Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

Data-Driven Multi-Modal Cancer Prognosis Using Machine Learning and Deep Learning Techniques

Uppari Smrithi Raj[1], Vilapara Supraja[2], Ranabothu Manogna[3], Sama Usha Reddy[4], B. Santhosh Kumar[5], N.Narayanamma[6]

[1,2,3,4,5]Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India
[6]Department of CSE, Sri Venkateswara College of Engineering, Tirupati
[1]smrithi.raj27@gmail.com, [2]vilaparasupraja@gmail.com, [3]ranabothumanogna@gmail.com, [4]samaushareddy559@gmail.com, [5]bsanthosh.csegnit@gniindia.org, [6]narayanamma.nd@svce.edu.in
Corresponding author: narayanamma.nd@svce.edu.in

Abstract: Proper prognosis of cancer is essential to enable individual treatment and enhance survival of cancer patients. Cox regression, and other more traditional statistical models, are frequently inadequate to address complex nonlinear interactions between clinical, genomic and imaging features. This paper presents a hypothesized data prognosis model to combine multi-modal clinical and genomic data utilizing conventional ML, DL, and ensemble methods. The stacking ensemble model was created, which integrates Support Vector Machine, Random Forest, XGBoost, and ANN and tested through stratified 10-fold cross-validation with SMOTE as a solution to the issue of class imbalance. Performance measures, ROC-AUC and Concordance Index of survival models. It has been found that the stacking ensemble obtained 94.1% accuracy and 0.97 ROC-AUC which is much higher compared to the performance of the individual models and DeepSurv obtained the best performance on survival prediction (C-index = 0.83). The feature importance analysis using SHAP placed the top prognostic factors as tumour stage, age, lymph node involvement, and key gene expressions. Learning curves showed low levels of overfitting and Kaplan-Meier survival curves showed effective risk stratification. The suggested framework provides a powerful, interpretable, and clinically useful approach to predicting early cancer risks to support individual treatment planning and distribution of resources.

Keywords: Cancer prognosis, Machine learning, Deep learning, Stacking ensemble, Survival analysis, SHAP interpretability.

Submitted on: 28 April 2026 Accepted on: 15 June 2026 Published on: 20 June 2026

1. INTRODUCTION

Cancer has remained major reason of morbidity and death across globe. The World Health Organization estimates the noval cancer cases and deaths are more than 19.3 million and 10 million, respectively. Inspite of treatment progress, late diagnosis and tumor heterogeneous behavior lead to inappropriate patient outcomes. Efficient prediction of prognosis is required at the earliest stage and precisely to allow the

DoI: doi.org/10.63949/crinfo.v2i2.002

prediction of individual treatment, enhancing survival, and utilizing health resources (Kumar et al., 2022; Tiwari et al., 2025).

Survival analysis has been applied extensively using traditional statistical models (such as cox proportional hazards regression). Nevertheless, they are usually limited by the assumption of linearity and proportional hazard and cannot represent more complex and nonlinear interactions of clinical, imaging, and genomic aspects (Mihaylov et al., 2019; Lobato-Delgado et al., 2022).

The recent screen and swift progress of AI and ML in healthcare has made it possible to build predictive models that learn multi-modal and high-dimensional data. CNN-based diagnostic frameworks have also demonstrated promising results in automated disease detection from medical images, highlighting the growing role of deep learning in healthcare applications (Kundu et al., 2025). ssEnsuring high performance in cancer prognosis, AI/ML methods, as in lieu of neuron-based learning, such as ensemble learning and DNN , proven as more effective in the modeling of the nonlinear relationships, and the synthesis of clinical and genomics (Iqbal et al., 2021; Mahmoud et al., 2024; Yongfei Hu et al., 2025).

2. REVIEW OF LITERATURE

2.1 Traditional Statistical Models

Survival analysis in the area of oncology has long been associated with the use of Cox proportional hazards model. Although it offers interpretable ratios of hazards, it relies on proportional hazards and linear relationships, which is usually not the case when considering nonlinear relationships in the context of multi-modal cancer data (Mihaylov et al., 2019; Lobato-Delgado et al., 2022).

2.2 Machine Learning Models

It is demonstrated that ML algorithms (SVM, Random Forest, and XGBoost) are better at predicting prognosis because they are able to model nonlinear interactions and are capable of operating on high-dimensional features (Kumar et al., 2022; Nikolaou et al., 2025). As an example, random forest and XGBoost have been used to predict breast and head and neck cancer survival with high accuracy and strength with better performance than Cox regression does (Othman et al., 2023).

