Evaluation of Medical Validity in Colorectal Cancer Clinical Synthetic Data and Enhancement through Data-Model-based Ensemble

장수영[1], 김종현[1], 이주현[1], 최영조[2]

[1] Department of Biomedical Systems Informatics, Yonsei University College of Medicine
[2] Digital Analytics, Yonsei University

Colorectal cancer (CRC)

- Colorectal cancer (CRC) is the third most common cancer, and the fourth cause of cancer-related death in Korea [1,2].

- Approximately 30-40% of patients experience a recurrence after surgery. Furthermore, the duration between the initial treatment and the onset of recurrence is closely linked to overall survival [3,4].

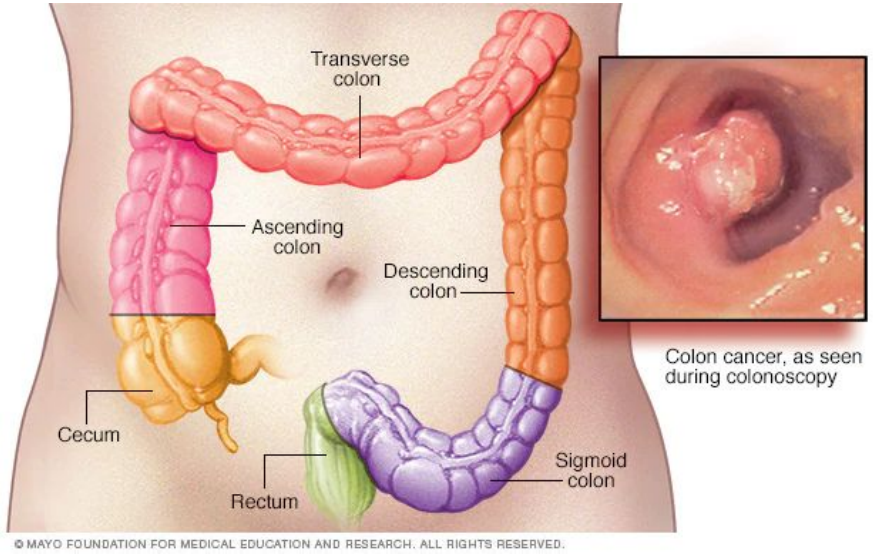


Figure 1. Colorectal cancer [5]

[1] https://www.cancer.go.kr/lay1/S1T639C641/contents.do
[2] https://www.cancer.go.kr/lay1/S1T645C646/contents.do
[3] Tjandra, J.J., Chan, M.K.Y. Follow-Up After Curative Resection of Colorectal Cancer: A Meta-Analysis. Dis Colon Rectum 50, 1783–1799 (2007). https://doi.org/10.1007/s10350-007-9030-5
[4] Walker, A. S., Johnson, E. K., Maykel, J. A., Stojadinovic, A., Nissan, A., Brucher, B., Champagne, B. J., & Steele, S. R. (2014). Future directions for the early detection of colorectal cancer recurrence. Journal of Cancer, 5(4), 272–280. https://doi.org/10.7150/jca.8871
[5] https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669

Machine learning & Synthetic data

- Machine learning is used to predict diagnosis, treatment response, recurrence, based on patient data [6].
- In the case of colorectal cancer, it is difficult to accurately predict the complex patterns of patient data due to various recurrence patterns and related factors, so technologies such as machine learning are utilized [7].
- However, real world medical data is sensitive and has limited accessibility [8].

[6] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542*(7639), 115-118.
[7] Huang, Y. Q., Liang, C. H., He, L., Tian, J., Liang, C. S., Chen, X., ... & Liu, Z. Y. (2016). Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *Journal of clinical oncology*, *34*(18), 2157-2164.
[8] Terry, N. P. (2012). Protecting patient privacy in the age of big data. *UMKC L. Rev.*, *81*, 385.

Machine learning & Synthetic data

- To overcome the limitations of real world medical data, synthetic data is used to protect privacy while maintaining the statistical properties of real world data [9].
- Synthetic data allows models to be trained on larger amounts of data, and the diversity of the data can improve the generalization performance of the model [10].

[9] DuMont Schütte, A., Hetzel, J., Gatidis, S., Hepp, T., Dietz, B., Bauer, S., & Schwab, P. (2021). Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, *4*(1), 141.
[10] Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28-45.

Objective

- To predict the recurrence and its timing in colorectal cancer patients, we develop and evaluate a medically valid prediction model using synthetic data.

  - Generate synthetic data for predicting recurrence and its timing in colorectal cancer patients.

  - Develop and evaluate a medically valid prediction model using the synthetic data.

  - Enhance the predictive accuracy with ensemble techniques.

Dataset

Table1. Dataset describe

column name	label name	column name	label name	column name	label name
PT_SBST_NO	환자번호	SGPT_VNIN_NM	외과병리정맥침명	SGPT_SRMG_DCTS_STAT_NM	외과병리수술절제면원위암조직크기상태명
BSPT_FRST_DIAG_CD	기본환자최초진단코드	SGPT_ANIN_NM	외과병리혈관조영침윤명	SGPT_SRMG_PCTS_STAT_NM	외과병리수술절제면근위암조직크기상태명
BSPT_IDGN_AGE	기본환자나이	SRGC_DIFF	외과병리접수일자	SGPT_SRMG_RCTS_STAT_NM	외과병리수술절제면방사형암조직크기상태코드
BSPT_SEX_CD	기본환자성별코드	IMPT_HM1E_RSLT_NM	면역병리HMLH1검사결과명	SGPT_SRMV_LN_CNT	외과병리절제림프절수
BSPT_STAG_VL	기본환자병기값	IMPT_HP2E_RSLT_NM	면역병리HP2E검사결과명	SGPT_TUMR_BUDD_NM	외과병리종양발아명
BSPT_T_STAG_VL	기본환자T병기값	IMPT_HS2E_RSLT_NM	면역병리HS2E검사결과명	CEA_DIAG	진단일과 가장 가까운 일자의 CEA Value
BSPT_N_STAG_VL	기본환자N병기값	IMPT_HS6E_RSLT_NM	면역병리HS6E검사결과명	CEA_MAX	최대 CEA Value
BSPT_M_STAG_VL	기본환자M병기값	MNTY_DIFF	면역병리접수일자	CEA_MIN	최소 CEA Value
MLPT_BRME_RSLT_NM	분자병리BRAF_MUTATION검사결과명	DEAD_DIFF	사망진단일자	CEA_RCNF	재발일과 가장 가까운 일자의 CEA Value
MLPT_KE2E_RSLT_NM	분자병리KRASMUTATION_EXON2검사결과명	OPRT_CLCN_OPRT_KIND_NM	수술대장암수술종류명	CEA_MED	CEA 중위값
MLPT_KRES_RSLT_NM	분자병리KRASMUTATION검사결과명	OPRT_CURA_RSCT_NM	수술근치적절제술명	CEA_OPRT	최초 수술과 가장 가까운 일자의 CEA Value
MLPT_MSIE_RSLT_NM	분자병리NRASMUTATION검사결과코드	OPRT_DIFF	수술일자	SGPT_PATL_M_STAG_VL	외과병리병리학적M병기값
MLPT_NREX_RSLT_NM	MLPT_NREX_RSLT_NM	ADJ_REGN_NM	Adjuvant항암치료명	SGPT_PATL_N_STAG_VL	외과병리병리학적N병기값
MLCR_DIFF	분자병리접수일자	ADJ_CYCL_VL_MAX	Adjuvant항암치료 최대 Cycle 수	SGPT_PATL_T_STAG_VL	외과병리병리학적T병기값
BPTH_BPSY_RSLT_CONT	생체검사병리생체검사결과내용	ADJ_CYCLE_PERIOD	Adjuvant항암치료일자	SGPT_PATL_STAG_VL	외과병리병리학적병기값
BPTH_CELL_DIFF_NM	생체검사병리세포분화명	PAL_REGN_NM_LINE	Palliative항암치료명	NEO_CYCLE_PERIOD	Neo-Adjuvant항암치료일자
BPSY_DIFF	생체검사병리접수일자	PAL_CYCL_VL_MAX_LINE	Palliative항암치료 최대 Cycle 수	DEAD_NFRM_DEAD	사망여부
SGPT_CELL_DIFF_NM	외과병리세포분화명	PAL_CYCLE_PERIOD_LINE	Palliative항암치료일자	SGPT_NERV_PREX_NM	외과병리신경주위침윤명
SGPT_MTST_LN_CNT	외과병리전이림프절수	NEO_REGN_NM	Neo-Adjuvant항암치료명	NEO_CYCL_VL_MAX	Neo-Adjuvant항암치료최대Cycle수

