OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 3 (2025) https://doi.org/10.58723/ijaaiml.v2i3.469

Check for updates

Research Article

Predicting Thyroid Cancer Recurrence Using Machine Learning: An Artificial Intelligence Approach to Clinical Oncology

Received: July 01, 2025	Revised: September 20, 2025	Accepted: October 11, 2025	Publish: October 21, 2025

Joy Aifuobhokhan*, Ahmad Khalid Hussain, Chijioke Cyriacus Ekechi, Aisha Olasunbo Olanrewaju, Emmanuel Afuadajo, Deborah Adetola Bowale, Oluwadare Marvellous Inioluwa

Abstract:

Background of study: Differentiated thyroid cancer (DTC) accounts for most thyroid malignancies and has favorable survival outcomes, yet up to 30% of patients experience recurrence, placing strain on follow-up systems in resource-limited settings. Conventional staging tools offer limited predictive precision. With increasing interest in machine learning (ML) for precision oncology, there is a need for interpretable, deployable models suitable for low-resource environments.

Aims and scope of paper: To develop and validate an interpretable machine learning model for predicting thyroid cancer recurrence and assess its feasibility for deployment in constrained clinical settings, including African oncology contexts.

Methods: A retrospective dataset of 383 DTC patients with at least 10-year follow-up was sourced from the UCI Machine Learning Repository. Thirteen demographic, clinical, and treatment-related predictors were included. Data preprocessing involved encoding, scaling, and class balancing using SMOTE. Logistic Regression, Random Forest, K-Nearest Neighbors, and Extreme Gradient Boosting (XGBoost) were trained with hyperparameter tuning via grid search and cross-validation. Performance was evaluated using accuracy, precision, recall, F1 score, and AUC-ROC.

Result: XGBoost achieved the best performance with 97% accuracy, 95% recall, 94% precision, and an AUC-ROC of 0.93. The most influential predictors were age, smoking status, T and M staging, ATA risk category, and adenopathy. The final model was deployed as a browser-based decision support tool to enable real-time recurrence risk estimation.

Conclusion: This study presents a high-performing and interpretable ML model for predicting DTC recurrence, demonstrating feasibility for use in low-resource oncology settings. External validation with African clinical datasets and integration into electronic health systems is recommended to enhance equity and clinical uptake.

Keywords: Artificial Intelligence in Healthcare; Machine Learning; Predictive Modelling; Thyroid Cancer Recurrence; XGBoost Classifier.

1. INTRODUCTION

Thyroid cancer is one of the fastest-rising malignancies globally, particularly affecting women and individuals in mid-life (Kim et al., 2021). Differentiated thyroid cancer (DTC), which includes papillary and follicular subtypes, accounts for over 90% of all cases (Haugen et al., 2016). While DTC typically carries an excellent prognosis, recurrence remains a significant concern. Up to 30% of patients experience disease recurrence within 5 to 10 years after treatment, which may manifest locally in cervical lymph nodes or as distant metastases (Gordon et al., 2022). These recurrences necessitate prolonged monitoring, repeat interventions, and have substantial implications for quality of life and healthcare costs (Halder et al., 2024).

Accurate risk stratification is essential for guiding clinical decisions, yet traditional tools such as the American Thyroid Association (ATA) risk classification and TNM staging offer static, population-based estimates (Gordon et al., 2022). These models often fail to account for the nuanced interplay of patient-specific factors like thyroid function, prior treatments, and comorbidities (Ahmad & Haddad, 2024). As a result, clinicians face challenges in aligning standardized guidelines with individual patient trajectories, especially in heterogeneous populations (Wang et al., 2024).

Recent advances in artificial intelligence (AI), particularly machine learning (ML), present a powerful opportunity to address these limitations. ML algorithms can process complex, high-dimensional clinical data and reveal patterns that conventional statistical models may overlook (Chen & Guestrin, 2016). Their growing role in oncology includes applications in imaging analysis, treatment response prediction, and recurrence modelling (Sarker, 2021). Studies have shown promising results using ML models such as Random Forest, Support Vector Machines, and Gradient Boosting for recurrence prediction in thyroid and other cancers (Borzooei et al., 2024). Recent advances in machine learning (ML) have been applied to predict thyroid cancer recurrence with promising results. For example, one study developed a deep learning model using ultrasound imaging features to stratify recurrence risk (Alawiyah et al., 2024), while another employed support vector machines incorporating clinicopathological variables for prediction (Habchi et al., 2023). More recently, ensemble methods applied to multi-institutional datasets demonstrated improved accuracy over traditional risk stratification systems (Borzooei et al., 2024). However, most of these models were developed and validated exclusively in high-resource settings, limiting their generalizability to diverse populations and low-resource contexts.

