SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Cancer Verse: An Explainable Digital-Twin Framework For Breast Cancer Survival Prediction Using Xgboost, SHAP, And Large Language Model

Meena Kumari Parigi[1], Ponnuru Sowjanya [2], Srinivasa Rao Dhanikonda[3], Ramakrishna Reddy K[4], Bharathi Kalva[5]

[1 3 4]Department of AI & DS, Faculty of Science and Technology (ICFAITech), ICFAI Foundation for Higher Education (Deemed to Be University), Hyderabad, Telangana, India.
[2]Dept. of Computer Science and Engineering, School of Technology, GITAM(Deemed to be University), Hyderabad, Telangana, India
[5]Department of Operations and Information Technology,ICFAI Business School (IBS),The ICFAI Foundation for Higher Education (IFHE), Hyderabad, Telangana, India

[1]parigimeenakumari@gmail.com, [2]sowjanya.ponnuru@gmail.com, [3]srinivasarao.dhanikonda@gmail.com, [4]krkreddy20@gmail.com, [5]kalvabharathi50@gmail.com

Abstract
Breast Cancer is still one of the main reasons for cancer mortality amongst women, and accurate and interpretable survival prediction can help with clinical decision-making. This paper introduces CancerVerse, an interpretable digital twin framework that combines a gradient-boosted classifier ensemble with post-hoc explanation generation and a generative language model that produces patient-specific prognostic narratives. A XGBoost classifier was trained on SEER Breast Cancer dataset (4,023 patients after pre-processing, 15 patient attributes with values converted into 26 predictor features) to predict binary survival status (Alive/Dead). The classifier demonstrated 90.68% accuracy, 80.77% precision, 51.22% recall, 62.69% F1 score and 0.8461 ROC-AUC on 20% (n=805) held-out test set. SHapley Additive exPlanations (SHAP) technique was used to generate both global feature attributions and local explanations per-patient and combine them in a structured 'digitaltwin' state, representing protective and risk factors of the prediction. Finally, a large language model (Gemini 2.5 Flash) was prompted with the digitaltwin state to generate a concise, clinician-readable summary without revealing the algorithmic details to the end user. This work demonstrates the potential of CancerVerse to produce trustworthy, interpretable, and communicable AI-assisted prognoses in oncology through a fully automated pipeline accessible via an interactive CancerVerse dashboard. The performance of the framework was discussed, along with its limitations regarding class imbalance and potential directions for clinical validation.

Keywords— Breast Cancer Survival Prediction, Explainable Artificial Intelligence, XGBoost, SHAP, Digital Twin, Large Language Models, Gemini AI, SEER Dataset.

I. Introduction

Breast cancer is one of the most frequently occurring malignant tumors and the leading cause of cancer deaths among women worldwide, accounting for a large number of new cases of cancer and cancer deaths annually [1]. The chances of survival depend on the interplay among demographic, histologic, and tumor-stage factors, making it hard to provide a personalized prognosis using population statistics alone. Various ML models, especially tree ensemble models, have been shown to perform well in predicting structured oncologic data[2]; however, their use in medical settings is limited by the "black-box" problem (i.e., lack of explainability)[3] . Clinicians and patients alike require not only an accurate prediction but also a transparent, interpretable justification for it.
XAI methods, most prominently SHapley Additive exPlanations (SHAP)[4][5], and Local Interpretable Model-agnostic Explanations (LIME)[6], attempt to fill this gap by assigning the contributions of individual input features to the model's output in a principled way. The growing adoption of such techniques for improving trust and transparency in clinical decision-making has been widely documented [7][8]. Nevertheless, SHAP visualizations, which are usually displayed to end users as bar graphs, waterfall plots, or beeswarm plots, are hard for non-expert

users to understand. Recent developments in large language models (LLMs) have created an opportunity to translate quantitatively sound interpretations into qualitative stories suitable for clinical communication.
In this paper, CancerVerse is presented which is a full-stack approach that (i) trains an XGBoost model using the SEER breast cancer data set and predicts survival status for patients; (ii) computes global and local SHAP explanations; (iii) structures these explanations in a "digital-twin" manner, representing each patient's clinical risk profile; and (iv) utilizes the generative language model (LLM) Gemini 2.5 Flash translates this structured representation of the patient's clinical state into a concise, clinically oriented narrative, presented through an interactive dashboard. The overall objective is to illustrate a reproducible framework for communicating interpretability in AI-assisted survival prognosis, where a human-understandable explanation, rather than a standalone probability score, is at the center of the output.

The key contributions of the current paper can be summarized as follows:

- Fully specified, replicable pipeline for the preprocessing and survival classification tasks on the SEER breast cancer dataset, resulting in 90.68% accuracy and 0.846 ROC-AUC by using the default hyperparameters of the XGBoost algorithm.
- Two-level SHAP explainability layer that creates the global feature importance within the whole cohort and exact patient-specific attributions with the help of TreeExplainer.

A "Digital Twin" structure that integrates the top SHAP-identified features for an individual patient into a single, easily interpretable representation, independent of the underlying predictive model.

- Clinical narrative generator based on LLM that is limited to the information included in the digital twin structure along with an idea of the interactive dashboard that integrates prediction, explanation, and clinical narrative.

The rest of the paper is structured as follows. Section II provides an overview of relevant literature in cancer prognosis using ML models, explainable AI, digital twins, and clinical narration using language models. Section III outlines the dataset, the pre-processing steps, the model architecture, and the digitaltwin/LLM narrative generation framework. Section IV presents quantitative results, including the classification accuracy, feature importance, and a case study involving a single patient. Section V provides a discussion and insights from the results.

