



GastroPredict: Gastric Cancer Survival Prediction and Explainable AI Support System

Vedant Pingat^a, Vinod Sule^a, Avishkar Mandhare^{a,*}

^aDept. of Computer Engineering, Smt. Indira Gandhi College of Engineering, Ghansoli, India

ABSTRACT:

Gastric (stomach) cancer is a leading cause of cancer death, and population registries such as the SEER Program hold decades of patient-level records that could support individual prognosis. Published studies show that machine learning (ML) and survival models can use registry variables competitively, but reported results depend on the cohort, the survival horizon and the evaluation protocol. Predictions are rarely accompanied by an individual survival curve, a per-patient explanation, or a question-answering component designed for safety. This review examines representative research on gastric-cancer prognosis models, SEER-based ML, survival-analysis methods and metrics, explainable AI, class imbalance and calibration, and clinical decision support. Based on these findings, it discusses the proposed GastroPredict system, built on 86,197 SEER gastric cancer records, which combines an XGBoost status classifier, a Cox proportional hazards survival curve, TreeSHAP explanations and a rule-based, non-generative assistant. The paper presents an initial targeted review rather than an exhaustive systematic review; its conclusions should be interpreted as a synthesis of selected literature, and the project results reported here have not been externally validated. They are modest: XGBoost's accuracy (91.75%) is only about six points above an always-"Dead" baseline (85.27%), and it identifies only 63.1% of surviving patients.

Keywords: Gastric cancer, SEER, survival prediction, XGBoost, Cox proportional hazards, SHAP, explainable AI, clinical decision support, rule-based assistant.

Introduction

Stomach cancer accounted for 4.9% of new cancer cases and 6.8% of cancer deaths worldwide in 2022, ranking fifth for both incidence and mortality [1]. Survival depends strongly on stage at diagnosis, and reasoning about an individual patient requires combining demographic, tumour and treatment variables; prognostic models for overall survival in primary gastric cancer have therefore been the subject of systematic review [2]. Registries are therefore an obvious basis for data-driven prognosis.

The Surveillance, Epidemiology, and End Results (SEER) Program of the U.S. National Cancer Institute is a widely used source. Studies built on it report competitive discrimination for gastric cancer survival [3], [4], and a recent systematic review of ML for gastric cancer survival finds promise alongside concerns about interpretability, reliance on retrospective data and generalisability [5]. A researcher who wants to explore an individual case still meets three practical problems: registry software reports descriptive statistics rather than individual predictions, high-scoring models rarely say why they predicted what they did, and charts, predictions and clinical look-up live in different programs.

This paper reviews selected literature on gastric-cancer prognosis models, SEER-based ML, survival-analysis methods, explainable AI (XAI), imbalance and calibration, and clinical decision support. It compares their methods and limitations, identifies research gaps, and discusses how the findings inform GastroPredict, a research prototype built by the authors. The aim is not to claim that GastroPredict has demonstrated clinical benefit, but to identify a literature-informed design and the evaluation it would require.

Review Methodology

This paper uses a targeted literature-review approach. Sources were identified through theme searches of publisher pages, PubMed Central and scholarly indexes, using concepts such as gastric cancer prognostic models, SEER machine learning, survival analysis, explainable AI in healthcare, class imbalance, calibration, dataset shift, reporting guidelines, clinical decision support and hallucination in medical language models. Each cited paper was confirmed on a publisher, repository or indexing record, and no source is cited from memory alone. The search was exploratory rather than exhaustive, and the exact search dates and result counts were not systematically recorded.

Studies were considered relevant when they addressed at least one of five areas: (1) gastric-cancer prognosis and SEER-based ML, (2) survival-analysis methods and metrics, (3) explainable AI, (4) imbalance, calibration and validation practice, or (5) decision support and safe question answering. Priority was given to journal articles and review papers from 2018-2026 with identifiable publication details. Older papers were kept when they are foundational (the Cox model, concordance index, random survival forests). Two conference papers that define widely used tools, XGBoost and SHAP, are cited as method sources only.

