a generous gift from the Apache Corporation, and support from the Tumor Measurement Initiative through the MD Anderson Strategic Initiative Development Program. Dr Van Dijk has received support from the Dutch Cancer Society (KWF-13529), Rubicon (NWO-452182317), and VENI (NWO-09150162010173). Dr Anderson received an Allied Scientist grant from the Society of Interventional Radiology. Abdallah Mohamed received support from the NIH through a NIH National Institute of Dental and Craniofacial Research (NIDCR) Academic Industrial Partnership Grant (R01DE028290), NIH/National Science Foundation NCI Smart Connected Health Program (R01CA257814), and an NIDCR Establish Outcome Measures for Clinical Studies of Oral and Craniofacial Diseases and Conditions award 1 (R01DE025248). Dr Fuller received an NCI Institutional Research Training Grant (T32CA261856) and National Institute of Biomedical Imaging and Bioengineering Grant for Research Education Programs for Residents and Clinical Fellows. Dr Lai has received support from the NIDCR (R01 DE025248).

Research data were acquired under NIH R01DE025248 and are stored on figshare at the following URL: https://figshare.com/articles/dataset/Dosevolume_histogram_DVH_parameters_of_the_mandible_for_Normal_Tissue_Complication_Probability_modelling/13568207

*Corresponding author: Brandon Reber.; E-mail: breber@mdanderson.org

https://doi.org/10.1016/j.adro.2022.101163

used each participant's dose cropped to the mandible. The effect of increasing the amount of available training data on the DL models' prediction performance was evaluated by training the DL models using increasing ratios of the original training data.

Results: The F1 score for the logistic regression model, the best-performing ML model, was 0.3. The best-performing ResNet, DenseNet, and autoencoder-based models had F1 scores of 0.07, 0.14, and 0.23, respectively, whereas the random classifier's F1 score was 0.17. No performance increase was apparent when we increased the amount of training data available for DL model training.

Conclusions: The ML models had superior performance to their DL counterparts. The lack of improvement in DL performance with increased training data suggests that either more data are needed for appropriate DL model construction or that the image features used in DL models are not suitable for this task.

Introduction

Head and neck cancers (HNCs) involve the oral cavity, sinuses, pharynx, larynx, and associated regions.[1] The global relative incidence rates for HNCs by region are 2.0% for the oral cavity, 1.0% for the larynx, 0.7% for the nasopharynx, 0.5% for the oropharynx, 0.4% for the hypopharynx, and 0.3% for the salivary glands.[2] Radiation therapy (RT) is a cornerstone treatment modality for HNC whether in the definitive or adjuvant setting.[3] Survival rates for head and neck squamous cell carcinoma have increased over the past few decades, with the Surveillance, Epidemiology, and End Results Program reporting 5-year survival rates of 54.7% in 1992 to 1996 and 65.9% in 2002 to 2006.[2] This is mainly attributed to the predominance of the prognostically better human papillomavirus −associated variants in recent decades.[4] This improvement in survival indicates the importance of reducing the incidence of HNC treatment late toxic effects to enhance both RT for these cancers and patient quality of life after treatment.

When treating HNCs with radiation, various treatment-related late toxic effects can occur afterward, including xerostomia, dysphagia, dysgeusia, trismus, and osteoradionecrosis (ORN).[5-7] Osteoradionecrosis is the persistent exposure of bone resulting from irradiation that does not heal over 3 months and can present as acute or delayed exposure after RT.[8] In RT for HNC, the mandible is the bone most affected by ORN; the maxilla also can be affected, but at a much lower prevalence (24:1).[9] The onset of ORN usually occurs within 4 months to 2 years after treatment.[10] The severity of ORN can be classified using various systems, with most distinguishing between higher and lower severity.[9] Management may include nonsurgical methods such as pentoxifylline and antibiotics or surgical procedures in which necrotic bone is resected.[10] Typically, earlier-stage ORN is treated with more conservative measures before moving to more-invasive strategies such as surgery.[11]

The ability to predict ORN risk before treatment would enable further optimization of treatment techniques (proton therapy, adaptive RT) and monitoring for early indications of ORN. Many studies have looked at risk factors for ORN, including clinical and dose-volume parameters. Identified risk factors include dosimetric parameters such as the D_{mean}, smoking, preradiation therapy surgery/tooth extraction, oral mucositis, dentist visits before RT, mandibular surgery, and tumor location. However, considerable variation remains regarding which of these parameters are significant for ORN development.[12-17]