II. Related Work

A. Machine Learning for Breast Cancer Survival Prediction.

Tree-based ensembles, including Random Forest[9], gradient boosting algorithms, and XGBoost[10], have found wide applications in analyzing structured medical datasets to predict outcomes in cancer patients due to their ability to model non-linear feature interactions, efficiency, and tolerance for mixed data types. In the past, studies were conducted using traditional classifiers like logistic regression, decision trees, support vector machines, naive Bayes, and k-nearest neighbor, thus establishing the viability of artificial intelligence (AI)-assisted cancer prediction, but were later surpassed by ensemble algorithms on bigger and more heterogeneous medical datasets. Among the ensembles mentioned, the XGBoost algorithm has proven more accurate and efficient. More specifically, breast cancer-focused studies have applied feature-selected XGBoost classifiers optimized by metaheuristic algorithms [11], hybrid ML-XAI pipelines for early-stage screening [12], integrated feature-selection frameworks for early detection [13], and explainable detection models [14] to improve predictive performance on structured clinical data, while other work has explored imaging-based radiomic approaches for characterizing breast tumors [15].

B. SEER-Based Survival Prediction and Statistical Baselines

The SEER Breast Cancer Dataset [16] has been widely used in prior work as a standard benchmark for survival and mortality classification due to its collection of demographic and clinical features[17], including tumor staging (T/N/6th stage), grading, hormonal receptor status, and nodal examination. Usually, survival outcomes were estimated using Cox proportional hazards models [18], a semiparametric approach that estimates the influence of covariates on the hazard rate, assuming proportional hazards and primarily linear covariate effects. In more recent works, the survival analysis formulation is combined with deep learning models, such as DeepSurv[19], in which, instead of the linear function of Cox regression, a neural network predicts nonlinear interactions among covariates. Other recent work has applied machine learning to analyze survival outcomes among breast cancer patients receiving combined treatment modalities such as chemotherapy, hormone therapy, surgery, and radiotherapy [20].

Despite these influential approaches in statistics and hybrid models, this work follows the more common trend and defines the task as binary classification of patients as alive or dead at a particular time period.

C. Explainable AI in Clinical Machine Learning

SHAP (based on cooperative game theory Shapley values[4]) has become a de facto standard approach to the post-hoc interpretation of tree ensembles since TreeExplainer exact and efficient computation of Shapley values for tree-based models[5]. On the other hand, there exists an alternative model-agnostic method, LIME, that locally fits surrogate models around specific predictions[6]. Previous clinical machine learning research works have employed SHAP summary plots to detect globally important features (nodal status, hormone receptor status, tumor grade) and waterfall plots for explanation of the specific prediction, but this information is usually delivered to technical audience (data scientists, biostatisticians), not clinicians and patients, and this motivates the narrative generation component of this work[21].

D. Digital Twins in Healthcare

The concept of "digital twin", initially used in the engineering/manufacturing domain to refer to a digital, continually updated twin of a real object, has been proposed in the healthcare sector as a patient-specific, interrogable, simulation-based computational model for decision-making[22][23]. Digital twins in healthcare have been applied to tasks such as monitoring patients, analyzing disease progression, and planning individualized treatment strategies by continually incorporating patient-specific clinical information. Yet, the combination of healthcare digital twins and explainable machine learning for predicting cancer survival remains underexplored. In particular, most studies only consider either one task or another, but not both, in the context of an interpretable decision support system, although recent oncology-specific multimodal digital-twin frameworks have begun to address this gap [24] CancerVerse uses the term "digital twin" to refer to the structured, feature-based view of an individual patient's risk profile, computed using SHAP attributions.

E. Large Language Models for Clinical Narrative Generation

Recent advances in generative AI have enabled large language models to produce clinical explanations that closely resemble human explanations, based on structured healthcare data[25]. General-purpose instruction-tuned LLMs have been shown to summarize clinical reports, explain lab results, and write educational information for patients, although literacy and validation gaps remain a concern for their safe clinical deployment [26]. The application of LLMs to describe structured outputs generated by models, rather than for prediction, represents a new research trend. This study does not involve fine-tuning a dedicated model but uses prompt engineering with a general-purpose instruction-tuned LLM, Gemini 2.5 Flash[27], to turn SHAP output table into text, with explicitly stated limitations that forbid any treatment [28] suggestions and do not let the reader know about the underlying algorithm used to generate it.

F. Comparative Summary and Research Gap

Approach	Data	Model	XAI	DT	LLM
Classical ML baselines	Wisconsin SEER	LogReg, SVM	✗	✗	✗
Ensemble tree learning	SEER	Random Forest	✗	✗	✗
Gradient-boosted survival	SEER	XGBoost	Part	✗	✗
Deep survival analysis	SEER	DNN/DeepSurv	✗	✗	✗
Post-hoc XAI on clinical ML	Institutional	XGBoost+SHAP	✓	✗	✗
Imagine-based deep learning	Hospital imaging	CNN	Part	✗	✗
Multi-center ensembles	Multi-center	Ensemble	✓	✗	✗
CancerVerse (proposed)	SEER	XGBoost	✓	✓	✓

XAI=Explainability; DT=Digital twin; LLM=Large-language-model narrative; Part=partial

