A LUPI distillation-based approach: Application to predicting Proximal Junctional Kyphosis

Yun Chao Lin YCL2112@COLUMBIA.EDU
Department of Computer Science
Columbia University
New York, NY, USA

Andrea Clark-Sevilla AOC2111@COLUMBIA.EDU
Department of Computer Science
Columbia University
New York, NY, USA

Rohith Ravindranath ROHITHR@STANFORD.EDU
Department of Ophthalmology, Byer Eye Institute
Stanford University
Palo Alto, CA, USA

Fthimnir Hassan FH2444@CUMC.COLUMBIA.EDU
Department of Orthopaedic Surgery Columbia University
New York, NY, USA

Justin Reyes JLR2268@CUMC.COLUMBIA.EDU
Department of Orthopaedic Surgery Columbia University
New York, NY, USA

Joseph Lombardi JML2285@CUMC.COLUMBIA.EDU
Department of Orthopaedic Surgery Columbia University
New York, NY, USA

Lawrence G. Lenke LL2989@CUMC.COLUMBIA.EDU
Department of Orthopaedic Surgery Columbia University
New York, NY, USA

Ansaf Salleb-Aouissi AS2933@COLUMBIA.EDU
Department of Computer Science
Columbia University
New York, NY, USA

Abstract

We propose a learning algorithm called XGBoost+, a modified version of the extreme gradient boosting algorithm (XGBoost). The new algorithm utilizes privileged information (PI), data collected after inference time. XGBoost+ incorporates PI into a distillation framework for XGBoost. We also evaluate our proposed method on a real-world clinical dataset about Proximal Junctional Kyphosis (PJK). Our approach outperforms vanilla XGBoost, SVM, and SVM+ on various datasets. Our approach showcases the advantage of using privileged information to improve the performance of machine learning models in healthcare, where data after inference time can be leveraged to build better models.

1. Introduction

Proximal junctional kyphosis (PJK) is a postoperative complication that occurs relatively frequently in the adult spinal deformity (ASD) population. While PJK is defined differently throughout literature, the most commonly used criterion is a change Δ in the proximal junctional sagittal angle (PJA) $> 10°$ postoperatively, in addition to having an absolute value of the postoperative PJA $> 10°$ Glattes et al. (2005) (see Figure 1). PJA is defined as the Cobb angle formed by the caudal endplate of the upper instrumented vertebrae (UIV) and the cephalad endplate of the vertebral body two levels cephalad to the UIV.

Depending on the definition used and the population studied, PJK is most often reported to occur in 17% to 46% of patients Glattes et al. (2005); Kim et al. (2007, 2008); Bridwell et al. (2013); Cho et al. (2013); Yagi et al. (2011); Kim et al. (2012); Lau et al. (2014); Kim and Iyer (2016). Not only does the incidence of PJK vary significantly, but so does the degree of symptoms patients report. While some patients will meet radiographic criteria for PJK and remain completely asymptomatic, others will experience significant impairment and require revision surgery. PJK requiring surgical intervention as a result of neurological deficits, pain, unacceptable kyphosis, or poor self-image is therefore termed proximal junctional failure (PJF). The rate of PJF is much lower than PJK and is reported in 1.4% and 5.6% of patients Hart et al. (2013); Yagi et al. (2014). Given the complexity of PJK, higher computational modeling is necessary to advance our understanding of its pathologic course.

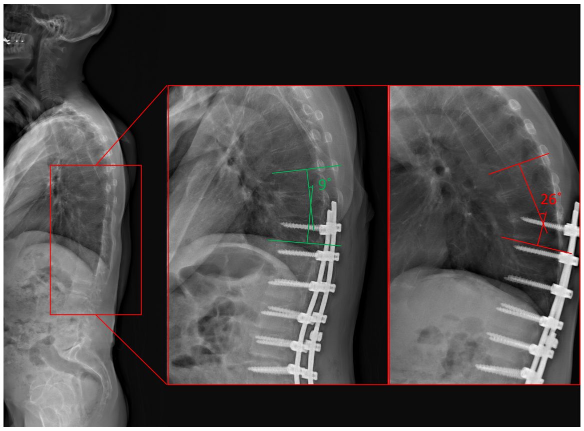

Figure 1: Example of patient with PJK. (Left) baseline (no hardware), (Middle) immediate postop image for patient with no PJK, and (Right) 2.5 years follow up image of patient with PJK.

Machine learning (ML) is becoming increasingly popular in spine surgery, given its capacity for nonlinear learning and predictive analytics (Lee et al. (2020); Lopez et al. (2022); Kang et al. (2013); Scheer et al. (2016); Peng et al. (2020)). Using ML to address the problem of PJK improves the surgical understanding of this problematic phenomenon and improves patient outcomes. However, addressing this problem is difficult for the following reasons: (i) PJK surgeries are infrequent every year; thus, there is limited patient data available; (ii) PJK is center-specific as the rate of PJK development also depends on the selection of the upper instrumented vertebra (UIV) and any prophylactic measures intra-operatively (i.e., transition rods, tethers, etc.), and (iii) only a subset of the measurements/features can be used for training ML models since only a limited number of features are available during inference time. Others have tried to tackle this problem with mild success. However, their models are trained against patients with a spinal deformity and thus need to be more generalizable to a broader range of patients. This work addresses the aforementioned problems by incorporating Vapnik and Vashist's learning using the privileged information (LUPI) paradigm first introduced in Vapnik and Vashist (2009). We utilize privileged information - data available at training time but not at inference time - to guide ML modeling to a better decision boundary. LUPI is further described in the subsequent sections, in particular in our PJK application. Using PI, we can incorporate all aspects of the available data. Furthermore, using privileged information as a "teacher" that guides the "learner" (i.e., our model) to a better decision boundary mitigates the need for abundant training examples.

In this paper, we propose a method for predicting PJK using LUPI by extending the XGBoost model to use the distillation framework (XGBoost+). Our results on several datasets demonstrate that XGBoost+ surpasses the baseline XGBoost model in performance. Furthermore, we conduct additional experiments with a well-known longitudinal dataset for predicting liver transplant or death in a cohort of primary biliary cirrhosis (PBC) patients. Our findings reveal that XGBoost+ outperforms its vanilla counterpart and the SVM+ framework (the original method incorporating privileged information) in both examples of PBC and PJK.

Our contributions are as follows: (i) the integration of privileged information into the XGBoost framework, using distillation and (ii) leveraging privileged information in PJK prediction with a significantly larger sample size (366 patients), as compared to other recent studies (e.g., 35 patients in "Chen et al. (2021) and 44 patients in Peng et al. (2020)). Compared to a cohort with a larger sample size, 510 patients in Scheer et al. (2016), our feature set includes a more comprehensive range of 62 features, in contrast to 13 features in Scheer et al. (2016). Our study is also more generalizable since our cohort includes patients who have not undergone Lenke type 5 adolescent idiopathic scoliosis (AIS) correction nor have already been categorized as having adult spinal deformity (ASD)[1].

Generalizable Insights about Machine Learning in the Context of Healthcare

A significant setback for the advancement of machine learning integration in healthcare settings has always been a need for more data at inference time, which we face with our PJK dataset. Our data features fall into several categories - demographics, surgical variables, pre-operative variables (pre-op), immediate postoperative variables (im-post-op), follow-up postoperative variables (fu-post-op), and descriptive variables regarding final classification. However, we only have access to demographics, surgical, pre-op, and im-post-op variables during inference time, a loss of 33 % of the features that are only available post-operatively.

To address this issue, we incorporate privileged features into the XGBoost modeling. Through the LUPI paradigm, we can incorporate the missing 33% of the features. Through our case study, we demonstrate the usefulness of the LUPI paradigm in machine learning

[1] Lenke 5 is a particular type of AIS defined as those patients that have a structural thoracolumbar or lumbar (TL/L) scoliosis (Lenke et al. (2001))

for healthcare applications. Although all healthcare applications can benefit from privileged information, those with limited data benefit the most. If we had an unlimited amount of data points, privileged information wouldn't be needed, as training with the provided data would suffice. However, we face the problem of having only 366 patients for our analysis, and collecting more patient data is difficult due to the limited number of PJK cases that occur in a single hospital. This is compounded by the fact that there is large dimensionality in the radiology domain. Although we are able to reduce the dimensionality, there is still a challenge due to the limited number of cases.

2. Related Work

Predicting Proximal Junctional Kyphosis A broad range of work has been attempted to investigate the risk factors and to develop predictive models for PJK. While the evidence is often mixed, the proposed risk factors associated with increased rates of PJK include older age Park et al. (2017); Liu et al. (2016), lower bone mineral density Hyun et al. (2016), UIV location Hart et al. (2013); Liu et al. (2016); Smith et al. (2013), fixation to the sacrum Liu et al. (2016); Smith et al. (2015), ligamentous resection Cahill et al. (2012); Cammarata et al. (2014), sagittal vertical alignment (SVA) correction > 50 mm Yagi et al. (2011), large preoperative PJA, and both large preoperative values and large changes in thoracic kyphosis Liu et al. (2016), to name a few. Zhao et al. (2018) focuses on identifying primary risk factors for PJK using a cohort of patients with ASD. A logistic regression model was constructed using variables found significant in a univariate analysis to find independent risk factors associated with PJK. The authors highlight that preoperative TLK (thoracolumbar kyphosis, the Cobb angle between the upper endplate of T10 and the lower endplate of L2), LL (lumbar lordosis, the Cobb angle between the upper endplate of L1 and superior end plate of S1) at follow-up, preoperative PT/SS, and PT/SS at follow-up were primary factors for PJK[2]. "Chen et al. (2021) also investigates the risks associated with PJK, using a different cohort of patients where the inclusion criteria was patients who have undergone Lenke type 5 AIS correction. Correlation and receiver operating characteristic curve analyses were performed to screen the parameters for significance and to calculate their thresholds. A survival analysis was performed to examine the differences between the two groups. The authors submit that the postoperative PJA and postoperative thoracic kyphosis (TK) can be used to effectively predict the occurrence of PJK in patients with Lenke type 5 AIS after corrective surgery. Despite the plethora of identified contributing factors, isolating strong individual risk factors remains challenging, especially considering the design of many studies that focus on only a few risk factors. Furthermore, risk factors are often examined in one of three siloed categories, including demographic, surgical, or radiographic parameters.

