A Novel Ensemble Framework for Comprehensive Early-Stage Colorectal Cancer Diagnosis, Prognosis, and Treatment: Integration of Gastroenterology-Specific Transformer Language Models and Multiple Decision Trees

Cem Simsek [1,2], Suayib Yalcin [3],Mete Ucdal[4], Derya Karakoc [1,5]

1. Medical and Surgical Research, Institute of Health Sciences, Hacettepe University, Ankara, Turkey
2. Division of Gastroenterology, Faculty of Medicine, Hacettepe University, Ankara, Turkey
3. Division of Medical Oncology, Faculty of Medicine, Hacettepe University, Ankara, Turkey
4. Division of Internal Medicine Hacettepe University, Ankara, Turkey
5. Department of General Surgery, Faculty of Medicine, Hacettepe University, Ankara, Turkey

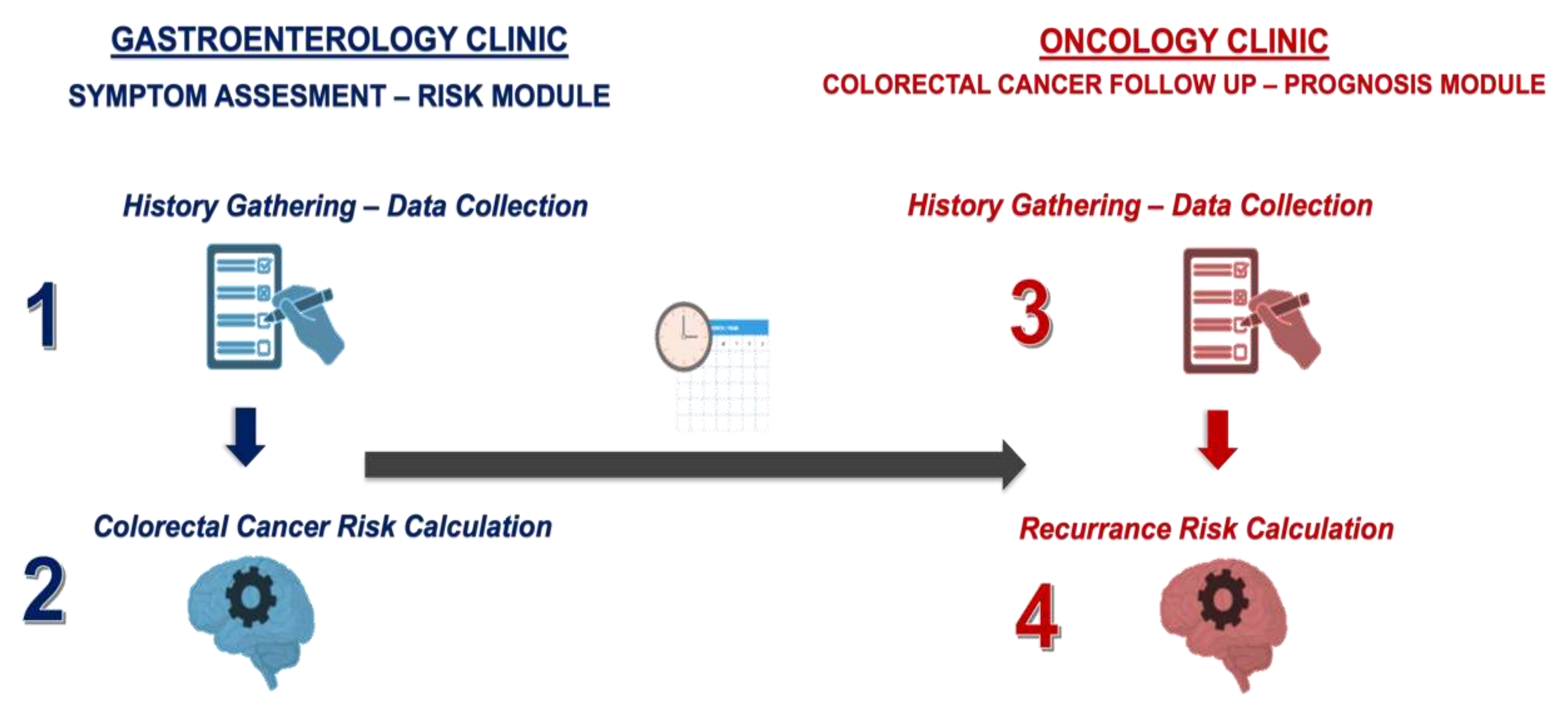
Background: Colorectal cancer (CRC) remains a significant global health burden, with early detection and intervention crucial for improving patient outcomes. This study aims to develop and evaluate a novel proof-of-concept ensemble framework combining transformer-based language models and decision tree-based models for early-stage CRC screening, diagnosis, and prognosis.