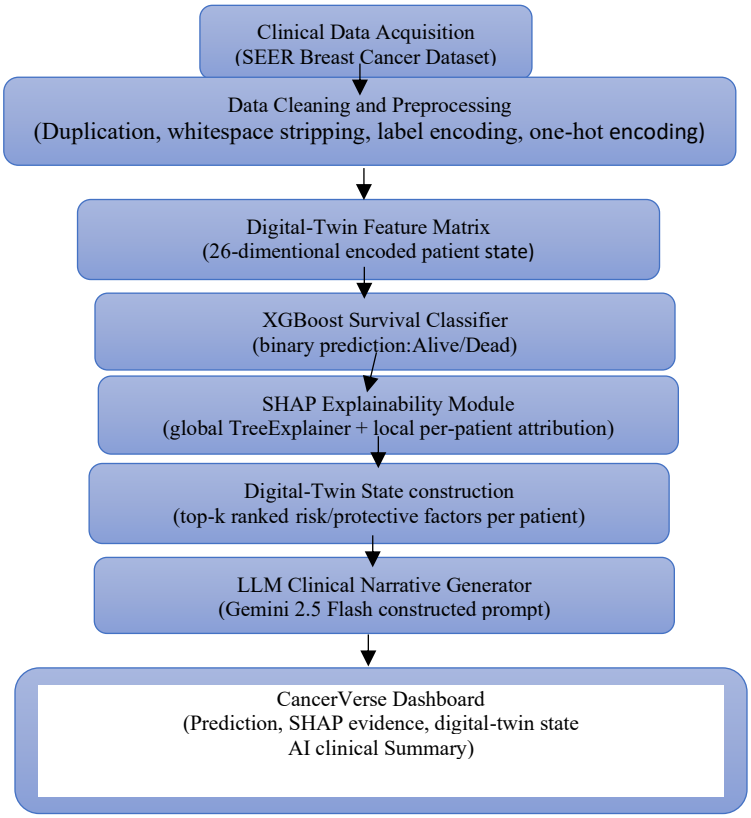
As shown in Table I, previous studies have individually considered tree-based prediction, statistical survival analysis, post-hoc explainability, and digital twin-based representation of patients, but not in combination of all of them within one clinical decision support pipeline. Specifically, (i) previous studies focused on predictive accuracy with minimal interpretability by clinicians; (ii) methods of post-hoc explainability are limited to feature importance visualization rather than patient-specific interpretation; (iii) digital twin-based approaches have not been used together with machine learning approaches to predict breast cancer survival; and (iv) only a few frameworks apply large language models to generate clinical interpretations based on computational predictions.

III. Materials And Methods

A. Framework Architecture

The CancerVerse architecture is comprised of three interdependent layers: a Data Orchestration Layer that ingests and curates the SEER clinical dataset; the Analytical Engine based on XGBoost model, which makes the survival prediction; and the Interpretability & Twin Interface, which calculates SHAP attributions and builds a per-patient digital twin, together with a text generation process driven by an LLM. Fig. 1 demonstrates the end-to-end pipeline starting from data ingestion to CancerVerse dashboard delivery.

Fig 1: CancerVerse end-to-end framework architecture from data acquisition to dashboard presentation



A. DATASET

The data set used here is the publicly accessible SEER Breast Cancer data set [16] that consists of 4,024 patient observations and 16 columns representing raw features related to demographic and tumor characteristics: Age, Race, Marital_Status, T_Stage, N_Stage, Sixth_Stage, Grade, A_Stage, Tumor_Size, Estrogen_Status, Progesterone_Status, Regional_Node_Examined, Regional_Node_Positive, Survival_Months, and the outcome feature Status (Alive/Dead). There was an auxiliary feature column with only empty cells “Unnamed: 3” in the raw export file.

Characteristic	Value
Number of Patients	4023
Number of Raw Features	15
Encoded Feature Count	26
Alive Patients	3407 (84.7%)
Dead Patients	616 (15.3%)
Missing Values (post-cleaning)	0
Duplicate Records Removed	1

B. Data Cleaning And Preprocessing

The following steps are included in preprocessing (i) deletion of the blank column; (ii) deletion of one duplicate patient entry to get a final clean list of 4,023 entries; (iii) removing the leading/trailing spaces for all string variables and column names; (iv) encoding the target variable Status using binary coding, where Alive = 0 and Dead = 1; and (v) checking that there were no missing values left in any column.

C. The Operational Logic Of The Cancerverse Framework:

Algorithm 1: CancerVerse, explainable breast cancer survival prediction
Input: SEER breast cancer dataset (D), with demographic, tumor, clinical attributes.
Output: predicted survival status (P), survival probability (Pr), patient specific Digital-Twin state (DT), and a clinical narrative (N).

Procedure:

i. **Data acquisition and cleaning**: Grab the SEER dataset. Remove the empty auxiliary column and also delete duplicate records, so you end up with the final cohort (n=4,023).

ii. **Preprocessing**: Trim leading/trailing spaces from every attribute name and categorical value. Then encode the target variable Status as,

a. Alive = 0
b. Dead = 1

iii. **Feature encoding**: Use one hot encoding for the 9 categorical fields, namely Race, Marital Status, T/N/6th Stage, Grade, A Stage , and Hormonal Status. After that, build the final 26-dimensional feature matrix (X).

iv. **Dataset partitioning**: Split data into training (80%) and testing (20%) sets, do stratified sampling so the class mix stays about 84.7% Alive , and 15.3% Dead.

v. **Model training**: Train an XGBoost classifier (M) on Xtrain and Ytrain, keep the default hyperparameters.

vi. **Survival prediction**: Feed a single test patient instance into model M to get the predicted binary class P and the survival probability Pr.

vii. **SHAP based explainability**: Set up a SHAP TreeExplainer from M. Compute local Shapley values so you can see the exact contribution from each feature, toward that individual patient's survival outcome.