Concerning predictive models for PJK, Scheer et al. (2016) uses an ensemble of decision trees using the C5.0 algorithm. Peng et al. (2020) utilizes random forest using SMOTE on a cohort of Lenke 5 adolescent idiopathic scoliosis (AIS) patients undergoing long posterior

[2] PT is the pelvic tilt, the angle between the vertical and the line through the midpoint of the sacral plate to femoral heads axis. SS is the sacrum slope, the angle between the horizontal and the sacral plate. The notation PT/SS is the ratio of PT and SS.

instrumentation and fusion surgery. To our knowledge, no work has been done leveraging privileged information for predicting PJK.

Privileged Information Privileged information was first proposed by Vapnik and Vashist (2009). The authors introduce a new relationship within the data: the teacher and the learner, wherein the teacher provides the learner with privileged information in the correcting space, called LUPI (Learning Using Privileged Information). The suggested paradigm was implemented into SVM and called SVM+. Privileged information was used to estimate the error-correcting slack term ζ (Pechyony and Vapnik (2010)). Their experimental results show that SVM+ has a lower error rate than SVM (Vapnik and Vashist (2009)). Further work explored incorporating privileged information into tree-based models and neural networks. Lambert et al. (2018) propose a new LUPI algorithm designed for Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) that uses a heteroscedastic dropout (i.e., dropout with a varying variance) and makes the variance of the dropout a function of privileged information. Their methodology can be applied to any neural network that uses a dropout function in its architecture. Lopez-Paz et al. (2015) unifies distillation with privileged information, illustrating how distillation techniques can address LUPI problems. Using privileged information has proven useful in many applications, such as biology Abbasi et al. (2018) and healthcare Alge et al. (2024).

Pasunuri et al. (2016) focuses on using privileged features to create additional labels for each example and using these privileged labels to guide tree-based learning algorithms. In Li et al. (2020), rather than the static manipulations of privileged information we see in current methods, the authors explore the idea of the "advice" learned from privileged information as being actively updated as the model is built. Both methodologies are implemented by extending vanilla Gradient Boosting. We compare our approach to those of Pasunuri et al. (2016); Li et al. (2020) in the method Section 4.

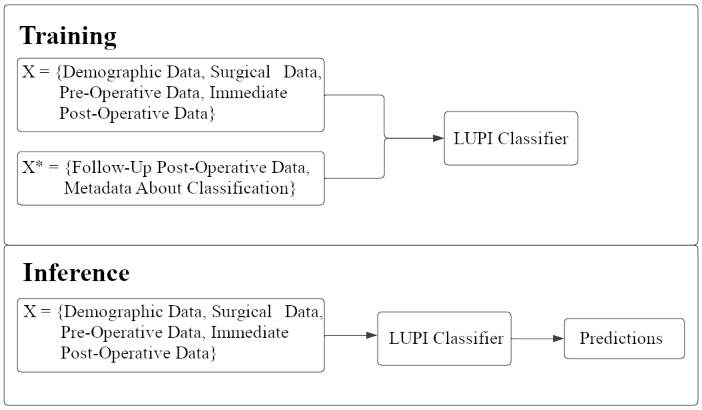
Training
X = {Demographic Data, Surgical Data, Pre-Operative Data, Immediate Post-Operative Data}
X* = {Follow-Up Post-Operative Data, Metadata About Classification}
LUPI Classifier
Inference
X = {Demographic Data, Surgical Data, Pre-Operative Data, Immediate Post-Operative Data}
LUPI Classifier
Predictions

Figure 2: Learning using privileged information for proximal junctional kyphosis prediction

One motivation for the utilization of privileged information is that convergence can be achieved by using fewer examples. If the error-correcting space (privileged data) is good, then the convergence (w.r.t. training set size) in the combined space could be on the order of $1/n$, as opposed to an algorithm operating only in the decision space (observable data), which converges on the order of $1/\sqrt{n}$, as demonstrated in Pechyony and Vapnik (2010); Vapnik and Vashist (2009). This is a very compelling property in our application given the limited number of patients in our PJK dataset.

Learning with privileged information (LUPI) for PJK Given the temporally evolving nature of the data, we believe we can bolster the performance of traditional machine learning models by training on both observable (pre-operative and operative) data and unobservable (future/postoperative) data while only predicting on the observable (Figure 2). In this manner, the unobservable, future data can be considered privileged, and thus we can apply the LUPI paradigm. This framework leverages the use of privileged information and shows significant promise in achieving higher performance.

Privileged information in our context comes in two flavors: features derived from a large set of X-ray post-operative images (e.g., proximal junctional angle and sagittal vertical alignment), along with structured data and measurements taken at different time points.

3. Study Cohort

This retrospective cohort study was conducted at a large urban hospital and approved by the Institutional Review Board (IRB). 366 consecutive patients who underwent posterior spinal fusion of five or more vertebral levels between 2015 and 2020 were included in the study. All patients were 18 years or older.

Data was extracted directly from the electronic health records of each patient. Demographic variables collected included age at surgery, gender, body mass index (BMI), smoking status, bone health status, baseline Scoliosis Research Society Questionnaire 22r (SRS) score, baseline Oswestry Disability Index (ODI) score, and the most recent follow-up SRS and ODI scores. These two scores are associated with the patient's experience: SRS, the "Scoliosis Research Society" questionnaire, which measures the pain levels of patients—the higher the score, the better, indicating less pain; and 2) Baseline ODI, the "Oswestry Disability Index," where a score of 0-20 reflects minimal disability, 21-40 moderate disability, 41-60 severe disability, 61-80 crippled, and 81-100 bed-bound. Demographic variables were collected at a time prior to the index surgery (baseline) or at the most recent follow-up. Surgical variables included fixation status to sacrum/pelvis, number of rods used, rod diameters, the use of hooks, surgical approach, upper instrumented vertebra (UIV), and lower instrumented vertebra (LIV).

Radiographic parameters were measured and collected at three time points: preoperative/baseline, immediate postoperative, and most recent follow-up (FU). Radiographic parameters were measured on full standing diagnostic radiographs of patients. At each time point, the radiographic parameters measured included the posterior cranial vertical line (PCVL)[3] Park et al. (2023). Measures include sacrum distance, acetabulum distance, medial malleolus distance, thoracic apex distance, sagittal vertical alignment (SVA), C2-pelvic angle (C2PA), proximal junctional angle (PJA), cervical lordosis (CL; C2-C7°), thoracic kyphosis (TK; T1-T12°), lumbar lordosis (LL; L1-L5°), sacral slope (SS), pelvic tilt (PT), pelvic incidence (PI), central sacral pelvic line (CSPL), and paraspinal fatty atrophy (FA).

[3] The PCVL is defined as a vertical plumb line drawn from the most posterior aspect of the occiput, with horizontal distances measured to the aforementioned anatomical locations.

Paraspinal fatty atrophy was determined using a grading system based on MRI scans at the three time points with grade 1 having 0-10% adipose infiltration of paraspinal musculature, grade 2 having 10-50% adiposity, and grade 3 having >50% adiposity Wen et al. (2023). Additional radiographic parameters collected at the immediate postoperative time point included the PCVL-UIV tulip distance, PCVL-UIV grade, and rod density. The UIV-tulip distance is measured from the PCVL to the pedicle screw tulip centroid. Rod density was defined as the sum of the diameters of the rod(s) at the UIV.

Additionally, PJK-related parameters were collected that included postoperative PJK status, pain/symptoms near the UIV, revision surgery due to PJK (termed proximal junctional failure, PJF), and revision surgery date. Radiographic PJK was defined as a final PJA $> 10°$ and Δ PJA $> 10°$. Unless otherwise indicated, Δ PJA will always refer to a preoperative to the latest follow-up change in PJA. Pain/symptoms near UIV were determined by examining follow-up notes postoperatively. Patients were considered symptomatic near their UIV if they exhibited pain or other neuromuscular abnormalities in close proximity to the UIV. The presence and date of revision surgery due to PJK (i.e., PJF) was found by examining operative and physician notes.

The PJK dataset contains numerical and categorical features. Categorical features were ultimately one-hot encoded. For continuous variables with missingness, mean imputation was used. For a full set of description of all features, please refer to the Appendix Tables 1 - 3.

4. Method: XGBoost with Privileged Information (XGBoost+)

In contrast to SVM, ensemble models offer the advantage of uncovering more intricate relationships between features, enhancing the model's ability to capture complex patterns in the data. In order to harness the advantages offered by ensemble models and the XGBoost package, we opted to extend the functionality of the package by making only modifications to the loss function.

The standard formulation of the LUPI framework Vapnik and Vashist (2009) is as follows: given a dataset $D = \{(x_i, x_i^*, y_i)\}_{i=1}^n$, where $(x, y) \in X \times Y$ and X is the decision space and Y is the label space. $x^* \in X^*$ is the privileged information in the correcting space X^*.

In the XGBoost model with K additive base learners, we can obtain the predictive output by taking their aggregated predictive values:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

f_k corresponds to a k-th regression tree. For each example, we use the decision of the regression trees to calculate the final score for x_i by summing up the scores in the corresponding leaves. This model is trained in an additive manner.