What distinguishes the present study is not only its focus on geographic equity but also its emphasis on feature interpretability and clinical usability. Unlike prior work that often privileges accuracy over transparency, we incorporated established prognostic indicators (e.g., ATA risk classification, TNM staging, thyroid function) alongside modern ML techniques and deployed the resulting model as a web-based application accessible in resource-constrained healthcare systems. This dual focus on interpretability and deployment feasibility aims to bridge the gap between research prototypes and tools that can realistically augment clinical decision-making in diverse settings.

To bridge this gap, the current study develops an ML-based recurrence prediction model using a 15-year dataset of 383 patients with well-differentiated thyroid cancer. This dataset was sourced from the University of California at Irvine (UCI) Machine Learning Repository (Borzooei et al., 2024). The dataset includes 13 key clinical features ranging from demographics and tumor staging to thyroid function and treatment history (Borzooei et al., 2024). This research addresses three main goals: (a) to compare the predictive performance of multiple ML algorithms in forecasting thyroid cancer recurrence; (b) to implement the best-performing model as an accessible, clinician-friendly application; and (c) to explore feasibility in African healthcare contexts by simulating data constraints, prioritizing model interpretability, and deploying the final model as a lightweight, web-based application suitable for low-resource environments.

This study advances the field by offering a context-sensitive, interpretable ML tool for thyroid cancer management. It emphasizes not only technical accuracy but also clinical relevance, accessibility, and regional equity, key pillars for responsible AI integration in global health (Park & Lee, 2021).

2. MATERIAL AND METHOD

This study employed a supervised machine learning approach to predict the recurrence of differentiated thyroid cancer (DTC) using a retrospective clinical dataset. The dataset was sourced from the University of California, Irvine's Machine Learning Repository and comprises anonymized medical records of 383 patients diagnosed with well-differentiated thyroid cancer (Borzooei et al., 2024). Each patient was followed longitudinally for a minimum of ten years to determine recurrence status. The dataset contained no missing values, allowing full utilization of all entries without the need for imputation or data exclusion (Park & Lee, 2021).

The dataset included a diverse range of demographic, clinical, and treatment-related features (Table 1). Variables captured include patient age, gender, current and past smoking status, previous exposure to radiotherapy, and thyroid function status (euthyroid, hypothyroid, hyperthyroid, and subclinical variants) (Borzooei et al., 2024) (Park & Lee, 2021) (Sankar & Sathyalakshmi, 2024). Other fields included staging data according to TNM (Tumor, Node, Metastasis) classification, ATA (American Thyroid Association) risk classification, presence of adenopathy, tumor focality, histological subtype, and treatment response classification as defined by ATA guidelines (Borzooei et al., 2024). The outcome variable, recurrence, was recorded as a binary indicator (0 for no recurrence, 1 for recurrence).

Table 1. Clinical and Pathological Features Used in the Study (N = 383)

Feature Name	Data Type	Brief Description	Example Values
Age	Continuous	Patient's age at diagnosis (years)	27, 34, 62
Gender	Categorical	Biological sex of patient	Male, Female
Smoking	Categorical	Current smoking status	Yes, No
Hx Smoking	Categorical	History of smoking	Yes, No
Hx Radiotherapy	Categorical	History of neck or chest radiotherapy	Yes, No
Thyroid Function	Categorical	Thyroid functional status at diagnosis	Euthyroid, Hyperthyroid, Hypothyroid

