Penn Medicine

Harnessing Pre- and Intra-Operative Predictors to Predict Postoperative Complications in Spine Deformity: Development of a Prognostic Artificial Intelligence Clinical Prediction Model

Mert Marcel Dagli, MD; Ali K Ozturk, MD; Jang W Yoon, MD; William C Welch, MD

June 10, 2024

Disclosures

None

Agenda

BACKGROUND

METHODS

RESULTS

CONCLUSIONS

Background

Posterior spinal fusion surgery (PSF) for adult spinal deformity correction (ASD) is highly complex, with patients at elevated risk of post-operative complications

Advanced machine learning (ML) algorithms have immense potential to support clinical decision-making and optimize outcomes prediction

Purpose: Development and Validation of predictive models using common machine learning approaches for post-operative complications following multi-level thoracolumbosacral PSF for ASD

Methods

PART 1

Guidelines: TRIPOD & STROBE

Data Source: retrospective review, institutional deformity database (DAC, charts)

Inclusion criteria: TLS PSF with 6+ vertebrae for ASD (n=646)

Data Cleaning and Handling of Missingness with Iterative Imputer with Random Forest and random states

Predictor Selection:

- Candidate predictors identified: literature, clinical importance, and analysis
- 12 predictors: age (years), gender, PSF levels, preop Hb {g/dL}, 3-column osteotomy (3CO) use, posterior column osteotomy (PCO) use, BMI, sacral involvement, ASA score, tranexamic acid (TXA) use, estimated blood loss (EBL) Total, operative time (minutes)

Methods

PART 2

Data split: train-test-val 70-15-15

Models: Artificial Neural Network (ANN), ensemble boosted ANN (ebANN), extreme gradient boosting (XGBoost), Support Vector Machine (SVM), Explainable Boosting Machine (EBM), Random Forest (RF), Logistic Regression (LR)

Customization, grid search, regularization, fine-tuning, etc.

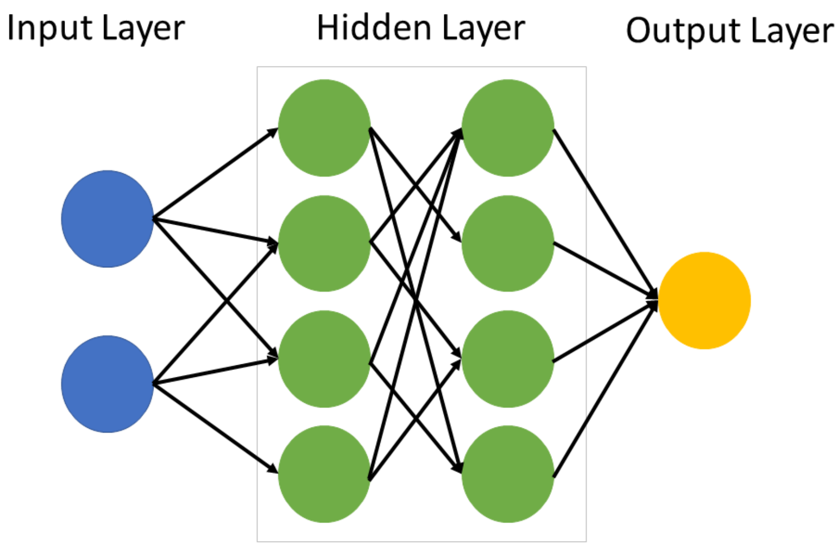
Performance metrices: accuracy, sensitivity, AUC-ROC (c-score), F1-score, precision (PPV) + Bootstrapping

Methods

PART 3 – MODEL ANN

ANN

- Package: PyTorch
- Standardization of features
- Batch size: 128
- Epochs: 200
- Patience: 20 (early stopping)
- Criterion: Binary Cross Entropy (BCE) loss
- Optimizer: AdamW (learning rate 0.01, weight decay 0.1)
- 4 layers: 2 layers input and output, 2 hidden layers, 12-128-64-1 with 20% and 10% dropout between layers 2-3 and 3-4