Researchers have applied machine-learning (ML) techniques to various problems related to cancer.[18] Traditional ML techniques use pre-extracted or hand-crafted features to infer a target class. In comparison, deep-learning (DL) techniques extract features within images, text, and other data without pre-extraction, creating features that may be hard to construct using traditional approaches. These low-level image features often include lines, curves, and gradients among other simple image components. Investigators have applied DL to several medical imaging tasks, such as segmentation, disease detection, and noise reduction.[19,20] Deep-learning also has been applied to outcome prediction for several anatomic sites and differing outcomes, but ORN prediction from HNC RT remains an ongoing problem of interest.[21] The DL models have progressed over the years, from the introduction of the convolutional layer to the skip connection, attention mechanisms, and recent transformer models.[22]

One problem that affects DL methods is a requirement for larger sample sizes than those used with traditional ML algorithms.[18] In medical imaging, obtaining large samples for DL can be difficult because of the relative smaller number of events and stronger privacy requirements compared with many natural image tasks. However, unlike traditional ML algorithms, which are limited to discretized variables, DL methods can use entire spatial gradients contained within images. Whereas different traditional ML algorithms have been compared for ORN prediction, to the best of our knowledge, no study has examined the viability of DL for this task, used full spatial dose information contained within images, or compared DL and ML performance for ORN prediction.[23] In this study, we compared the performance of traditional ML algorithms with DL algorithms for the prediction of binary ORN outcome using HNC patient radiation dose distributions. With full 3-dimensional (3D) dose information, we believe that DL should outperform the ML models for this prediction task.

Methods and Materials

Data

After institutional review board approval (RCR03-0800), retrospective subject data from 2005 to 2015 at the University of Texas MD Anderson Cancer Center was obtained and evaluated. The subjects' eligibility included patients with head and neck squamous cell carcinoma treated with RT alone or in conjunction with surgery or chemotherapy with curative intent. Initially, 1789 subjects were identified for inclusion; however, 530 were excluded as the result of having previous HNC irradiation, a survival time shorter than 1 year, a history of salivary gland cancer, or unavailable treatment plans. Figure E1 shows the exclusion criteria, and Table E1 shows the treatment prescription. The 1259 remaining evaluable subjects were followed for a minimum of 12 months after RT. This minimum follow-up time was chosen to maximize the number of cases followed while still allowing time posttreatment for ORN cases to develop. Most cases received a splitfield that matched a larynx midline block and lower anterior neck field for primary tumors and upper nodal neck disease. Intensity modulated RT was used when tumors were inferiorly positioned. There were not changes to the dose calculation algorithm throughout the study. Full 3D dose maps were readily available for 1236 of the subjects. The 23 3D dose maps not included could have their dose−volume histogram parameters extracted for the ML approaches, but the images themselves were not available for the DL methods. The ORN grading scheme used was the one defined by Tsai et al[24]: grade 1, minimal bone exposure with conservative management only; grade 2, minor debridement; grade 3, hyperbaric oxygen therapy; and grade 4, major invasive mandible surgery.

Computed tomography images of the head and neck used for treatment planning were obtained for each subject. A multiatlas-based segmentation of the mandible on each computed tomography image was performed using ADMIRE software (research version 1.1; Elekta). Dose grids were obtained using 1 of 2 treatment planning systems: Pinnacle (version 6.2b or later; Philips Medical Systems) or CORVUS (version 4.0; Nomos Corporation). Spacing of 4 mm × 4 mm × 4 mm was ensured for the dose fields and mandible contours. The Python package SimpleITK (version 2.1.1) was used to resample the images using nearest neighbor interpolation to ensure correct spacing, if necessary.[25] The SimpleITK package also was used to ensure that the mandible contours and dose maps had the same physical origin for each patient. The mandible contour for each subject was used to crop the corresponding 3D dose grids to the pixel dimensions of 32 × 128 × 128 around the mandible using a Python script.

All cropped images were inspected to ensure that the entire mandible fit within the 32 × 128 × 128 cropping. Mandibles smaller than the cropping window had additional adjacent voxels included to ensure the cropped image met the required size. Including voxels not solely within the region of interest was needed to ensure that all input images had the same size. In addition, including voxels outside the region of interest, in this case the mandible, is common when applying convolutional neural networks to medical image tasks.[20] An example of the dose and cropping is given in Fig. E2.