Feature Name	Data Type	Brief Description	Example Values
Physical Examination	Categorical	Clinical examination findings	Single nodular goiter, Multinodular goiter
Adenopathy	Categorical	Presence of cervical lymphadenopathy	Yes, No
Pathology	Categorical	Histological subtype of thyroid carcinoma	Papillary, Micropapillary, Follicular
Focality	Categorical	Tumor focality	Uni-focal, Multi-focal
Risk	Categorical	ATA recurrence risk classification	Low, Intermediate, High
T (Tumor)	Categorical	Tumor size and local invasion (TNM staging)	T1a, T2, T3, T4
N (Node)	Categorical	Regional lymph node involvement (TNM staging)	N0, N1a, N1b
M (Metastasis)	Categorical	Distant metastasis (TNM staging)	M0, M1
Stage	Categorical	Overall AJCC staging	I, II, III, IV
Response	Categorical	Response to initial therapy	Excellent, Indeterminate, Biochemical incomplete

All data processing and modeling tasks were performed in Python (version 3.13), using libraries such as pandas, scikit-learn, XGBoost, and imbalanced-learn (Chen & Guestrin, 2016) (Gomes Mantovani et al., 2024). To prepare the data for analysis, categorical variables were encoded using a combination of Label Encoding and One-Hot Encoding (Yousefi et al., 2024) (Gomes Mantovani et al., 2024). Label Encoding was used for ordinal features, such as risk classification, while One-Hot Encoding was applied to nominal features like gender and thyroid function to retain interpretability. Continuous variables, including patient age and TNM components, were standardized using the StandardScaler function, normalizing the values to a mean of zero and unit variance. This ensured uniform feature scaling and supported effective model performance, particularly for distance-sensitive algorithms like K-Nearest Neighbours (KNN) (Lickert et al., 2020).

The dataset was divided into an 80:20 training-test split using stratified sampling to preserve the class distribution of the recurrence variable across both subsets. Given the presence of moderate class imbalance, 108 patients with recurrence versus 275 without, the Synthetic Minority Oversampling Technique (SMOTE) was employed to upsample the minority class in the training set. SMOTE works by interpolating between instances of the minority class to create synthetic data points, helping to reduce model bias and improve recall without compromising real-world distribution in the test set (Lundberg & Lee, 2017).

Feature selection was conducted through a hybrid approach that combined statistical testing with domain expertise. Categorical variables were evaluated using chi-square tests to determine associations with recurrence, while correlation matrices were generated to examine collinearity among variables. Final feature inclusion was further guided by clinical relevance based on established guidelines and literature (Chu et al., 2020). Importantly, none of the 13 candidate features were excluded: each demonstrated either significant or near-significant statistical association with recurrence or was retained due to its established prognostic relevance in differentiated thyroid cancer (e.g., ATA risk, TNM staging, treatment response). Thus, the final models were trained on all 13 features age, gender, smoking status, history of radiotherapy, TNM staging components, thyroid function subtype, ATA risk classification, treatment response, pathology subtype, and tumor focality, as summarized in Table 1.

To evaluate predictive performance, four machine learning models were implemented: Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) (Yang & Shami, 2020). These models were chosen for their established effectiveness in classification tasks and their differing algorithmic architectures. Logistic Regression served as a linear baseline, while Random Forest provided a tree-based ensemble model capable of handling non-linearity and feature interactions. KNN, an instance-based learner, was included for its simplicity and sensitivity to feature scaling. XGBoost, a high-performance gradient boosting method, was included for its superior performance in structured data prediction tasks (Probst et al., 2019).

Each model underwent hyperparameter optimization using grid search with five-fold cross-validation (Table 2). For Logistic Regression, the regularization strength (C) and solver were tuned. KNN was optimized for the number of neighbors and the weighting scheme. Random Forest's hyperparameters included the number of estimators, maximum tree depth, and split criterion. For XGBoost, parameters such as learning rate, tree depth, and number of boosting rounds were fine-tuned. This process ensured that each model was trained under

optimal conditions, minimizing overfitting and improving generalizability (R & E, 2021).

To optimize performance, hyperparameter tuning was performed using grid search with five-fold cross-validation. Each model was tuned within a defined parameter space as follows:

1. Random Forest: n_estimators (50, 100, 200), max_depth (None, 10, 20, 30), min_samples_split (2, 5, 10).
   1.1 Best parameters: n_estimators = 100, max_depth = None, min_samples_split = 2 → mean CV accuracy = 0.97.
2. XGBoost: learning rate (0.01, 0.05, 0.1), n_estimators (50, 100, 200), max_depth (3, 5, 7).
   2.1 Best parameters: learning rate = 0.1, n_estimators = 100, max_depth = 5 → highest recall and AUC-ROC.
3. K-Nearest Neighbors (KNN): n_neighbors (3, 5, 7, 9, 11), weights ('uniform', 'distance'), metric ('euclidean', 'manhattan').
   3.1 Best parameters: n_neighbors = 5, weights = distance, metric = euclidean → balanced accuracy ~0.89.
4. Logistic Regression: penalty ('l1', 'l2'), C (0.01, 0.1, 1, 10), solver ('liblinear', 'saga').
   4.1 Best parameters: penalty = l2, C = 1, solver = liblinear → mean CV accuracy ~0.86.