Data preprocessing

- Removed columns with Nan values greater than or equal to 90
- Remove variables that cannot be clinically present at the time of prediction

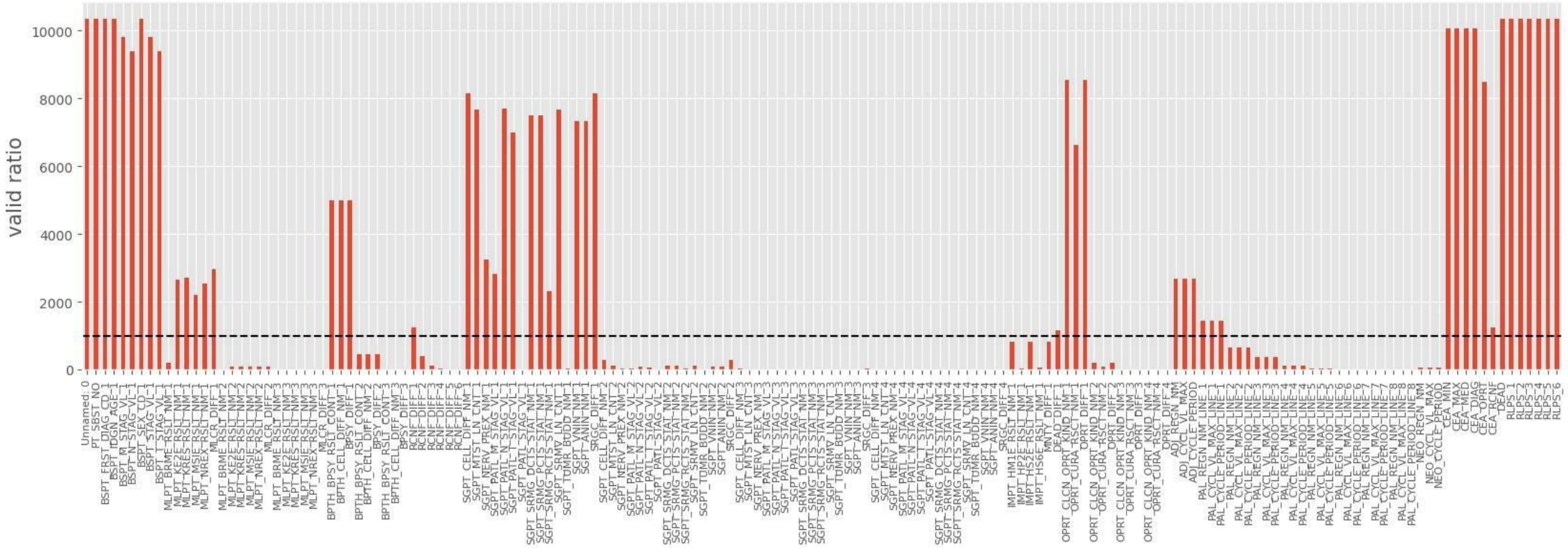


Figure 2. Number of values for all variables

Study design

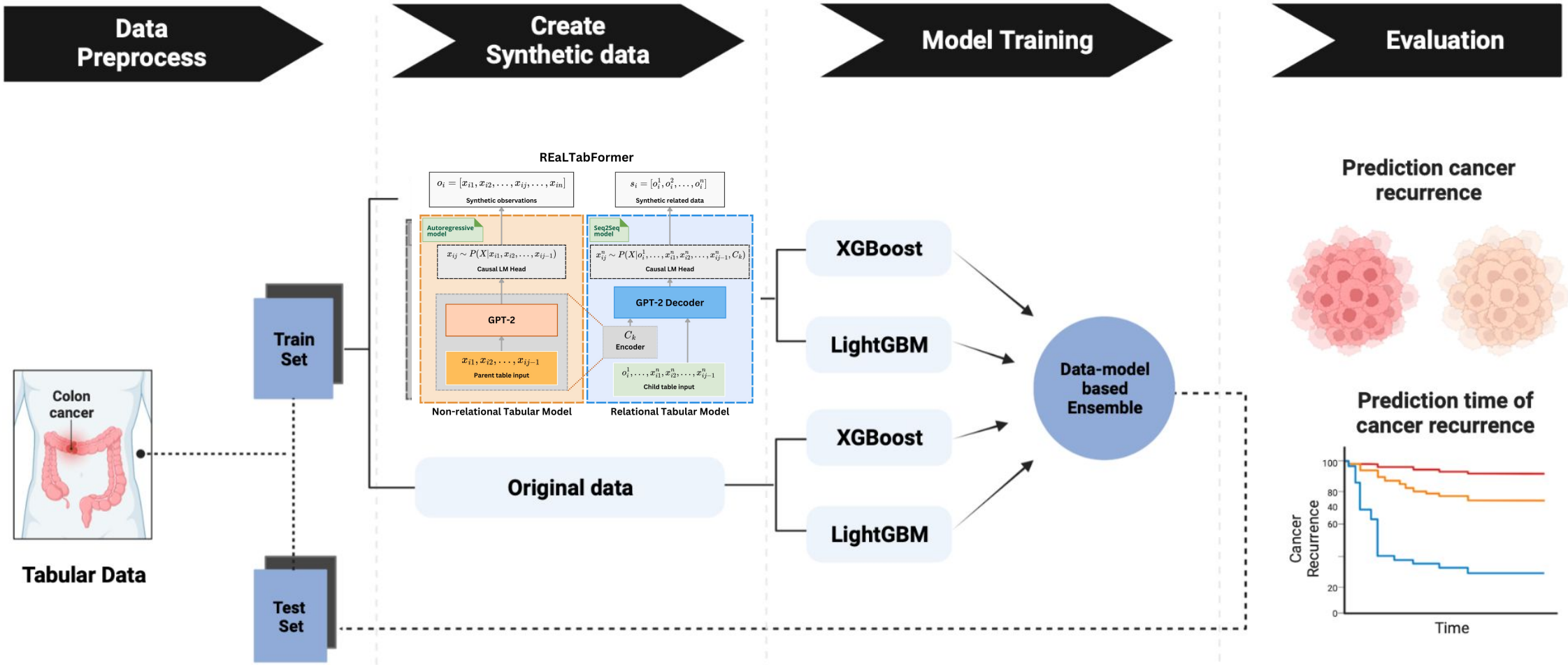


Figure 3. Study design of Data-Model-based Ensemble model for predicting colorectal cancer

REaLTabFormer Model for Tabular Data

- Based on: GPT-2 architecture, optimized for handling tabular data.

- Designed for immediate application on any table with independent observations, offering versatility in modeling diverse datasets.