The subject data were split into training and withheld test sets for the ML and DL models. A total of 1236 subjects were available for use with the DL models, with 171 ORN + cases and 1065 ORN− cases. In comparison, a total of 1259 subjects were available for use with the traditional ML models, with 173 ORN+ and 1086 ORN− cases. The same cases for the test set were withheld from all ML and DL models: 369 subjects with 48 ORN+ cases. Although the total number of cases were different between the ML and DL approaches, the test sets had the same cases for both, which allowed for final performance comparisons. For the traditional ML methods, the remaining data were used in a nested cross-validation. For the DL models, the remaining data were split into training and validation sets. The validation set was used during training to select the best set of hyperparameters for each DL model type. The final data split was 650, 217, and 369 subjects in the training, validation, and test sets, respectively. The number of ORN+ cases was 111, 12, and 48 in the training, validation, and test sets, respectively. A random number generator was used to split the data into the different groups so that the incidence rate of ORN+ cases in the test set was approximately similar to the incidence rate of ORN+ in the overall data set. The training data was selected to be 75% of the remaining data not included in the test set. A larger proportion of ORN+ cases in the training data set compared with the validation data set was allowed to maximize the number of ORN+ cases seen during training.

All data sets for the DL models were z-score−standardized using the mean and SD voxel values from the training data set. All voxel values for all training subject data from the cropped 3D dose maps were used to calculate the mean and SD voxel values. Standardizing the convolutional neural network model input instead of using the original voxel values is common in medical image DL.[20] To account for data imbalance, the class with a smaller number of samples within the data set (ORN+) was oversampled randomly with replacement to match the number of samples of the class with a larger number of cases within the data set (ORN−). This random oversampling was only applied to the training set. The oversampled training set had 1078 subjects.

Standard ML

The standard ML techniques used were logistic regression, random forest, and support vector machine. R (version 4.0.4; R Foundation for Statistical Computing, Vienna, Austria) was used for the logistic regression model with the package caret to construct the model.[26,27] The Python package Scikit-learn (version 0.24.2) was used to construct the random forest and support vector machine models.[28] A random classifier was created to establish a reference ORN prediction model. The random classifier randomly classifies a case as ORN+ or ORN− with equal probability. The dose-volume histogram parameters of the mandible used in the models were the following: V_5-V_{70} in 5-Gy increments, D_5-D_{95} in 5% increments, D_2, D_{97}, D_{98}, D_{99}, mean dose, min dose, and max dose. The Pearson correlation coefficient was used to remove collinear variables. Variables were removed if the Pearson correlation coefficient was >0.90. A nested cross-validation was used to compare the ML techniques. The inner loop performs a hyperparameter grid search for the random forest and support vector machine models. The inner loop is replaced by a stepwise feature selection method for the logistic regression model. The outer loop is used to compare the performance of the ML models. Both inner and outer loops use a 10-fold stratified cross-validation with 10 repeats. The withheld test set was not used in the nested cross-validation.

Data were z-score−standardized using the mean and SD of training data within each cross-validation iteration. A description of the hyperparameters used in the grid search can be found in Appendix E1. A backward stepwise feature selection using the Bayesian information criterion was used to select features for the logistic regression model to then use in the corresponding outer loop iteration. The accuracy, balanced accuracy, recall, precision, F1 score, area under the receiver operating characteristic curve (AUROC), and area under the precision recall curve (AUPRC) were evaluated for each outer loop iteration withheld fold. The mean (±SD) values for the metrics from all outer loop iterations' withheld cross-validation folds were collected. Next, the best-performing ML algorithm was identified by the largest AUROC and AUPRC from the cross-validation. This identified ML model was then trained on the entire training data set and evaluated on the withheld test set.

DL models

The DL models used were 3D versions of the residual neural network (ResNet) and densely connected convolutional network (DenseNet) architectures.[29,30] In addition, a model using an autoencoder as a feature extractor and a series of convolution layers using the bottleneck features was constructed. Diagrams and descriptions of the DL models can be found in Appendix E2. A grid search to select the best hyperparameters for each of the 3 architecture types also was completed. The grid search procedure is also described in the Appendix E3.