To formulate the loss function, let $\hat{y}_i^{(t)}$ be the prediction of the i-th instance at the t-th iteration and f_t be a regression tree used to predict the residual error at iteration t (Chen and Guestrin (2016)). XGBoost minimizes the following objective:

$$\begin{gathered} L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \\ \text{where } \Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \end{gathered} \tag{2}$$

Here, l is a twice-differentiable convex loss function that tries to minimize the difference between target y_i and the prediction result in $\hat{y}_i^{(t-1)}$ plus the estimated residual error $f_t(x_i)$. Ω is the penalty for model complexity, T is the number of leaves in the tree, and w represents a vector of scores on the leaf nodes. The objective reduces to the traditional gradient tree boosting one when this regularization term is zero.

Inspired by Lopez-Paz et al. (2015), we apply the distillation method to incorporate privileged information into the XGBoost model. We achieve this by introducing a soft label s_i for each example x_i. s_i that serves as a teacher to the XGBoost model and is determined by building a model in the privileged space. We incorporated s_i into the loss function defined in Equation 2 as follows:

$$\begin{gathered} L^{(t)} = \sum_{i=1}^{n} l(y_i, s_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \\ \text{where } \Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \end{gathered} \tag{3}$$

In a naive implementation of Gradient Boosting, the loss function is computationally expensive. One computational advantage of XGBoost is that it approximates the loss function using a second-order Taylor expansion. Now, when we incorporate the soft label (s_i) into the loss function and compute the second-order approximation using Taylor expansions, we derive the following approximation of Equation 3:

$$\begin{aligned} L^{(t)} \simeq \sum_{i=1}^{n} l(y_i, s_i, \hat{y}_i^{(t-1)}) &+ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \\ &+ \Omega(f_t) \\ \text{where } g_i &= \partial_{\hat{y}^{(t-1)}} l(y_i, s_i, \hat{y}^{(t-1)}) \\ h_i &= \partial^2_{\hat{y}^{(t-1)}} l(y_i, s_i, \hat{y}^{(t-1)}) \end{aligned} \tag{4}$$

where g and h represent the gradient and the Hessian, respectively. Working with binary classification, as is the case for all of our experiments, we chose the loss l to be the logistic loss. We use a distilled loss that we define as follows:

$$l(y_i, s_i, \hat{y}^{(t-1)}) = (1 - \alpha) l_S(s_i, \hat{y}^{(t-1)}) + (\alpha) l_T(y_i, \hat{y}^{(t-1)}) \tag{5}$$

where

$$\begin{aligned} l_S(s_i, \hat{y}^{(t-1)}) = s_i \log(1 + \exp(-\hat{y}^{(t-1)})) \\ + (1 - s_i) \log(1 + \exp(\hat{y}^{(t-1)})) \end{aligned} \tag{6}$$

and

$$\begin{aligned} l_T(y_i, \hat{y}^{(t-1)}) &= y_i \log(1 + \exp(-\hat{y}^{(t-1)})) \\ &+ (1 - y_i) \log(1 + (\exp \hat{y}^{(t-1)})) \end{aligned} \tag{7}$$

Here, $\alpha \in [0, 1]$ is defined as the imitation parameter that controls the contribution of the privileged information. The smaller the value of α, the higher the contribution of the privileged information. When α is 1, the model performs exactly like the XGBoost algorithm, but when α is 0, the model follows exactly the teacher's advice and disregards the true label (y_i) altogether. Calculating the gradient statistics for Equation 4, we arrive at the following:

$$g_i = \frac{\exp(\hat{y}^{(t-1)})}{1 + \exp(\hat{y}^{(t-1)})} - (1 - \alpha)s_i - \alpha y_i \tag{8}$$

and

$$h_i = \frac{\exp(\hat{y}^{(t-1)})}{(1 + \exp(\hat{y}^{(t-1)}))^2} \tag{9}$$

This concludes the modification of the loss function. However, we still need to provide the set of soft labels (s_i). As in Hinton et al. (2015), we define $s_i = \sigma(f_*(x_i^*)/T^*)$, where σ is the softmax function and $T^* > 0$ is a temperature parameter that provides smoothing of the class-probability prediction. Although f_* can be any classifier of our choosing, a simple model allows for the discovery of a general set of rules in the privileged space. This is to avoid overfitting, since generalized rules can be more easily transferable to the decision space. In our experiments, we chose f_* to be a logistic regression classifier. Lastly, for prediction on the test set, it is the same as native in the XGBoost package, so no additional modifications are needed.

4.1. Comparison to other gradient boosting privileged information methods

Two other studies that utilize privileged information in boosting are Pasunuri et al. (2016) (GB+) and Li et al. (2020) (IPL). IPL used a similar approach to SVM+ where a slack variable is introduced to XGBoost. At each iteration of the boosting round, a linear model is used to estimate the value of this slack. Essentially, at the i^{th} iteration, this linear model is updated by finding a model that correctly estimates differences between the output from the i^{th} regression tree to the residual from the $(i - 1)$ regression tree.

The GB+ method shares more similarities with our method compared to the IPL method. Both our method and GB+ perform better in scenarios where privileged information provides knowledge that can be transferred. The major distinction between our method and GB+ is that for GB+, both the privileged information model and the decision space model are based on a gradient-boosting model trained simultaneously. We chose logistic regression, as the privileged model because it is less complex than gradient boosting. This allows it to provide more generalized information, and the privileged model remains unaffected by the decision space model.

Lastly, with a modification solely to the loss function of the XGBoost algorithm, we can still leverage the inherent functionality of the XGBoost package. This facilitates code reproducibility and application to other datasets without concerns about the dataset's scale.

5. Results

5.1. Synthetic datasets

We conducted experiments with synthetic datasets that defined relevant features as privileged information and increased the difficulty of predicting the test set. This is done by introducing features that have a high correlation to the label in the training set, but this correlation ceases to exist in the test set. The rest of this subsection describes how the dataset was created.

Table 1: Performance on synthetic datasets

Method	Accuracy[1]	ROC-AUC[1]
Decision Tree	0.76 ±0.12	0.79 ±0.13
Random Forest	0.77 ±0.14	0.85 ±0.13
SVM	0.78 ±0.10	0.86 ±0.10
SVM+	0.78 ±0.10	0.87 ±0.10
IPL	0.80 ±0.13	0.86 ±0.16
GB+	0.81 ±0.13	0.90 ±0.11
XGBoost	0.78 ±0.16	0.83 ±0.17
XGBoost+	**0.92** ±0.07	**0.98** ±0.04

[1] Average over 100 trials. Result are reported with mean value ± standard deviation.

Privileged information can be categorized into three types: (1) clean label as privileged information, (2) clean feature as privileged information, and (3) relevant features as privileged information (Lopez-Paz et al. (2015)). The former two types of privileged information imply noise in the labels and the features, which is not the case for PJK prediction; curated data collection was carried out by multiple professional clinicians and meticulously cross-referenced. Therefore, we decided to focus on the relevant features as privileged information and built synthetic datasets around it.

For the synthetic dataset, the training set is denoted by data triplets $(x_i^{train}, x_i^{train*}, y_i^{train})$, and the test set by (x_i^{test}, y_i^{test}). The regular features x_i^{train} and x_i^{test} have a dimensionality of d, and the separating hyper-planes $\alpha \in \mathbb{R}^d$ follow a standard normal distribution $\mathcal{N}(0, I_d)$. We define J as the subset of indices of feature columns $\{1, ..., d\}$.

We also define H as a (privileged-indexed) subset of J; The formulation for the training set is as follows:

$$\begin{aligned} x_i^{train} &\sim \mathcal{N}(0, I_d) \\ x_i^{train*} &\leftarrow x_{i,H}^{train} \\ y_i^{train} &\leftarrow \mathbb{I}(\langle \alpha_{Ji}, x_{i,J} \rangle > 0) \end{aligned} \tag{10}$$

and the test set as

$$\begin{aligned} x_i^{test} &\sim \mathcal{N}(0, I_d) \\ y_i^{test} &\leftarrow \mathbb{I}(\langle \alpha_{Hi}, x_{i,H} \rangle > 0) \end{aligned} \tag{11}$$

The privileged information (x_i^*) is equal to x_i^H. For example, we can consider a dataset predicting diabetes, and $x_{i,J}$ corresponds to the feature set including gender, BMI, and height and $x_{i,H}$ contains just BMI and height, so $J - H$ is gender. Since the label y_t^{test} is determined by the inner product of α and x_i^H and the train set α and x_i^J. This means that gender ceases to be predictive of the test label y_{test}. The scenario can be common to datasets that are relatively small in size, as with healthcare datasets. For instance, during phases of cross-validation, we might coincidentally pick a set of training examples where all females have diabetes.

For Table 1, we set $|J| = 3$ and $|H| = 2$, randomly selecting values for them. We sampled 200 examples from the training distribution and 1000 samples from the test distribution. We performed 100 trials and recorded the accuracy and ROC-AUC in Table 1.

We repeated the experiment by varying the values of $|J|$ and $|H|$, reporting the results in Appendix Table 4. In all experiments, XGBoost+ outperforms XGBoost. We noticed that as we increased the size of $J - H$, the performance of XGBoost substantially deteriorated, but XGBoost+ maintained excellent performance.