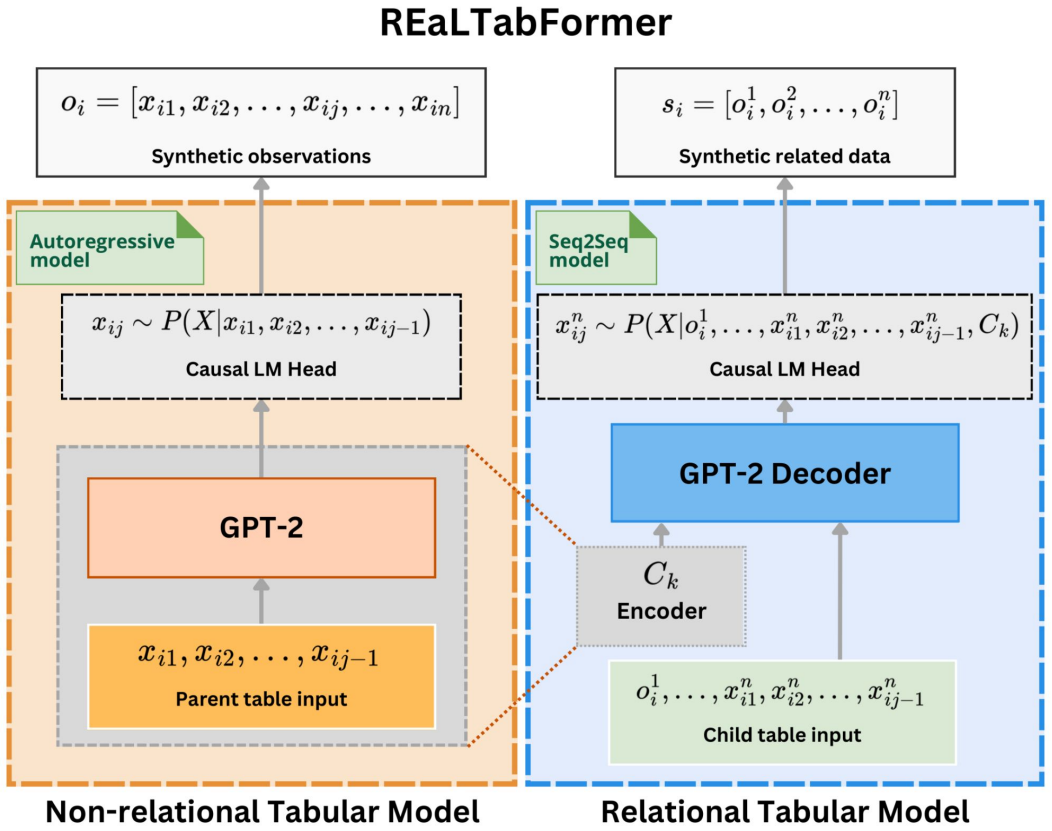


Figure 4. Illustration of the REaLTabFormer model [11]

[11] Solatorio, A. V., & Dupriez, O. (2023). REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers. arXiv preprint arXiv:2302.02041.

REaLTabFormer Model for Tabular Data

- **stratifiedkfold(n_splits=5) -> Create synthetic data**

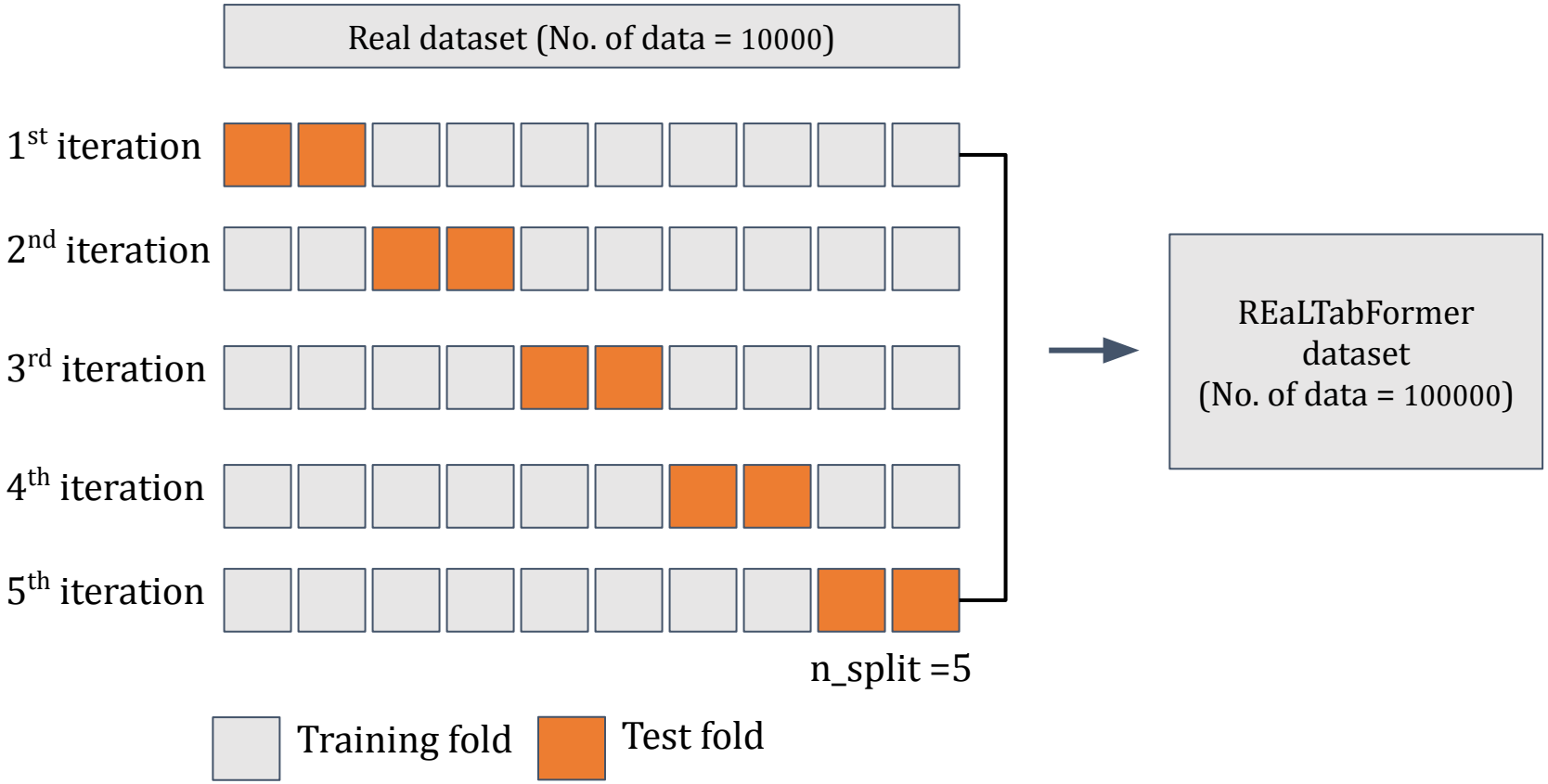


Figure 5. Apply stratifiedkfold (n_splits=5) to generate synthetic data

Prediction model (Data-Model based Ensemble)