Studies were compared by objective, method, data, reported findings and limitations. Summaries rest on abstracts and visible record text rather than full papers, so numerical results are quoted only as reported by the source and are not treated as comparable across different cohorts, horizons and tasks.

Where a record did not show author names, this is stated in the reference list. This is a targeted narrative review; its selection process does not support a claim of complete coverage.

Gastric Cancer Prognosis and SEER-Based Machine Learning

Reviews of prognostic modelling are consistently cautious. A systematic review of overall-survival models for primary gastric cancer, following the CHARMS checklist, examined how such models are developed and reported [2]. A 2025 systematic review of ML for gastric cancer survival reports encouraging performance but flags data quality, interpretability and generalisability as open problems [5]. Reviews of AI in gastric cancer cover mainly image-based diagnosis and prognosis [6], and a meta-analysis of radiomics-based ML addresses chemotherapy response and survival [7]. These reviews agree on direction but not on a ready tool: they call for better validation, transparency and interpretability.

SEER-based studies show that ML is competitive on registry variables. An improved weighted random forest achieved a minor performance gain over standard configurations on 110,697 gastric cases [3]. Random survival forests and DeepSurv were compared with Cox regression for 14,177 patients with gastric adenocarcinoma, and DeepSurv reached AUC values of 0.825-0.871 for 1-, 3-, 5- and 10-year survival [4]. A 2026 study of early gastric cancer built interval classifiers and used SHAP to rank predictors; its best AUC fell from 0.8142 for 1-3 years to 0.6362 for 3-5 years and 0.692 for 5-10 years [8]. A nomogram study added an external validation cohort [9].

Three lessons follow. First, performance depends on the prediction horizon: the same pipeline can discriminate well over short windows and much less well over long ones [8]. Second, most studies restrict the cohort to a subtype, stage group or surgical population and use detailed clinical variables, which differ from the broad eight-variable view available in an all-stomach registry export. Third, explanation is rare: among these abstracts only [8] reports SHAP, and a nomogram offers an additive but model-specific explanation [9].

Table 1 - Comparison of SEER-based gastric cancer prediction research

Study / source	Task and method	Evidence reported	Limitation
Weighted random forest [3]	Weighted random forest on SEER gastric cases	110,697 cases; +0.79% over standard forest	Gain is small; no individual survival curve
DeepSurv comparison [4]	RSF and DeepSurv vs Cox, gastric adenocarcinoma	AUC 0.825-0.871 for 1/3/5/10 years	Single subtype; explanation not reported
Early-gastric-cancer study [8]	Interval classifiers with SHAP ranking	AUC 0.8142, 0.6362, 0.692 by interval	Performance drops at longer horizons
Nomogram study [9]	Cox-based nomogram by grade; external cohort	Externally validated survival nomogram	Model-specific; not an interactive tool
GastroPredict (proposed)	XGBoost status + CoxPH curve + TreeSHAP	Own test set; see Section 10	Not externally validated; calibration unassessed

Survival Analysis Methods and Evaluation

The Cox proportional hazards model remains the reference method: it models the hazard as a baseline function multiplied by an exponential of covariates and yields a full survival function for an individual patient [10]. ML surveys describe how censoring, the defining difficulty of time-to-event data, is handled by trees, ensembles and neural networks [11]. Random survival forests extend random forests to right-censored data [12], and DeepSurv is a Cox-based neural network that removes the need to hand-craft interactions [13].

Evaluation differs from ordinary classification. The concordance index, which originates in the evaluation of medical tests, asks whether the patient predicted to be at higher risk fails first, with 0.5 indicating chance [14]. A model with a moderate concordance value and an unchecked proportional-hazards assumption should be described as a baseline, not a final answer. Survival models should also be compared on the same data split, against simple baselines, and at stated horizons.