viii. **Digital-Twin state construction**:

a. Rank the features by the absolute magnitude of their SHAP attributions (like importance, but strictly from SHAP).
b. Pick the top 10 contributing features.
c. Attach tags: if SHAP is positive use "Increases Risk" , if SHAP is negative use "Decreases Risk".

ix. **Clinical Narrative Generation:**

a. Transform the structured Digital-Twin state, into a sort of natural language prompt, for the Gemini 2.5 Flash model . Keep it bound by: produce roughly 120–150 words, use clinical terminology, and do not mention any specific algorithms like SHAP or XGBoost. Also, do not provide, or suggest, treatment recommendations.

x. **Result Presentation**: Bundle the outputs into the CancerVerse dashboard as
{P, Pr, SHAP, DT, N}

where P represents the prediction, Pr is probability, SHAP is the evidence summary, DT is the digital-twin state , and N is the clinical narrative.

End Algorithm

D. Feature Encoding And Train/Test Split

The categorical attributes (Race, Marital Status, T Stage, N Stage, 6th Stage, Grade, A Stage, Estrogen Status, Progesterone Status) were processed by one-hot encoding with the first attribute category dropped in order to prevent dummy variable trap and thus increasing the dimensionality of the feature space from 14 initial attributes to 26 attributes. Stratified sampling technique was applied to create training (80%, n=3,218) and test (20%, n=805) datasets based on target variable (random_state = 42), maintaining the same proportion of classes in both sets (train: 2,725 Alive / 493 Dead; test: 682 Alive / 123 Dead). A full vector of encoded features, along with the class label, serves as raw input for the step of building a digital twin described in Section III-G.

E. Model Development

An XGBoost Classifier (XGBClassifier, scikit-learn-compatible API, logloss scoring function, random_state = 42) was used to fit the encoded training dataset with default parameters. This model choice was made because XGBoost performs well on tabular datasets and can handle feature interactions. Moreover, SHAP uses the exact TreeExplainer, which requires XGBoost.

F. Explainability Via Shap

After training, an shap.TreeExplainer object was built from the trained XGBoost model so as to calculate the exact Shapley values for each of the instances in the test data set. The Shapley values were applied in two distinct ways:
(1) the calculation of global Shapley values for all 805 test cases, with features ranked according to how much they contribute to the model output; and
(2) the calculation of a local Shapley value for one specific test case.

G. Digital-Twin State Construction

The top ten contributing attributes to the selected patient based on the absolute value of their SHAP value have been picked out and tagged with "Increases Risk" (in case of positive SHAP value) and "Decreases Risk" (in case of negative SHAP value), along with their raw feature value and level of contribution. This tabulated information forms the state of the digital twin of the selected patient – a compressed representation of the reasoning behind that particular decision, regardless of the model architecture.

H. Llm-Based Clinical Narrative Generation

The digital twin state was converted into a natural language prompt and inputted into Gemini 2.5 Flash[27], with specific instructions to: (1) provide an explanation of the prediction in 120–150 words; (2) list both risk and protective factors; (3) keep the wording clinical and without recommendations; (4) avoid any references to SHAP, XGBoost, or machine learning algorithms by name; and (5) end with one concluding sentence about prognosis. It is important that the digital twin's originator be kept secret during the generation process because the goal of such an approach is to generate a clinically-oriented text, whereas numeric attributions will still be visible in the table of the digital twin. Prediction, SHAP evidence, digital twin, and text explanation are all available within the CancerVerse dashboard, allowing the user to switch from narrative to numeric attributions at any point.

IV. Result

A. Exploratory Data Analysis

The descriptive statistics (Table III) reveal an average age of patients of 53.97 years (Standard deviation 8.96), an average tumor size of 30.48 mm (standard deviation 21.12), an average number of regional nodes examined of 14.36, an average number of positive regional nodes of 4.16, and an average survival follow-up of 71.30 months. Figure 2 shows that the age distribution of patients ranges from 45 to 65 years, peaking in the mid-50s. The data set is unbalanced with respect to the outcome, with 3,407 patients (84.7%) classified as alive and 616 (15.3%) as dead, as shown in Figure 3. The tumor size distribution (Figure 4) is positively skewed, with most tumors smaller than 40mm.

Fig 2. Age distribution of patients in the cleaned SEER breast cancer dataset.

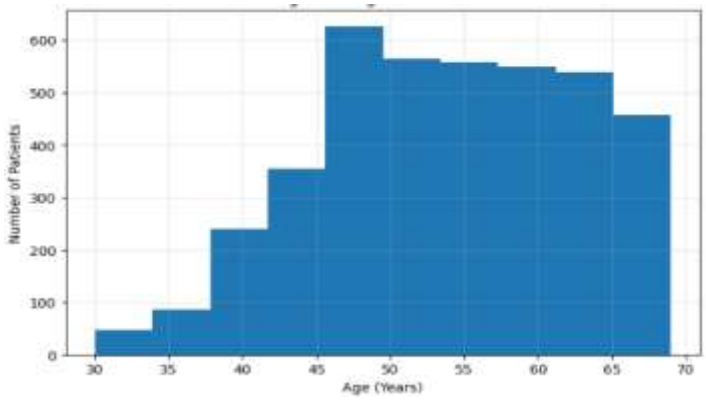


Fig 3. Survival status distribution in the cleaned SEER breast cancer dataset (N=4.023)

