Robust Models for Esopohageal Toxicity Prediction from Radiation Dose Maps

Matthew Gil[1,2], William H Nailon[2,3], Stephen Harrow[2], Paul Murray[1], and Stephen Marshall[1]

[1]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom
[2]Edinburgh Cancer Centre, NHS Lothian, Edinburgh, Scotland
[3]School of Engineering, University of Edinburgh, Edinburgh, Scotland

Abstract
Esophageal toxicity is a common group of adverse events caused by radiotherapy treatment in the chest region. Previous work has shown that dose based prediction models are possible but there has been a lack of testing of the robustness of these models. We use patient dose information and outcomes (N=397) from the RTOG-0617 clinical trial to create esophageal toxicity prediction models using an ANN and LSBoost achieving an AUC of 0.707 and 0.691 for the prediction of grade ≥ 3 and grade ≥ 2 esophageal toxicity respectively. A MAE of 0.884 was achieved when predicting exact toxicity grades. The robustness of the models was evaluated by applying random noise to the test sets and observing the degradation of performance. Methods to improve the robustness including L2 regularisation, SMOTE, random noise-based data augmentation and ensemble model approaches were shown to be effective for different models.

1 Introduction

Esophageal toxicity often occurs after the treatment of lung cancer, or other cancers, by external beam radiotherapy (RT) with esophagitis being the most common esophageal toxicity. Many studies have been produced aiming to predict and limit the occurrence and severity of esophageal toxicity with older studies generally aiming to establishing radiation dose volume effects [1] and more recent studies generally aiming to build machine-learning based models to predict esophageal toxicity [2]. These recent studies generally report the AUC ROC for grade ≥ 2 or ≥ 3 toxicity which usually falls in the range of 0.6 to 0.75 when only using dose features [3].
This study applies two popular models, artificial neural networks (ANNs) and boosted decision trees, for the prediction of esophageal toxicity from dose and clinical information using data available from the RTOG-0617 clinical trial [4]. The robustness of these models was evaluated by applying random noise to the test data with incrementally increasing magnitude. Data augmentation by SMOTE and the application of random noise during training were used to increase model robustness as well as increasing the L2 regularisation of the ANN and combining the models as an ensemble to produce final recommendations for building robust toxicity prediction models.

2 Materials and Methods

Unless mentioned otherwise, all computational methods were applied using MATLAB R2022b v9.13.0 (The Mathworks Inc).

2.1 Dataset

This study uses data from the RTOG-0617 clinical trial [4], a multi-centre trial investigating the effects of standard (60 Gy) versus high (74 Gy) doses of radiation to treat lung cancer patients as well as the effects of using cetuximab during treatment. This is, to the authors knowledge, currently the largest public RT dataset with the necessary information for our analysis. Patients in the RTOG-0617 trial received either IMRT or 3DCRT from one of 185 different institutions in the USA and Canada. Of the 544 patients initially recruited to the RTOG-0617 trial, we excluded 147 patients from our study due to non-completion during the RTOG-0617 trial (N=49), an overall survival of less than 6 months post RT (N=57) or due to issues with the radiotherapy data (N=41) leaving N=397 patients for our study. The data from the RTOG-0617 study is available and was accessed through The Cancer Imaging Archive (TCIA) [5], [6].

Toxicity Grade	0	1	2	3	4	5
Number of Patients	84	99	148	62	2	2

Table 1: The number of patients that developed a maximum toxicity for each toxicity grade (as defined by CTCAE v3 [7]).

The following esophageal toxicities, as defined by CTCAE v3 [7] (the version used in the RTOG-0617 study), were considered; esophagitis, acquired tracheo-esophageal fistula, dysphagia, dyspepsia, esophageal ulcer, esophageal stenosis and esophageal perforation. The number of patients that developed an esophageal toxicity for each CTCAE grade is displayed in table 1.

2.2 Dose and Clinical Features

Dose-volume histogram features were calculated from the RT planning dose maps isolated to the esophagus volume that was available from the RT planning contours. The dose features calculated were the V_x (the percent of the total volume receiving over x Grays), the Vol_x (the total volume receiving over x Grays), the mean esophageal dose and the max esophageal dose. The V_x and Vol_x were calculated for the range 5-80 Gy in steps of 5 Gy. Available clinical features were also included, these features were; age, gender, Zubrod performance, carcinoma type, RT technique (IMRT or 3DCRT), smoking status and the primary tumour volume.