Methods: The ensemble framework consists of four key components: (1) GastroGPT, a transformer-based language model for extracting relevant data points from patient histories; (2) A decision tree-based model for assessing CRC risk and recommending colonoscopy; (3) GastroGPT for extracting data points from early CRC patients' histories; and (4) A suite of decision tree-based models for predicting survival outcomes in early-stage CRC patients. The study employed a retrospective, observational, methodological design using simulated patient cases.(Figure 1)

Results: GastroGPT demonstrated high accuracy in extracting relevant data points from patient histories. The decision tree-based model for CRC risk assessment achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.85 (95% CI: 0.78-0.92) in predicting the need for colonoscopy(Graph 1) . The decision tree-based models for survival prediction showed strong performance, with C-indices ranging from 0.71 to 0.75 for overall survival and disease-free survival at 24, 36, and 48 months(Graph 2).

Conclusion: The novel ensemble framework demonstrates promising performance in early-stage CRC screening, diagnosis, and prognosis. Further research is needed to validate the models using larger, real-world datasets and to assess their clinical utility in prospective studies.

ENSEMBLE COLORECTAL CANCER DIAGNOSIS, TREATMENT AND FOLLOW UP MODULE



GASTROENTEROLOGY CLINIC
SYMPTOM ASSESMENT – RISK MODULE

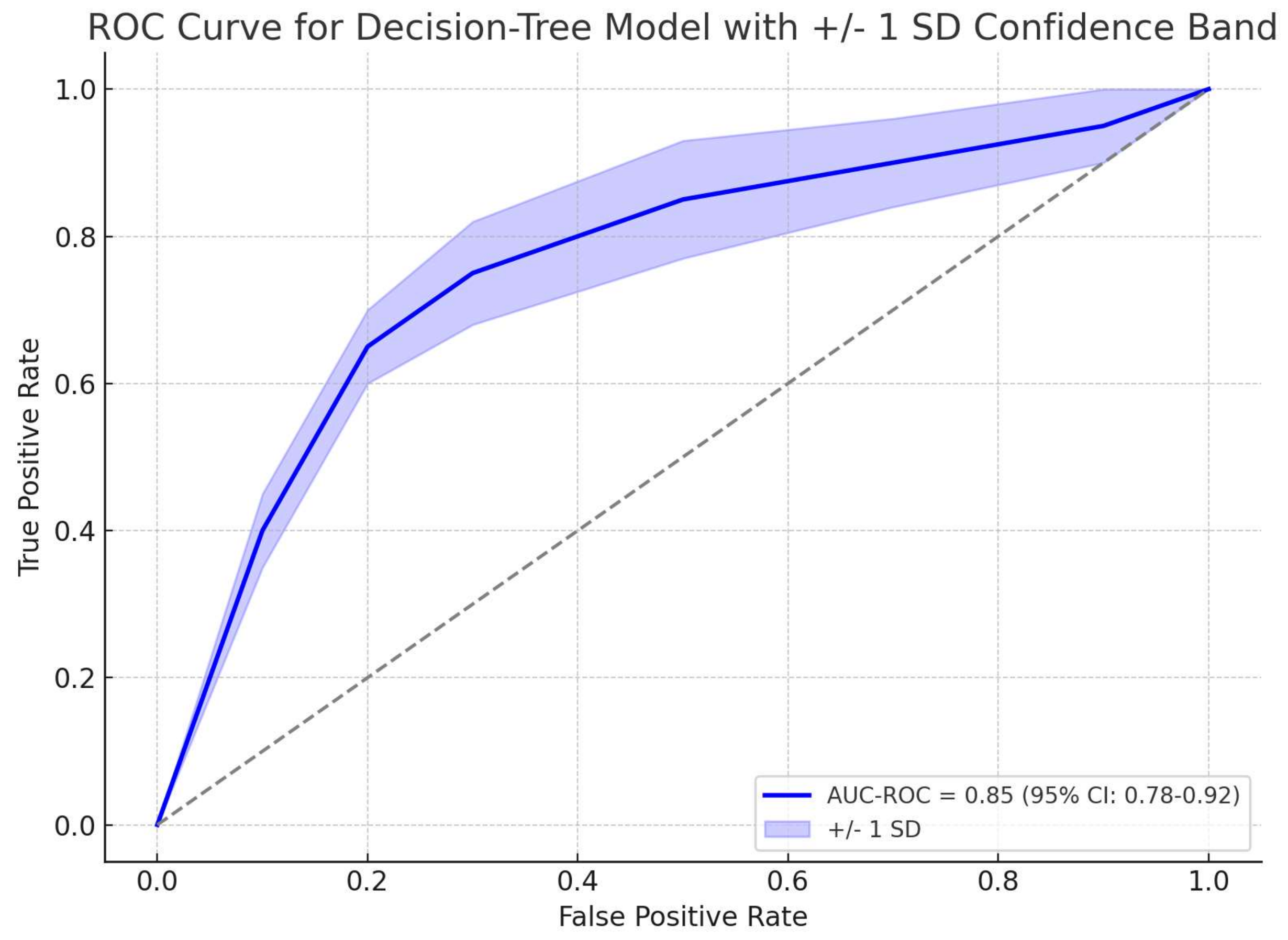
History Gathering – Data Collection
1
Colorectal Cancer Risk Calculation
2