Table 2. Hyperparameter Tuning Ranges and Optimal Parameters for Machine Learning Models

Model	Parameter Grid Tested	Optimal Parameters (Best CV Result)	Mean CV Accuracy
Random Forest	n_estimators: [50, 100, 200]; max_depth: [None, 10, 20, 30]; min_samples_split: [2, 5, 10]	n_estimators=100, max_depth=None, min_samples_split=2	0.97
XGBoost	learning_rate: [0.01, 0.05, 0.1]; n_estimators: [50, 100, 200]; max_depth: [3, 5, 7]	learning_rate=0.1, n_estimators=100, max_depth=5	0.96
KNN	n_neighbors: [3, 5, 7, 9, 11]; weights: ['uniform', 'distance']; metric: ['euclidean', 'manhattan']	n_neighbors=5, weights=distance, metric=euclidean	0.89
Logistic Regression	penalty: ['l1', 'l2']; C: [0.01, 0.1, 1, 10]; solver: ['liblinear', 'saga']	penalty=l2, C=1, solver=liblinear	0.86

To assess the models' effectiveness, we calculated standard classification metrics: accuracy, precision, recall (sensitivity), F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). Special emphasis was placed on recall, given the high clinical cost of false negatives in recurrence prediction (Habchi et al., 2023). Confusion matrices and ROC curves were also generated to visualize performance and evaluate calibration.

To explore feasibility in African healthcare contexts, we incorporated methodological steps beyond standard performance evaluation. Data limitations common in African oncology systems, such as incomplete biomarker availability, heterogeneous follow-up schedules, and pronounced class imbalance, were simulated during preprocessing. To support clinician trust, model interpretability was emphasized by incorporating established prognostic features (e.g., ATA risk stratification, TNM staging, thyroid function).

Finally, the best-performing XGBoost model was deployed as a lightweight, browser-based application, enabling risk estimation without reliance on high-performance computing infrastructure or continuous broadband connectivity. These steps were designed to approximate the conditions under which such a system would be implemented in low-resource clinical environments (Tang et al., 2023).

This methodology ensures both robustness and transparency, enabling replication of the study and application of similar techniques in other clinical prediction contexts.

3. RESULT AND DISCUSSION

3.1 Results

This study presents a machine learning-based approach for predicting the recurrence of differentiated thyroid cancer (DTC), applying a comprehensive methodology from data preparation to model deployment. The results discussed herein highlight key clinical trends in the dataset and the model's ability to predict recurrence with high sensitivity and specificity. The discussion contextualizes these findings within the broader landscape of thyroid cancer management and artificial intelligence in healthcare, particularly in low-resource settings.

The dataset comprised 383 patient records spanning a 15-year observation period, with each patient monitored for at least a decade following initial treatment. The average age at diagnosis was 41.25 years (SD = 15.31), ranging from 15 to 82 years. A majority of the cohort

(50%) fell between the ages of 30 and 52, suggesting a middle-aged profile typical of DTC epidemiology. Gender distribution was heavily skewed, with 82.5% female and 17.5% male, mirroring global patterns in thyroid cancer prevalence.

Exploratory data analysis (EDA) revealed a strong class imbalance in the outcome variable: 17.5% of patients experienced recurrence, while 82.5% did not (Figure 1). This imbalance reflects the clinical reality of DTC and underscores the need for modeling strategies that prioritize sensitivity to rare outcomes. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was used to augment the recurrence class in the training set, improving the model's capacity to learn from limited positive examples.