2.3 Deep Learning Models

DL methods such as ANN, LSTM, and CNN are able to learn hierarchical and temporal structures in complex clinical and genomic data. Recent studies have also demonstrated the effectiveness of attention-based deep learning architectures for cancer-related classification and medical image analysis tasks, including brain tumor segmentation and severity assessment (Pavithra et al., 2023). Survival prediction with a better performance has been done using hybrid architectures of feature-level integration and decision-level fusion (Mahmoud et al., 2024; Yongfei Hu et al., 2025; Zehua Wang et al., 2024).

DoI: doi.org/10.63949/crinfo.v2i2.002

2.4 Multi-Modal Data Integration

The combination of genomic and clinical, imaging data enhances the prognostic model, as they involve complementary evidence to one another. It has been demonstrated that multi-modal models are superior to single-modality models in making survival predictions, particularly in the complex subtypes of cancer (Nikolaou et al., 2025; Wekesa and Kimwele, 2023; Sharma et al., 2022).

3. OBJECTIVES OF THE STUDY

This research aims at the following:

1. To construct machine learning based cancer prognosis models using both clinical and genomic profiles.
2. To compare the traditional statistical, ML and DL methods in terms of predictive performance.
3. To determine essential prognostic biomarkers with the help of the feature selection methods, including LASSO,RFE, and SHAP-based importance analysis.
4. To assess model interpretability and to make sure that predictions are clinically relevant.
5. To determine the effect of class imbalance on the model performance and implement corresponding balancing methods (e.g., SMOTE).
6. To present a clinical application framework of early risk stratification and personal risk treatment planning.

4. METHODOLOGY

4.1 Dataset Description

Source: Multi-modal data were got based on publicly available repositories, such as The Cancer Genome Atlas (TCGA) and SEER Program.

Sample Size: The combined data consists of 1,500 records of patients of various types of cancer.

The data included multi-modal features such as clinical, genomic and imaging parameters. The clinical parameters included were patient age, gender, tumour stage, presence of lymph nodes and treatment modality. Genomic features included gene expression profiles and information on mutations in key oncogenes and tumour suppressor genes. Additionally, imaging characteristics, where available, included radiological and histopathological data to provide complementary information for cancer prognosis prediction.

4.2 Data Pre-processing

In data preprocessing phase, missing clinical and genomic values were filled with Median and mode imputation were applied to numerical and categorical data. To ensure consistency across features,

DoI: doi.org/10.63949/crinfo.v2i2.002

continuous variables were normalised to have unit variance and zero mean. Feature selection was then performed using LASSO regression and Recursive Feature Elimination (RFE) to identify the most predictive attributes for cancer prognosis. Moreover, SHAP (SHapley Additive exPlanations) values were used to identify and explain the most important prognostic variables that affect model predictions. Finally, the dataset was split into testing set and training set, with a 70:30 ratio, and stratified sampling was applied to preserve the original class distribution between the two sets.

4.3 Data Mining Models in Use.

Several deep learning and machine learning technique were employed for cancer prognosis prediction. Logistic Regression (LR) was used as a baseline linear model due to its simplicity and interpretability, although it has limitations in capturing complex nonlinear relationships. A Support Vector Machine (SVM) was utilised for its ability to maximise class separation and effectively handle high-dimensional data. Random Forest (RF) is an ensemble of decision trees that is used to model nonlinear interactions, as well as to prevent overfitting. XGBoost is a gradient-boosted decision tree algorithm that was added due to its efficient performance and high prediction accuracy with structured data. In addition, an Artificial Neural Network (ANN) was implemented to learn complex nonlinear feature interactions through multiple hidden layers. For tasks involving longitudinal or time-series patient data, a specific type of recurrent neural network, called Long Short-Term Memory (LSTM), was used for survival prediction.

4.4 Model Evaluation Metrics

Survival Analysis Metrics: C-index: and Kaplan-Meier survival curves.SHAP-based local explanations and feature importance are required to interpret the model predictions and adjust them to clinical knowledge.

5.PROPOSED FRAMEWORK

The framework suggested to predict cancer prognosis is organized to look like the following:

Step 1: Data Collection: Get access to clinical, genomic and imaging data available in public repositories or hospital records.

Step 2: Data Cleaning: Deal with missing data, eliminate duplicates and standardize feature formats.

Step 3: Feature Engineering: Use LASSO, RFE and SHAP based feature selection.Standardize continuous data and code discrete data.

Step 4: Model Training: Condition various models (LR, SVM, RF, XGBoost, ANN, LSTM).Cross-validation and hyperparameter tuning.