Table 2 - Survival-analysis methods and evaluation sources

Study / source	Method	Role	Implication for GastroPredict
Cox model [10]	Semi-parametric hazard model	Reference survival model	Basis of the survival-curve component
ML survival survey [11]	Survey of ML survival methods	Method taxonomy	Map of alternatives to CoxPH
Random survival forest [12]	Random survival forest	Non-parametric ensemble	Candidate comparison model
DeepSurv [13]	DeepSurv neural network	Deep Cox model	Future model option
Concordance index [14]	Concordance-type evaluation	Ranking metric	Context for the reported C-index

Explainable AI for Clinical Prediction

TreeSHAP computes exact Shapley-value explanations for tree ensembles in polynomial time and combines many local explanations into global importance [15]. It builds on the SHAP framework [16] and pairs naturally with XGBoost [17] for tabular clinical data. Explanation is valuable because clinicians need to see which inputs drove a prediction, but it has limits. It has been argued that explaining a black box is not the same as using an interpretable model, and that post-hoc explanations can mislead in high-stakes decisions [18]. SHAP values describe the model's behaviour, not causal effects in the patient, and should be presented that way.

Class Imbalance, Calibration and Validation

Medical data sets are often imbalanced, which biases standard learners toward the majority class [19]. Accuracy alone can then look strong while the minority class is poorly identified, so recall and precision per class should be reported. Calibration, whether predicted risks match observed frequencies, is under-reported yet can make a model misleading or harmful when used for decisions [20]. Temporal dataset shift degrades models over time, mostly through calibration (11 of 15 studies in one review) rather than discrimination (3 of 15) [21], which matters for registries spanning decades. The TRIPOD+AI statement gives a 27-item checklist for reporting prediction-model studies that use regression or ML [22]. Together these sources define what a trustworthy evaluation of a registry model should contain.

Decision Support and Safe Question Answering

An overview of clinical decision support systems documents benefits but also common pitfalls and potential harms, and recommends risk-aware design and maintenance [23]. For question answering, generative language models produce fluent text but can be wrong: the EVAL framework for gastroenterology was proposed because inaccurate responses pose risks in medical decisions and need expert verification [24]. A deterministic keyword-and-knowledge-base assistant cannot invent content, but its coverage is limited by what the knowledge base contains, so it needs its own evaluation and clinician-reviewed content. The reviewed work evaluates generative models; it does not report the coverage or acceptability of a rule-based assistant for gastric cancer questions.

Comparative Analysis

The reviewed work addresses complementary parts of a potential decision-support tool. SEER-based ML studies establish that registry variables carry prognostic signal; survival methods supply individual curves; XAI methods explain tree models; validation literature defines honest evaluation; and safety work warns against unverified generated answers. Table 3 summarises what each stream contributes and where it stops. The entries describe what the reviewed abstracts report; "not reported" means the record does not mention a feature, not that it is absent.

Table 3 - Comparative summary of research areas relevant to GastroPredict

Research area	Common method	Potential strength	Key limitation
SEER-based gastric ML	RF, DeepSurv, nomograms, interval classifiers	Competitive discrimination on registry variables	Subtype-specific cohorts; explanation and interface rarely reported
Survival analysis	Cox, RSF, DeepSurv	Individual survival curves with censoring	Needs assumption checks and more than one metric
Explainable AI	TreeSHAP, nomogram points	Per-patient feature attribution	Describes the model, not causes; can mislead
Validation practice	Calibration, TRIPOD+AI, temporal splits	Credible, reproducible evaluation	Under-reported in registry gastric studies
Safe clinical Q&A;	Verified LLM frameworks; rule-based look-up	Controlled, reviewed answers	Rule-based coverage not yet evaluated

Research Gaps and Future Directions

First, integration remains an engineering and evaluation challenge. Among the reviewed papers, none combines registry-scale individual prediction, an individual survival curve, per-patient explanation and a non-generative question-answering component in one interface. This is a finding from a limited search, not proof that no such system exists.