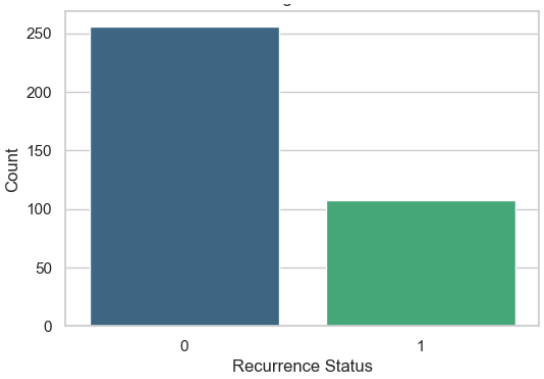
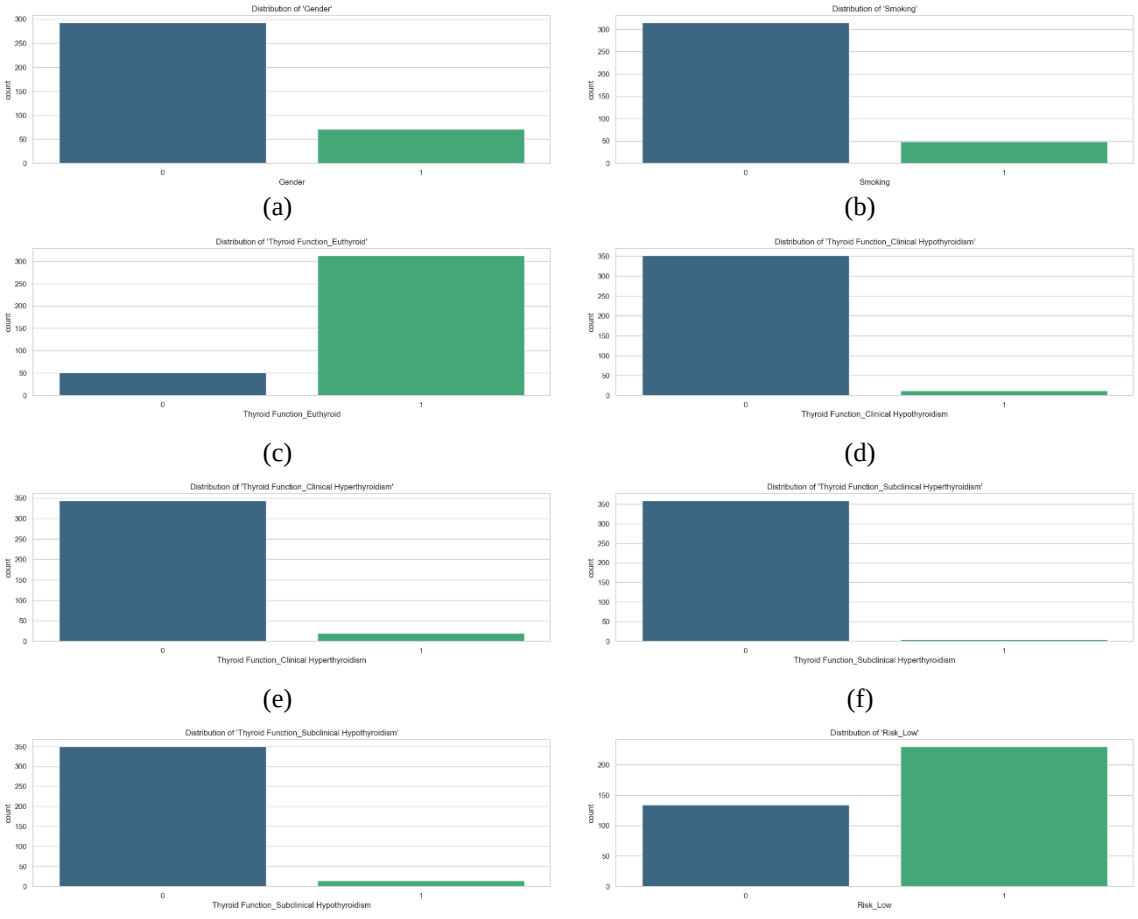
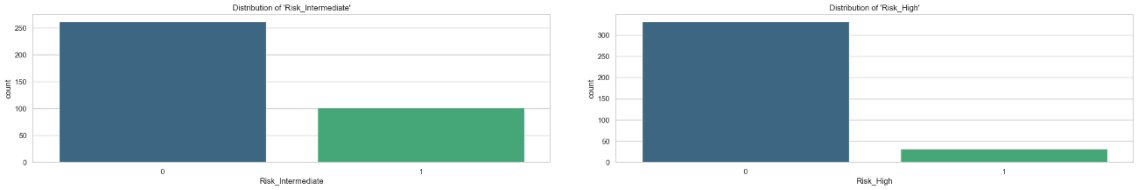