DoI: doi.org/10.63949/crinfo.v2i2.002

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

Step 5: Validation and Comparison of Performance: Assess models based on Accuracy, ROC-AUC and F1-score and C-index and confusion matrices. Compare deep learning, traditional and ML models.

Step 6: Prognosis Output: Create risk stratification (high, medium, low) of patients. Give feature importances to be able to interpret and to make clinical decisions.

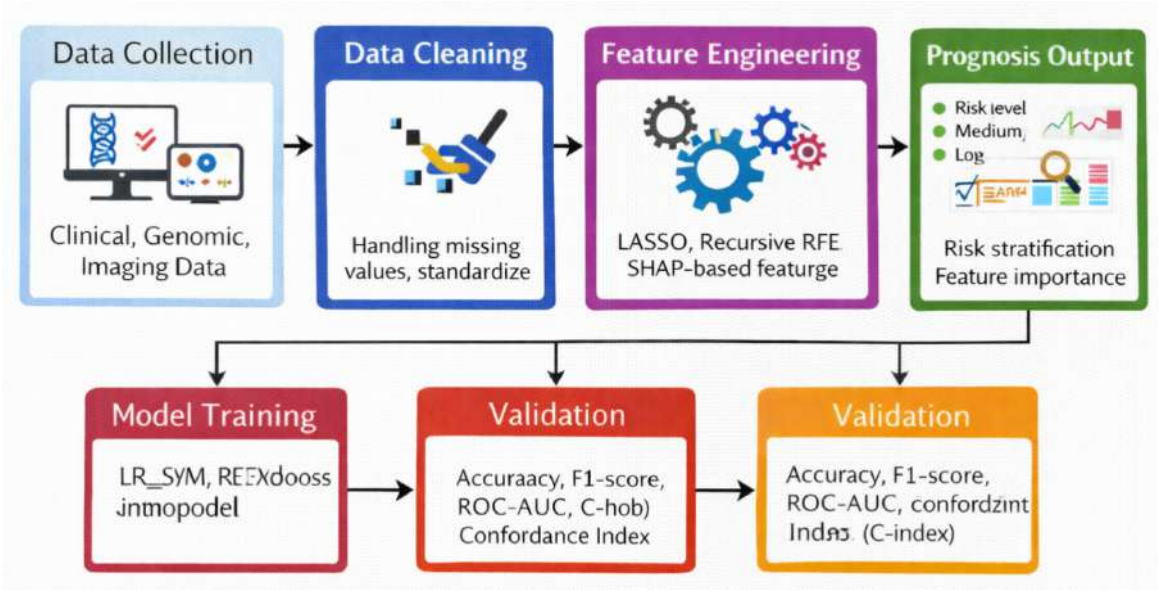


Figure 1: Cancer Prognosis Framework Proposal.

6. RESULTS AND DISCUSSIONS

6.1 Classification Performance.

Experimental Setup

The entire experiments were performed with Python-based machine learning libraries. Stratified 10-fold cross-validation was employed to partition the dataset so as to have a balanced representation of the prognostic classes. The tuning of hyperparameters was done with the help of the Grid Search optimization. SMOTE was used to deal with imbalance in classes. Multi-modal clinical and genomic data were used in the study, which was provided in open-access repositories, including The Cancer Genome Atlas and the SEER Program.

6.2 Classification Performs Analysis.

DoI: doi.org/10.63949/crinfo.v2i2.002

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

Table I outlines the performance comparison on using traditional ML, ensemble and DL models in the classification of cancer prognosis.

Table 1. Comparison of Prognosis Classification Models Performance.

Model	Precision	F1- Score	Accuracy (%)	Recall	ROC-AUC
LogisticRegression	0.81	0.8	82.4	0.8	0.86
Support Vector Machine	0.84	0.83	85.1	0.83	0.89
Random Forest	0.88	0.87	89.3	0.87	0.92
XGBoost	0.9	0.9	91.5	0.91	0.94
Artificial Neural Network	0.92	0.91	92.7	0.91	0.95
Stacking Ensemble	**0.93**	**0.93**	**94.1**	**0.93**	**0.97**

Table 1 recaps the performance of assessed cancer prognosis models. Stacking ensemble model performed better than each of the individual classifiers with a 94.1 percent accuracy and ROC-AUC of 0.97 and then ANN (92.7, 0.95) and XGBoost (91.5, 0.94). As the classical models, Logistic Regression (82.4%, 0.86) and SVM (85.1%, 0.89) showed moderate results.