5.2. Primary Biliary Cirrhosis Dataset

In addition to PJK prediction, as a proof of concept, we also experimented with XGBoost+ on another real-world dataset. The primary biliary cirrhosis (PBC) dataset was obtained through the Survival Package in R Terry M. Therneau and Patricia M. Grambsch (2000); Therneau (2023). This dataset focuses on the progression of primary biliary cirrhosis in 312 patients seen at the Mayo Clinic between January 1974 and May 1984. PBC is a chronic liver disease characterized by the progressive destruction of small bile ducts within the liver. PCB is an autoimmune disease where the immune system mistakenly attacks and damages the bile duct, leading to inflammation and scarring of the liver tissue. The early stages of PBC may be asymptomatic, but in the later stages, PBC can lead to complications such as cirrhosis, liver failure, and portal hypertension. Additionally, it increases the risk of hepatocellular carcinoma (liver cancer). There is no cure for PBC, but various medications can be prescribed to slow its progression. In severe cases, a liver transplant is considered for individuals with advanced cirrhosis and liver failure. Early detection is crucial for identifying patients who require a transplant. Features of the dataset include patient age at the first diagnosis, physical symptoms such as ascites or hepatomegaly, and blood values related to liver function such as bilirubin, albumin, and alkaline phosphatase collected from multiple visits Lin and Zelterman (2002). For a complete list of features, please refer to Therneau (2023). This dataset is usually used for survival analysis to predict the survival rate of PBC patients. However, we transformed the dataset into a privileged information problem where the decision feature (X) is the set of information available at the earliest visit, and (X^*) is the set of information available at the patient's last visit. The label Y is defined as those who received a liver transplant or died after the first visit.

We compared our XGBoost+ model to XGBoost, SVM, and SVM+. The reason we want to utilize privileged information in boosting models is that these models have been shown to work well with tabular data and in our previous healthcare works. Specifically, boosting models have been demonstrated to be among the best non-neural network supervised models Caruana and Niculescu-Mizil (2006). This is why we chose methods that are boosting in

Table 2: Performance on PBC dataset

Method	**Accuracy**[1]	**ROC-AUC**[1]
Decision Tree	0.68 ±0.04	0.68 ±0.04
Random Forest	0.73 ±0.04	0.78 ±0.04
SVM	0.72 ±0.04	0.79 ±0.04
SVM+	0.74 ±0.05	0.81 ±0.04
IPL	0.74 ±0.04	0.80 ±0.04
GB+	0.75 ±0.04	0.80 ±0.04
XGBoost	0.73 ±0.05	0.79 ±0.04
XGBoost+	**0.76** ±0.04	**0.82** ±0.04

[1] Average over 100 trials. Result are reported with mean value ± standard deviation.

nature and utilize privileged information. We also added SVM+ since it is one of the original models introduced by Vapnik. Results are from 100 cross-validation trials, where we retained 30% of the samples for training and reserved 70% for testing. Accuracy and ROC-AUC are both reported for the four models. The LUPI-based methods performed better than their vanilla counterparts, demonstrating the benefits of utilizing privileged information in prediction. XGBoost+ is the best-performing model (Table 2). Since the SVM and XGBoost models exhibit similar performance, it is noteworthy that XGBoost+ demonstrates a higher increase in performance in this experiment. The XGBoost model trained and tested with knowledge of privileged information resulted in a 0.79 AUC. The performance increase from vanilla XGBoost (73%) to XGBoost+ (76%) leads to an error recovery of $(76 - 73)/(79 - 73) = 0.50$. Hence, we are able to close 50% of the AUC loss between the vanilla XGBoost model using only the standard features by using the XGBoost+ model with privileged information.

5.3. PJK Dataset

We show the performance of our PJK prediction model, XGBoost+, compared to other methods that also utilize privileged information and their vanilla counterparts (SVM, SVM+) in Table 3, highlighting the importance and advantage of using privileged information. Out of the 366 patients, some patients have a large amount of missingness, reaching as high as 50%. Imputing data for these patients with extensive missing data can significantly degrade the model's performance. In the reduced dataset, there are 67 PJK versus 201 non-PJK patients, which still maintains a 24% target ratio compared to the whole dataset. The results in Table 3 showcase the model's performance when retaining only patients with 3% or less missing data. This inclusion enables the consideration of samples with only 1-2 features missing. In the same table, we present the performance of the models without removing patients with a large amount of missing data. Additionally, in Figure 3, we also included a performance graph of XGBoost+ and XGBoost using leave-one-out cross-validation at each missingness level, for which we observe that ROC-AUC significantly increases until the missingness drops to 3%. Patients with large missingness in the decision features are more likely to have missing privileged information. Removing patients with missing values does not affect the proportion of positive to negative examples as shown in Figure 3 (Right).

5.3.1. Study Design

We run our model over multiple trials ($n = 50$) and compute the average of the metrics of the best model in each trial. Within each trial, we split our data into train and test sets. We use stratified sampling to ensure that the proportion of each class is reflected appropriately in both the train and test sets. Furthermore, we tune the model over a pre-determined hyper-parameter space using grid-search. Once we have our tuned model, we predict over the test set and retrieve our metrics based on the predicted values.

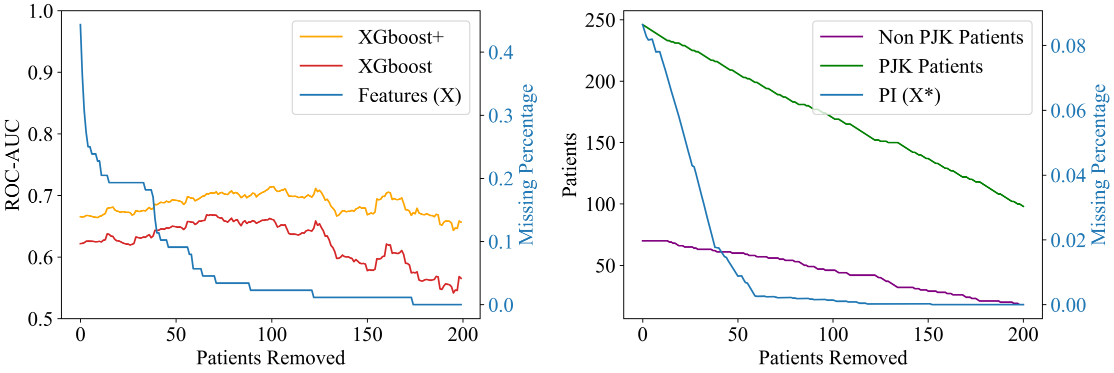

Figure 3: (Left) Model performance of XGBoost and XGBoost+ as a function of patient removal. Feature missingness of removed patients is plotted with the blue curve. (Right) Number of patients in each class as a function of patient removal. Privileged information average missingness is plotted with the blue curve.

5.3.2. Results on PJK dataset

Table 3 presents our results for XGBoost+ model on the PJK dataset. For three of the four metrics we considered, we noticed an improvement from the vanilla XGBoost model, alluding to the importance of privileged information. Our XGBoost+ model has the best ROC-AUC score out of all the methods presented, with 5% improvement over the vanilla XGBoost. We compare our model to other LUPI methods and their associated vanilla models. The XGBoost model trained and tested with knowledge of privileged information resulted in a 0.76 AUC. The performance increase from the vanilla XGBoost (67%) to XGBoost+ (72%) leads to a recovery of 55.5% of performance loss by including privileged information. We have constructed statistical tests comparing the result of the XGBoost+ method with the other methods using the Wilcoxon test and found all p-values < 0.03. Therefore, we can conclude that the results of the XGBoost+ method are statistically significantly different from those of the other methods. These results have been added to the Appendix.

Table 3: Peformance on the PJK Dataset

Method	**ROC-AUC** [1]	**Precision** [1]	**Sensitivity** [1]	**Specificity**[1]
Decision Tree	0.48 ± 0.02	0.38 ± 0.16	0.28 ± 0.1	0.76 ± 0.09
Random Forest	0.67 ± 0.03	0.17 ± 0.11	0.04 ± 0.02	**0.96 ± 0.02**
SVM	0.61 ± 0.04	0.35 ± 0.09	0.24 ± 0.05	0.84 ± 0.04
SVM+	0.63 ± 0.03	0.46 ± 0.15	0.26 ± 0.05	0.85 ± 0.06
IPL	0.69 ± 0.06	**0.46 ± 0.15**	0.29 ± 0.09	0.89 ± 0.04
GB+	0.69 ± 0.04	0.41 ± 0.12	0.2 ± 0.01	0.89 ± 0.03
XGBoost	0.67 ± 0.05	0.43 ± 0.08	0.32 ± 0.11	0.86 ± 0.05
XGBoost+	**0.72 ± 0.05**	0.44 ± 0.05	**0.42 ± 0.14**	0.82 ± 0.06
Decision Tree*	0.58 ± 0.07	0.33 ± 0.1	0.31 ± 0.12	0.85 ± 0.04
Random Forest*	**0.66 ± 0.03**	0.1 ± 0.1	0.01 ± 0.01	**0.98 ± 0.01**
SVM*	0.63 ± 0.05	0.42 ± 0.15	0.19 ± 0.07	0.9 ± 0.04
SVM+*	0.62 ± 0.04	0.27 ± 0.05	0.24 ± 0.09	0.82 ± 0.04
IPL*	0.61 ± 0.03	0.37 ± 0.08	**0.43 ± 0.14**	0.76 ± 0.14
GB+*	0.63 ± 0.05	0.40 ± 0.06	0.24 ± 0.04	0.82 ± 0.05
XGBoost*	0.64 ± 0.04	0.46 ± 0.12	0.2 ± 0.04	0.9 ± 0.04
XGBoost+*	0.64 ± 0.03	**0.46 ± 0.1**	0.31 ± 0.04	0.85 ± 0.06

[1] Average over 50 trials. Result are reported with mean value ± standard deviation.

*Performance of the models without removing patients with a large amount of missing data.