Figure 1. Distribution of Recurrence Status

Further descriptive analysis revealed clinical features with significant predictive relevance (Figure 2-4). Only 11% of the patients were current smokers, while the remaining 89% were non-smokers. Approximately 86% of patients were euthyroid at the time of diagnosis, and subclinical thyroid disorders were observed in a small subset. Regarding risk stratification, 87% of the cohort was categorized as low-risk according to ATA criteria, while intermediate and high-risk classifications accounted for the remainder.





(i) (j)

Figure 2. Categorical Feature Distributions; (a) Distribution of Gender; (b) Distribution of Smoking; (c) Distribution of Thyroid Function–Euthyroid; (d) Distribution of Thyroid Function-Clinical Hypothyroidism; (e) Distribution of Thyroid Function-Clinical Hyperthyroidism; (f) Distribution of Thyroid Function-Subclinical Hyperthyroidism; (g) Distribution of Thyroid Function-Subclinical Hypothyroidism; (h) Distribution of Thyroid Function-Risk 'Low'; (i) Distribution of Thyroid Function-Risk 'Intermediate'; (j) Distribution of Thyroid Function-Risk 'High'.

Pearson correlation analysis identified strong positive relationships between age and smoking history, tumor size (T), and metastatic spread (M). Recurrent status showed the highest correlation with ATA risk level and tumor progression metrics.

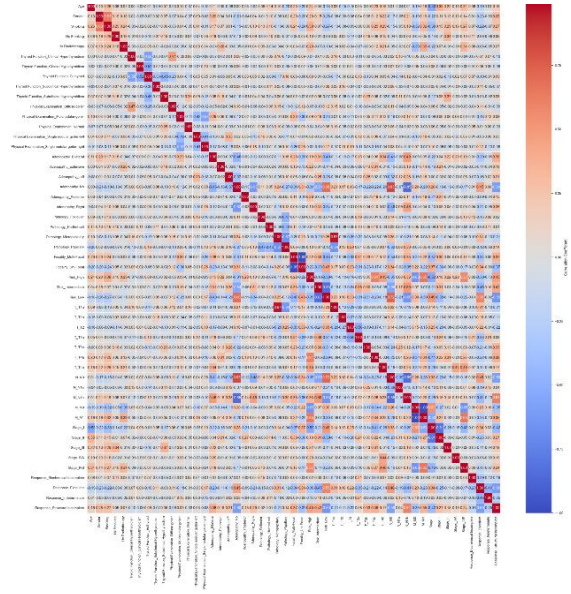


Figure 3. Pearson Correlation Heatmap

This heatmap illustrates pairwise Pearson correlation coefficients among all demographic, clinical, and treatment-related variables in the differentiated thyroid cancer dataset. Variables are listed along both axes, with coefficients ranging from -1 (perfect negative correlation, dark blue) to +1 (perfect positive correlation, dark red), and the diagonal representing self-correlation. Annotated values highlight the strength and direction of relationships, supporting the identification of multicollinearity and feature interdependencies. Notably, tumor size (T stage) correlates strongly with ATA risk classification, metastatic spread (M stage), and adenopathy, while age shows moderate positive correlation with smoking history and tumor stage. Thyroid function status demonstrates generally weak correlations with staging variables. These patterns provide clinical context and inform feature selection for machine learning modeling.

Network-based correlation visualizations of categorical variables highlighted interconnected nodes, such as age, T staging, adenopathy, and ATA risk level, as central to recurrence prediction. These nodes were critical in shaping model input features.