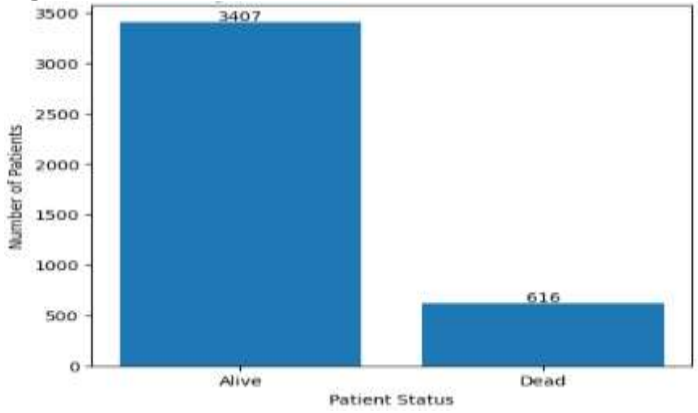


Fig. 4. Distribution of tumor size (mm) across the cohort.

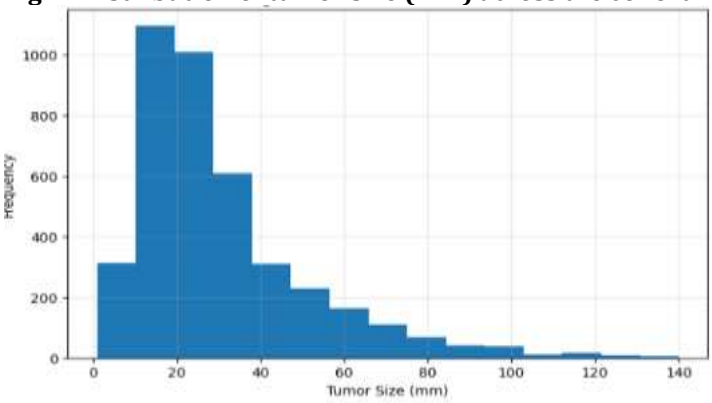


Table III. Descriptive Statistics (Numeric Features)

Variable	Mean	SD	Range
Age (yrs)	53.97	8.96	30-69
Tumor Size (mm)	30.48	21.12	1-140
Reg. Nodes Examined	14.36	8.10	1-61
Reg. Nodes Positive	4.16	5.11	1-46
Survival Months	71.30	22.92	1-107

B. Classification Performance

The results of testing the trained model using the held-out test set of 805 patients are presented in Table IV and Fig. 6. The accuracy, precision, recall, F1-score, and ROC-AUC of the model are 90.68%, 80.77%, 51.22%, 62.69% and

0.8461, respectively. The difference between precision and recall is caused by the class imbalance in the dataset (84.7% Alive vs. 15.3% Dead) and the conservative prediction of the minority "Dead" class – the model predicts 63 of 123 dead patients (51.2%) as "Dead" but keeps the low rate of false positive predictions for the "Alive" class (15 out of 682). The confusion matrix (Fig. 5) shows that there are 667 true negatives, 15 false positives, 60 false negatives, and 63 true positives. The high rate of false negatives (60 dead patients predicted as "Alive") is the cause of moderate recall and the most significant error type.

Fig. 5. Confusion matrix for the XGBoost classifier on the test set (n=805)

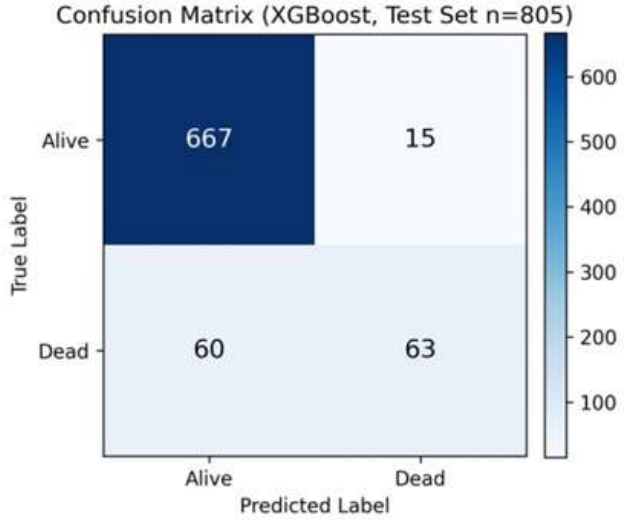


Fig. 6. Summary of classification performance metrics.

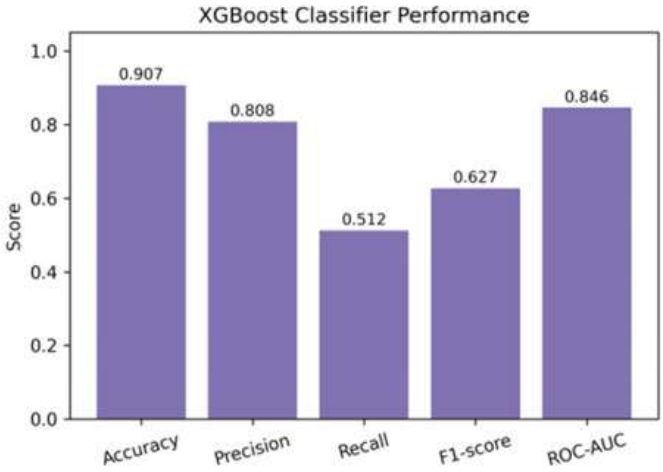


TABLE IV. XB Boost Classification Performance

METRIC	VALUE
Accuracy	0.9068
Precision	0.8077
Recall	0.5122
F1-score	0.6269
ROC-AUC	0.8461