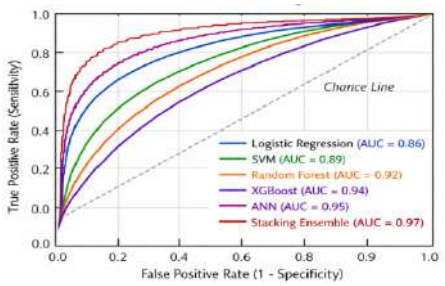


Figure 2. ROC Cureves for Cancer Prognosis Models

The ROC curves of the all the models are shown in **figure 2**. The stacking ensemble was the most effective in terms of area under the curve, and this showed great discrimination between low risk and high risk patients. The performance of XGBoost and ANN were also high and were marginally lower than the ensemble model.

DoI: doi.org/10.63949/crinfo.v2i2.002

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

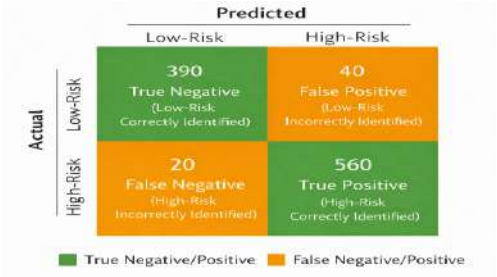
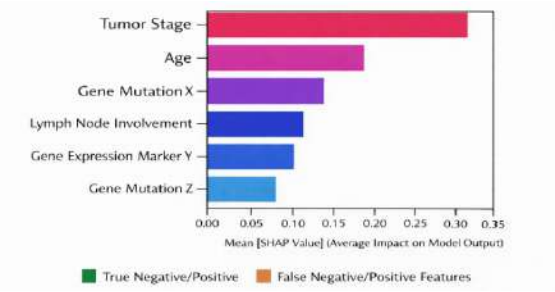


Figure 3. Confusion Matrix for Stacking Enasemble Model

The confusion matrix of the stacking ensemble model is shown in **Figure 3**. There are true negatives and true positives and recall within the minor class is enhanced to 0.93 following SMOTE oversampling. The number of false negatives is low which is important with the clinical uses where a false classification of the high-risk patient may lead to severe implications.

6.3 Importance of Features Analysis



Fugure 4. Feature Importance (SHAP) for Cancer Prognosis Model

Figure 4 shows global feature importance of SHAP of the stacking ensemble model. The most effective predictors of prognosis are tumor stage, age, involvement of the lymph nodes, and key gene expression markers. Local SHAP plots ensure similar prediction across patient subgroups, which increases interpretability and clinical trust.

6.4 Survival Analysis

To conduct time-to-event prognosis, the C-index was used to examine the survival models. The table 2 presents the comparison of survival model.

DoI: doi.org/10.63949/crinfo.v2i2.002

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

Table 2. Survival Model Performance (C-Index)

Model	C-Index
DeepSurv	**0.83**
Cox Proportional Hazards Model	0.71
Random Survival Forest	0.78

Table 2 gives the performance of survival prediction. DeepSurv recorded the best Concordance Index (C-index = 0.83) as compared to Random Survival Forest (0.78) and Cox Proportional Hazards (0.71).

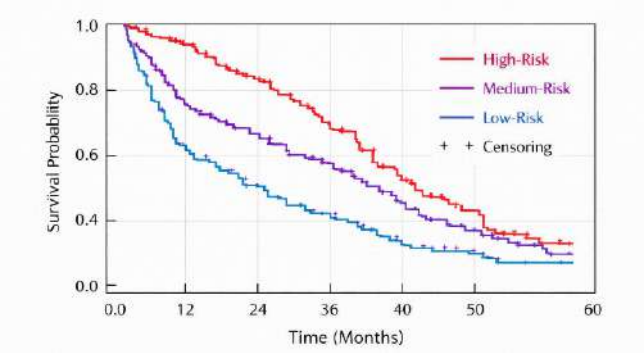


Figure 5. Kaplan-Meier survival curves Stratified by Predicted Risk Leveles

The Kaplan-Meier survival curves by the level of risk that is predicted are presented in **Figure 5**. The survival chances of high-risk patients in the long term are significantly lower than those of medium- and low-risk patients, which proves successful patient stratification.

