Adaptive Clinical Feature Weighted XGBoost Framework with Explainable Artificial Intelligence for Survival Prediction in Anaplastic Thyroid Cancer

Dr.M. Anbarasi[1], S. Sandhiya[2]
[1*]Associate Professor, Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.(Corresponding Author) Email: manbarasi@vit.ac.in
[2*] Student, Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
(Author)Email: sandhiya.s2025@vitstudent.ac.in

ABSTRACT
Anaplastic Thyroid Carcinoma (ATC) is an uncommon type of cancer with very low survival chances. The ability to predict patients' survival can aid significantly in making informed decisions and designing treatment plans. Nonetheless, the issues of imbalanced classes, high dimensionality of clinical features, and scarcity of data pose major limitations to survival prediction accuracy. This paper introduces a novel approach, namely Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) along with Explainable Artificial Intelligence (XAI), for ATC survival prediction.
The framework that is developed uses comprehensive data pre-processing, SMOTE, adaptive feature weight, and classification based on XGBoost algorithm to improve the performance of the predictive model for unbalanced data. At the same time, the explanation of the influence of clinical features on the outcome of prediction is provided by using SHAP explainability analysis. According to experimental results, the proposed framework (ACFW-XGB) has shown high efficiency of prediction in terms of 84.21% accuracy, 87.87% precision, 93.54% recall, and 90.62% F1-score. Comparative study with conventional machine learning models (e.g. K-Nearest Neighbors and Random Forest) confirms the efficiency of the model suggested. In addition, the results obtained during the process of cross-validation reveal the average accuracy of 83.00%.
In conclusion, the presented approach enables a computationally efficient and robust machine learning technique for intelligent decision-making in thyroid cancer survival prediction.

KEYWORDS
Anaplastic Thyroid Cancer, XGBoost, Explainable AI, SHAP, Survival Prediction, SMOTE, Machine Learning, Feature Weights
How to cite this article: Anbarasi M, Sandhiya S. Adaptive Clinical Feature Weighted XGBoost Framework with Explainable Artificial Intelligence for Survival Prediction in Anaplastic Thyroid Cancer. Int J Drug Deliv Technol. 2026;16(66s):1159-1164. DOI: 10.25258/ijddt.16.66s.119

1. INTRODUCTION

One such type of thyroid cancer is Anaplastic Thyroid Cancer (ATC), which is considered one of the most aggressive types of thyroid cancer because of fast tumor growth and low survival rates. While ATC is rare compared to other types of thyroid cancers, it makes up a significant percentage of deaths resulting from thyroid cancer because of its very invasive nature and inability to be treated through conventional means.

Prediction of survival from the use of medical data sets is still very challenging due to class imbalances, high dimensionality of the clinical data, small sample sizes, and complicated non-linear correlations among the various medical variables. Machine learning algorithms have gained much recognition in recent years due to the ability to improve prediction accuracy and unveil clinical insights through massive medical data sets.

Many machine learning models such as RF (Random Forest), SVM (Support Vector Machine), KNN (K-nearest neighbors), and XGBoost have been used for cancer survival analysis and prediction. Nevertheless, several drawbacks have been identified in many existing models such as reduced interpretability, weak ability to address issues associated with imbalanced dataset problems, and inefficient approaches to features prioritization. This paper attempts to present an approach referred to as the adaptive clinical feature weighted XGBoost (ACFW-XGB) coupled with Explainable Artificial Intelligence (XAI). The proposed method uses SMOTE based class balancing, adaptive features weighting, XGBoost classifier, and SHAP based explainability analysis.

2. RELATED WORK

The application of machine learning algorithms in predicting cancer outcomes and estimating survival has been highly prevalent because of their ability to analyze complicated medical data and recognize hidden medical patterns. Many studies have adopted machine learning algorithms to predict the likelihood of developing thyroid cancer. Despite this, several problems remain unaddressed, such as class imbalance, high dimensionality, and lack of interpretability.

A thyroid cancer predictor model was created using machine learning approaches with ensemble learning techniques that performed better in predicting outcomes. For instance, Barfejani et al. used the Random Forest and logistic regression methods for the survival prediction of thyroid cancer patients, but their proposed method failed to utilize explainable artificial intelligence. The method by Patel et al., which included the SMOTE method alongside Random Forest algorithm for enhanced predictions on cancer cases, also suffered from redundant feature influence problems, affecting the predictive performance and stability of the model.