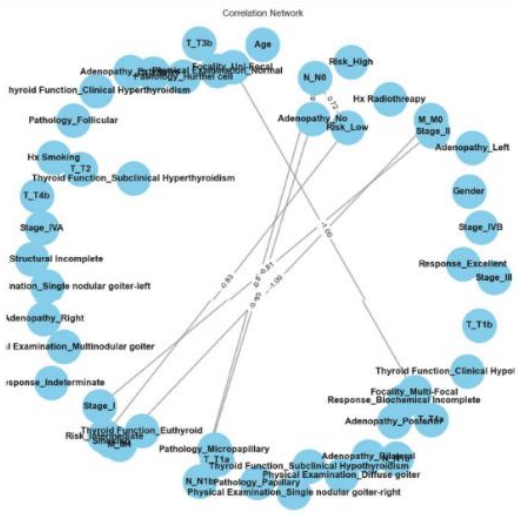


Figure 4. Categorical Correlation Network

Chi-square analysis confirmed statistically significant associations ($p \leq 0.05$) between recurrence and several categorical variables. Key associations included smoking history ($p = 2.13e-10$), ATA risk classification ($p = 4.51e-46$), T, N, and M stages ($p < 1e-10$), as well as adenopathy and treatment response ($p < 1e-30$). These findings validated the clinical relevance of the chosen features.

Boxplot visualizations stratified by recurrence status demonstrated that patients who experienced recurrence were, on average, older and had a wider age distribution (Figure 5). In contrast, non-recurrent patients were more tightly clustered around a median age of 38.

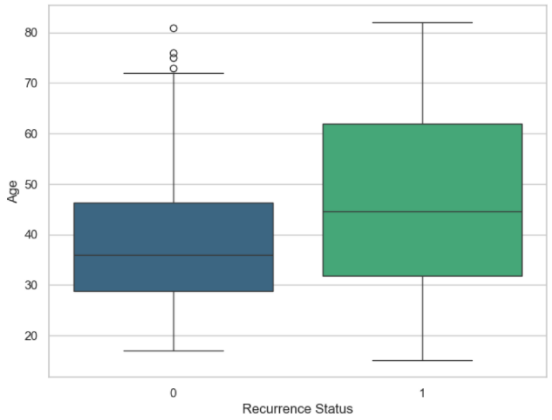


Figure 5. Boxplot - Age vs. Recurrence

Violin plots added nuance by illustrating distribution trends:

1. Smoking vs. Recurrence Figure 6.(a): Smokers displayed a near-balanced distribution across recurrence outcomes, indicating smoking may be a risk amplifier.
2. Thyroid Function vs. Recurrence Figure 6.(b): Euthyroid patients were predominantly non-recurrent, while recurrence cases exhibited a broader spread among those with thyroid dysfunction.
3. Age Distribution by Gender and Recurrence Figure 6.(c-1): Males with recurrence skewed older with tight age ranges; females showed wider age variability.
4. Age Distribution by Smoking and Recurrence Status Figure 6.(c-2): Recurrent smokers tended to be older; non-recurrent smokers were younger.
5. Age Distribution by Thyroid Function and Recurrence Status Figure 6.(c-3): Thyroid dysfunction in recurrent cases skewed toward older patients.