C. Global Feature Importance and SHAP Analysis

The gain-based feature importance (Fig. 8) ranks Survival Months (0.121), grade I of well-differentiated tumor grade (0.092), progesterone-receptor positivity (0.064), nodal stage N3 (0.061), and estrogen-receptor positivity (0.048) as the top five most important predictors, as shown in Table V below. The

ranking is supported by existing oncological prognostic criteria: hormone receptor positivity and tumor well-differentiation are known to be protective factors, while nodal stage is a risk factor. As indicated in the Pearson correlation between numeric features (Fig. 7), Survival Months is moderately negatively correlated with mortality status (r =
-0.48), while Regional Node Positive is positively correlated with mortality (r = 0.26) and Regional Node Examined (r = 0.41), supporting the clinical knowledge that a higher nodal stage is linked with poorer prognosis. This clinical knowledge is directly confirmed by the raw outcome data: as shown in Fig. 9 below, the mortality rate increases from nodal stage N1 to N3.

TABLE V. Top 10 Features By Gain-Based Importance

Feature	Importance
Survival Months	0.1212
Grade: Well differentiated (I)	0.0918
Progesterone Status: Positive	0.0639
N Stage : N3	0.0614
Estrogen Status:Positive	0.0481
6th Stage: IIB	0.0443
Regional Node Positive	0.0442
A Stage: Regional	0.0401
Grade: Poorly differentiated (III)	0.0401
Race: Other	0.0392

Fig 7. Pearson correlation matrix of numeric clinical features.

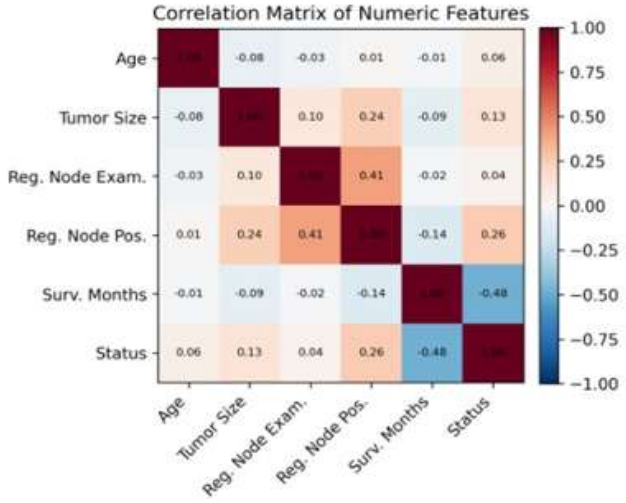


Fig. 8. Top 12 features by XGBoost gain-based importance

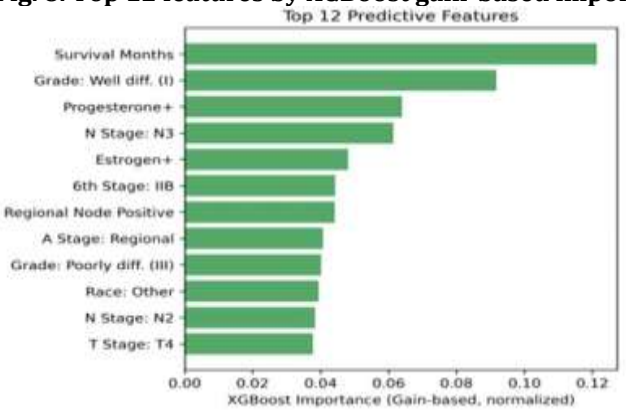
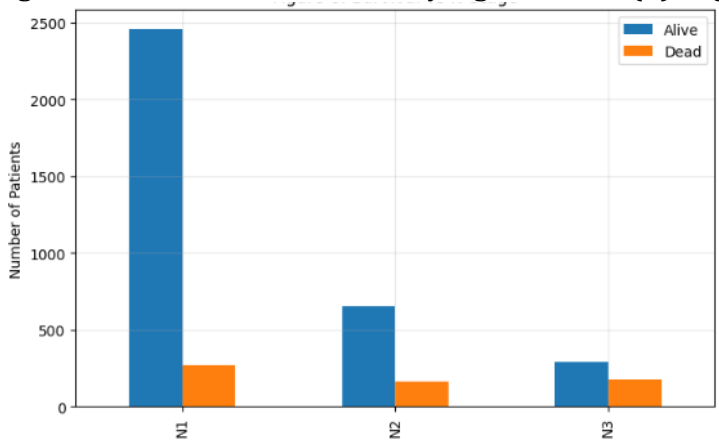


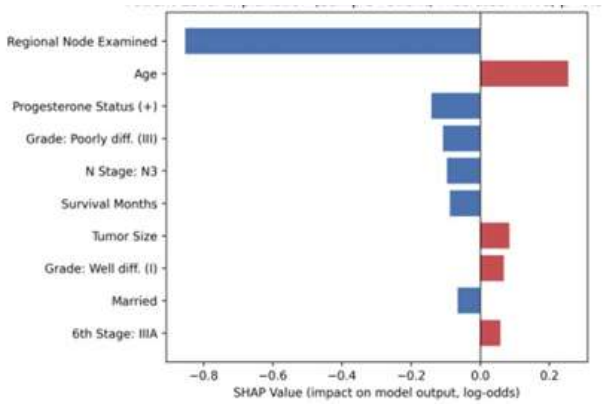
Fig. 9. Observed survival outcome by regional nodal (N) stage.