All features were converted to z-scores (mean = 0, std dev =1) before model training.

2.3 Model Training and Hyperparameter Tuning

The models investigated were an ANN and the LSBoost algorithm [8] for boosted decision trees, both using the features from section 2.2 as the model input. The ANN consists of an input layer, fully connected layer, activation layer, fully connected layer and an output layer connected in order. The models were trained as regression models where the output is a continuous variable that correlates to the expected toxicity grade. After training, both model outputs were corrected for variance bias by setting the mean and variance of the model predictions on the training set equal to the mean and variance of the training set ground truth by applying the following equation:

$$\hat{y}_C = (\hat{y} - mean(\hat{Y}_{train})) * \sqrt{\frac{var(Y_{train})}{var(\hat{Y}_{train})}} + mean(Y_{train})$$

Where Y_{train} and $\hat{Y}_{train}$ are the set of all ground truth and model predictions for the training set respectively, $\hat{y}$ is a new prediction and $\hat{y}_C$ is the corrected new prediction. The models were evaluated as an exact grade prediction model and as a binary classifier by defining a cutoff value, 1.5 for grade ≥ 2 and 2.5 for grade ≥ 3 classification. The models were evaluated using the AUC for both grade ≥ 2 and grade ≥ 3 binary classification and the mean average error (MAE) for the exact grade output. Additionally, the ANN and LSBoost models were combined as an ensemble model by taking a weighted average of the two model outputs where the weighting was a tunable parameter.

Test Data Split A 4-fold data split was taken where one fold was held back for testing and the other three folds were used for training and hyperparameter tuning. The training and hyperparameter tuning was repeated four times, leaving out a separate fold for testing each time so that testing could be repeated for the full dataset.

Training and Hyperparameter Tuning Data Split The data in the three folds left out for training and hyperparameter tuning were again split into four folds. One of these folds was used for hyperparameter tuning and the others were used for training. The loss function used during training for both models was the mean square error (MSE). Cross-validation was performed so the hyperparameter tuning was repeated on each fold and the average hyperparameters were used to train the final model on all of the folds (not including the test data). Hyperparameter tuning was applied by Bayesian optimisation to minimise the MSE on the validation set. The hyperparameters tuned and their average values are given in section 3.1.

2.4 Robustness Tests

To test the robustness of the models, random noise was incrementally added to the features of the test set and the degradation of model performance in terms of the AUC for grade ≥ 3 toxicity and the MAE was recorded. For all features in the test set, random noise, X, was added to feature F converting it to feature F_N by $F_N = F + X$. Here $X \sim U(-\alpha, \alpha)$ where $U(-\alpha, \alpha)$ is a uniform distribution with maximum and minimum values α and $-\alpha$. As the features had been converted to z-scores, a value of $\alpha = 1$ means that the noise has a range of 1 standard deviation in the positive and negative directions. The noise was added incrementally with α increasing in steps of 0.1 from 0.0 to 5.0. For all robustness tests, no hyperparameter tuning was performed and the average hyperparameter values given in section 3.1 were used.

2.4.1 Improving Robustness

Techniques were employed to improve the robustness of the ANN and LSBoost models. Again the AUC for grade ≥ 3 classification and the MAE for the exact grade prediction were calculated.

L2 Regularisation: L2 regularisation is known to increase model robustness by forcing a model to rely on features more evenly, reducing overfitting. The robustness test was repeated for the ANN with different values of λ, the L2 regularisation term.

SMOTE: Synthetic minority oversampling technique (SMOTE) [9] is commonly applied in radiotherapy prediction models [2] to improve model performance by generating synthetic training data. SMOTE was applied to the training data to oversample the underrepresented grades. Grade 2 was the most common grade for esophageal toxicity so SMOTE was applied to oversample grade 0, 1 and 3 cases to be as represented as grade 2 cases. SMOTE was not applied to the grade 4 and 5 cases as there were only two of each case in the full dataset so it is not possible to oversample these cases. The performance of the ANN and LSBoost model was recorded as noise was applied to the model after training with SMOTE.