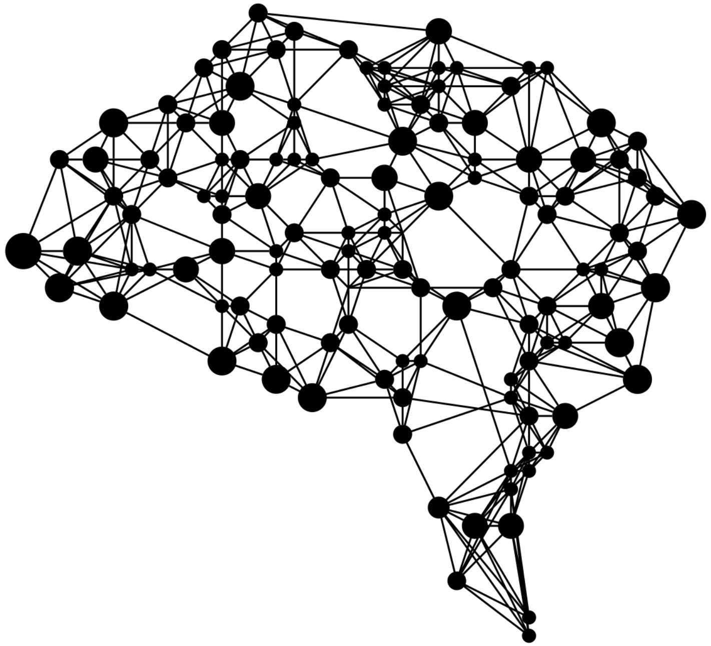
Artificial Neural Networks

Methods

PART 4 – MODEL EBANN

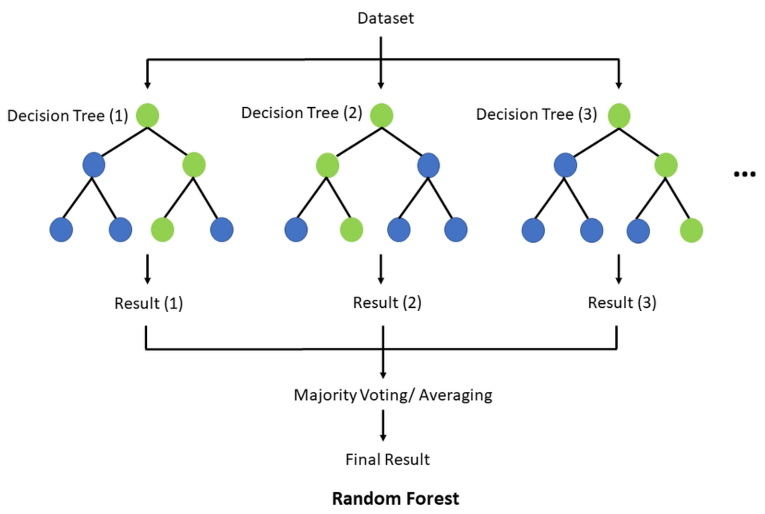
ebANN

- Package: PyTorch
- Ensemble ANNs with previous architecture
- Boosting rounds: 5 (iterative grid search range)



Methods

PART 4 – MODELS XGBOOST AND RF



Random Forest

XGBoost

- Max depth: 30
- Learning rate: 0.1
- Number of rounds: 1000
- Automatic early stopping (XGBoost)

RF

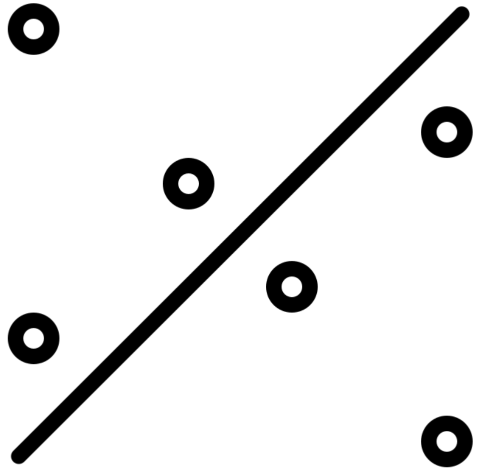
- 3 folds each 216 different parameter combination totaling grid search of 648 models
- Samples for each tree randomly (bootstrap: True).
- Didn’t limit how deep the trees could grow (max_depth: None).
- At least 2 samples to be at a leaf node before stopping (min_samples_leaf: 2).
- At least 10 samples to split a decision node (min_samples_split: 10).
- Forest with 300 trees (n_estimators: 300).

Methods

PART 5 – MODEL SVM

SVM

- Tested 720 different setups across 5 rounds (folds) each, totaling 3600 checks.
- Best setup found:
- 'C' set to 0.01, adjusting the model's sensitivity to the data.
- 'class_weight' set to 'balanced', ensuring equal focus across all classes.
- 'coef0' at 0.5, influencing the model's learning function.
- 'degree' at 4, setting the complexity of the model.
- 'gamma' on 'scale', allowing automatic adjustment based on the data.
- 'kernel' chosen as 'poly', determining how the model views data relationships.



Methods

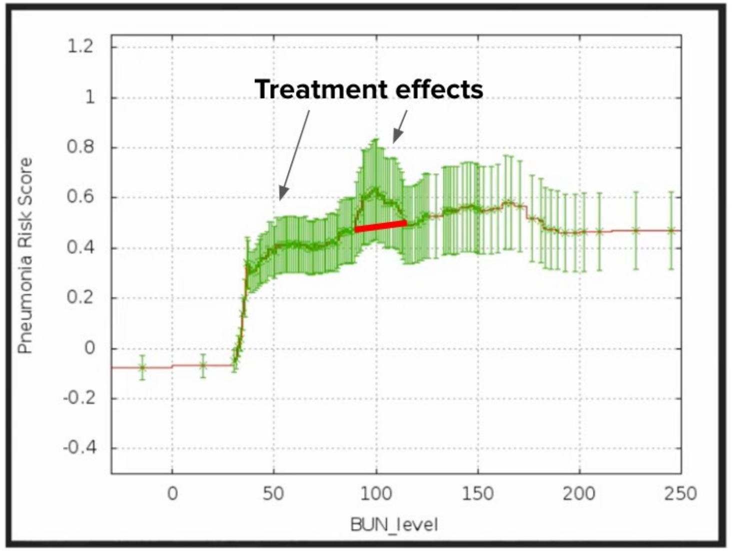
PART 6 – MODEL EBM AND LR

EBM

- Setup simple, minimal need for hyperparameter tuning (to some extent)
- Automatic out of the box: missing values, categorical feature handling, monotonically constrained features, etc.