ONCOLOGY CLINIC
COLORECTAL CANCER FOLLOW UP – PROGNOSIS MODULE

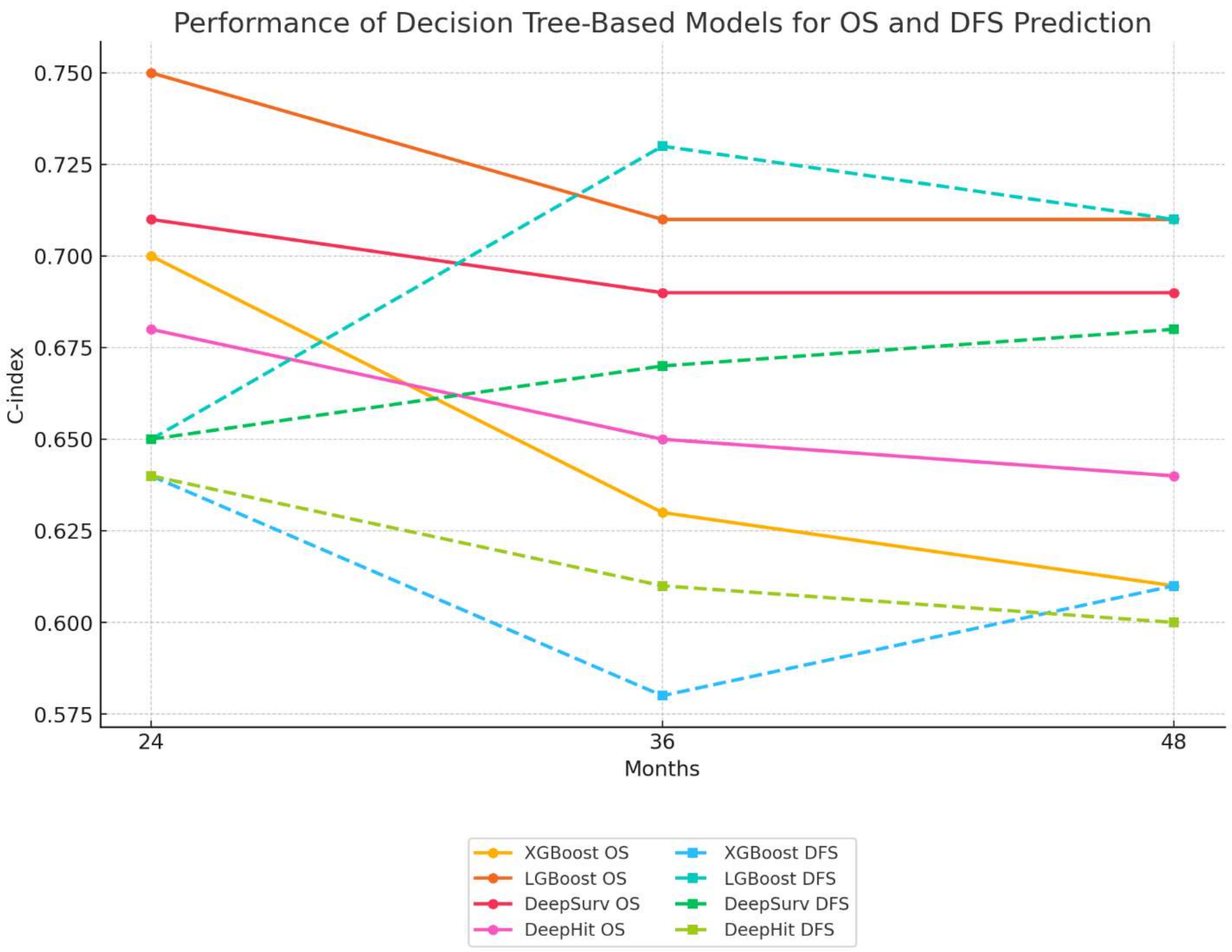
History Gathering – Data Collection
3
Recurrance Risk Calculation
4

Ensemble Colorectal Cancer Diagnosis, Treatment and Follow-up Module

- Gastroenterology Clinic - Symptom Assessment and Risk Module a. History Gathering – Data Collection (Step 1) b. Colorectal Cancer Risk Calculation (Step 2)
- Oncology Clinic - Colorectal Cancer Follow-up and Prognosis Module a. History Gathering – Data Collection (Step 3) b. Recurrence Risk Calculation (Step 4)



Graph 1: ROC curve showing the performance of the decision-tree model in predicting the need for colonoscopy, with an AUC-ROC of 0.85 (95% CI: 0.78-0.92).]



Graph 2: Line graph showing the performance (C-index) of different decision tree-based models (XGBoost, LGBoost, DeepSurv, DeepHit) for overall survival (OS) and disease-free survival (DFS) prediction at 24, 36, and 48 months. (p-values: XGBoost = 0.04, LGBoost = 0.03, DeepSurv = 0.05, DeepHit = 0.07).