Second, explanations and survival output may disagree. SHAP explains the classifier's death probability, while the survival curve comes from a different model; none of the reviewed work examines whether the two stay coherent when shown side by side.

Third, target definition and temporal drift need attention. Recent studies use fixed horizons of 1, 3, 5 or 10 years [4], [8], whereas a vital-status target defined at the study cut-off mixes follow-up length with year of diagnosis, and registry shift mostly affects calibration [21].

Fourth, imbalance and calibration are under-reported [19], [20], and the proportional-hazards assumption and concordance pitfalls need explicit checks.

Finally, a rule-based assistant should be tested with representative users and clinicians for coverage, wrong-rule matching and acceptability.

Table 4 - Research gaps and evaluation priorities for GastroPredict

Research gap	Why it matters	Recommended evaluation
Combined prediction, curve, explanation and Q&A;	Capabilities are covered by different studies, not together	Scenario-based tests of the complete application
Target definition and drift	Cut-off status mixes follow-up length with diagnosis year	Fixed-horizon targets; train on earlier years, test on later
Imbalance and calibration	Accuracy can hide poor minority-class recall; risks may be miscalibrated	Per-class recall and precision; calibration curves
Cox assumptions and metrics	Non-proportional hazards and single C-index can mislead	Assumption tests; stratified models; compare RSF and DeepSurv
Safe Q&A; coverage	Unanswered or wrong-rule queries could mislead users	Query-log coverage test; clinician review of rules

1. Proposed System: GastroPredict

GastroPredict is a research-oriented decision-support prototype built on 86,197 gastric cancer records from SEER Research Data (8 Registries, November 2025 submission), diagnosed 1975-2023. Eight inputs are used: age group, sex, race, year of diagnosis, histology, summary stage, surgery and radiation. Three classifiers (logistic regression, random forest and XGBoost) were trained on an 80/20 stratified split (68,957 training and 17,240 test records) to predict vital status. A Cox model fitted on the 84,854 records with known survival time supplies individual survival curves and median survival, and TreeSHAP ranks the factors behind each XGBoost prediction. A separate deterministic rule-based engine answers clinical questions from a predefined knowledge base. The system estimates probabilities from historical groups and does not predict a time left to live.

On the held-out test set XGBoost reached accuracy 0.9175, recall 0.9670 for the Dead class, F1 0.9523 and ROC-AUC 0.9385, and the Cox model reached a concordance index of 0.7006 on unseen data. These numbers need careful reading (Table 5). Because 85.3% of patients are deceased, a rule that always answers Dead already scores 85.27% accuracy, so XGBoost's gain is about six points, and its recall for the Alive class is only 0.63. Overall, the model remains weak at identifying survivors: XGBoost finds only about 63% of Alive patients, so its headline accuracy should not be read as strong prognostic performance. A second caveat is a data-tracking limitation, not a model-performance result. Vital status is recorded at a fixed study cut-off rather than over a standardised horizon such as 3 or 5 years, and year of diagnosis is an input, so patients diagnosed long ago have simply been followed for far longer. A probability such as 99.76% for a patient diagnosed in 1981 therefore mainly reflects more than 40 years of follow-up, not prognostic skill. The proportional-hazards check flagged several Cox covariates, and calibration and external validation have not yet been assessed.

Table 5 - Project test-set results (n = 17,240), read against the trivial baseline

Model	Accuracy	Recall, Alive class	Reading
Always answer "Dead"	0.8527	0.00	Trivial baseline
Logistic regression	0.8498	0.868	Class-weighted; below baseline accuracy
Random forest	0.8856	0.659	Moderate gain
XGBoost	0.9175	0.631	Best accuracy; minority class still weak

Limitation of the current design. GastroPredict itself has the disjointed-evaluation gap described in Section 9. The XGBoost classifier, whose output SHAP explains, and the Cox model, which supplies the survival curve and median, are fitted independently, and no analysis yet checks that they agree for the same patient. For example, a high SHAP-explained probability of death could appear beside a long median survival. Users could therefore see apparently conflicting outputs, and a coherence check between the two models is the first

evaluation priority.

