# Trustworthy Transformer Models for Thyroid Cancer Recurrence Prediction through Explanation Stability

**Flavia COSTI[a,*], Ana CORLAN[b], Emanuel COVACI[a], and Darian ONCHIS[a]**

[a] Computer Science Department, Faculty of Computer Science, West University of Timişoara, Bd. Vasile Pârvan no. 4, 300223 Timişoara, Romania.
[b] Endocrinology Department, University of Medicine and Pharmacy, Eftimie Murgu Square no. 2, 300041 Timişoara, Romania.
E-mails: (*) flavia.costi@e-uvt.ro; ana.corlan@umft.ro; emanuel.covaci@e-uvt.ro; darian.onchis@e-uvt.ro

* Author to whom correspondence should be addressed;

## Abstract

*Background*: Accurate prediction of postoperative recurrence in differentiated thyroid cancer (DTC) is important for risk stratification and follow-up planning. Beyond predictive performance, clinically useful models should also provide explanations that remain stable across training runs and plausible input variation. We therefore study DTC recurrence prediction with a Transformer-based tabular model and explicitly assess explanation stability as a criterion for trustworthy clinical AI. *Methods*: We used the public Differentiated Thyroid Cancer Recurrence dataset, including 383 patients described by 16 clinical and demographic variables, with binary recurrence status as the outcome variable (108 recurrence, 275 non-recurrence). Features were one-hot encoded and used to train a Transformer-based classifier, evaluated with stratified five-fold cross-validation. Explanations were generated with SHAP. To assess stability, the model was retrained with different random initializations and under small perturbations of numerical features simulating plausible clinical variability. Stability-aware feature selection retained variables with high explanatory importance and low cross-run variability, and the model was retrained on the reduced feature set. *Results*: The full model achieved a mean accuracy of 96.6%, with similarly strong macro-F1 performance across folds. SHAP analysis showed high consistency in feature rankings across repeated runs, with Spearman correlation coefficients approaching 1, indicating near-identical ordering of the most influential variables. Prediction outputs were minimally affected by small perturbations of numerical inputs, supporting robustness under realistic feature variation. The reduced model obtained through stability-aware feature selection preserved overall predictive performance and yielded slight improvements in some folds, while reducing input dimensionality and model complexity. *Conclusions*: Transformer-based tabular learning can support accurate prediction of postoperative recurrence in DTC while also providing highly reproducible post hoc explanations. Incorporating explanation stability into feature selection improves model parsimony without compromising predictive quality, thereby supporting the development of more transparent and trustworthy clinical decision support systems.

**Keywords:** Explainable Artificial Intelligence; Transformer Models; Thyroid Cancer Recurrence.