All DL model training and evaluation was performed using TensorFlow software (version 2.4.1).[31] Dose map images were augmented using the random rotation of images by ±90° in the transverse plane and reflections in the median plane. A batch size of 1 was used for all models. The binary cross-entropy loss was used for all models except the autoencoder component of the autoencoder-based approach. The autoencoder-based approach was trained in 2 stages. In the first stage, the autoencoder was trained using the mean squared error loss between the input and reconstructed dose input. In the second stage, the ORN classification layers were trained using the binary cross-entropy loss. A cosine decay learning rate schedule was used with 200 epochs, with the learning rate starting at 1×10^{-5} using the Adam optimizer. Training continued until the loss did not improve on the validation set after 20 epochs. The saved weights for each model hyperparameter combination were the weights that had the lowest binary cross-entropy loss on the validation set. The best-performing ResNet, DenseNet, and autoencoder-based models were used to predict ORN in the test set, and the performance of each model was measured by calculating the accuracy, balanced accuracy, recall, precision, F1 score, AUROC, and AUPRC.

DL model performance with increased available training data

An additional study was completed to gauge the usefulness of increasing the amount of training data available to the best performing DL models. The architectures of the best-performing DenseNet and ResNet models were trained using smaller subsets of the total training data set (10%-100% in 10% increments in addition to 25% and 75%) to look for changes in performance on the test set. The validation and test sets were not changed. The models were trained 5 times for each subset of the total training data using random weight initialization and shuffling of the available training data. The model training was performed using the same training strategy as the prior models. The 5 models trained for each subset of the total training data were used to create a majority votes (3 of 5) prediction of ORN and were evaluated on the test set that was withheld from model training. An additional ensemble was created using the 5 DenseNet and the 5 ResNet models trained on the entire data set. A majority votes prediction of ORN status (5 of 10, with a tie predicting ORN negativity) was then completed on the cases in the withheld test set. The metrics of accuracy, balanced accuracy, recall, precision, and F1 score were then calculated for each training set ratio.

Statistical analysis

The performance of the models was evaluated using receiver operating characteristic (ROC) analysis and precision-recall analysis. The metrics derived from these analyses include AUROC, accuracy, recall, precision, balanced accuracy, F1 score, and AUPRC. Balanced accuracy is the arithmetic mean of recall and specificity, and the F1 score is the harmonic mean of precision and recall. The Scikitlearn package (version 0.24.2) was used to calculate model metrics for the random forest and support vector machine models, and Tensorflow (version 2.4.1) was used to calculate model metrics for the DL models.[28,31] The R package MLmetrics was used to obtain the accuracy, balanced accuracy, recall, precision, and F1 score for the logistic regression model.[32] The R package pROC was used to calculate the area under the AUROC and AUPRC for the logistic regression models.[33] The best-performing prediction model between the ML and DL models was identified by greater metric values on the test set. For examining DL model performance with increasing amounts of training data, the accuracy, balanced accuracy, recall, precision, and F1 score were calculated manually for each training set ratio. Increases in metric values with training ratio were used to determine whether there was increasing DL model performance as more training data was available for model development.

Results

Standard ML

The demographics of all subjects are summarized in Table 1. After we filtered out correlated variables with the Pearson correlation coefficient, 4 DVH parameters remained: the mean dose to the mandible, the minimum dose to the mandible, the maximum dose to the mandible, and V_{65}. These 4 features were used in the ML models. The logistic regression feature selection selected in the cross-validation the mean dose 100% of the time and added the minimum dose 85% of the time. No other variables were selected. For the random forest, the most selected hyperparameter combination in the cross-validation was 2 max features per split, 10 minimum samples per leaf, and 10 minimum samples per split. This combination was selected 25% of the time. The most important feature for the random forest model was the mean dose. For the support vector machine, the most selected hyperparameter combination for the cross-validation was the radial basis function kernel, a value of 10 for C, and a value of 0.01 for γ. This combination was selected 37% of the time. The frequency of all hyperparameter combinations from the grid search and a plot of the variable importance for the random forest model is available in the supporting documentation.

The outer cross-validation loop determined the best-performing ML ORN prediction model. The mean values for the metrics on the withheld fold for each iteration and their SDs are shown in Table 2. Specifically, the results for the 3 traditional ML models and the random classifier used for ORN prediction with the dose-volume histogram parameters are shown in the table.