Adding Noise During Training: The training dataset was expanded by adding random noise to the training data features to generate new training examples. This was applied using the equation $F_N = F + X$, described in section 2.4, with an α value of 0.2. The training dataset was expanded to 5 times its original size using this technique. This was done for both the ANN and LSBoost models both on its own and after the application of SMOTE.

Ensemble Model: Using a weighted ensemble, as detailed in section 2.3, was additionally investigated as a method for increasing robustness to noise.

3 Results

3.1 Performance Metrics and Hyperparameter Tuning

The results of the model training with hyperparameter tuning are displayed in Figure 2. The hyperparameter values most

Model	AUC (Grade $\geq$ 3)	AUC (Grade $\geq$ 2)	MAE
ANN	**0.709**	0.690	0.894
LSBoost	0.682	0.669	0.925
Ensemble	0.707	**0.691**	**0.884**

Table 2: Results for the hyperparameter tuning of the ANN and LSBoost models.

commonly selected by the Bayesian hyperparameter tuning for both the ANN and LSBoost models are given below. **ANN Hyperparameters:** L2 regularisation $\lambda = 0.1$, fully connected layer size = 20, activation function = sigmoid, loss gradient tolerance = $1e^{-4}$. **LSBoost Hyperparameters:** Maximum number of splits = 2, minimum leaf size = 2, minimum parent size = 5, number of learning cycles = 400, learning rate = 0.01. **Ensemble Weighting:** The weighted average ensemble weighted the ANN 4 times higher than the LSBoost model on average. Further details regarding the hyperparameters can be found in the MATLAB R2022b files for the functions *'fitensemble()'* and *'fitrnet()'* for the LSBoost and ANN models respectively.

3.2 Robustness Tests

The results of the robustness tests are displayed in figures 1 to 5. Figures 2 and 1 show that increasing L2 regularisation increases robustness in terms of both MAE and AUC for the ANN model. Figure 3 shows that using SMOTE or added noise reduces the performance of the ANN in terms of the MAE. Figure 4 shows an improvement to the robustness of the LSBoost model in terms of the MAE. The best-performing ANN and LSBoost models were selected as the final models. No observable change was observed in the robustness of the ANN or LSBoost model in terms of the AUC when applying SMOTE or noise during training so the figures are not included here. The final ANN model uses a high L2 regularisation term of $\lambda = 0.95$ and only the original data during training. The final LSBoost model uses both SMOTE and added noise during training. The AUC and MAE responses of these final models as well as their weighted ensemble are displayed in figures 5 and 6 respectively.

4 Discussion

We have shown that a model for the prediction of esophageal toxicity from dose maps can be produced to achieve an AUC of 0.707 and 0.691 for the classification of grade ≥ 3 and grade ≥ 2 respectively with a MAE of 0.884 when trying to predict exact toxicity grade. The ANN beats the LSBoost model on all metrics when no test set noise is added. The ANN also outperforms the LSBoost model for all robustness tests provided there is a high level of L2 regularisation present. However, using a weighted ensemble of the two models provides benefit from the LSBoost model by increasing the model robustness to noise in terms of the MAE while not impacting the AUC. To improve the robustness of the models it was observed that a high level of L2 regularisation was beneficial for the ANN and, for the LSBoost model, using SMOTE and adding noise during training was beneficial. L2 regularisation was not available for the LSBoost model but may also be beneficial if implemented.

Many of the improvements to the robustness of the models would not be apparent if the tests to add noise were not applied. For example, in Figure 1, where $\alpha = 0$ and no noise is added, all of the models perform equally. A sharp divergence is then observed for lower values of λ with added noise. This highlights the benefit of these robustness tests for reducing overfitting and building robust models.

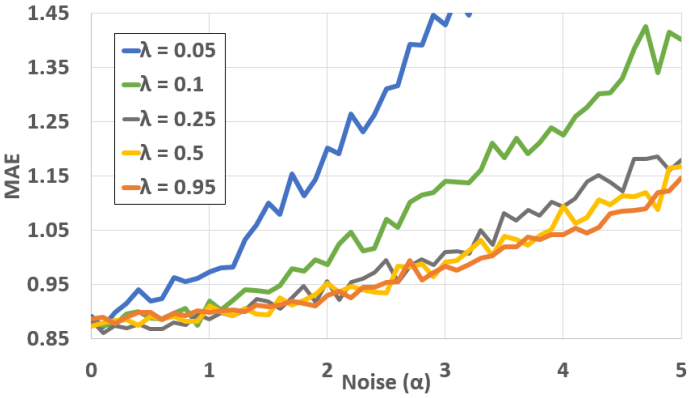

Figure 1: MAE plots for the ANN L2 regularisation tests.