Finally, we also present the variable importance of the standard features (available during training and inference) in Figure 4. Variable importance is determined by calculating the SHAP values for each feature Lundberg and Lee (2017). The bar graph shows that the SVA (sagittal vertical axis) has the highest importance as a variable. SVA is measured as the distance between the C7 plumb line and the posterior-superior corner of S1 (Jackson and McManus (1994); Gelb et al. (1995)). Figure 5 is a beeswarm plot that shows the impact each feature (y-axis) has on the XGBoost+ model's output. Points in pink indicate a high feature value, while points in blue indicate a low feature value, and the position of the points on the x-axis represents the SHAP value, showing how positively or negatively impactful the feature is on the model's prediction. For instance, high values of feature SVA positively impact the model's output, while lower values have a negative impact.

6. Discussion

Technical Context In this novel approach, we combine existing models (XGBoost) and distillation framework. We notice that SVM+ shows slight improvement compared to its vanilla counterpart, most likely because SVM models cannot discover intricate relationships between features. Hence, the predictive power for SVM and SVM+ is significantly lower. However, ensemble tree methods such as XGBoost - in tandem with the algorithm's specific optimization techniques - tackle this problem as each new weak learner is added. Comparing XGboost with XGboost+, our XGBoost+ model can achieve the same results as the XGBoost model in approximately 43% of its runtime.

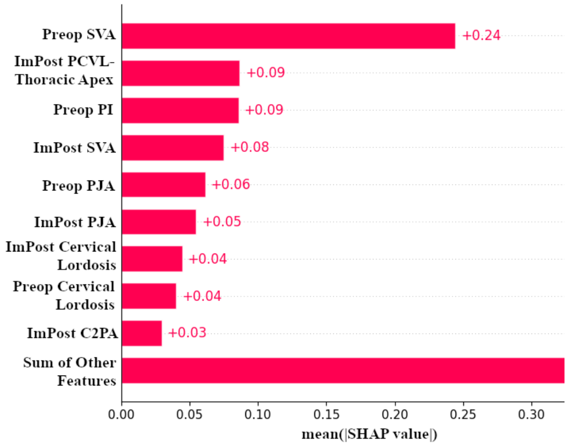

Figure 4: Top 10 feature importance of XGBoost+ model

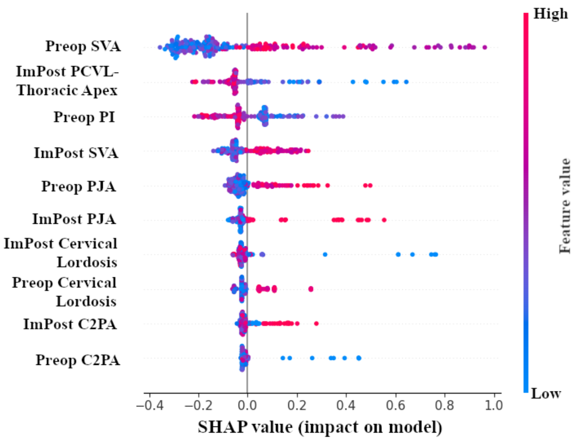

Figure 5: SHAP waterfall chart of top 10 features

Clinical Context From a medical aspect, our proposed model performs best in terms of AUC (0.72). We also observe a high specificity (0.82) and lower sensitivity (0.42) evaluated at a 0.5 threshold, highlighting that the model can predict the negative class but still has better predictive power for the positive class, as compared to other methods. From a feature importance perspective, we highlight the top five features with the highest importance - preoperative and preoperative and ImPost SVA (sagittal vertical alignment), ImPost PCVL-Thoracic Apex, PI (pelvic incidence), and PJA (proximal junctional angle). Previous work has also highlighted all top features as primary risk factors associated with predicting PJK (Zhao et al. (2018); Yagi et al. (2011)).

Our study cohort comprises 366 patients, unlike other studies like Peng et al. (2020), which involve 44 patients. In both cohorts, the female-to-male ratio is greater than 3:1. While Peng et al. (2020) has identified gender as the primary feature in their model, our analysis does not support this finding. This can result of sampling bias since the cohort only has 10 PJK patients. Similarly to our findings, Peng et al. (2020) identifies SVA as one of the top predictive features in their model.

Our second prominent feature in Figure 4, ImPost PCVL-Thoracic Apex, aligns with the observations in Lee et al. (2020); ”Chen et al. (2021). The study suggests that the subsequent postoperative assessment of thoracic kyphosis can be anticipated based on prior measurements of thoracic kyphosis. Elevated kyphosis in the non-instrumented thoracic spine may amplify anterior compressive forces and Upper Instrumented Vertebra (UIV), contributing to the onset of Proximal Junctional Kyphosis (PJK) or Proximal Junctional Failure (PJF).

It is also important to note the relative contributions of privileged information to the model. Pain UIV, Overall PRO Decline, ODI Change, and SRS Change can significantly contribute to refining our model's decision boundary. Specifically, we hypothesize that with insight into the post-surgical features, the XGBoost model can identify finer subcategories of PJK, thereby shaping a better decision boundary for prediction.

Limitations As an initial step in incorporating privileged information for XGBoost, our study has a few limitations to consider. Our proposed model utilizes privileged information, but the exact effect of the privileged information on the model's decision is still unknown. This is something we wish to investigate in future work.

Concerning our model's predictive power, our data cohort is more inclusive than previous studies (Scheer et al. (2016); Peng et al. (2020)). Previous studies have inclusion criteria where patients already have ASD or Lenke 5 AIS. Not having such inclusion criteria broadens our cohort, making the machine learning task much more complex but allowing our model to be more applicable to a broader range of patients in a clinical setting.

7. Conclusion

In this paper, we incorporate the distillation framework in XGBoost to improve the model's ability to predict PJK. Our results show an increase of 5% as compared to the vanilla XGBoost. We also showcase the variable importance of both our features and the privileged information, in which our results are supported by previous literature, highlighting the interpretability and reliability of our model. Our primary goal with this work is to showcase the importance of utilizing privileged information in a healthcare setting. With respect to predicting PJK, we wish to continue our work in three directions: (i) investigating the effect of privileged information on the prediction made by our proposed XGBoost+ model, (ii) incorporating a causal model that can propose counterfactual explanations – fine-grained modifications of some actionable features in order to decrease the risk of PJK/PJF in patients, and thus change the classifier's output, and lastly (iii) developing a graphical user interface and a tool deploying the models in (i) and (ii) so that physicians can use them in a clinical setting. We hope our work will provide physicians with the tools necessary to make informed and reliable decisions to prevent PJK/PJF.

References

Wajid Arshad Abbasi, Amina Asif, Asa Ben-Hur, and Fayyaz ul Amir Afsar Minhas. Learning protein binding affinity using privileged information. *BMC bioinformatics*, 19:1–12, 2018.

Olivia P. Alge, Jonathan Gryak, J. Scott VanEpps, and Kayvan Najarian. Sepsis trajectory prediction using privileged information and continuous physiological signals. *Diagnostics*, 14(3), 2024. ISSN 2075-4418. doi: 10.3390/diagnostics14030234. URL https://www.mdpi.com/2075-4418/14/3/234.

K. H. Bridwell, L. G. Lenke, S. K. Cho, J. M. Pahys, L. P. Zebala, I. G. Dorward, W. Cho, C. Baldus, B. W. Hill, and M. M. Kang. Proximal junctional kyphosis in primary adult deformity surgery: evaluation of 20 degrees as a critical angle. *Neurosurgery*, 72(6): 899–906, Jun 2013.

P. J. Cahill, W. Wang, J. Asghar, R. Booker, R. R. Betz, C. Ramsey, and G. Baran. The use of a transition rod may prevent proximal junctional kyphosis in the thoracic spine after scoliosis surgery: a finite element analysis. *Spine (Phila Pa 1976)*, 37(12):E687–695, May 2012.

M. Cammarata, C. É. Aubin, X. Wang, and J. M. Mac-Thiong. Biomechanical risk factors for proximal junctional kyphosis: a detailed numerical analysis of surgical instrumentation variables. *Spine (Phila Pa 1976)*, 39(8):E500–507, Apr 2014.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

J. "Chen, H. Fan, Sui W., J. Yang, Y. Deng, Z. Huang, and J." Yang. Risk and predictive factors for proximal junctional kyphosis in patients treated by lenke type 5 adolescent idiopathic scoliosis correction. *World Neurosurgery*, 147:e315–e323, 2021. ISSN 1878-8750. doi: https://doi.org/10.1016/j.wneu.2020.12.044. URL https://www.sciencedirect.com/science/article/pii/S1878875020326103.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

K. J. Cho, S. I. Suk, S. R. Park, J. H. Kim, and J. H. Jung. Selection of proximal fusion level for adult degenerative lumbar scoliosis. *Eur Spine J*, 22(2):394–401, Feb 2013.

D. E. Gelb, L. G. Lenke, K. H. Bridwell, K. Blanke, and K. W. McEnery. An analysis of sagittal spinal alignment in 100 asymptomatic middle and older aged volunteers. *Spine (Phila Pa 1976)*, 20(12):1351–1358, Jun 1995.

R. C. Glattes, K. H. Bridwell, L. G. Lenke, Y. J. Kim, A. Rinella, and C. Edwards. Proximal junctional kyphosis in adult spinal deformity following long instrumented posterior spinal fusion: incidence, outcomes, and risk factor analysis. *Spine (Phila Pa 1976)*, 30(14):1643–1649, Jul 2005.

R. Hart, I. McCarthy, M. Obrien, S. Bess, B. Line, O. B. Adjei, D. Burton, M. Gupta, C. Ames, V. Deviren, K. Kebaish, C. Shaffrey, K. Wood, and R. Hostin. Identification of decision criteria for revision surgery among patients with proximal junctional failure after surgical treatment of spinal deformity. *Spine (Phila Pa 1976)*, 38(19):E1223–1227, Sep 2013.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

S. J. Hyun, Y. J. Kim, and S. C. Rhim. Patients with proximal junctional kyphosis after stopping at thoracolumbar junction have lower muscularity, fatty degeneration at the thoracolumbar area. *Spine J*, 16(9):1095–1101, 09 2016.