We selected the logistic regression model based on the superior cross-validation balanced accuracy, AUROC, and AUPRC results. We trained the logistic regression model using the full training data set and subsequently evaluated the model on the test data set. The resulting metrics are as follows: 0.64 accuracy, 0.63 balanced accuracy, 0.61 recall, 0.20 precision, 0.31 F1 score, 0.70 AUROC, and 0.24 AUPRC. The mean dose to the mandible was selected.

DL model

The best ResNet model constructed after the hyperparameter search had 5 stages with 2 blocks per stage and 64 starting filters. The best DenseNet model had 4 stages with 6 repetitions for each dense block and 64 starting filters. The best autoencoder-based approach had 3 downsampling stages and 1 convolutional layer per stage. Table 3 shows the performance of the best DL models on the test set that is withheld during the training process and only used for best model evaluation. Table 3 shows the results from single model evaluations, so the reported metrics are not averages with associated standard deviations.

DL model performance with increased available training data

The DL models underperformed compared with the traditional ML models in examining the metrics most sensitive to data set imbalance, such as AUPRC, F1 score, and balanced accuracy. Subpar DL model performance compared with the traditional ML models motivated the examination of how different amounts of training data available for model creation affect the final performance of the DL models. Figure 1 shows how the DL model performance changed with increased amounts of training data available for model creation. Figure 1 shows results from a single evaluation on a test set, so there are no

Table 1 Summary of subject demographics*

	ORN−	ORN+
Number of subjects	1086	173
Age, y, median	61	60
Sex, male, n (%)	894 (82%)	150 (87%)
Smoking, current, n (%)	153 (14%)	27 (16%)
Smoking, pack-years, median	7	8
Postoperative RT	172 (16%)	44 (25%)
Dental extraction pre-RT	270 (25%)	72 (42%)
Tumor site		
Oral cavity	146 (13%)	44 (25%)
Oropharynx	703 (65%)	123 (71%)
Hypopharynx/larynx/nasopharynx/unknown-primary	237 (22%)	6 (3%)

Abbreviations: ORN = osteoradionecrosis; RT = radiation therapy.

* Percent signs within cells indicate the percent of the subject cohort for the ORN− and ORN + cases separately that have each row attribute.

average and standard deviations of metrics to report. The ensemble combining the best performing DenseNet and ResNet architectures and using the entire training data set had the following results: accuracy 0.87, balanced accuracy 0.58, recall 0.25, precision 0.15, and F1-score 0.19. These results are a slight performance increase compared with the models in Table 3 but underperformed compared with the traditional ML models in Table 2. The ensemble approach using just the DenseNet or ResNet architecture did not result in significant performance improvement in majority voting after increasing the data amount from 10% to 100%.

Discussion

Overall, the ML algorithms outperformed the DL ORN prediction models. Most of the traditional ML algorithms performed similarly to each other according to the cross-validation metrics. Because of the imbalance between ORN− and ORN + cases in the data set, metrics less influenced by data set imbalance should be prioritized, such as the F1 score, the AUPRC, and balanced accuracy. In this HNC data set, the AUPRC of a random classifier in the test set would have a value of 0.13 (P / [N + P] = 48/ 369 = 0.13). The logistic regression model evaluated on the test set surpassed this value. The traditional ML methods also produced balanced accuracy values greater than the balanced accuracy of 0.5 that a random classifier would produce, as shown in the cross-validation results. The F1 score is the harmonic mean of the recall and precision and gives a good indication of model performance on data sets with imbalance between classes. The ML models' have relatively greater F1 scores compared to the random classifier reference model.

Overall, the DL models performed worse than the logistic regression model evaluated on the test set. Metrics sensitive to data set imbalance (ie, F1 score, balanced accuracy, and AUPRC) were lower for the DL models than for the logistic regression model. In particular, the F1 score was greater for the logistic regression (0.31) than the ResNet (0.07), DenseNet (0.14), autoencoder-based (0.23), and random classifier (0.13) models. The ResNet and DenseNet models performed better than a random

Table 2 Mean (±SD) metric values for the cross-validation withheld folds for the ML models*