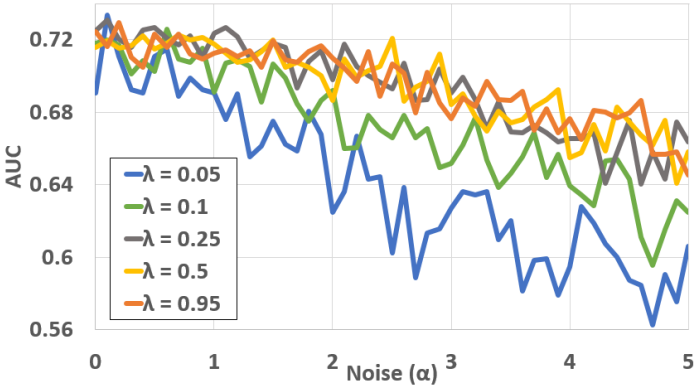

Figure 2: AUC plots for the ANN L2 regularisation tests.

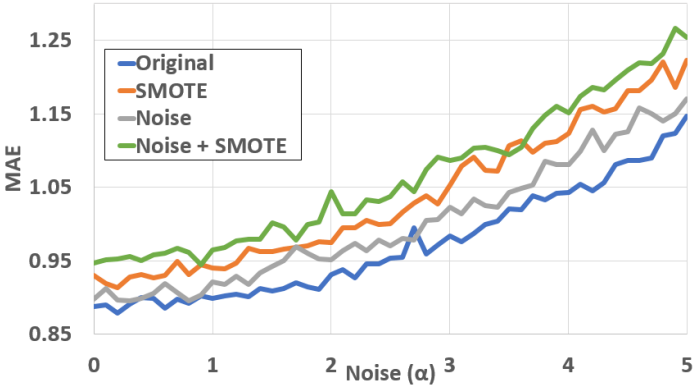

Figure 3: MAE plots for the ANN with SMOTE and noise added.

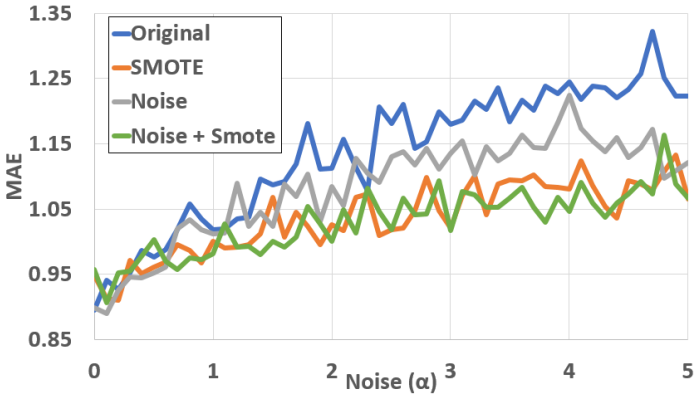

Figure 4: MAE for the LSBoost tests with SMOTE and noise.

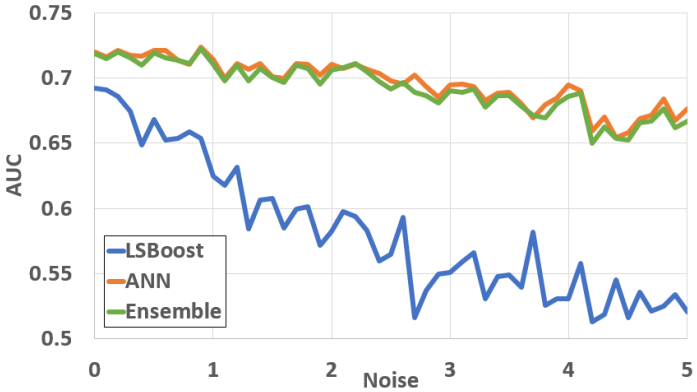

Figure 5: Final LSBoost, ANN and ensemble AUC plots.