D. Patient-Level Case Study

To demonstrate the digital twin and narrative generation process, only one test patient sample (age 67 years, tumor size 25 mm, nodal stage N1, number of regional lymph nodes 17, 50 months of survival at final follow-up) was used. The model has shown a probability of "Alive" at 94.5%. The resulting local SHAP value is presented in Fig. 10. High numbers of examined regional lymph nodes (17) have been the main protective contributor (SHAP =−0.853), along with positive progesterone-receptor (−0.141) and poor-grade absence (−0.108,
−0.096). Advanced age (67 years, SHAP = +0.255) and tumor size (25 mm, SHAP = +0.085) have been the leading risk contributors.

Fig.10. patient-level local SHAP explanation for the case-study patient



The digital-twin table for this patient was fed into Gemini 2.5 Flash, which generated the following representative story line, paraphrased below for conciseness: the story line revealed extensive lymph-node assessment, hormone-receptor positivity, and good tumor differentiation to be the primary protective factors, whereas patient age and tumor size to be the primary
risk-increasing factors, finishing off with an overall prognosis that is described as guarded but favorably leaning towards the protective factors. The output above meets all of the design criteria – it mentions the protective and risk factors from the digital twin, avoids algorithm terminology, and results in a single prognostic sentence.

V. Discussion

The CancerVerse pipeline shows that, using a typical gradient-boosted classifier, along with exact SHAP explanations and a constraint LLM prompting technique, one can achieve personalized and understandable prognostic explanations for patients without training an explanation model. The discriminative performance of the model (ROC-AUC = 0.846) is comparable with the previous SEER-based survival analysis studies that use tree

ensembles[10][16][20] and the features selected – nodal status, hormone-receptor status, tumor grade, and survival months – correspond with the breast cancer prognostic literature.

The most significant limitation from a clinical standpoint is the 51.2% recall rate for the "Dead" class which belongs to the minority class. False negatives (cases where the algorithm predicts survival when in actuality the patient's prognosis is unfavorable) incur more clinical costs than false positives. Thus, efforts to mitigate class imbalance should be undertaken before applying the model in a clinical setting.

This digital-twin/LLM layer of the story solves a different yet complementary challenge: even the most accurate model has very little utility if its outputs are not communicated in a way that clinicians and patients can understand. The restriction of the LLM in summarizing only the SHAP-derived digital-twin state (not in generating anything else freely) minimizes (yet does not completely eliminate) the chance of hallucinating clinical facts, since every sentence in the story is linked to some specific feature attribution. Having all of the predictions, SHAP evidence, digital-twin state, and the story presented in a single dashboard view enhances transparency, since one can easily switch between the natural-language story and the numeric evidence.

VI. Limitations

A number of constraints affect the generalizability of the results. Firstly, the data are retrospective, registry-based, and drawn from a single nationwide cancer surveillance program; external validation in new cohorts is necessary. Secondly, the class imbalance prevents high recall for the minority mortality class, and no sampling method or cost-sensitive learning was used in this particular application. Thirdly, the model hyper parameters were chosen using library defaults rather than tuned via cross-validation; therefore, the obtained metrics most likely underestimate the model's potential performance. Fourthly, the story generated by an LLM, although limited by the prompt, is not deterministic across generations and has not been validated by healthcare professionals for factual accuracy; it must be considered a communication tool, not a diagnosis or treatment option. Lastly, the "digital twin" in this case refers to a one-time point set of features, rather than a dynamically updated and simulated patient model as in some definitions of a digital twin.

VII. Conclusion And Future Work.

CancerVerse was proposed as an explainable digital twin framework for predicting the survival probability of breast cancer patients using the combination of XGBoost classification, global and local explanations using SHAP, construction of the digital twin state with structured data, and clinical narratives using language model within the context of an interactive dashboard. The framework performed at an accuracy of 90.68% and an ROC-AUC of 0.846 on the SEER breast cancer dataset while demonstrating attributions of the features aligned with known factors used by oncologists for prognosis.

Future directions include: (1) class-imbalance aware training and hyper-parameter optimization for better recall for the minority class; (2) independent multi-institutional validation of external validity; (3) formal clinical evaluation of the LLM-generated narratives for accuracy and safety; (4) expanding the digital twin approach to longitudinal and multimodal clinical data, including imaging, histopathology, and genomics; (5) use of graph neural networks, transformers, and federated learning approaches to enhance prediction performance without compromising the privacy of patients from other institutions; and (6) a usability trial evaluating clinician trust and understanding of CancerVerse compared to raw SHAP visualizations.

References

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021. DOI: 10.3322/caac.21660.
2. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015. DOI: 10.1016/j.csbj.2014.11.005
3. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, pp. 1347–1358, 2019. DOI: 10.1056/NEJMra1814259