Model	Accuracy	Balanced accuracy	Recall	Precision	F1 score	AUROC	AUPRC
Logistic regression	0.69 ± 0.05	0.70 ± 0.07	0.72 ± 0.14	0.27 ± 0.05	0.39 ± 0.07	0.74 ± 0.07	0.28 ± 0.08
Random forest	0.65 ± 0.05	0.69 ± 0.07	0.74 ± 0.14	0.25 ± 0.04	0.37 ± 0.06	0.69 ± 0.07	0.23 ± 0.04
Support vector machine	0.69 ± 0.04	0.70 ± 0.07	0.71 ± 0.13	0.27 ± 0.04	0.39 ± 0.06	0.70 ± 0.07	0.24 ± 0.04
Random classifier	0.52 ± 0.04	0.49 ± 0.08	0.45 ± 0.14	0.14 ± 0.04	0.21 ± 0.07	0.50 ± 0.00	0.14 ± 0.01

Abbreviations: AUPRC = area under the precision recall curve; AUROC = area under the receiver operating characteristic curve; ML = machine learning.

* Each cell shows the mean (±SD) of the metrics from the withheld folds of the stratified 10-fold cross-validation with 10 repeats.

Table 3 Performance of the best DL models for each architecture type*

Architecture	Accuracy	Balanced accuracy	Recall	Precision	F1 score	AUROC	AUPRC
ResNet	0.87	0.69	0.04	0.50	0.07	0.57	0.23
DenseNet	0.83	0.54	0.10	0.21	0.14	0.58	0.17
Autoencoder	0.71	0.53	0.33	0.18	0.23	0.59	0.15
Random	0.49	0.46	0.46	0.11	0.17	0.49	0.13

Abbreviations: AUPRC = area under the precision recall curve; AUROC = area under the receiver operating characteristic curve; DL = deep learning.

* The reported metrics are from the withheld test set not used during model training or selection. Metrics sensitive to data imbalance, such balanced accuracy, F1 score, and AUPRC, were lower than those for the logistic regression model using the test set.

classifier when we compared the AUPRC and balanced accuracy but performed worse than the traditional ML methods. Unlike the traditional ML models, the DL models tend to misclassify ORN+ cases as ORN−. This is also reflected in the greater accuracy scores for the DL models than for the traditional ML methods.

Ensembles of DL models can be used to improve the performance of DL prediction models versus the use of a single model alone. Using the entire training data set, we found that the ensemble of the best ResNet and DenseNet models did not outperform the logistic regression model performance on the test set according to metrics such as balanced accuracy and the F1 score. To further examine the performance of the DL models, we constructed various ensembles of models using various ratios of the total training data. The performance of the classifiers should improve as more data becomes available for training. In addition, using ensembles of models helps limit prediction variability owing to random weight initialization. However, trends of improvement in performance with more data, as shown in Fig 1, do not occur. The increasing and decreasing changes in performance while increasing training data size shown in Fig 1 suggests that there is insufficient training data in total for establishing a meaningful DL prediction model. If there is sufficient training data, the results could suggest that the low-level features of the dose maps used by the DL models are not as powerful as the dose-volume histogram associations used by the ML models for ORN prediction.

The results for the DL models highlight the challenges of data set size for medical imaging data sets. Relative complication rates should be considered before attempting DL approaches, with rarer complications increasingly requiring larger amounts of total data than more common complications.

A common issue with the application of DL models to medical imaging tasks is limited testing of the models using data from external institutions. A DL model that performs similarly on the internal test set and external institutional data are more robust compared with DL models training exclusively on a single institution data set. The original intent of this study had the DL ORN prediction models proven superior to the traditional ML models was to use an external data set from a different