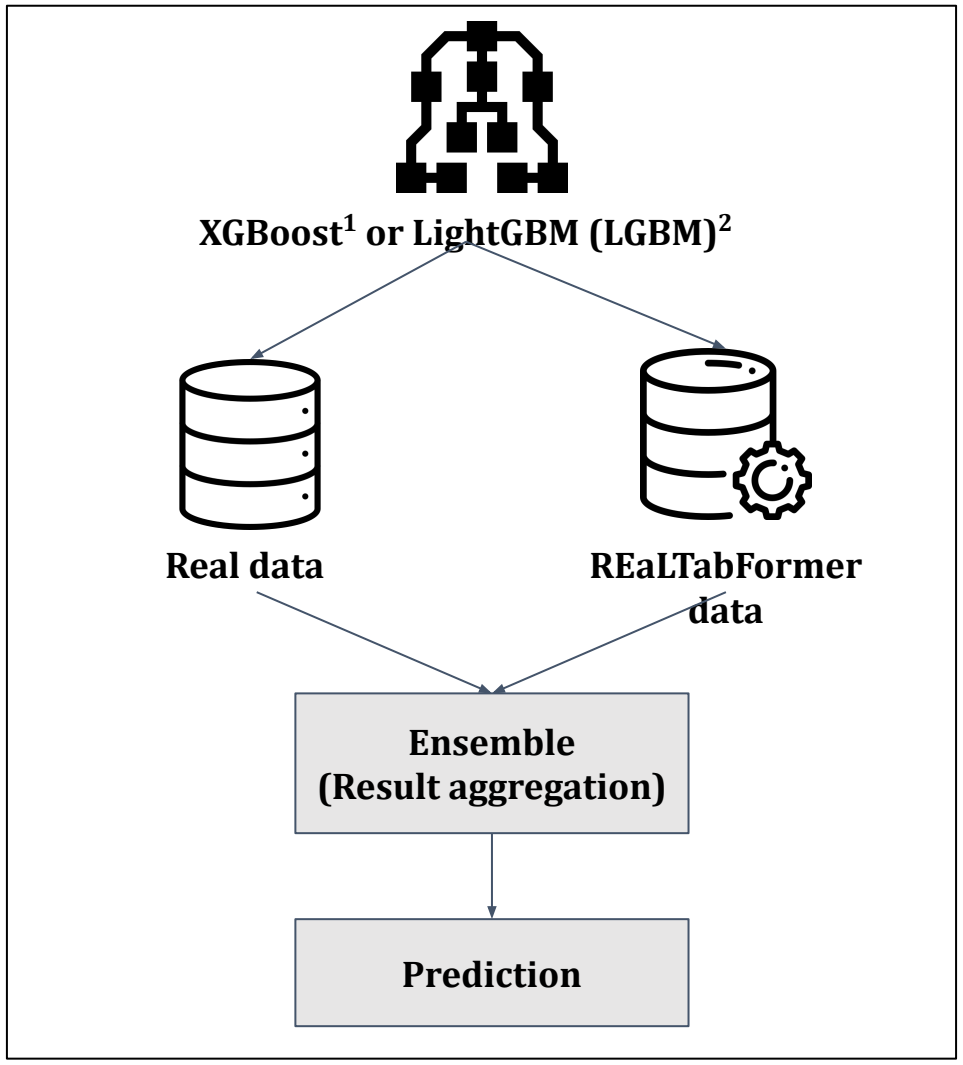


Figure 6. Framework for Ensemble model 1[1] (using XGBoost) and Ensemble model 2[2] (using LightGBM)

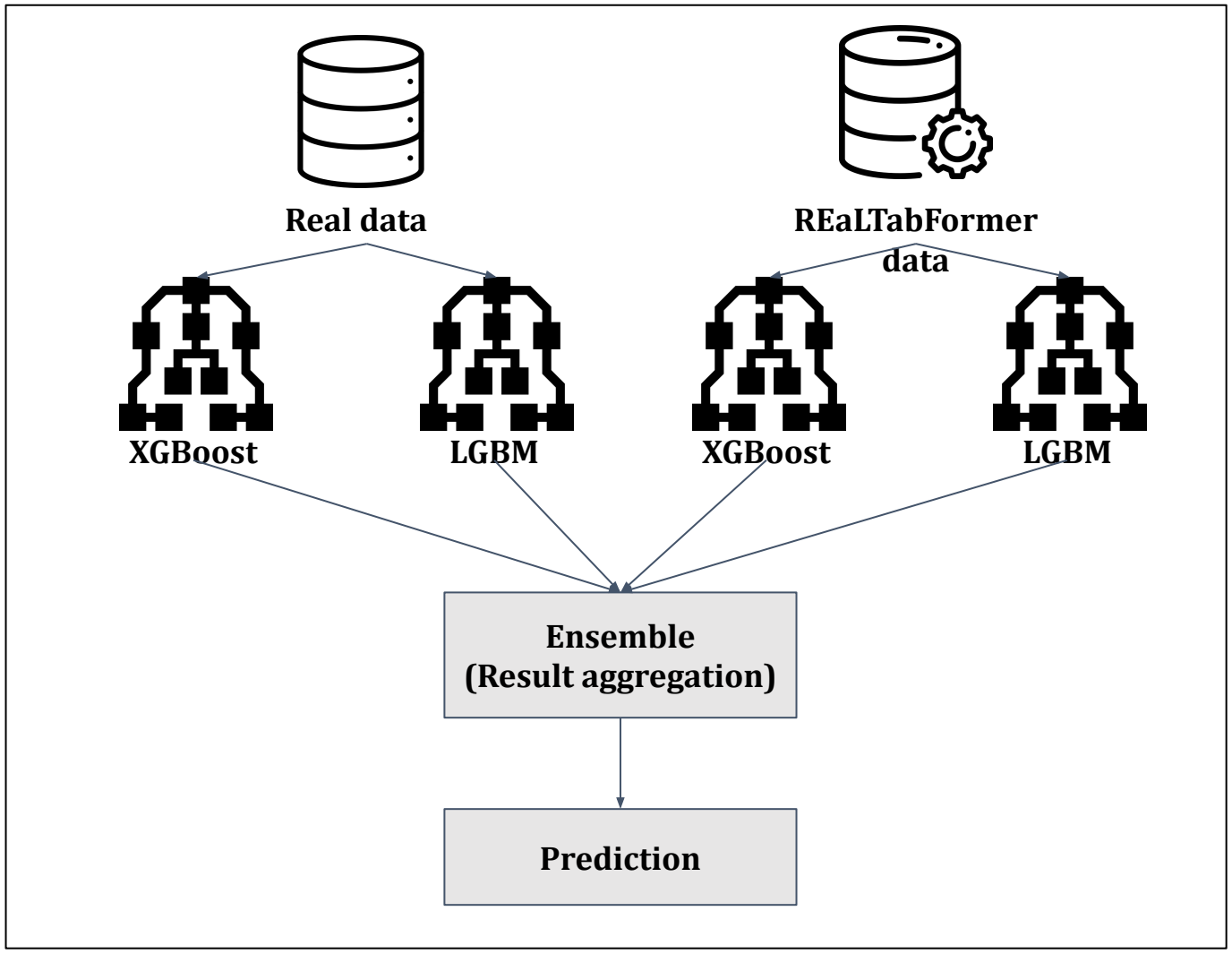


Figure 7. Framework for Ensemble model 3

Prediction model for time series data (colorectal cancer recurs)

- We designed a model to predict colorectal cancer recurrence in 6-month intervals.

  : As a result, a model was designed to predict relapse within 6 months, relapse within 1 year, and etc.

- We predict colon cancer recurrence using real data and given synthetic data with xgboost.

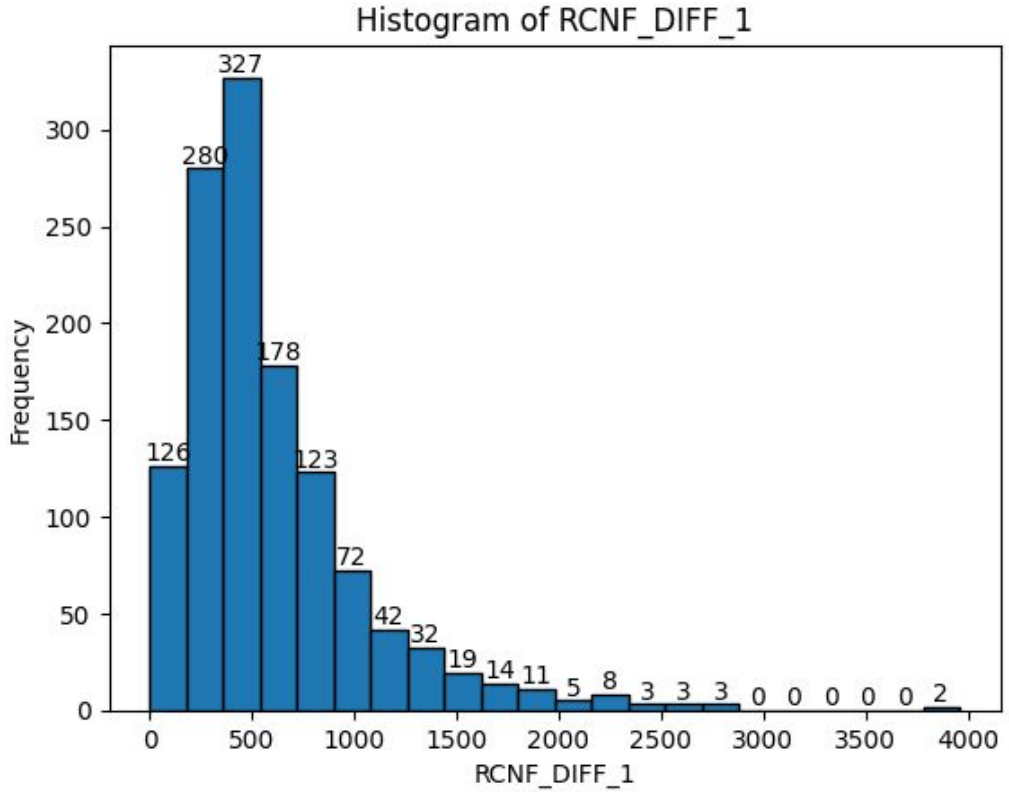


Figure 8. Histogram of recurrence time intervals (days)

External validation using genomic data

- We developed a model for predicting sepsis severity using single-cell RNA-seq data from sepsis patients.
- Filtered 136 septic shock marker gene sets from single cell RNA seq data.
- Validated model performance using real data and synthesised data based on real data
- Identified potential improvements in model performance using a previously developed data/model based ensemble model.

