Predicting Cellular Response to Ionizing Radiation through Machine Learning: Random Forest and Feed-Forward Neural Network Models

Ali Abu Shqair , Dong-Seok Lee, Eun-Hee Kim*
Department of Nuclear Engineering, Seoul National University, Seoul 08826, Republic of Korea
**Corresponding author: eunhee@snu.ac.kr*

***Keywords :** Ionizing radiation, Machine Learning, Surviving fraction model, PIDE,

1. Introduction

The cellular response to radiation exposure is a complex process involving critical mechanisms like DNA damage, damage repair, and cell survival or death. The Linear-Quadratic Model (LQM) is widely used to describe the relationship between radiation dose and cell surviving fraction (SF), characterized by the α and β coefficients for the linear and quadratic terms, respectively. However, applying this model across different types of radiation and varying cellular conditions remains a significant challenge.

Advances in data-driven methodologies, particularly machine learning, have opened new possibilities for addressing these challenges. The Particle Irradiation Data Ensemble (PIDE) database [1, 2], provides a comprehensive collection of experimental data on cell SF after exposure to different types of ionizing radiation. The PIDE database was utilized to train ML algorithms for predicting cell SF after exposure to different types of ionizing radiation under various physical and biological conditions of experiments.

Two ML algorithms were trained using the PIDE database: Random Forest (RF), and Feed-Forward Neural Network (FFNN). The RF model, an ensemble learning method, enhances prediction accuracy by averaging multiple decision trees [3]. RF mitigates overfitting by utilizing bootstrap aggregation (bagging) where the ensemble uses the same training algorithm for different predictors, each trained on random subsets sampled with replacement, allowing overlap. Additionally, RF provides feature importance rankings, allowing insight into the relative contribution of different variables to the prediction.

The FFNN, a fully connected deep learning model, processes data sequentially through input, hidden, and output layers [4]. FFNN learns complex patterns and improves its prediction performance through backpropagation and weight adjustments during the training process. The use of hidden layers and neurons and their activation functions in FFNN determines its performance of capturing nonlinear relationships.

2. Methods

2.1 Data preparation

The PIDE dataset was restructured to maintain a consistent format so that features and the cell SF measurement of each experiment are represented in a single row. The resulting dataset was processed to remove inconsistencies such SF values exceeding 1, and negative radiation dose values or α and β coefficients. The dataset was further processed by removing outliers using the z-score method, eliminating rows with feature values that deviated more than 2.65 standard deviations from the mean.

Before training, categorical features were transformed to numerical values, and all features were subsequently normalized to scale their values between 0 and 1. This preprocessing step mitigated the dominance of features with larger magnitudes and facilitated faster model convergence during training.

Table 1 summarizes the selected features and labels used in the model training process. The final dataset included about 728 experiments with varying SF(D) data points for each experiment.

Table 1. Physical and biological features of experiments in PIDE database used for training the ML models.

Biological features	**Cells**	cell lines
	Cell class	tumor (t) or normal (n) cells
	Cell origin	human (h) or rodent (r) cells
	Cell cycle	unsynchronized (u) or synchronized (s) cells
	DNA content	DNA content of cell
Physical features	**Radiation source**	ion mass in amu
	Radiation energy	monoenergetic (m) or SOBP (s)
	LET	linear energy transfer of radiation
	Cellular dose	absorbed dose in Gy
Biological response parameter	**SF**	cell surviving fraction

2.2. ML model structure

2.2.1. Random Forest

The optimal Random Forest model consisted of 300 decision trees, with no limit on tree depth and a minimum of two samples required to split an internal node. Each leaf node contained at least one sample, preventing empty terminal nodes. The algorithm employed bootstrap sampling (sampling with replacement), allowing individual data points to be selected multiple times within different subsets. Additionally, the algorithm considered all available input features for determining the best split at each node.

2.2.2. Feed-Forward Neural Network

Several model architectures were tested, and hyperparameters were iterated to determine the optimal model structure that achieves the highest prediction precision. The optimal NN was a fully connected network consisting of an input layer with 9 neurons for the 9 input features, two hidden layers each consisting of 2048 neurons with the *Relu* activation function *(max(0,x))*, and an output layer consisting of a neuron that used the *Sigmoid* activation function *(1/(1+exp(-x))*. The architecture of the final FFNN model is shown in Figure 1.

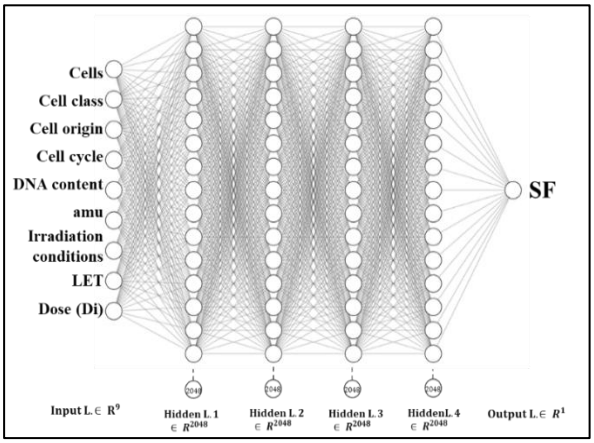

Fig. 1. Architecture of the FFNN developed and trained on the PIDE data.