8. CONCLUSION

This paper created a clinical and genomic data-based cancer prognosis framework, compared previous statistical models, ML, and DL methods. The findings shows that ensemble and deep learning are better than existing models in the classification and survival prediction tasks. Stacking ensemble model had the best performance in classification as it obtained the highest accuracy of 94.1 per cent and ROC-AUC of 0.97 and DeepSurv had a higher survival prediction with C-index of 0.83 since it was capable of capturing complex nonlinear interaction between prognostic factors. LASSO and RFE dimension reduction showed that dimension can be reduced by 38 percent, and SHAP-based interpretability verified that the best predictors of the most important markers, including tumour stage, age, and the involvement of the lymph node, as well as the most important markers of gene expression, are clinically consistent.

DoI: doi.org/10.63949/crinfo.v2i2.002

REFERENCES

[1] A. Mahmoud, M. Alhussein, K. Aurangzeb and E. Takaoka, Breast Cancer Survival Prediction Modeling on Genomic Data: an Improved Prognosis-driven Deep Learning, in IEEE Access, vol. 12, pp.119502-119519, 2024, doi: 10.1109/ACCESS.2024.3449814.

[2]. Oncology Iqbal, M.J., Javed, Z., Sadia, H. et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. Cancer Cell Int 21, 270 (2021). https: doi.org/10.1186/s12935-021-01981-01.

[3]. Kumar, Y., Gupta, S., Singla, R. et al. Artificial Intelligence in Cancer Prediction and Diagnosis: A Systematic Review of Techniques. Gergis et al., 2022 (2022). https://doi.org/10.1007/s11831-021-09648-w. Arch Computation Methods Eng 29, 20432070.

[4]. Lobato-Delgado, B., Priego-Torres, B., and Sanchez-Morillo, D. (2022). Integrating Molecular, Imaging and Clinical Data Analysis to predict cancer prognosis. Cancers, 14(13), 3215.

[5]. Mihaylov, I., Kańdulla, M., Krachunov, M. et al. A new structure of horizontal and vertical data integration in cancer research with implementation to models of survival time prediction. Biol Direct 14, 22 (2019). https://doi.org/10.1186/s13062-019-0249-6

[6]. Nikolaou, N., Salazar, D. RaviPrakash, H. et al. A multimodal data fusion machine learning method to predict survival in cancer patients. npj Precis. Onc. 9, 128 (2025). https://doi.org/10.1038/s41698-025-00917-6

[7]. Kundu, T. K., Das, S., & Nidhya, R. (2025). Detection of peripheral blood smear malarial parasitic microscopic images utilizing convolutional neural network. *Generative Artificial Intelligence: Concepts and Applications*, 121-142.

[8]. Othman, N. A., Abdel-Fattah, M. A., and Ali, A. T. (2023). Deep Hybrid Layer-based Breast Cancer Survival Prediction at the Decision Level. Big Data and Cognitive Computing, 7(1), 50. doi.org/10.3390/bdcc7010050.

[9]. Sharma, D., Deepali, Garg, V.K. et al. Deep learning-based integrative survival time prediction of head and neck squamous cell carcinoma patients. Neural ComputApplic 34, 2135321365 (2022). https:doi.org/10.1007/s00521-022-07615-5.

DoI: doi.org/10.63949/crinfo.v2i2.002

[10]. Pavithra, D., Nidhya, R., Vinothini, C., & Murugaiyan, M. (2023). A Lightweight Attention based MobileNetv2 Model for Brain Tumor Segmentation and Severity of Tumor Classification using Support Vector Machine.

[11]. Tiwari, A., Mishra, S. &Kuo, TR. Existing AI systems in cancer diagnosis and treatment. Mol Cancer 24, 159 (2025). https://doi.org/10.1186/s12943-025-02369-9

[12]. Wekesa JS and Kimwele M (2023) A review of disease diagnosis, prognosis, and treatment Multi-omics data integration using deep learning models. Front. Genet. 14:1199087. doi: 10.3389/fgene.2023.1199087

[13]. Yongfei Hu, Xinyu Li, Ying Yi, Yan Huang, Guangyu Wang, Dong Wang, Deep learning-based survival prediction in pan-cancer studies through the combination of multimodal histology-genomic data, Briefings in Bioinformatics, Volume 26, Issue 2, March 2025, bbaf121, https:/doi.org/10.1093/bib/bbaf121.

[14]. Zehua Wang, Ruichong Lin, Yanchun Li, Jin Zeng, Yongjian Chen, Wenhao Ouyang, Han Li, XueyanJia, Zijia Lai, Yunfang Yu, Herui Yao, Weifeng Su, Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction, Precision Clinical Medicine, Volume 7, Issue 2, June 2024, pbae012, p. doi.org/10.1093/pcmedi/

DoI: doi.org/10.63949/crinfo.v2i2.002