R. P. Jackson and A. C. McManus. Radiographic analysis of sagittal plane alignment and balance in standing volunteers and patients with low back pain matched for age, sex, and size. A prospective controlled clinical study. *Spine (Phila Pa 1976)*, 19(14):1611–1618, Jul 1994.

D. G. Kang, R. A. Lehman, and L. G. Lenke. Challenges in the classification of adolescent idiopathic scoliosis and the utility of artificial neural networks. *Spine J*, 13(11):1534–1537, Nov 2013.

H. J. Kim and S. Iyer. Proximal Junctional Kyphosis. *J Am Acad Orthop Surg*, 24(5): 318–326, May 2016.

H. J. Kim, L. G. Lenke, C. I. Shaffrey, E. M. Van Alstyne, and A. C. Skelly. Proximal junctional kyphosis as a distinct form of adjacent segment pathology after spinal deformity surgery: a systematic review. *Spine (Phila Pa 1976)*, 37(22 Suppl):S144–164, Oct 2012.

Y. J. Kim, K. H. Bridwell, L. G. Lenke, S. Rhim, and Y. W. Kim. Is the T9, T11, or L1 the more reliable proximal level after adult lumbar or lumbosacral instrumented fusion to L5 or S1? *Spine (Phila Pa 1976)*, 32(24):2653–2661, Nov 2007.

Y. J. Kim, K. H. Bridwell, L. G. Lenke, C. R. Glattes, S. Rhim, and G. Cheh. Proximal junctional kyphosis in adult spinal deformity after segmental posterior spinal instrumentation and fusion: minimum five-year follow-up. *Spine (Phila Pa 1976)*, 33(20):2179–2184, Sep 2008.

J. Lambert, Sener O, and S. Savarese. Deep learning under privileged information using heteroscedastic dropout. *CoRR*, abs/1805.11614, 2018. URL http://arxiv.org/abs/1805.11614.

D. Lau, A. J. Clark, J. K. Scheer, M. D. Daubs, J. D. Coe, K. J. Paonessa, M. O. LaGrone, M. D. Kasten, R. A. Amaral, P. D. Trobisch, J. H. Lee, D. Fabris-Monterumici, N. Anand, A. K. Cree, R. A. Hart, L. A. Hey, and C. P. Ames. Proximal junctional kyphosis and failure after spinal deformity surgery: a systematic review of the literature as a background to classification development. *Spine (Phila Pa 1976)*, 39(25):2093–2102, Dec 2014.

N. J. Lee, Z. M. Sardar, V. Boddapati, J. Mathew, M. Cerpa, E. Leung, J. Lombardi, L. G. Lenke, and R. A. Lehman. Can Machine Learning Accurately Predict Postoperative Compensation for the Uninstrumented Thoracic Spine and Pelvis After Fusion From the Lower Thoracic Spine to the Sacrum? *Global Spine J*, page 2192568220956978, Oct 2020.

L. G. Lenke, R. R. Betz, J. Harms, K. H. Bridwell, D. H. Clements, T. G. Lowe, and K. Blanke. Adolescent idiopathic scoliosis: a new classification to determine extent of spinal arthrodesis. *J Bone Joint Surg Am*, 83(8):1169–1181, Aug 2001.

Xue Li, Bo Du, Yipeng Zhang, Chang Xu, and Dacheng Tao. Iterative privileged learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2805–2817, 2020. doi: 10.1109/TNNLS.2018.2889906.

Haiqun Lin and Daniel Zelterman. Modeling survival data: extending the cox model, 2002.

F. Y. Liu, T. Wang, S. D. Yang, H. Wang, D. L. Yang, and W. Y. Ding. Incidence and risk factors for proximal junctional kyphosis: a meta-analysis. *Eur Spine J*, 25(8):2376–2383, 08 2016.

C. D. Lopez, V. Boddapati, J. M. Lombardi, N. J. Lee, J. Mathew, N. C. Danford, R. R. Iyer, M. D. Dyrszka, Z. M. Sardar, L. G. Lenke, and R. A. Lehman. Artificial Learning and Machine Learning Applications in Spine Surgery: A Systematic Review. *Global Spine J*, page 21925682211049164, Feb 2022.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

Paul J Park, Fthimnir M Hassan, Xavier E Ferrer, Cole Morrissette, Nathan J Lee, Meghan Cerpa, Zeeshan M Sardar, Michael P Kelly, Stephane Bourret, Kazuhiro Hasegawa, et al. The posterior cranial vertical line: a novel radiographic marker for classifying global sagittal alignment. *Neurospine*, 20(3):790, 2023.

S. J. Park, C. S. Lee, S. S. Chung, J. Y. Lee, S. S. Kang, and S. H. Park. Different Risk Factors of Proximal Junctional Kyphosis and Proximal Junctional Failure Following Long Instrumented Fusion to the Sacrum for Adult Spinal Deformity: Survivorship Analysis of 160 Patients. *Neurosurgery*, 80(2):279–286, 02 2017.

R. Pasunuri, P. Odom, T Khot, K Kersting, and S Natarajan. Learning with privileged information: Decision-trees and boosting. pages 1–7, Cham, 2016. Proc. Int. Joint Conf. Artif. Intell. Workshop.

D. Pechyony and V. Vapnik. On the theory of learnining with privileged information. *Advances in neural information processing systems*, 23, 2010.

Li Peng, Lan Lan, Peng Xiu, Guangming Zhang, Bowen Hu, Xi Yang, Yueming Song, Xiaoyan Yang, Yonghong Gu, Rui Yang, and Xiaobo Zhou. Prediction of proximal junctional kyphosis after posterior scoliosis surgery with machine learning in the lenke 5 adolescent idiopathic scoliosis patient. *Frontiers in Bioengineering and Biotechnology*, 8, 2020. ISSN 2296-4185. doi: 10.3389/fbioe.2020.559387. URL https://www.frontiersin.org/article/10.3389/fbioe.2020.559387.

J. K. Scheer, J. A. Osorio, J. S. Smith, F. Schwab, V. Lafage, R. A. Hart, S. Bess, B. Line, B. G. Diebo, T. S. Protopsaltis, A. Jain, T. Ailon, D. C. Burton, C. I. Shaffrey, E. Klineberg, and C. P. Ames. Development of Validated Computer-based Preoperative Predictive Model for Proximal Junction Failure (PJF) or Clinically Significant PJK With 86 *Spine (Phila Pa 1976)*, 41(22):E1328–E1335, Nov 2016.

M. W. Smith, P. Annis, B. D. Lawrence, M. D. Daubs, and D. S. Brodke. Early proximal junctional failure in patients with preoperative sagittal imbalance. *Evid Based Spine Care J*, 4(2):163–164, Oct 2013.

M. W. Smith, P. Annis, B. D. Lawrence, M. D. Daubs, and D. S. Brodke. Acute proximal junctional failure in patients with preoperative sagittal imbalance. *Spine J*, 15(10):2142–2148, Oct 2015.

Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. *Medical decision making*. John Wiley & Sons, 2024.

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

Terry M Therneau. *A Package for Survival Analysis in R*, 2023. URL https://CRAN.R-project.org/package=survival. R package version 3.5-7.

V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130. Advances in Neural Networks Research: IJCNN2009.

Gang Wen, Wanmei Hou, and Guangwei Xu. Enhanced grading methods for lumbar paraspinal fat infiltration and its prognostic value in predicting lumbar disc herniation. *Journal of Orthopaedic Surgery and Research*, 18(1):752, 2023.

M. Yagi, K. B. Akilah, and O. Boachie-Adjei. Incidence, risk factors and classification of proximal junctional kyphosis: surgical outcomes review of adult idiopathic scoliosis. *Spine (Phila Pa 1976)*, 36(1):E60–68, Jan 2011.

M. Yagi, M. Rahm, R. Gaines, A. Maziad, T. Ross, H. J. Kim, K. Kebaish, and O. Boachie-Adjei. Characterization and surgical outcomes of proximal junctional failure in surgically treated patients with adult spinal deformity. *Spine (Phila Pa 1976)*, 39(10):E607–614, May 2014.

J. Zhao, M. Yang, Y. Yang, X. Yin, C. Yang, L. Li, and M. Li. Proximal junctional kyphosis in adult spinal deformity: a novel predictive index. *Eur Spine J*, 27(9):2303–2311, 09 2018.

Appendix

Data and Code Availability The PJK dataset was collected from a large hospital, and due to a confidentiality policy, we are unable to release it to the public. For interested users, the synthetic datasets, the primary biliary cirrhosis (PBC) dataset, and the source code for replicating the experiments presented in this work is available on GitHub[4]. The PBC dataset is also available from the R survival package (Therneau (2023)).

Institutional Review Board (IRB) This retrospective cohort study was conducted at the New York Presbyterian-Columbia University Irving Medical Center (NYP-CUIMC) and the Och Spine Hospital at the Allen Hospital. The study was approved by the NYP-CUIMC's Institutional Review Board (IRB AAAT6670).

Model result using ideal threshold determined by harm and benefit ratio We acknowledge that the 0.5 threshold used to report precision, specificity, and sensitivity might not be adequate for real-world applications. To have a more adequate threshold, we use the formula $Threshold = \frac{1-p(PJK)}{p(PJK)} * \frac{Harm}{Benefit}$ (Sox et al. (2024)). The harm to benefit ratio that the our clinical team has decided is 15 to 85. So, there are more benefits when issuing a treatment. The probability of the disease is 0.24, so the cutoff point to selected is 0.56 = (1 - 0.24) /(0.24) * (15/85). Using this threshold, Table 3 is updated and the new result is reported in Appendix Table 5.