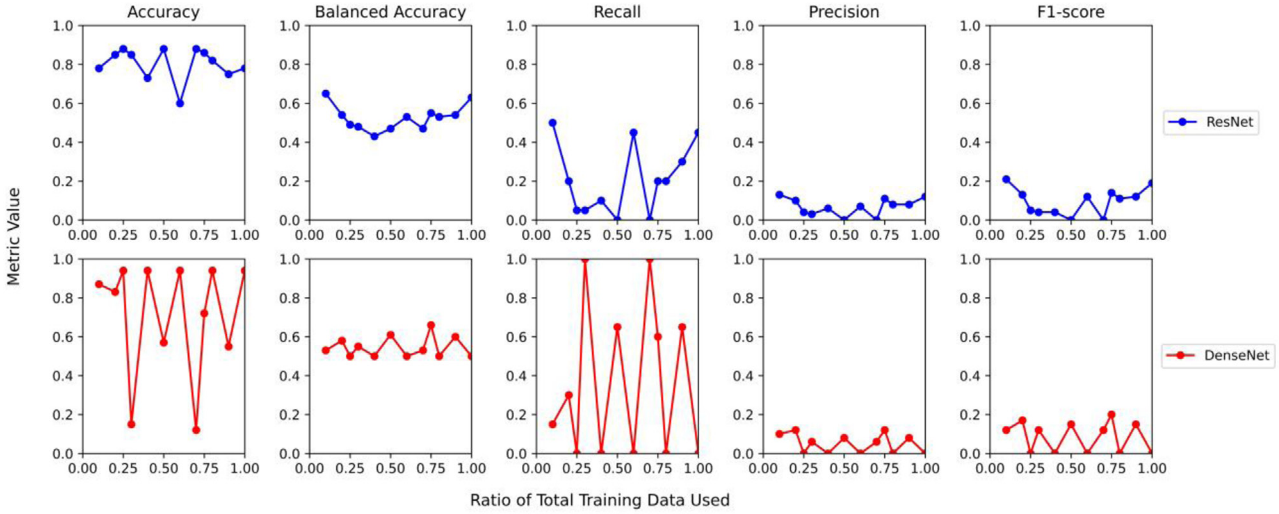

Figure 1 Deep-learning model performance with increasing amounts of training data.

institution to evaluate the DL models' generalizability. Because of the low performance of the DL models, this step was not needed. A test set typically should not be used to evaluate different model iterations such as completed when examining how the DL model performance changes with an increase in the available training data. However, the low performance of the DL models motivated this exploration to determine whether a meaningful prediction model was created.

To our knowledge, this is the first study to examine the feasibility of DL for ORN prediction. Humbert-Vidan et al[23] previously studied the viability of ML techniques for mandibular ORN prediction. In their study, they similarly concluded that the ML models performed similarly for ORN prediction.[23] Both studies have similar test accuracy metrics, but their study had slightly greater recall and precision values.[23] However, direct comparisons are difficult because of the smaller sample size and the different case occurrence rates between the 2 data sets. In this study, for the logistic regression model, the mean dose was the most selected variable. The mean dose was found to be highly associated with ORN development in other studies as well.[12,13]

There are several limitations to this study. First, only dose was used in the models, and additional imaging modalities such as functional magnetic resonance imaging or computed tomography could be included in the future. Furthermore, the population used to construct the models were obtained from a single geographic region and may not be representative of populations in other communities. Finally, an external validation set should be used in the future to determine the generalizability of the ML models.

In the future, more imaging data can be collected for model construction that could potentially benefit the DL approaches. Moreover, future DL architectures may improve the performance of DL on ORN prediction tasks. The use of additional imaging modalities such as functional magnetic resonance imaging also can be explored.

Conclusion

In this work, we compared traditional ML algorithms to DL algorithms for the prediction of mandible ORN resulting from HNC RT. The traditional ML algorithms performed similarly to each other when using cross-validation and were successful at predicting ORN. The performance of the ML models shows promise in clinical integration for future studies. Despite our use of different architectures and model ensembles, the DL models continued to underperform compared to the best-performing ML algorithm identified by cross-validation, logistic regression, when evaluated on the test set. When we used additional training data, no performance improvement trends were evident, suggesting that more data are needed despite the relatively large HNC patient cohort. In further work, researchers could use more subjects, additional imaging data, more imaging modalities, and future DL architectures to improve on this ORN prediction task.

Acknowledgments

We thank the University of Texas MD Anderson Cancer Center Research Medical Library for manuscript editing.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.adro.2022.101163.

