Dataset assembling example: Given the condition 'Number of arcs', the first of the six datasets was constituted by 100% of plans having the characteristic {a} (one arc) plus 0% of plans with the characteristic {b} (two arcs), expressed by $\{a\%/b\%\}$ = 100%/0%. And subsequently the other five datasets had 80%/20%, 60%/40%, 40%/60%, 20%/80%, and 0%/100% proportions. (Each dataset has 210 plans)

Figure 1. ROC-AUC results and its standard deviations for (a) RF, (b) XG-Boost, and (c) NN models, considering the reference dataset, each heterogeneity source (Number of arcs // Treatment unit), and its different proportions [a%-b%]. The datasets with more dominant condition presented higher AUC values, and NN models showed lower variability.

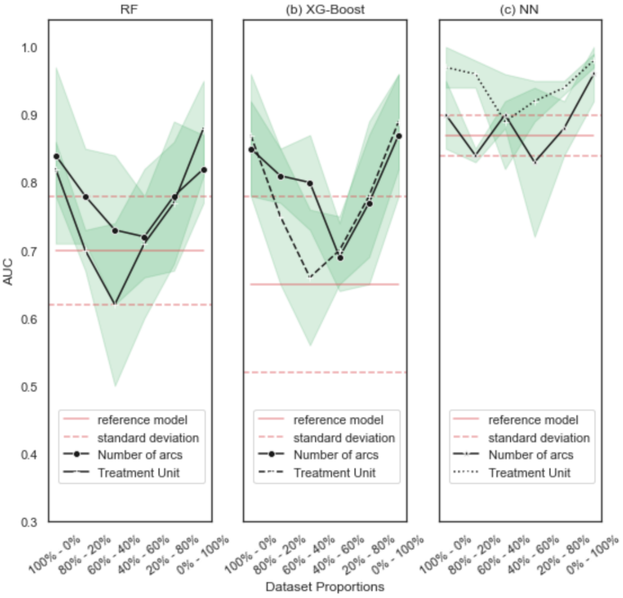
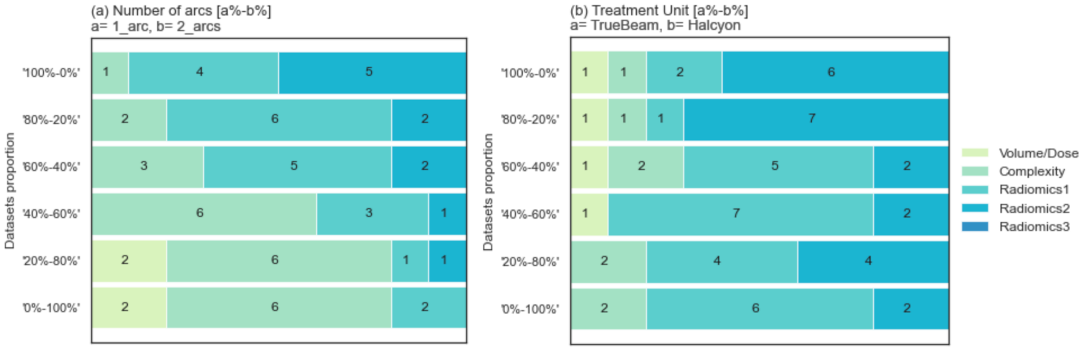


Figure 2. Features variations for each dataset and heterogeneity source considering the predictor features (plan parameters based on volume or dose metrics, modulation complexity metrics, and radiomic metrics extracted from the MLC movements, Radiomics1, the dose distribution, Radiomics2, and the dose blended images used for portal dosimetry, Radiomics3). (a) The number of arcs: models based on datasets with two-arcs plans rely more on complexity metrics than models with plans with a single plan. (b) Treatment Unit: models based on plans optimized for Halcyon (dual-layer MLC) rely more on the modulation maps than models optimized for TrueBeam (single-layer MLC). This figure shows how the same model (RF) relies on different predictor factors depending on the proportion of plans with specific treatment conditions. Indeed, the more heterogeneous datasets generate predicting models based on random associations rather than physical aspects, reducing the reliability and interpretability when applied in practice.


(a) Number of arcs [a%-b%]
a= 1_arc, b= 2_arcs
(b) Treatment Unit [a%-b%]
a= TrueBeam, b= Halcyon
Datasets proportion
'100%-0%'
'80%-20%'
'60%-40%'
'40%-60%'
'20%-80%'
'0%-100%'
Volume/Dose
Complexity
Radiomics1
Radiomics2
Radiomics3