Ablation Study for All Experiments For each of the experiments (Synthetic, PBC, and PJK), we performed a grid search to determine the optimal $\alpha^* \in [0, 1]$ with increments of 0.05. The α^* identified through this grid search were 0.85 for the Synthetic dataset, 0.7 for the PBC dataset, and 0.55 for the PJK dataset. To understand the impact of α on each model, we evaluated performance under three conditions: without the contribution of privileged information soft labels ($\alpha = 0$), at the optimal mixture of soft labels and true labels ($\alpha = \alpha^*$), and with only soft labels considered as true labels ($\alpha = 1$). For both the Synthetic and PBC datasets, the AUC values were comparable at $\alpha = \alpha^*$ and $\alpha = 1$. However, for the PJK dataset, the AUC was lower at $\alpha = 1$ (Appendix Table 6). This suggests that for the Synthetic and PBC datasets, the privileged information provides substantial guidance as a teacher, whereas for the PJK dataset, some information might not be fully captured by the privileged information.

[4] https://github.com/adamlin859/xgboost_plus.git

Table 1: Description of surgical and demographic variables for the whole dataset and dataset where patients with large missingness are removed. Data is expressed by the Median (IQR) or N (%).

	Whole Dataset		**Reduced Dataset**	
Variables	**Non-PJK**	**PJK**	**Non-PJK**	**PJK**
No. of Cases	278	88	201	67
Surgical and Demographic				
Age At Surgery	56.4 (34.4-65.1)	60.8 (49.9-68.1)	56.0 (34.1-64.3)	62.3 (53.2-69.0)
Gender				
Female	187 (67.3%)	60 (68.2%)	146 (72.6%)	44 (65.7%)
Male	91 (32.7%)	28 (31.8%)	55 (27.4%)	23 (34.3%)
BMI	24.3 (21.7-29.3)	25.3 (21.9-28.9)	24.0 (21.7-28.3)	25.2 (22.0-28.9)
Smoker				
Never smoked	200 (71.9%)	62 (70.5%)	144 (71.6%)	46 (68.7%)
1-2 cigarettes per week	6 (2.2%)	3 (3.4%)	3 (1.5%)	2 (3.0%)
2 cigarettes everyday	72 (25.9%)	23 (26.1%)	54 (26.9%)	19 (28.4%)
Bone Density	0.0 (0.0-0.0)	0.0 (0.0-1.0)	0.0 (0.0-0.0)	0.0 (0.0-1.0)
Baseline SRS	67.5 (55.0-79.2)	58.0 (53.0-71.0)	67.0 (55.2-79.0)	58.0 (53.0-71.0)
Baseline ODI	38.0 (17.0-51.0)	38.0 (30.0-52.0)	38.0 (15.0-50.0)	40.0 (30.0-51.5)
PJK ($\geq 10\Delta \geq 10$)				
No	239 (86.0%)	57 (64.8%)	168 (83.6%)	40 (59.7%)
Yes	39 (14.0%)	31 (35.2%)	33 (16.4%)	27 (40.3%)
Pain/Sx Near UIV				
No	278 (100.0%)	7 (8.0%)	201 (100.0%)	6 (9.0%)
Yes	0 (0.0%)	81 (92.0%)	0 (0%)	1 (61.0%)
Fixation to Sacrum or Pelvis				
No	78 (28.1%)	16 (18.2%)	55 (27.4%)	9 (13.4%)
Yes	200 (71.9%)	72 (81.8%)	146 (72.6%)	58 (86.6%)
Number of Rods	3.0 (2.0-4.0)	3.0 (2.0-3.2)	3.0 (2.0-4.0)	3.0 (2.0-4.0)
Hooks				
No	259 (93.2%)	82 (93.2%)	184 (91.5%)	62 (92.5%)
Yes	19 (6.8%)	6 (6.8%)	17 (8.5%)	5 (7.5%)
UIV				
C1 - C6	9 (3.2%)	6 (6.8%)	2 (0.9%)	4 (6.0%)
C7 - T4	158 (56.8%)	36 (40.9%)	129 (64.2%)	26 (38.8%)
T5 - T7	10 (3.6%)	36 (4.5%)	7 (3.5%)	3 (4.5%)
T8 - T12	(24.5%)	32 (36.3%)	59 (29.4%)	32 (47.8%)
L1 - L2	8 (2.9%)	2 (2.2%)	4 (2.0%)	2 (3.0%)
LIV				
L1 - L3	42 (15.1%)	7 (7.9%)	30 (14.9%)	3 (4.5%)
L4 - L5	18 (6.5%)	6 (6.8%)	16 (8.0%)	4 (6.0%)
T2 - T4	2 (0.7%)	2 (2.2%)	0 (0.0%)	1 (1.5%)
T6 - T12	12 (4.3%)	3 (3.3%)	7 (3.5%)	3 (4.5%)
S1	26 (9.4%)	4 (4.5%)	23 (11.4%)	3 (4.5%)
ilium	178 (64.0%)	66 (75.0%)	125 (62.2%)	53 (79.1%)
Total Instrumented Levels	12.0 (8.0-15.0)	10.0 (8.0-15.0)	12.0 (8.0-15.0)	13.0 (8.0-15.0)

Table 2: Description of preoperative and immediate post-operative radiographic variables for the whole dataset and dataset where patients with large missingness are removed. Data is expressed by the Median (IQR) or N (%).

	Whole Dataset		**Reduced Dataset**	
Variables	**Non-PJK**	**PJK**	**Non-PJK**	**PJK**
No. of Cases	278	88	201	67
Preoperative Radiographic Measurements				
PCVL - Sacrum	64.5 (5.6-96.0)	37.7 (6.9-85.2)	68.5 (21.7-99.6)	37.8 (6.9-86.2)
PCVL - Acetabulum	110.6 (62.1-140.3)	93.4 (67.8-134.9)	118.7 (77.5-144.2)	94.7 (64.8-136.3)
PCVL - Medial Maleolus	74.3 (38.9-95.4)	62.0 (37.8-84.0)	78.0 (43.0-96.7)	60.4 (36.8-81.7)
PCVL - Thoracic Apex	11.8 (-21.9-40.7)	3.5 (-29.7-33.7)	12.2 (-17.2-40.5)	2.9 (-30.4-28.0)
PCVL TA and Sacrum Grade	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)
Thoracic Apex Level	8.0 (6.0-10.0)	7.0 (6.0-9.0)	8.0 (7.0-11.0)	7.0 (7.0-9.0)
SVA	36.6 (-2.9-78.3)	71.3 (37.3-90.3)	30.5 (-6.0-77.8)	67.6 (39.9-92.2)
C2PA	7.6 (3.6-11.9)	7.0 (3.2-12.1)	7.5 (3.7-11.9)	7.7 (4.0-12.5)
PJA	6.4 (3.6-10.8)	7.3 (4.0-13.8)	6.6 (3.9-10.7)	8.3 (3.8-14.7)
Cervical Lordosis (C2-C7)	19.3 (9.0-32.6)	22.8 (10.1-33.4)	20.6 (10.7-33.2)	23.4 (16.4-35.0)
Thoracic Kyphosis (T1-T12)	34.7 (21.5-47.5)	38.5 (25.8-51.6)	35.2 (21.8-48.6)	38.6 (27.8-51.4)
Lumbar Lordosis (L1-L5)	30.8 (17.3-44.2)	27.5 (16.5-40.4)	31.7 (17.1-46.1)	27.9 (16.6-40.8)
SS	29.1 (20.9-38.6)	26.6 (21.2-35.2)	28.7 (20.1-38.0)	25.5 (19.9-32.1)
PT	23.6 (15.7-31.8)	26.9 (18.7-33.5)	22.6 (16.0-32.1)	26.1 (20.3-33.4)
PI	52.8 (43.0-63.1)	51.5 (46.0-61.7)	53.0 (43.1-61.9)	50.5 (44.3-61.5)
Central Sacral Pelvic Line	22.2 (12.0-38.4)	29.7 (13.6-46.0)	24.6 (13.1-39.1)	22.7 (12.8-44.9)
Fat Artophy	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)
Immediate Post-Operative Radiographic Measurements				
PCVL - Sacrum	54.2 (25.9-84.8)	37.3 (11.6-62.3)	59.9 (30.3-87.9)	35.9 (10.2-69.0)
PCVL - Acetabulum	99.4 (65.6-135.4)	90.4 (52.9-115.0)	106.8 (72.0-136.5)	86.1 (49.5-115.5)
PCVL - Medial Maleolus	62.4 (34.4-86.0)	48.8 (19.2-75.5)	65.1 (40.2-87.3)	43.5 (12.5-74.9)
PCVL - UIV Tulip	19.4 (-7.7-36.4)	3.4 (-22.9-25.8)	20.1 (-3.8-36.8)	-1.1 (-28.6-25.8)
PCVL - UIV Grade	0.0 (0.0-1.0)	1.0 (0.0-1.0)	0.0 (0.0-1.0)	1.0 (0.0-1.0)
PCVL - Thoracic Apex	15.8 (-10.2-38.1)	3.8 (-23.8-29.5)	18.1 (-6.8-39.8)	3.4 (-24.1-25.8)
PCVL TA and Sacrum Grade	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)
SVA	30.9 (4.0-65.7)	56.1 (31.6-79.7)	25.3 (1.4-61.5)	60.3 (36.0-81.3)
C2PA	11.2 (7.0-16.1)	12.2 (6.3-17.4)	11.5 (7.1-16.5)	11.2 (6.2-17.2)
PJA	7.5 (3.5-12.4)	9.8 (5.0-15.2)	7.5 (3.6-12.5)	8.9 (4.8-16.0)
Cervical Lordosis (C2-C7)	20.4 (12.1-30.9)	24.0 (13.6-31.0)	21.0 (12.6-30.5)	25.3 (14.1-32.7)
Thoracic Kyphosis (T1-T12)	42.0 (33.3-51.0)	44.2 (36.1-50.4)	42.1 (34.2-49.7)	44.8 (37.2-50.4)
Lumbar Lordosis (L1-L5)	35.7 (27.0-45.1)	33.5 (24.6-42.8)	36.6 (27.3-46.4)	33.5 (24.2-43.5)
SS	30.9 (24.9-40.2)	29.7 (24.3-36.5)	30.7 (24.4-39.6)	29.6 (23.4-35.8)
PT	19.0 (12.6-25.5)	20.5 (14.5-29.3)	18.8 (13.2-25.2)	20.4 (13.6-26.7)
PI	51.5 (40.6-62.5)	49.3 (42.5-59.0)	51.9 (40.3-60.7)	48.8 (42.3-57.1)
Central Sacral Pelvic Line	17.5 (8.6-28.8)	19.2 (9.1-28.2)	17.4 (9.0-30.9)	19.7 (9.2-27.7)
Rod Density	11.5 (11.0-12.0)	12.0 (11.0-12.0)	11.5 (11.0-12.0)	12.0 (11.0-12.0)

Table 3: Description of follow-up post-operative radiographic and metadata about classificaiton variables and for the whole dataset and dataset where patients with large missingness are removed. Data is expressed by the Median (IQR) or N (%).