The survey changes how several design claims should be stated. That XGBoost is the best classifier is supported on the team's test set but with the imbalance caveat above [19]. That SHAP explains each prediction is partly supported: TreeSHAP gives exact attributions for tree models [15], but they describe the model rather than causes [18]. That the rule engine avoids hallucination is supported by design, in contrast to the risks reported for generative models [24], but its coverage is untested. Probability calibration [20] and usability remain unverified.

Evaluation should follow the gaps: fixed-horizon targets; a temporal split with calibration curves; minority-class recall and a threshold study; proportional-hazards tests with comparison against random survival forests and DeepSurv; a check that SHAP factors and the Cox curve tell a consistent story; a coverage test and clinician review of the rule engine; and reporting in line with TRIPOD+AI [22].

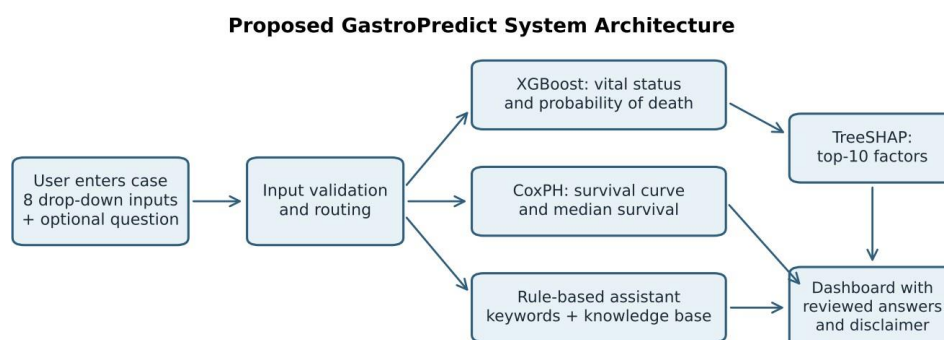
Conclusion

Machine learning and survival analysis can use SEER registry variables for gastric cancer prognosis, and exact explanations for tree models are available. The reviewed literature supports each capability separately, but reported results depend on cohort, horizon and evaluation, and none of the reviewed studies, within the limits of this targeted search, combines individual prediction, a survival curve, per-patient explanation and a safe non-generative assistant in one interface.

GastroPredict is proposed as an integrated research prototype that brings these capabilities together. This paper does not report clinical effectiveness. On the team's own test set the best model is only about six accuracy points above a trivial baseline and still misses over a third of surviving patients, so it must be regarded as weak for the minority class. Agreement between the classifier and the survival curve, calibration, fixed-horizon targets and temporal robustness must be established before any claim of usefulness, and external validation and a usability study with clinicians are needed. GastroPredict outputs are research estimates only and must not replace professional diagnosis.

Proposed GastroPredict Flowcharts

Figure 1. Proposed GastroPredict system architecture (conceptual; not clinically validated).



Proposed Survival-Prediction Workflow**Figure 2. Proposed survival-prediction workflow.****Proposed Rule-Based Assistant Workflow****Figure 3. Proposed rule-based assistant workflow.****Acknowledgements**

The authors thank their project guide, Prof. Vaishali Yeole, and the Department of Computer Engineering, Smt. Indira Gandhi College of Engineering, Ghansoli, for guidance and support.