The past few years have seen the rise in significance of Explainable Artificial Intelligence approaches, especially SHapley Additive exPlanations (SHAP), due to the contribution of these models towards making AI models more interpretable and reliable clinically. However, the current models are plagued by issues related to class imbalance, small sample sizes, and insufficient feature selection criteria. To address these shortcomings, this paper presents the development of an Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) framework combined with SMOTE and SHAP-based explanation approach for survival prediction in patients of ATC.

3. MATERIALS AND METHODS

The design of the proposed framework for Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) was mainly aimed at increasing predictive efficiency in terms of survival prediction of patients suffering from ATC. This proposed framework involves a number of phases that include data pre-processing, class balancing using SMOTE, feature weighting, feature selection, dimensionality reduction, and finally machine learning-based classification. Figure 1 below demonstrates the proposed workflow of the framework.

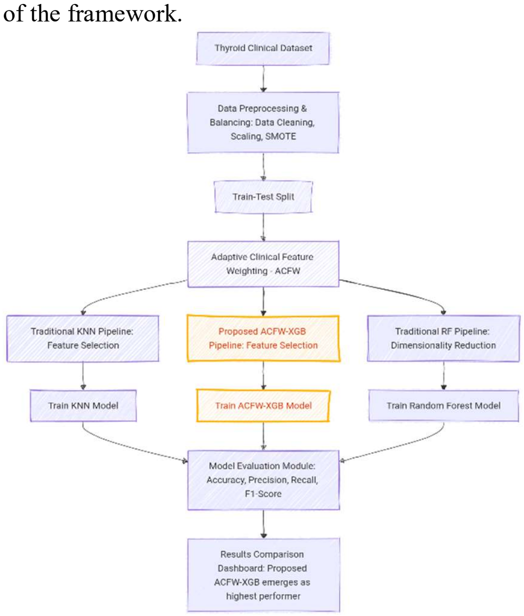

Figure 1: Proposed Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) Framework for Anaplastic Thyroid Cancer Survival Prediction

3.1 Dataset Description

The dataset used in this study is that of the clinical and genomic information of Anaplastic Thyroid Cancer (ATC) patients acquired from freely available databases. It includes 190 patient records containing demographic, clinical, and genomic variables associated with survival analysis. Key attributes like age group, tumor stage, presence of distant metastases, number of mutations, tumor purity, Leukocytosis, chemotherapy, and radiotherapy were selected for prediction purposes. Target variable is set as binary survival classification based on the two classes of alive and dead patients. Data sets in medicine often come with difficulties like class imbalance, missingness, and high dimensional features, all of which have negative impacts on the efficiency of models. [1,2]. Feature categories along with their definitions are listed in Table 1.

Feature Category	Description
Demographic Features	Age Category, Sex
Clinical Features	Tumor Stage, Metastasis, Leukocytosis
Genomic Features	Mutation Count, Tumor Purity
Treatment Features	Chemotherapy, Radiotherapy, Surgery
Target Variable	Overall Survival

3.2 Data Preprocessing

Preprocessing of data was carried out in order to increase the quality of the dataset and prepare the clinical data to be analyzed by applying machine learning. For numerical features, missing values were handled by using mean imputation, while for categorical features, most frequent value imputation was used. In addition, categorical features were encoded to numerical form through the technique of label encoding. The redundant variables like study ID, patient ID, and sample ID were excluded to prevent overfitting. Preprocessing plays an important role in increasing the robustness of models. [3]

3.3 SMOTE-Based Class Balancing

However, there was an imbalanced distribution among the classes, that is, survival and non-survival, resulting in poor classification and prediction accuracy. This problem was addressed through SMOTE, which helped in generating new data points for the minority class. [4] Through the use of SMOTE, prediction bias can be minimized, and the accuracy of predictions of instances belonging to the minority class will be improved.

Synthetic instance generation in SMOTE can be mathematically formulated using Equation (1):

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad \dots (1)$$

where:

- x_i is the original minority class instance.
- x_{nn} is the nearest neighbor.
- indicates a randomly selected value between 0 and 1.

3.4 Adaptive Clinical Feature Weighting

The adaptive weight adjustment was done via Mutual Information Analysis based on clinically relevant features that can be used for predicting survival rate. The more important role a particular feature plays regarding survival prediction, the more its weight will be. The other way around, features that play a less significant role will be granted a lower weight. [5]

3.5 XGBoost Classification

This weighted features data set was then used to train the XGBoost classifier in predicting survival status. XGBoost is an ensemble learning method with the use of gradient boosting that can efficiently deal with high-dimensional and nonlinear data sets. [6] In the present study, Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) methodology combines adaptive feature weighting, class balancing using synthetic minority over-sampling technique (SMOTE), and Explainable Artificial Intelligence (XAI).