Table2. Single-cell RNA gene expression matrix

Cellular barcode	N4BP2L2	NAA10	ACAP1	NFKBIA
5_1_ATCACGAAGTGGGAAA-1	0.878078650648452	0	0.878078650648452	0
5_1_GATCACAGTCCTTGTC-1	1.95258258018174	0	1.95258258018174	0.920966104239639
5_1_TCATCATGTAGAGACC-1	1.392703431126	0	0.921421269761609	1.392703431126
5_1_AACAAGATCTGGGTCG-1	1.71646472012716	0	0.924896324040134	0
5_1_ACCAACAAGGGCCTCT-1	1.74956480449501	0	1.42745211425187	0.949351336568677
5_1_CGGAATTGTAATTAGG-1	1.75574247070183	0	0.953938847214584	0

Results of model exploration

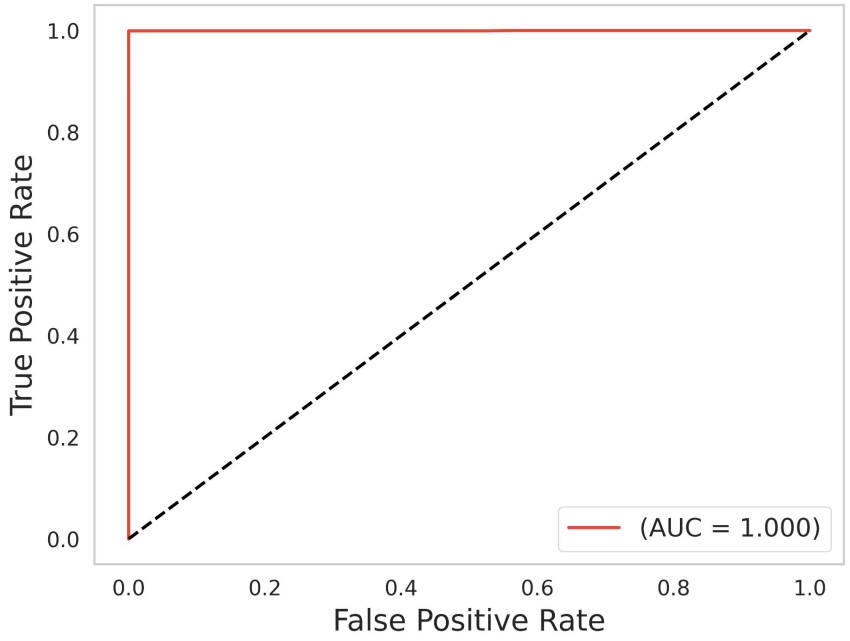


Figure 9. Result of real data (AUC = 1.00)

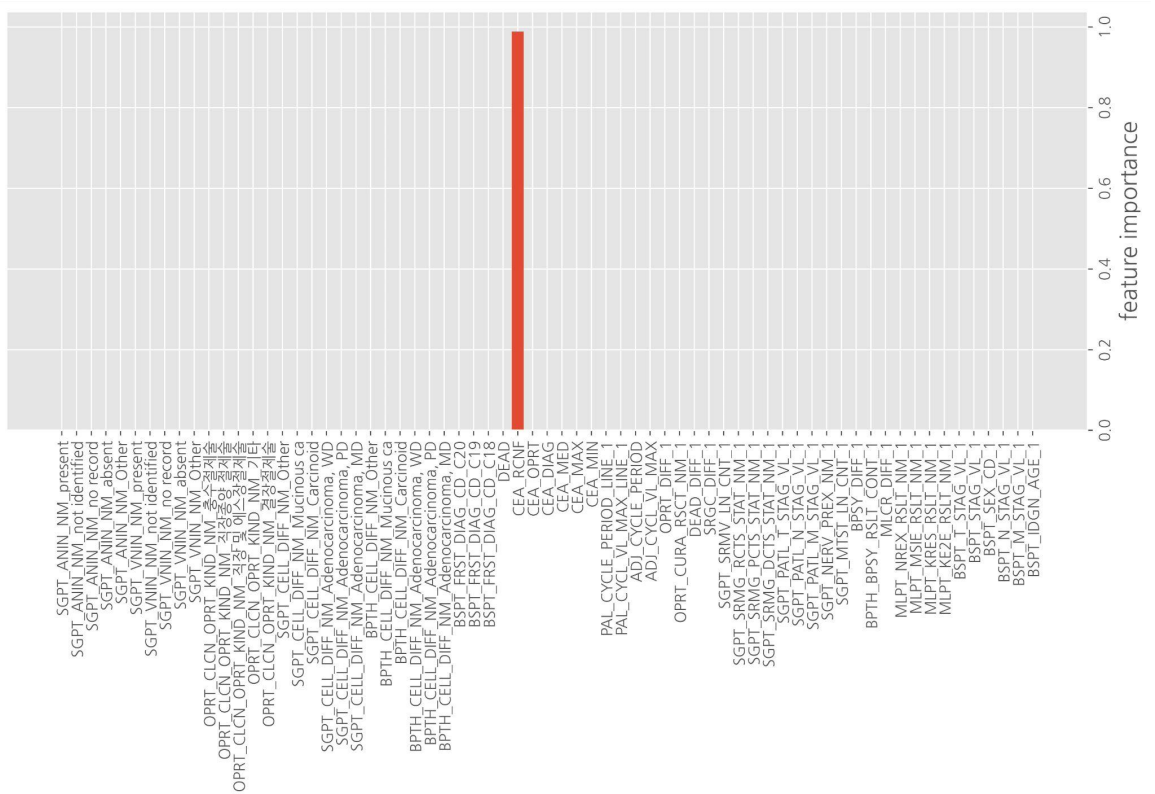


Figure 10. Feature importance of real data

- We found the AUC value to be 1.0 when developing the model using all variables.
- However, we did not exclude variable names with information unobtainable at the time of prediction.

Results of model exploration

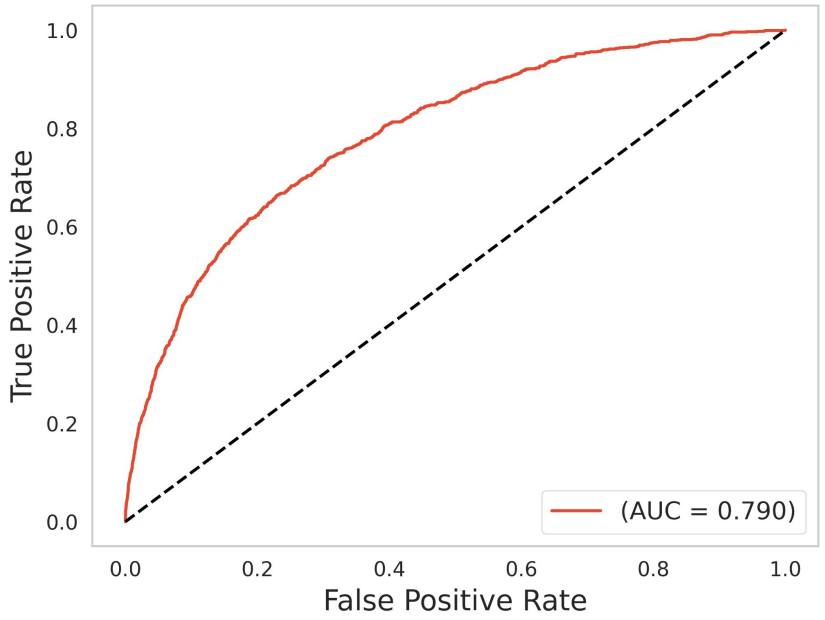


Figure 11. Result of real data (removed columns) (AUC = 0.79)