REFERENCES

- [1] F. Bray, M. Laversanne, H. Sung *et al.*, “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229-263, 2024, doi: 10.3322/caac.21834.
- [2] Q. Feng, M. T. May, S. Ingle, M. Lu, Z. Yang and J. Tang, “Prognostic models for predicting overall survival in patients with primary gastric cancer: A systematic review,” *BioMed Research International*, vol. 2019, Art. no. 5634598, 2019, doi: 10.1155/2019/5634598.
- [3] C. Xu, J. Wang, T. Zheng, Y. Cao and F. Ye, “Prediction of prognosis and survival of patients with gastric cancer by weighted improved random forest model: An application of machine learning in medicine,” *Archives of Medical Science*, vol. 18, no. 5, pp. 1208-1220, 2021, doi: 10.5114/aoms/135594.
- [4] J. Zeng, K. Li, F. Cao and Y. Zheng, “Development and validation of survival prediction model for gastric adenocarcinoma patients using deep learning: A SEER-based study,” *Frontiers in Oncology*, vol. 13, Art. no. 1131859, 2023, doi: 10.3389/fonc.2023.1131859.
- [5] Authors not shown on source record, “Predicting gastric cancer survival using machine learning: A systematic review,” *World Journal of Gastrointestinal Oncology*, vol. 17, no. 5, 2025. Available: <https://www.wjgnet.com/1948-5204/full/v17/i5/103804.htm>
- [6] P.-H. Niu, L.-L. Zhao, H.-L. Wu, D.-B. Zhao and Y.-T. Chen, “Artificial intelligence in gastric cancer: Application and future perspectives,” *World Journal of Gastroenterology*, vol. 26, no. 36, pp. 5408-5419, 2020, doi: 10.3748/wjg.v26.i36.5408.
- [7] D. Adili, A. Mohetaer and W. Zhang, “Diagnostic accuracy of radiomics-based machine learning for neoadjuvant chemotherapy response and survival prediction in gastric cancer patients: A systematic review and meta-analysis,” *European Journal of Radiology*, vol. 173, Art. no. 111249, 2024, doi: 10.1016/j.ejrad.2023.111249.
- [8] Authors not shown on source record, “Building and validating a machine learning-based survival prediction model for early gastric cancer: A SEER database analysis,” *Surgical Oncology*, vol. 66, Art. no. 102416, 2026.
- [9] Authors not shown on source record, “Survival nomogram for different grades of gastric cancer patients based on SEER database and external validation cohort,” *Frontiers in Oncology*, 2022, doi: 10.3389/fonc.2022.951444.
- [10] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187-220, 1972.
- [11] P. Wang, Y. Li and C. K. Reddy, “Machine learning for survival analysis: A survey,” *ACM Computing Surveys*, vol. 51, no. 6, Art. no. 110, 2019.
- [12] H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841-860, 2008, doi: 10.1214/08-AOAS169.
- [13] J. L. Katzman *et al.*, “DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network,” *BMC Medical Research Methodology*, vol. 18, Art. no. 24, 2018.
- [14] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee and R. A. Rosati, “Evaluating the yield of medical tests,” *JAMA*, vol. 247, no. 18, pp. 2543-2546, 1982.
- [15] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56-67, 2020, doi: 10.1038/s42256-019-0138-9.
- [16] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765-4774. (Conference method source.)
- [17] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794. (Conference method source.)
- [18] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, 2019, doi: 10.1038/s42256-019-0048-x.

- [19] Authors not shown on source record, "Class imbalance in medical datasets [title abbreviated on source record]," *Artificial Intelligence Review*, 2024, doi: 10.1007/s10462-024-10884-2.
- [20] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.
- [21] L. L. Guo *et al.*, "Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine," *Applied Clinical Informatics*, vol. 12, no. 4, pp. 808-815, 2021, doi: 10.1055/s-0041-1735184.
- [22] G. S. Collins *et al.*, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.
- [23] R. T. Sutton *et al.*, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, Art. no. 17, 2020, doi: 10.1038/s41746-020-0221-y.
- [24] M. Giuffre *et al.*, "Expert-of-Experts Verification and Alignment (EVAL) framework for large language models safety in gastroenterology," *npj Digital Medicine*, vol. 8, Art. no. 242, 2025.

* E-mail addresses: 2023ce44f@sigce.edu.in (Vedant Pingat); 2023ce53f@sigce.edu.in (Vinod Sule); 2023ce32f@sigce.edu.in (Avishkar Mandhare).