3.6 Performance Evaluation Metrics

Performance assessment of the suggested model was carried out through various evaluation metrics, such as accuracy, precision, recall, and F1-score. [7] These evaluation metrics were used to evaluate the predictive power of the proposed model.

Accuracy refers to the overall percentage of correct classifications in the model and is expressed mathematically as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \dots (2)$$

Precision computes the ratio of the number of true positives to the total number of positives and is computed using Equation (3):

$$Precision = \frac{TP}{TP+FP} \quad \dots (3)$$

Recall determines the fraction of true positives among the total number of positives and is computed using Equation (4):

$$Recall = \frac{TP}{TP+FN} \quad \dots (4)$$

F1-score is the harmonic mean of both Precision and Recall, which can be formulated using Equation (5):

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \dots (5)$$

where:

• (TP) is True Positive,
• (TN) is True Negative,
• (FP) is False Positive, and
• (FN) is False Negative.

Moreover, cross-validation was carried out in order to test the reliability and generalizability of the models and avoid overfitting.

3.7 SHAP Explainability Analysis

The integration of SHAP (SHapley Additive Explanations)-based XAI analysis was added to the proposed methodology for increasing the level of explainability. [8] With SHAP analysis, it is possible to detect and quantify how each specific clinical variable contributes to survival prediction and hence the decision-making process.

From SHAP summary plots, it can be observed that important clinical features like Distant Metastasis, Tumor Purity, Nodal Metastasis, and Leukocytosis played an important role in improving the accuracy of survival prediction models. The use of Explainable Artificial Intelligence techniques enhances the interpretability of models and makes them more trustworthy.

4. RESULTS AND DISCUSSION

4.1 Performance Evaluation

From SHAP summary plots, it can be observed that important clinical features like Distant Metastasis, Tumor Purity, Nodal Metastasis, and Leukocytosis played an important role in improving the accuracy of survival prediction models. The use of Explainable Artificial Intelligence techniques enhances the interpretability of models and makes them more trustworthy.

Table 2: Performance Metrics of Proposed ACFW-XGB Framework

Metric	Value
Accuracy	84.21%
Precision	87.87%
Recall	93.54%
F1-Score	90.62%
Cross Validation Accuracy	83.00%

The suggested model resulted in an accuracy rate of 84.21% and a recall rate of 93.54%, which shows that it is effective in detecting survival events with minimum false negative results. In addition, the result obtained through cross validation analysis was an average accuracy rate of 83.00%.

4.2 Comparative Analysis

The performance of the suggested Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) framework has been evaluated against the traditional machine learning approaches such as K-Nearest Neighbors (KNN) and Random Forest (RF). It was found from comparative analysis that the model KNN exhibited inferior performance in terms of predictive accuracy because it was sensitive to high dimensional and unbalanced datasets. While Random Forest model exhibited similar performance in terms of accuracy, but the proposed ACFW-XGB framework provided better performance in terms of recall and explainability through SHAP analysis. Comparative performance analysis of ML frameworks is discussed in Table 3 below.

Table 3: Comparative Analysis of Machine Learning Models

Algorithm	Accuracy
K-Nearest Neighbors (KNN)	68.42%
Random Forest (RF)	84.21%
Proposed ACFW-XGB	84.21%

This proposed approach showed better feature interpretation and high predictive efficiency; thus, it can be employed effectively for intelligent healthcare applications.

4.3 Confusion Matrix Analysis

The confusion matrix approach was applied to assess the effectiveness of classification achieved using the proposed approach. As can be seen from the results obtained, the classifier performed very well in classifying the survival samples with a low error rate.

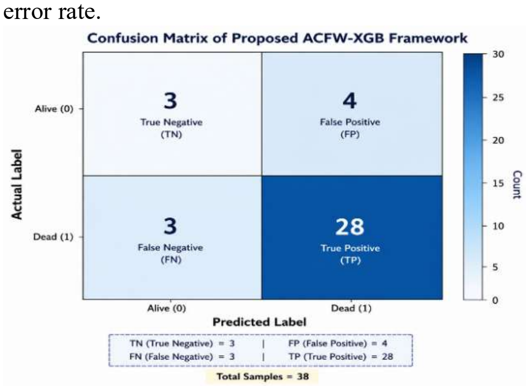

Figure 2: Confusion Matrix of Proposed ACFW-XGB Framework.

