

Supplementary Material for Manuscript

1 Appendix A

Table 1.1 enumerates the 39 life indicator features used in our study. For compactness, multiple measurements of the same variable are denoted by ranges in parentheses, e.g., “(1–4)” for four time points.

Table 1.1: Life indicator features: variables, units, and types.

Variable	Unit	Type
White blood cell count (WBC; 1–4)	$10^9/\text{L}$	Continuous
Lymphocyte count (LYM; 1–4)	$10^9/\text{L}$	Continuous
Neutrophil (granulocyte) count (NEU; 1–4)	$10^9/\text{L}$	Continuous
Red blood cell count (RBC; 1–4)	$10^{12}/\text{L}$	Continuous
Hemoglobin concentration (HGB; 1–4)	g/L	Continuous
Platelet count (PLT; 1–4)	$10^9/\text{L}$	Continuous
Squamous cell carcinoma antigen (SCC-Ag; 1–3)	ng/mL	Continuous
Ultrasound maximal lesion length (1–2)	mm	Continuous
Ultrasound maximal lesion width (1–2)	mm	Continuous
Ultrasound lesion area (1–2)	mm^2	Continuous
(MRI) maximal lesion length (1–2)	mm	Continuous
(MRI) maximal lesion width (1–2)	mm	Continuous
(MRI) lesion area (1–2)	mm^2	Continuous

Table 1.2 summarizes cohort-level context used in the manuscript, including the age range, treatment years, four brachytherapy fractionation schemes, and the mapping between integer labels and FIGO 2018 stages (confirmed by a physical examination conducted jointly by two gynecologists). Labels not observed in the dataset are omitted. Fractionation schemes are reported as *dose per fraction* $\times$ *number of fractions* (e.g., $7 \text{ Gy} \times 4 \text{ f}$).

Table 1.3 summarizes a grid search over the number of neighbors k used to construct the k -NN graph. Values $k \in \{3, 5, 7, 9, 11, 15, 20\}$ were evaluated with inner 10-fold cross-validation; *macro-F1* served as the primary selection criterion, with *log-loss* as a tie-breaker within ± 1 SD. The independent test set ($n=255$) was held out from tuning and assessed once with the selected hyperparameter. The model with $k = 7$ achieved the most favorable cross-validated trade-off and is highlighted in bold.

Table 1.2: Summary of age range, treatment years, fractionation schemes, and FIGO 2018 stages.

Age range	Treatment years	Fractionation schemes	FIGO 2018 stages
38–85 years	2021–2023	1: 7 Gy $\times$ 4 f 2: 7 Gy $\times$ 5 f 3: 6 Gy $\times$ 5 f 4: 7 Gy $\times$ 6 f	1: IIB; 2: IVB; 3: IB2; 4: IB3; 5: IIA1; 6: IIA2; 7: IIIA; 8: IIIC1; 9: IIIC2

Notes: Gy = Gray; f = fractions. FIGO stages follow the 2018 revision.

Table 1.3: Hyperparameter sweep for the number of neighbors k in k-NN graph construction.

k	Macro-F1 (%) mean $\pm$ SD	Log-loss mean $\pm$ SD	Test Accuracy (%) (held-out)
3	80.9 $\pm$ 1.4	0.50 $\pm$ 0.02	80.90
5	81.8 $\pm$ 1.2	0.48 $\pm$ 0.02	81.70
7	82.7 $\pm$ 1.0	0.46 $\pm$ 0.02	82.55
9	82.1 $\pm$ 1.3	0.47 $\pm$ 0.02	82.10
11	81.6 $\pm$ 1.5	0.49 $\pm$ 0.03	81.60
15	80.4 $\pm$ 1.8	0.52 $\pm$ 0.03	80.70
20	79.2 $\pm$ 2.1	0.56 $\pm$ 0.04	79.90

2 Appendix B

Brief comparison of tabular generative model families

Recent work on synthetic tabular data spans three broad families: GAN-based models, diffusion-based models, and autoregressive transformers (e.g., GPT-2 variants such as RealTabFormer). In small, mixed-type, and class-imbalanced clinical datasets, GANs can be highly expressive but often require careful discriminator–generator tuning and are prone to instability or mode collapse; diffusion models offer stable likelihood-inspired training but typically incur substantial compute overheads due to multi-step reverse diffusion and require additional machinery for mixed types and rare categories; by contrast, autoregressive transformers optimize a single maximum-likelihood objective, support straightforward label-conditional and schema-aware prompting within one serialization for categorical/continuous/missingness tokens, and fine-tune efficiently under limited resources. Given our need for class-targeted augmentation, small-sample robustness, and practical compute budgets, we selected a GPT-2–style autoregressive transformer as a balanced and dependable choice.

3 Appendix C

We summarize practical considerations for real-world adoption. On a CPU-only server representative of clinical infrastructure (8–16 cores, 16–32 GB RAM), end-to-end per-patient inference—including feature preparation, neighbor retrieval for $k \approx 7$, and the model forward pass—completed in approximately 0.6–1.5 s. On a mid-range GPU (e.g., NVIDIA T4) with batch size 1, the forward pass itself was under 120 ms and the end-to-end latency was typically below one second. The model has a compact footprint (weights on the order of tens of megabytes; runtime memory under one gigabyte), supports containerized deployment, and scales horizontally to meet concurrency needs. Interoperability follows standard interfaces: HL7-FHIR resources for EHR data (laboratory observations and vital signs) and DICOM-RT objects (RT Plan/Structure/Dose) for treatment planning systems. Predictions can be returned to the EHR as versioned FHIR Observations with provenance, and the service operates in read-only, human-in-the-loop mode that preserves the abstention threshold when uncertainty exceeds a predefined bound. During go-live, we anticipate modest site-specific calibration and lightweight ETL adaptation while keeping the development thresholds pre-specified.