References

1. Cramer JD, Burtness B, Le QT, et al. The changing therapeutic landscape of head and neck cancer. *Nat Rev Clin Oncol*. 2019;16:669-683.
2. Pulte D, Brenner H. Changes in survival in head and neck cancers in the late 20th and early 21st century: A period analysis. *Oncologist*. 2010;15:994.
3. Oosting SF, Haddad RI. Best practice in systemic therapy for head and neck squamous cell carcinoma. *Front Oncol*. 2019;9:815.
4. Ang KK, Harris J, Wheeler R, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 2010;363:24.
5. Brook I. Late side effects of radiation treatment for head and neck cancer. *Radiat Oncol J*. 2020;38:84-92.
6. Ortigara GB, Schulz RE, Soldera EB, et al. Association between trismus and dysphagia related quality of life in survivors of head and neck cancer in Brazil. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2019;128:235-242.
7. Togni L, Mascitti M, Vignigni A, et al. Treatment-related dysgeusia in oral and oropharyngeal cancer: A comprehensive review. *Nutrients*. 2021;13:3325.
8. Khojastepour L, Bronoosh P, Zeinalzade M. Mandibular bone changes induced by head and neck radiotherapy. *Indian J Dent Res*. 2012;23:774-777.
9. Chronopoulos A, Zarra T, Ehrenfeld M, et al. Osteoradionecrosis of the jaws: Definition, epidemiology, staging and clinical and radiological findings. A concise review. *Int Dent J*. 2018;68:22-30.
10. Nadella KR, Kodali RM, Guttikonda LK, et al. Osteoradionecrosis of the jaws: Clinicotherapeutic management: A literature review and update. *J Maxillofac Oral Surg*. 2015;14:891.
11. Frankart AJ, Frankart MJ, Cervenka B, et al. Osteoradionecrosis: Exposing the evidence not the bone. *Int J Radiat Oncol Biol Phys*. 2021;109:1206-1218.
12. Aarup-Kristensen S, Hansen CR, Forner L, et al. Osteoradionecrosis of the mandible after radiotherapy for head and neck cancer: Risk factors and dose-volume correlations. *Acta Oncol*. 2019;58:1373-1377.
13. Pereira IF, Firmino RT, Meira HC, et al. Osteoradionecrosis prevalence and associated factors: A ten years retrospective study. *Med Oral Patol Oral Cir Bucal*. 2018;23:e633.
14. Kubota H, Miyawaki D, Mukumoto N, et al. Risk factors for osteoradionecrosis of the jaw in patients with head and neck squamous cell carcinoma. *Radiat Oncol*. 2021;16:1-11.

15. Kuhnt T, Stang A, Wienke A, et al. Potential risk factors for jaw osteoradionecrosis after radiotherapy for head and neck cancer. *Radiat Oncol*. 2016;11:1-7.
16. van Dijk LV, Abusaif AA, Rigert J, et al. Normal Tissue Complication Probability (NTCP) prediction model for osteoradionecrosis of the mandible in patients with head and neck cancer after radiation therapy: Large-scale observational cohort. *Int J Radiat Oncol Biol Phys*. 2021;111:549-558.
17. Rosenfeld E, Eid B, Masri D, et al. Is the risk to develop osteoradionecrosis of the jaws following IMRT for head and neck cancer related to co-factors? *Medicina*. 2021;57:468.
18. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: A survey. *Evol Intell*. 2021;1:1-22.
19. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: A survey. *IET Image Process*. 2022;16:1243-1267.
20. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE*. 2020;109:820-838.
21. Appelt AL, Elhaminia B, Gooya A, et al. Deep learning for radiotherapy outcome prediction using dose data�A review. *Clin Oncol*. 2022;34:e87-e96.
22. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 2021;8:1-74.
23. Humbert-Vidan L, Patel V, Oksuz I, et al. Comparison of machine learning methods for prediction of osteoradionecrosis incidence in patients with head and neck cancer. *Br J Radiol*. 2021;94: 20200026.
24. Tsai CJ, Hofstede TM, Sturgis EM, et al. Osteoradionecrosis and radiation dose to the mandible in patients with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys*. 2013;85:415-420.
25. Beare R, Lowekamp B, Yaniv Z. Image segmentation, registration and characterization in R with SimpleITK. *J Stat Softw*. 2018;86:1-35.
26. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: F Foundation for Statistical Computing; 2021.
27. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28:1-26.
28. Pedregosa F, Michel V, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Machine Learning Res*. 2011;12:2825-2830.
29. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. arXiv:1512.03385 [cs.CV].
30. Huang G, Liu Z, van der Maaten L, et al. Densely Connected Convolutional Networks. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 20162261-2269. 2017-January.
31. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Available at: http://download.tensorflow.org/paper/whitepaper2015.pdf. Accessed July 26, 2022.
32. GitHub - yanyachen/MLmetrics: Machine Learning Evaluation Metrics. Available at: https://github.com/yanyachen/MLmetrics. Accessed August 15, 2022.
33. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:1-8.