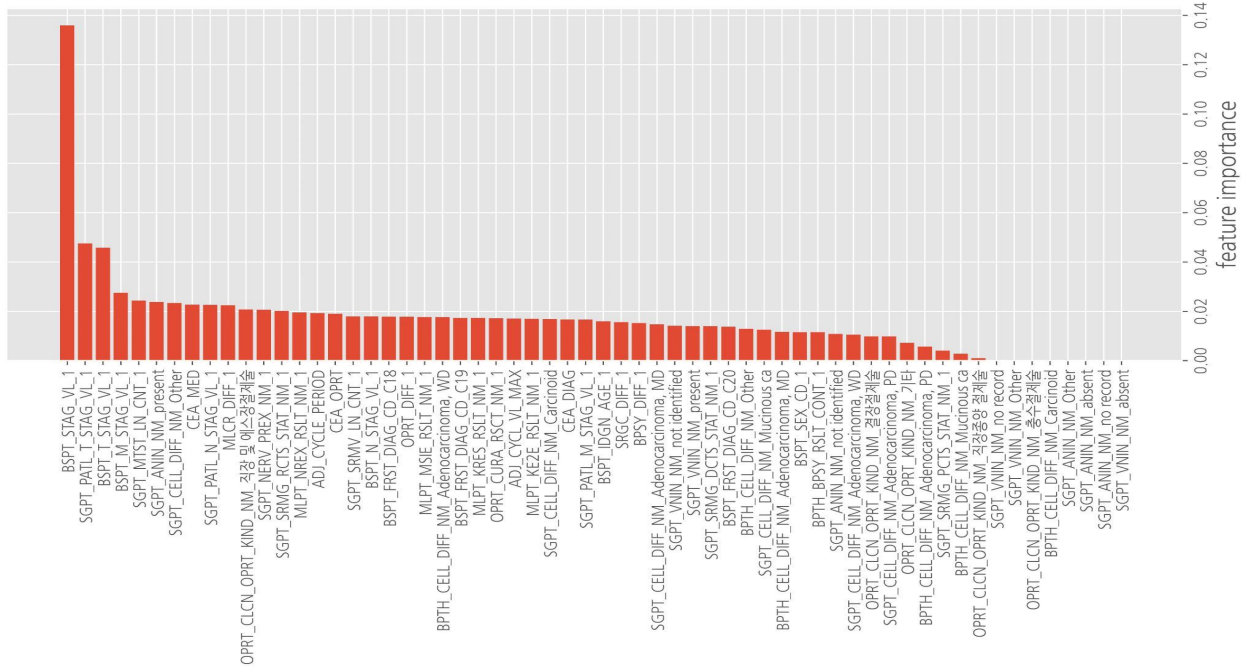
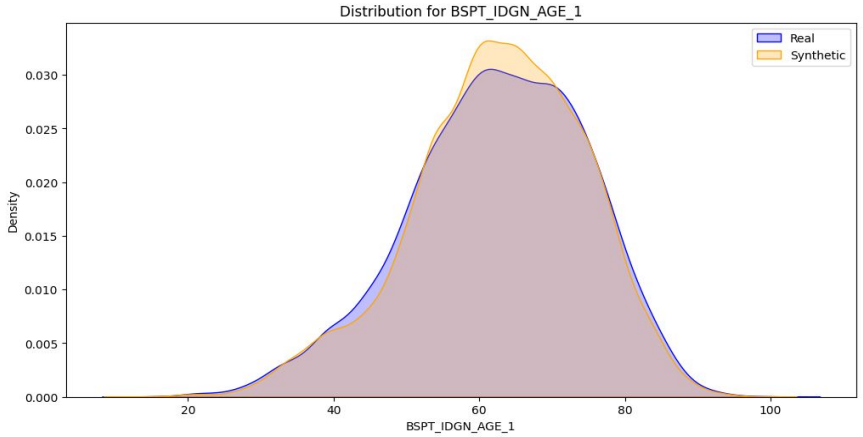


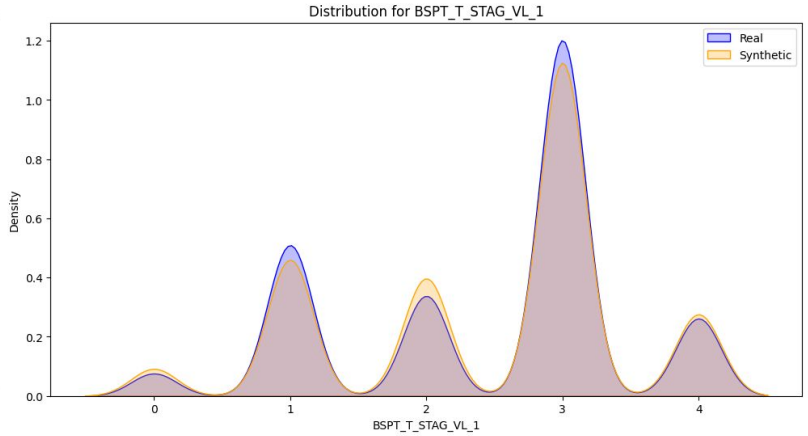
Figure 12. Feature importance of real data (removed columns)

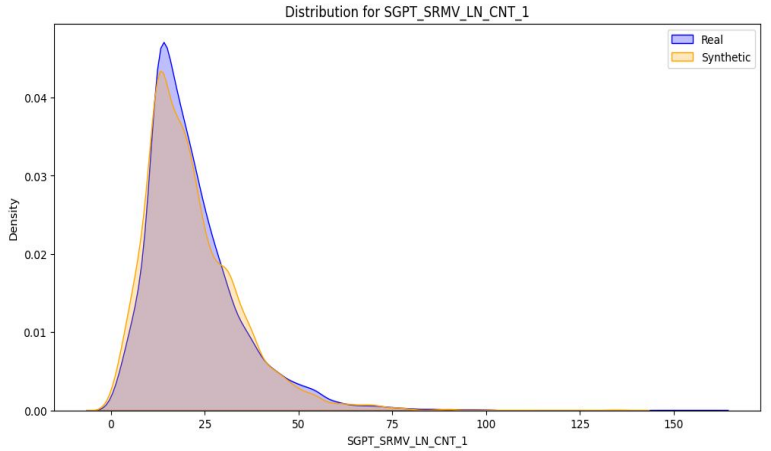
- Excludes variables that have information not available at the time of prediction
  - CEA_RCNF(CEA value at the time of recurrence), PAL_CYCLE_VL_MAX_LINE_1(Maximum number of cycles of palliative chemotherapy) PAL_CYCLE_PERIOD_LINE_1(Date of palliative chemotherapy) CEA_MAX(Maximum value of CEA value from diagnosis to recurrence), CEA_MIN(Maximum value of CEA value from diagnosis to recurrence), DEAD'(Death status), DEAD_DIFF_1(Duration from diagnosis to death)