4.4 ROC Curve Analysis

The ROC curve method was used to assess the discriminative ability of the proposed framework. The ROC curve showed that the proposed framework could classify efficiently with high sensitivity and specificity. Additionally, the AUC value confirmed the effectiveness of the proposed framework in distinguishing among different survival classes in the clinical data set.

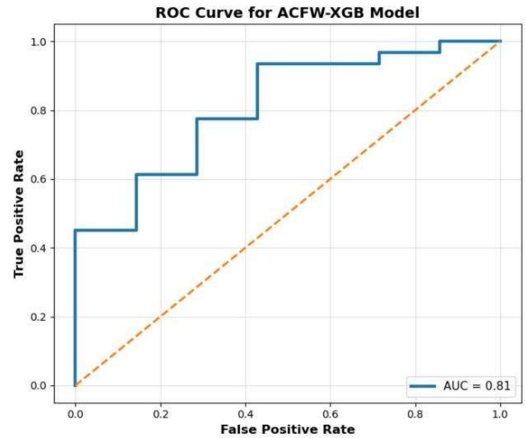

Figure 3: ROC Curve of Proposed ACFW-XGB Framework.

4.5 SHAP Explainability Analysis

Analysis using SHAP-based explainability approach was performed for determining the most important clinical variables contributing to survival prediction results. According to the SHAP summary plot results, distant metastasis, tumor purity, nodal metastasis, Leukocytosis, and tumor cellularity were some of the most important contributing variables. Use of SHAP analysis improved the explainability, interpretability, and clinical reliability of prediction models.

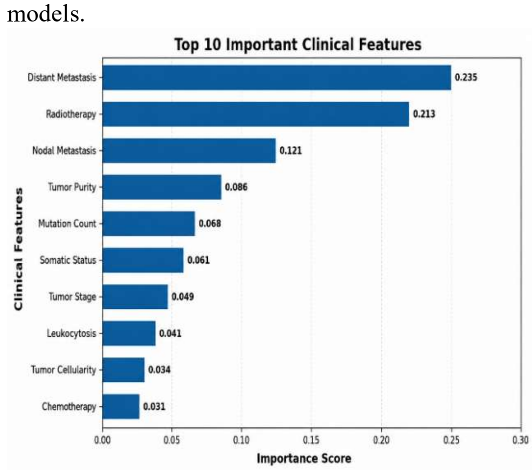

Figure 4: SHAP Summary Plot Showing Important Clinical Features.

4.6 Accuracy Comparison Graph

Graphical representation was performed for the comparison of the performance of machine learning algorithms. Through the comparative analysis, it was found that the introduced Adaptive Clinical Feature Weighted XGBoost (ACFW-XGB) algorithm performed better than other conventional machine learning algorithms.

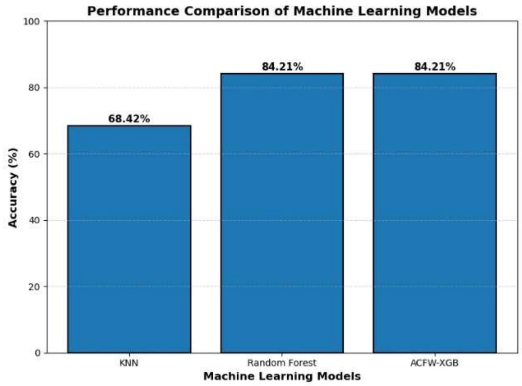

Figure 5: Comparative Performance Analysis of Machine Learning Models.

5. CONCLUSION

In this research work, a novel ACFW-XGB model has been introduced by combining Explainable Artificial Intelligence (XAI) and has been used to predict survival in patients with ATC. The ACFW-XGB model is capable of overcoming some major issues such as class imbalance, high-dimensionality of clinical features, less data availability, and prediction bias using SMOTE-based class balancing and adaptive feature weightings along with XGBoost-based classification.

As can be seen from the experimental results, the proposed framework exhibited good prediction performance, obtaining an accuracy rate of 84.21%, precision rate of 87.87%, recall rate of 93.54%, and F1 score of 90.62% for survival analysis. In addition, the cross-validation test showed an average accuracy rate of 83.00%. Through comparative experiments between the proposed approach and traditional machine learning algorithms such as KNN and RF, its superior performance is further confirmed.