4. R. Bommasani, "NeurIPS should lead scientific consensus on AI policy," 2022, DOI: 10.52202/085713-1197.
5. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020. DOI: 10.1038/s42256-019-0138-9
6. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778
7. A. Adeniran, A. P. Onebunne, and P. William, "Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making," DOI: 10.30574/wjarr.2024.23.3.2936
8. J. Inukonda, V. R. R. Tetala, and J. Hallur, "Explainable Artificial Intelligence (XAI) in Healthcare: Enhancing Transparency and Trust," International Journal For Multidisciplinary Research, 2024, DOI: 10.36948/ijfmr.2024.v06i06.30010
9. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324
10. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16), 2016, pp. 785–794. DOI: 10.1145/2939672.2939785
11. P. Sarker, A. Ksibi, M. M. Jamjoom, K. Choi, A. Al Nahid, and M. A. Samad, "Breast cancer prediction with feature-selected XGB classifier, optimized by metaheuristic algorithms," Results in Engineering, vol. 19, 101340, 2023. DOI: 10.1186/s40537-025-01132-7
12. S. Shubhangi, A. Husain, G. Agarwal, and D. R. Ojha, "A data-driven hybrid ML–XAI approach for early stage breast cancer screening," Naveen International Journal of Multidisciplinary Sciences, vol. 1, no. 6, 2025, DOI: 10.71126/nijms.v1i6.79
13. J. Zhu, Z. Zhao, B. Yin, C. Wu, C. Yin, R. Chen, and Y. Ding, "An integrated approach of feature selection and machine learning for early detection of breast cancer," Scientific Reports, vol. 15, 2025. DOI: 10.1038/s41598-025-97685-x
14. T. Arravalli, K. Chadaga, H. Muralikrishna, N. Sampathila, D. Cenitta, R. Chadaga, and K. S. Swathi, "Detection of breast cancer using machine learning and explainable artificial intelligence," Scientific Reports, vol. 15, Art. no. 26931, 2025. DOI: 10.1038/s41598-025-12644-w
15. X. Jiang, D. Zhu, X. Luo, S. Fang, N. Song, Q. Yu, S. Huang, and S. Wang, "Multiparametric MRI-based radiomics of whole-tumor and habitat regions for predicting HER2 status in young breast cancer: a two-center study," Frontiers in Oncology, vol. 16, Art. no. 1760589, 2026. DOI: 10.3389/fonc.2026.1760589.
16. SEER Program, National Cancer Institute, "SEER Breast Cancer Dataset," seer.cancer.gov. [Accessed 2026].
17. K. L. Vangsness, A.-P. Sam, N. Munabi, M. Chu, A. Wong, J. Chang, and A. L. Carre, "A review of the SEER database: An analysis on barriers of access to breast reconstruction in California," Plastic and Reconstructive Surgery–Global Open, vol. 12, no. 1 Suppl. 1, pp. 34–34, 2024. DOI: 10.1097/01.GOX.0001006028.53744.ea
18. D. R. Cox, "Regression models and life-tables," J. Royal Statistical Society: Series B, vol. 34, no. 2, pp. 187–220, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x
19. J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," BMC Medical Research Methodology, vol. 18, no. 1, Art. no. 24, 2018. DOI: 10.1186/s12874-018-0482-1
20. E. M. Tegaw and B. B. Asfaw, "Machine learning analysis of survival outcomes in breast cancer patients treated with chemotherapy, hormone therapy, surgery, and radiotherapy," Scientific Reports, vol. 15, Art. no. 24981, 2025. DOI: 10.1038/s41598-025-97763-0
21. S. Mujawar, M. Sayed, and A. Chaware, "Empowering Decision Making in Healthcare—A Comparative Analysis of eXplainable AI Techniques Using SHAP and LIME for Cancer Patients Data," in Proc. 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON), New Delhi, India, 2024, pp. 1–5. DOI:10.1109/DELCON64804.2024.10867049.
22. M. Peshkova, V. Yumasheva, E. Rudenko, N. Kretova, P. Timashev, and T. Demura, "Digital twin concept: Healthcare, education, research," Journal of Pathology Informatics, 2023, DOI:10.1016/j.jpi.2023.100313.
23. S. Kanakaprabha and A. Prasanth, "Digital Twins in Oncology: AI-based Multimodal Data Framework for Cancer Patient Care," in Proc. 2025 International Conference on Sustainable Communication Networks and Application (ICSCN), 2025, pp. 325–330, DOI:10.1109/ICSCN67106.2025.11308608.
24. S. Reddy, "Generative AI in healthcare: An implementation science informed translational path on application, integration and governance," Implementation Science, Mar. 15, 2024, DOI: 10.1186/s13012-024-01357-9.

25. D. Rodger, S. P. Mann, B. Earp, J. Savulescu, C. Bobier, and B. P. Blackshaw, “Generative AI in healthcare education: How AI literacy gaps could compromise learning and patient safety,” Nurse Education in Practice, vol. 87, Art. no. 104461, 2025, DOI:10.1016/j.nepr.2025.104461
26. Padmalal, S., Arumugam, P., Anusha, M. Baby, Bamane, Kalyan Devappa, Yamsani, Nagendar and Muppavaram, Kireet. "Machine Learning for Early Detection of Chronic Diseases: A Case Study in Diabetes Prediction". Machine Learning in Healthcare: Data-Driven Decisions, Predictive Modelling, Personalized Medicine, edited by Saurav Mallik, Sandeep Kumar Mathivanan, S.K.B. Sangeetha, Ben Othman Soufiene, Saravanan Srinivasan and Aimin Li, De Gruyter, 2026, pp. 171-192. https://doi.org/10.1515/9783111541235-009.
27. Aparna, S., Muppavaram, K., Ramayanam, C. C. V., & Ramani, K. S. S. (2021). Mask RCNN with RESNET50 for Dental Filling Detection. International Journal of Advanced Computer Science and Applications, 12(10). https://doi.org/10.14569/IJACSA.2021.0121079.
28. Kireet Muppavaram, Aparna Shivampeta, Sudeepthi Govathoti, Deepthi Kamidi, Kiran kumar mamidi, Manyam Thaile, "Investigation of Omnidirectional Vision and Privacy Protection in Omnidirectional Cameras," International Journal of Electronics and Communication Engineering, vol. 10, no. 5, pp. 105-116, 2023. Crossref, https://doi.org/10.14445/23488549/IJECE-V10I5P110.