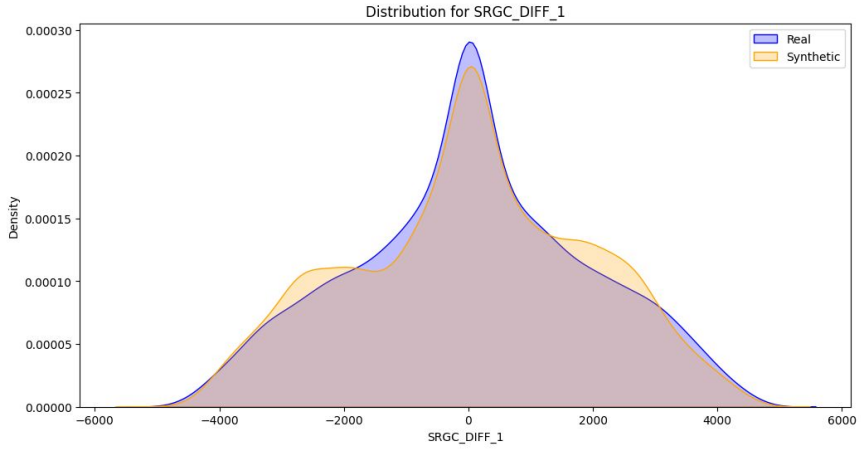
Compare synthetic and real data

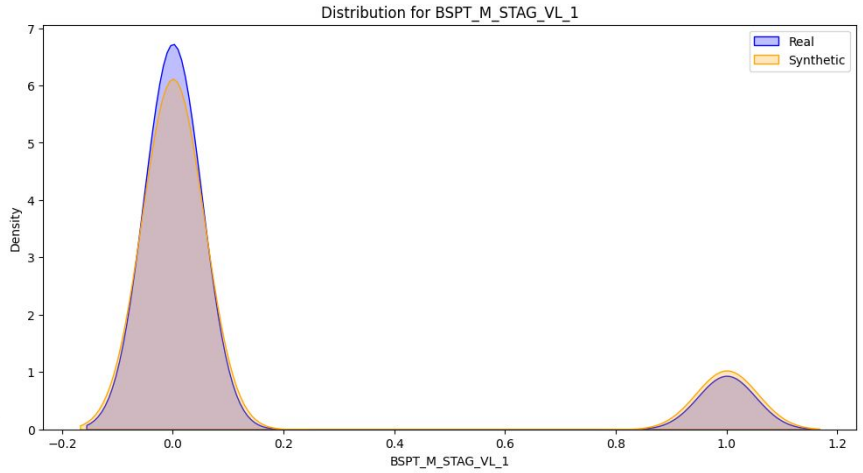
Figure 13. Distribution of Real data and REaLTabFormer data