2.3. Cross validation evaluation

The cross-validation process in this study was configured to divide the data into five folds, each containing mutually exclusive sets of groups, and to derive five distinct models. Four folds of data among five were used for training and deriving a prediction model with the remaining fold reserved for testing the prediction model. Five different choices of testing fold resulted in five different prediction models.

3. Results

The RF and the FFNN models predicted the cell SF for validation data sets with a reasonable precision. Evaluation metrics of predictions (R^2 and MSE) were similar with a slightly better R^2 of SF predictions by the RF than FFNN model. The relative biological effectiveness (RBE_{10} denoting RBE at SF=0.1) values were derived from the predicted SF curves for different experiments. The derived RBE_{10} values were comparable to the predictions based on the local effect model (LEM) [5] with a slightly better R^2 by FFNN than RF model. (Figs. 2 and 3).

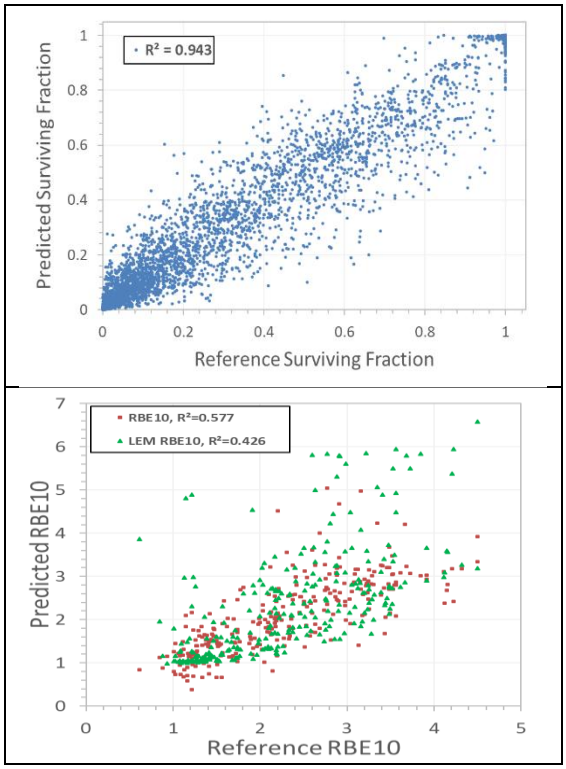

Fig. 2. Correlations between RF-based SF predictions and the reference SFs (upper) and between the RBE_{10} values derived from SF predictions or from LEM with the reference RBE_{10} values (lower), for testing the developed RF model.

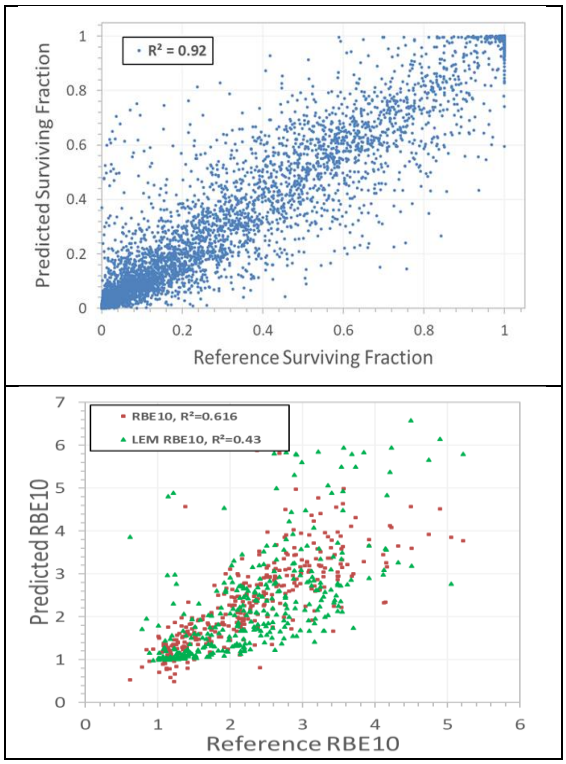

Fig. 3. Correlations between FFNN-based SF predictions and the reference SFs (upper) and between the RBE_{10} values derived from SF predictions or from LEM with the reference RBE_{10} values (lower), for testing the developed FFNN.

4. Conclusion

Both RF and FFNN models were able to predict with reasonable accuracy the SF curves and RBE_{10} from experiments, which were not realized at the stage of model training. Compared to traditional machine learning models such as RF, the FFNN required extensive tuning of hyperparameters, architecture, and optimization strategies to improve its predictive accuracy on unseen data (generalization). The FFNN will be the main scheme for modelling data of complexity in radiation biology study due to its scalability in handling more complex problems, and adaptability for customization.

REFERENCES

[1] T. Friedrich, U. Scholz, T. ElsäSser, M. Durante, and M. Scholz, "Systematic analysis of RBE and related quantities using a database of cell survival experiments with ion beam irradiation," Journal of radiation research, vol. 54, no. 3, pp. 494-514, 2013.

[2] T. Friedrich, T. Pfuhl, and M. Scholz, "Update of the particle irradiation data ensemble (PIDE) for cell survival," Journal of Radiation Research, vol. 62, no. 4, pp. 645-655, 2021.

[3] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.

[4] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural networks, vol. 2, no. 5, pp. 359-366, 1989.

[5] T. Friedrich, M. Durante, and M. Scholz, "The local effect model-principles and applications," The Health Risks of Extraterrestrial Environments, vol. 1, pp. 1-14, 2013.