LR

- Simple logistic regression



Methods

STEP 7 – EVALUATION OF INTERNAL VALIDATION AND PERFORMANCE

Evaluation of confusion matrix to evaluate predictions (TN, TP, FN, FP) is a form of internal validation.

Internal validation with testing set

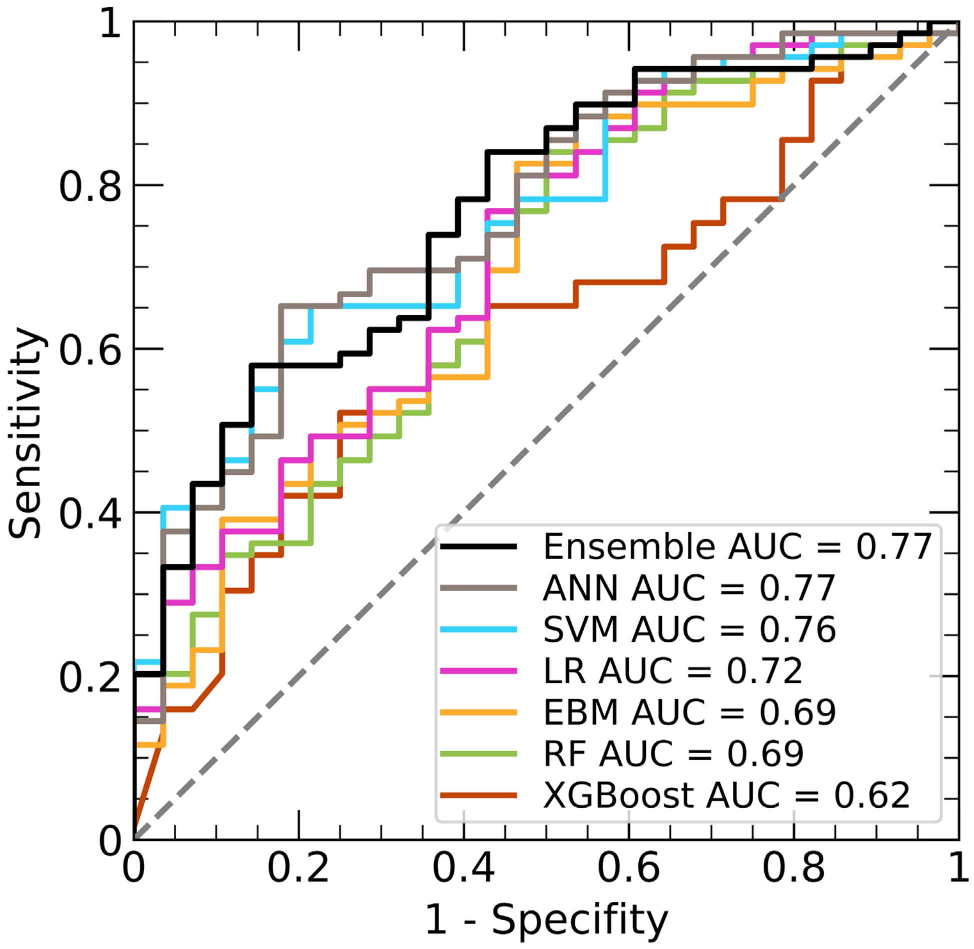
Selection of standard performance metrices derived from testing set:

- Sensitivity: True Positive Rate
- Specificity: True Negative Rate
- Discrimination AUC-ROC: model’s ability to discriminate between classes
- F1-Score: Harmonic mean between precision and sensitivity (minimizing FP, max. TP)
- Precision (PPV): ratio of TP predictions : all positive predictions

Results

Best balanced model: ebANN

	Value (95% CI)
Accuracy	0.7732 (0.6905-0.8553)
Sensitivity	0.942 (0,8823-0.9863)
AUC-ROC	0.766 (0.6568-0.8553)
F1-Score	0.8553 (0.7938-0.9104)
Precision (PPV)	0.7831 (0.6897-0.8706)



Conclusions

A spectrum of explainable and opaque models to develop the best predictive model, while also including information on reasoning, is important and guides future research

Not all AI models automatically perform better than simple LR (multifactorial)

AI Prediction Models outperformed simple LR and can accurately predict postoperative complications after multi-level TLS PSF for ASD

The role of AI Prediction Models in surgery is very promising

Penn Medicine