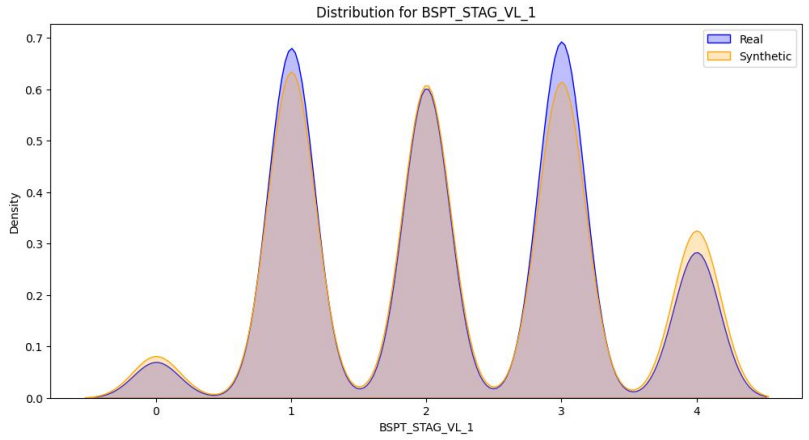






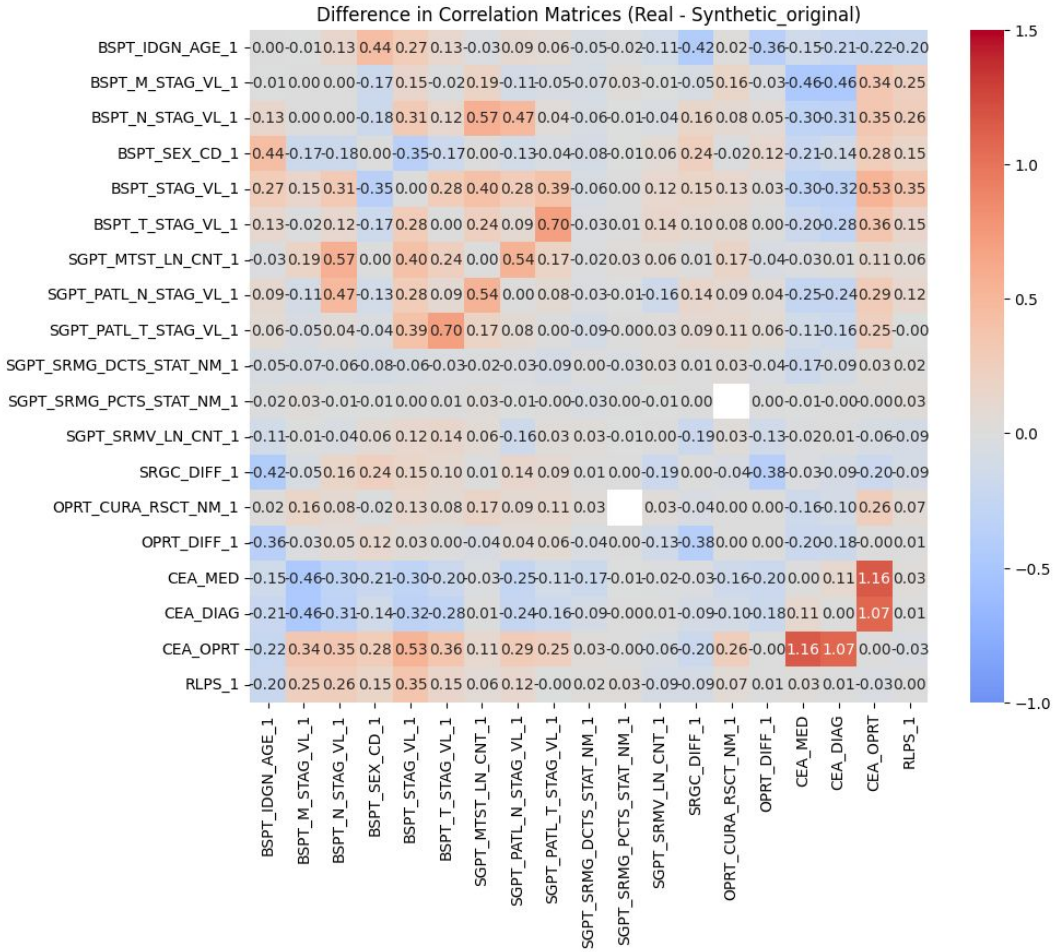






Correlations between synthetic and real data

A. Correlation of Given synthetic data and real data



B. Correlation of REaLTabForme-synthetic data and real data

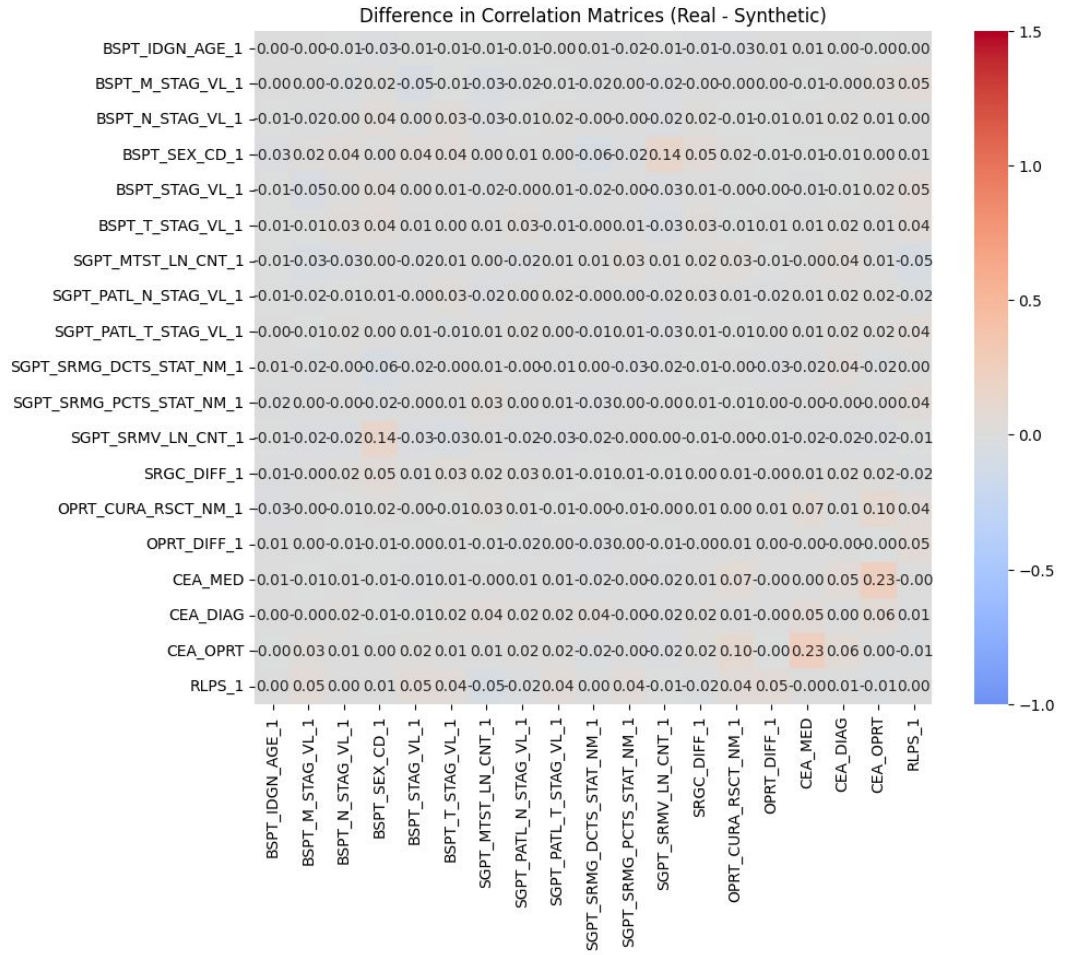


Figure 14. Heatmap of correlation between synthetic and real data

Prediction model (w/o Ensemble)

Table2. Result of colorectal cancer recurrence prediction model without ensemble

Model		Accuracy	AUROC	Confusion matrix
XGBoost	**Real data**	0.8805	0.7901	[[8813 277] [958 290]]
	Given synthetic data	0.8467	0.5368	[[8636 454] [1131 117]]
	REaLTabFormer synthetic data	0.8703	0.7619	[[8667 423] [918 330]]
	Real data + given synthetic data	0.8844	0.7913	[[8847 243] [952 296]]
	Real data + REaLTabFormer synthetic data	0.8814(0.9250*)	0.7888(0.9034*)	[[8818 272] [954 294]]

* These values are from REaLTabFormer, synthesized using the entire set of real data without avoiding train-test overlap.

Prediction model (Ensemble)

Table 3. Result of colorectal cancer recurrence prediction model with ensemble

Model		Accuracy	AUROC	Confusion matrix
Ensemble model 1 (using XGBoost)	**Real data**	0.8793	0.7737	[[8806 284] [964 284]]
	REaLTabFormer synthetic data	0.8703	0.7619	[[8667 423] [918 330]]
	Ensemble	**0.8801**	**0.7844**	[[8813 277] [969 279]]
Ensemble model 2 (using LGBM)	**Real data**	0.8825	0.7987	[[8872 218] [997 251]]
	REaLTabFormer synthetic data	0.8830	0.7880	[[8924 166] [1044 204]]
	Ensemble	**0.8831**	**0.8047**	[[8931 159] [1022 226]]
Ensemble model 3 (using XGBoost +LGBM)	**Real data**	0.8814	0.7989	[[8883 207] [995 253]]
	REaLTabFormer synthetic data	0.8691	0.7862	[[8914 176] [1040 208]]
	Ensemble	**0.8863**	**0.8046**	[[8926 164] [1011 237]]

Prediction model (when colorectal cancer recurs)

Table 4. Comparison of Time-to-Recurrence Predictions in Colorectal Cancer (Real data Vs Synthetic Data) (5-fold)

Data	Metrics	Time series (day)								
		180	360	540	720	900	1080	1260	1440	1620
Real data	Accuracy	0.9875	0.9583	0.9264	**0.9109**	**0.8994**	**0.8937**	**0.8915**	**0.8880**	**0.8875**
	AUROC	**0.7142**	**0.7520**	**0.7830**	**0.7994**	**0.7888**	**0.7948**	**0.7911**	**0.7899**	**0.8002**
	Confusion matrix	[10207, 4] [125, 2]	[9897, 34] [397, 10]	[9497, 108] [653, 80]	[9279, 148] [773, 138]	[9103, 201] [839, 195]	[9008, 224] [875, 231]	[8978, 212] [910, 238]	[8909, 249] [909, 271]	[8891, 248] [915, 284]
Given synthetic data	Accuracy	**0.9877**	**0.9606**	**0.9270**	0.8825	0.8776	0.8617	0.8519	0.8504	0.8481
	AUROC	0.5000	0.5000	0.5359	0.4909	0.5471	0.5481	0.5355	0.5222	0.5359
	Confusion matrix	[10211, 0] [127, 0]	[9931, 0] [407, 0]	[9579, 26] [729, 4]	[9075, 352] [863, 48]	[9006, 298] [967, 67]	[8805, 427] [1003, 103]	[8695, 495] [1036, 112]	[8693, 465] [1082, 98]	[8669, 470] [1100, 99]

External validation using genomic data

Table 5. Result of sepsis severity prediction model with ensemble

Model	Data	Accuracy	AUROC	Confusion matrix
Ensemble model 1 (using XGBoost)	**Real data**	0.813	0.815	[[39 77] [21 388]
	REaLTabFormer synthetic data	0.803	0.742	[[34 82] 21 388]]
	Ensemble	**0.794**	**0.782**	[[30 86] 22 387]]
Ensemble model 2 (using LGBM)	**Real data**	0.8457	0.837	[[48 68] [13 396]]
	REaLTabFormer synthetic data	0.7924	0.784	[[26 90] 19 390]]
	Ensemble	**0.813**	**0.817**	[[33 83] [15 394]]
Ensemble model 3 (using XGBoost +LGBM)	**Real data**	0.843	0.837	[[55 61] [21 388]]
	REaLTabFormer synthetic data	0.800	0.786	[[39 77[[28 381]]
	Ensemble	**0.8267**	**0.839**	[[40 76] 15 394]]

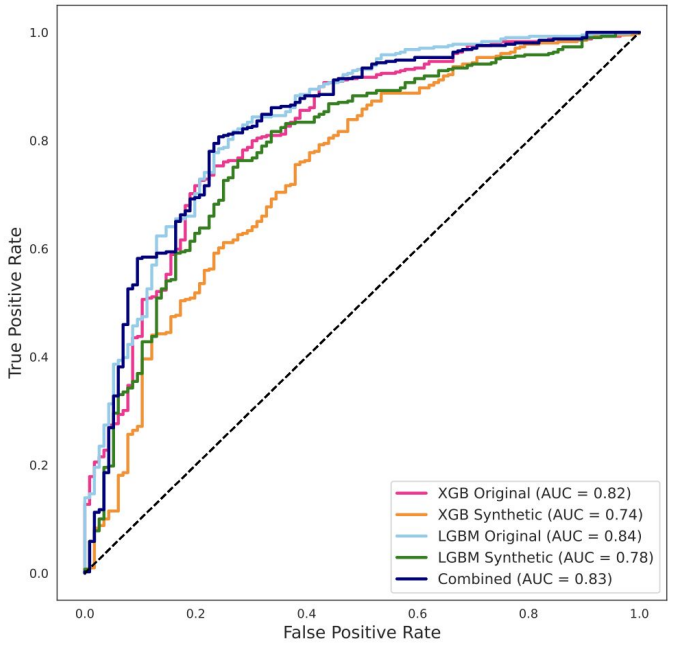


Figure 15. ROC curve of sepsis severity prediction model with ensemble

Discussion

- Variable selections is vital for both model performance and medical validity. So, we avoided 1) train, test set contamination, and 2) data target leakage focused on meaningful variables.
- We developed a prediction model for colorectal cancer recurrence using synthetic data. Enhanced performance was observed with XGBoost and LGBM ensemble model. Synthetic data proved crucial for model performance improved.
- We completed the external validation by applying the data-model based ensemble model to a genomic dataset, which shows that our model is applicable to other datasets as well.

YONSEI UNIVERSITY
COLLEGE OF MEDICINE

DHLab
DIGITAL HEALTHCARE LABORATORY