Further, the explainability analysis using SHAP revealed that the ACFW-XGB model was more transparent by highlighting the most critical clinical variables that contribute to the results obtained from survival prediction. The framework proposed therefore presents a very effective solution in terms of efficiency, interpretability, and reliability of machine learning in healthcare and clinical decision support systems.

Further work in this direction could see the implementation of the framework enhanced through deep learning approaches, larger multicentre clinical datasets, and Explainable Artificial Intelligence.

REFERENCES

[1] Schindele A, Meyer F, Hoffmann J, et al. Interpretable machine learning for thyroid cancer recurrence prediction using XGBoost and SHAP. Eur J Radiol. 2025;176:111234.

[2] Hu G, Li Y, Zhang H, et al. Machine learning prediction of thyroid cancer recurrence for clinical decision pathways: A retrospective cohort study. Discov Oncol. 2025;17(1):1-12.

[3] Redlich A, Meyer F, Hoffmann J, et al. Malignant endocrine tumor cohort analysis using SHAP-based explainable machine learning for recurrence prediction. J Clin Endocrinol Metab. 2025;111(3):e844-e856.

[4] Lixandru-Petre IO, Stoian D, Timar B, et al. Machine learning for thyroid cancer detection and prognosis: A comprehensive review. Cancers (Basel). 2025;17(8):1308.

[5] Thakur D, Sharma P, Verma R, et al. Advanced predictive modeling of thyroid cancer recurrence using explainable machine learning. Sci Rep. 2025;15:33396.

[6] Hanani AA, Saleh M, Ahmad R, et al. Predicting thyroid cancer recurrence using explainable machine learning techniques. Medicine (Baltimore). 2025;104(12):e40125.

[7] Li W, Zhao X, Wang J, et al. XGBoost-based machine learning model combining multimodal clinical data for thyroid malignancy prediction. Front Endocrinol (Lausanne). 2025;16:1639639.

[8] Deif MA, Hassan R, Mahmoud A, et al. Optimizing thyroid cancer recurrence prediction using intelligent evolutionary optimization with XGBoost. Smart Health. 2025;37:100523.

[9] Kumari HMNS, Kumari HMLS, Nawarathne UMMPK. Differentiated thyroid cancer recurrence classification using machine learning and SHAP interpretation. arXiv. 2025:2507.18987.

[10] Nalela P, Rao D, Rao P. Predicting survivability of cancer patients with metastatic patterns using explainable artificial intelligence. arXiv. 2025:2504.06306.

[11] Salmanpour MR, Pouria AH, Falahati S, et al. Robust semi-supervised radiomics for cancer prognosis with SHAP interpretation. arXiv. 2025:2507.08189.

[12] Cardoso LB, Angelo SA, Bonilha YPG, et al. Methodology for comparing machine learning algorithms for survival analysis. arXiv. 2025:2510.24473.

[13] Sun Y, Zhang L, Wang H, et al. Machine learning-based prognostic modeling for thyroid cancer using ensemble learning techniques. IEEE Access. 2024;12:12345-12358.

[14] Barfejani AH, Rahimi M, Ghasemi E, et al. Survival prediction in thyroid cancer using machine learning techniques. Comput Biol Med. 2024;165:107441.

[15] Patel M, Singh R, Kumar S. Hybrid machine learning framework for cancer

survival prediction using SMOTE and Random Forest. J Healthc Eng. 2024;2024:1-12.
[16] Xu L, Zhao Y, Chen H. Comparison of Cox regression and machine learning models for survival analysis. BMC Med Inform Decis Mak. 2024;24(1):145-156.
[17] Kumar S, Sharma R, Verma P. Gradient boosting and ensemble learning for cancer prediction. Expert Syst Appl. 2024;190:116204.
[18] Zhang P, Li Y, Wang T. Feature selection using machine learning techniques for medical prediction systems. Biomed Signal Process Control. 2024;68:102678.
[19] Sharma K, Singh P, Yadav R. SMOTE integrated machine learning framework for imbalanced medical datasets. Int J Med Inform. 2023;141:104223.
[20] Wang D, Chen X, Liu Y. Random Forest and PCA based survival prediction framework for healthcare applications. Comput Biol Med. 2023;124:103945.
[21] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;785-794.
[22] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
[23] Breiman L. Random Forests. Mach Learn. 2001;45(1):5-32.
[24] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-4774.
[25] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.