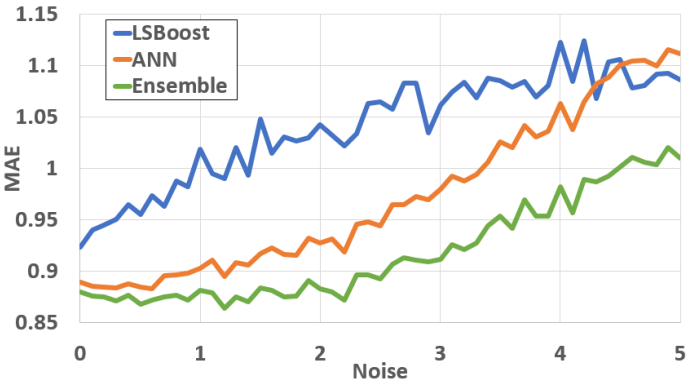

Figure 6: Final LSBoost, ANN and ensemble MAE plots.

This work may be expanded in the future to investigate the application of systematic errors to the test data to study robustness by altering the esophagus volume to simulate contouring errors or by randomly translating the esophagus volume to simulate positioning errors. Additional data augmentation techniques, such as the use of GANs to generate data, and their effect on robustness may also be an interesting avenue of further research.

5 Conclusion

ANN and LSBoost models were trained to predict esophageal toxicity from dose maps using the RTOG-0617 dataset. A weighted ensemble model formed of these two models achieved an AUC of 0.707 and 0.691 for classification of grade ≥ 3 and grade ≥ 2 toxicity respectively and a MAE of 0.884 when predicting exact toxicity grades. Model robustness was evaluated by adding random noise to the test data. This showed that for the ANN model, it is beneficial to include a high level of L2 regularisation and for the LSBoost model it is beneficial to use SMOTE and the application of random noise to augment the training dataset. Finally, it was shown that a weighted ensemble of the LSBoost and ANN approaches can further improve the model robustness.

Acknowledgements

This manuscript was prepared using data from Datasets (RTOG-0617; NCT00533949-D1, D2, D3) from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data was originally collected from clinical trial NCT number NCT00533949, titled "A Randomized Phase III Comparison of Standard-Dose (60 Gy) Versus High-Dose (74 Gy) Conformal Radiotherapy With Concurrent and Consolidation Carboplatin/Paclitaxel +/- Cetuximab (IND #103444) in Patients With Stage IIIA/IIIB Non-Small Cell Lung Cancer". All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, or the NCI.
This work was funded by the Engineering & Physical Sciences Research Council (EPSRC) (grant reference: EP/S022821/1) and the Beatson Cancer Charity (grant reference: 19-20-043).

References

[1] M. Werner-Wasik, E. Yorke, J. Deasy, et al. "Radiation dose-volume effects in the esophagus". *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S86–S93.

[2] L. J. Isaksson, M. Pepa, M. Zaffaroni, et al. "Machine learning-based models for prediction of toxicity outcomes in radiotherapy". *Frontiers in Oncology* 10 (2020), p. 790.

[3] X. Zheng, W. Guo, Y. Wang, et al. "Multi-omics to predict acute radiation esophagitis in patients with lung cancer treated with intensity-modulated radiation therapy". *European Journal of Medical Research* 28.1 (2023), pp. 1–10.

[4] J. D. Bradley, R. Paulus, R. Komaki, et al. "Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study". *The lancet oncology* 16.2 (2015), pp. 187–199.

[5] K. Clark, B. Vendt, K. Smith, et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". *Journal of digital imaging* 26 (2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7.

[6] J. Bradley and K. Forster. "Data from NSCLC-Cetuximab". *The Cancer Imaging Archive* (2018). DOI: 10.7937/TCIA.2018.jze75u7v.

[7] A. Trotti, A. D. Colevas, A. Setser, et al. "CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment". *Seminars in radiation oncology*. Vol. 13. 3. Elsevier. 2003, pp. 176–181.

[8] T. Hastie, R. Tibshirani, J. H. Friedman, et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al. "SMOTE: synthetic minority over-sampling technique". *Journal of artificial intelligence research* 16 (2002), pp. 321–357.