	Whole Dataset		**Reduced Dataset**	
Variables	**Non-PJK**	**PJK**	**Non-PJK**	**PJK**
No. of Cases	278	88	201	67
Most Recent FU Post-Operative Measurements				
PCVL - Sacrum	71.3 (30.9-94.0)	43.5 (4.6-83.5)	75.7 (41.7-97.0)	44.0 (6.8-84.8)
PCVL - Acetabulum	121.2 (88.6-144.6)	99.2 (58.4-135.3)	125.2 (96.9-147.4)	105.1 (61.8-137.1)
PCVL - Medial Maleolus	78.7 (52.8-96.2)	58.1 (27.5-86.6)	80.6 (54.3-98.9)	58.5 (27.5-87.0)
PCVL - UIV Tulip	20.8 (-5.8-37.8)	-6.5 (-34.4-24.0)	23.0 (-0.5-37.8)	-6.5 (-37.3-20.4)
PCVL - UIV Grade	0.0 (0.0-1.0)	1.0 (0.0-2.0)	0.0 (0.0-1.0)	1.0 (0.0-2.0)
PCVL - Thoracic Apex	22.0 (-7.0-40.9)	-5.8 (-32.4-18.0)	23.9 (-1.7-41.6)	0.9 (-29.6-18.0)
PCVL TA and Sacrum Grade	1.0 (1.0-2.0)	2.0 (1.0-2.0)	1.0 (1.0-2.0)	1.5 (1.0-2.0)
SVA	23.2 (-4.4-60.9)	54.0 (17.6-79.3)	15.8 (-6.5-56.5)	55.7 (22.4-78.7)
C2PA	12.6 (7.2-17.0)	15.2 (6.5-21.1)	13.1 (8.1-17.4)	15.2 (6.5-21.0)
PJA	7.8 (4.1-13.4)	16.6 (7.2-21.5)	7.8 (4.0-13.6)	17.5 (9.2-22.6)
Cervical Lordosis (C2-C7)	23.1 (13.4-32.5)	25.9 (12.2-34.8)	23.6 (13.5-32.8)	26.7 (13.5-35.1)
Thoracic Kyphosis (T1-T12)	45.0 (34.9-52.3)	47.5 (38.8-55.7)	45.3 (35.4-52.3)	48.7 (39.5-56.8)
Lumbar Lordosis (L1-L5)	34.2 (25.1-45.5)	31.5 (24.7-40.0)	34.4 (25.8-46.1)	31.9 (25.4-39.0)
SS	30.4 (24.1-40.0)	27.4 (20.2-34.9)	30.2 (24.2-39.5)	26.8 (20.1-34.0)
PT	23.0 (15.9-30.6)	25.2 (19.1-31.4)	22.3 (15.3-30.3)	25.0 (19.2-30.6)
PI	53.0 (42.2-64.6)	50.6 (43.5-63.1)	53.2 (43.9-65.0)	52.1 (44.4-62.8)
Central Sacral Pelvic Line	18.4 (7.2-32.1)	17.3 (9.6-31.1)	18.6 (7.4-32.5)	17.2 (9.1-31.1)
Metadata About Classificaiton				
Final Postop PJA	7.7 (3.7-13.2)	13.8 (6.1-20.8)	7.8 (4.0-13.6)	17.5 (9.2-22.6)
Final Δ PJA	1.4 (-2.8-6.6)	6.5 (-1.0-13.0)	1.0 (-3.3-7.2)	8.6 (1.6-14.3)
SRS Change	0.0 (0.0-0.2)	0.0 (0.0-0.0)	0.1 (0.0-0.2)	0.0 (0.0-0.1)
ODI Change	-0.2 (-0.4–0.0)	-0.3 (-0.4-0.0)	-0.2 (-0.4–0.1)	-0.3 (-0.4-0.0)
Overall PRO Change	0.2 (0.0-0.5)	0.2 (-0.0-0.4)	0.3 (0.1-0.5)	0.2 (-0.0-0.4)
PRO Decline No	222 (79.9%)	64 (72.7%)	166 (82.6%)	48 (71.6%)
Yes	56 (20.1%)	24 (27.3%)	35 (17.4%)	19 (28.4%)

Table 4: ROC-AUC average over 100 trials of XGBoost model and XGBoost+ model by varying the size of H and J

J -H	**Method**	**J**								
		1	2	3	4	5	6	7	8	9
0	XGBoost	0.997	0.944	0.912	0.881	0.855	0.841	0.835	0.819	0.803
	XGBoost+	**0.997**	**0.959**	**0.927**	**0.897**	**0.871**	**0.856**	**0.846**	**0.830**	**0.817**
1	XGBoost		0.748	0.782	0.808	0.806	0.791	0.807	0.783	0.788
	XGBoost+		**0.916**	**0.924**	**0.917**	**0.890**	**0.865**	**0.860**	**0.838**	**0.829**
2	XGBoost			0.675	0.705	0.740	0.757	0.762	0.763	0.773
	XGBoost+			**0.877**	**0.904**	**0.893**	**0.886**	**0.861**	**0.853**	**0.843**
3	XGBoost				0.614	0.667	0.732	0.723	0.728	0.727
	XGBoost+				**0.827**	**0.897**	**0.899**	**0.883**	**0.861**	**0.847**
4	XGBoost					0.607	0.657	0.681	0.706	0.720
	XGBoost+					**0.827**	**0.889**	**0.879**	**0.871**	**0.865**
5	XGBoost						0.605	0.624	0.661	0.686
	XGBoost+						**0.855**	**0.875**	**0.889**	**0.871**
6	XGBoost							0.574	0.635	0.651
	XGBoost+							**0.813**	**0.863**	**0.873**
7	XGBoost								0.560	0.631
	XGBoost+								**0.775**	**0.873**
8	XGBoost									0.565
	XGBoost+									**0.796**

Table 5: Performance on the PJK Dataset based on 15 to 85 harm to benefit ratio

Method	**ROC-AUC** [1]	**Precision** [1]	**Sensitivity** [1]	**Specificity**[1]
Decision Tree	0.48 ± 0.02	0.41 ± 0.18	0.26 ± 0.15	0.79 ± 0.16
Random Forest	0.67 ± 0.03	0.23 ± 0.18	0.04 ± 0.05	0.96 ± 0.02
SVM	0.61 ± 0.04	0.40 ± 0.15	0.2 ± 0.05	0.88 ± 0.05
SVM+	0.63 ± 0.03	0.42 ± 0.16	0.19 ± 0.10	0.90 ± 0.09
IPL	0.69 ± 0.06	0.48 ± 0.03	0.24 ± 0.11	0.91 ± 0.08
GB+	0.69 ± 0.04	0.45 ± 0.15	0.1 ± 0.05	**0.98 ± 0.02**
XGBoost	0.67 ± 0.05	**0.53 ± 0.11**	0.29 ± 0.14	0.88 ± 0.08
XGBoost+	**0.72 ± 0.05**	**0.53 ± 0.10**	**0.30 ± 0.16**	0.91 ± 0.07

[1] Average over 50 trials. Result are reported with mean value ± standard deviation.

Table 6: Ablation study for different value of α on all datasets

Experiment	$\alpha = 0$	$\alpha = \alpha^*$	$\alpha = 1$
Synthetic Dataset[1]	0.83	0.98	**0.98**
PBC Dataset[1]	0.79	0.82	**0.82**
PJK Dataset[1]	0.67	**0.72**	0.69

[1] Optimal alpha (α^*) for each experiment base on gridsearch are: Synthetic Dataset (0.85), PBC Dataset (0.7) and PJK dataset (0.55)

Table 7: Performance on Synthetic dataset comparison for XGBoost+to other methods

Method	**p-value**
Decision Tree	< 0.001
Random Forest	< 0.001
SVM	< 0.001
SVM+	< 0.001
IPL	< 0.001
GB+	< 0.001
XGBoost	< 0.001

Table 8: Performance on PBC dataset comparison for XGBoost+to other methods

Method	**p-value**
Decision Tree	< 0.001
Random Forest	< 0.001
SVM	< 0.001
SVM+	0.006
IPL	< 0.001
GB+	< 0.001
XGBoost	< 0.001

Table 9: Performance on PJK dataset comparison for XGBoost+to other methods

Method	**p-value**
Decision Tree	< 0.001
Random Forest	< 0.001
SVM	< 0.001
SVM+	< 0.001
IPL	< 0.001
GB+	< 0.001
XGBoost	0.028