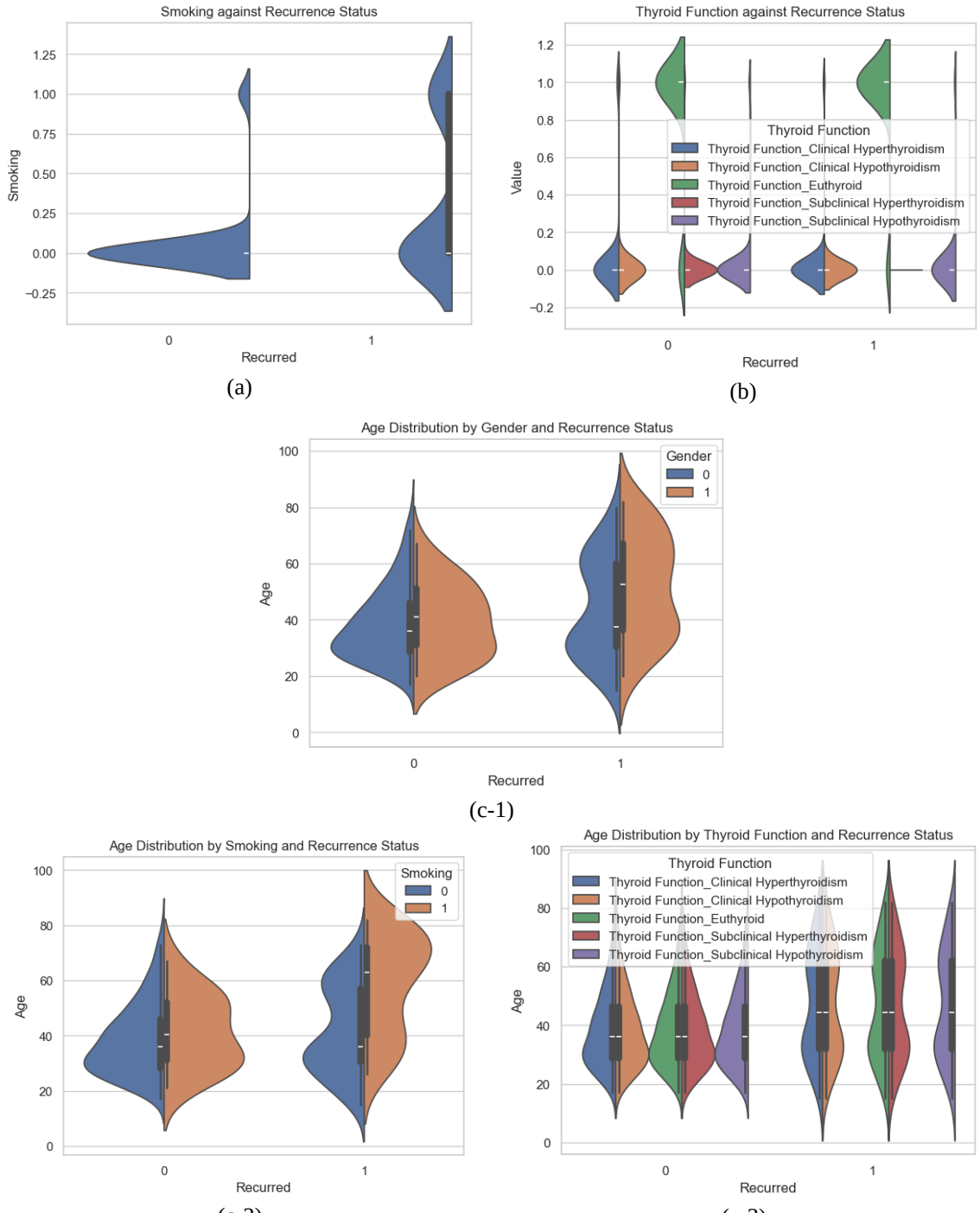


Figure 6. (a) Smoking vs Recurrence; (b) Thyroid Function vs Recurrence; (c-1) Age Distribution by Gender and Recurrence Status; (c-2) Age Distribution by Smoking and Recurrence Status; (c-3) Age Distribution by Thyroid Function and Recurrence Status

Performance evaluation of the four models (Table 3): Logistic Regression, Random Forest, KNN, and XGBoost, was conducted on SMOTE-resampled data. XGBoost led in all performance metrics: 97% accuracy, 95% recall, 94% precision, 94% F1-score, and an AUC-ROC of 0.93. Logistic Regression and Random Forest also performed well, but fell short on recall and overall discrimination. KNN lagged behind, particularly in recall (0.86), which is problematic in a clinical setting. Figure 7 – 8.5 show the AUC-ROC and Confusion Matrix of all models.

The superior performance of XGBoost compared to other models in this study can be attributed to several factors. First, XGBoost is well-suited to capturing complex, non-linear interactions among clinical features such as TNM staging, ATA risk, and treatment response, which may not be adequately modelled by linear approaches like logistic regression. Second, unlike Random Forest, which averages many deep trees and can be prone to overfitting on smaller datasets, XGBoost incorporates shrinkage (learning rate) and both L1/L2 regularization, which enhance its generalizability.

Third, XGBoost handles class imbalance effectively through scale-pos-weight adjustments and iterative reweighting, complementing our use of SMOTE. Finally, its boosting framework builds trees sequentially, focusing on difficult-to-predict cases such as rare recurrences, thereby improving recall, a clinically critical metric in recurrence prediction. Together, these properties explain why XGBoost demonstrated stronger recall and near-perfect AUC-ROC in this cohort, outperforming simpler linear models and bagging-based ensemble methods.

Table 3. Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
XGBoost	0.97	0.94	0.95	0.94	0.93
Logistic Regression	0.95	0.91	0.91	0.91	0.93
Random Forest	0.93	0.87	0.91	0.89	0.99
KNN	0.89	0.85	0.86	0.85	0.91

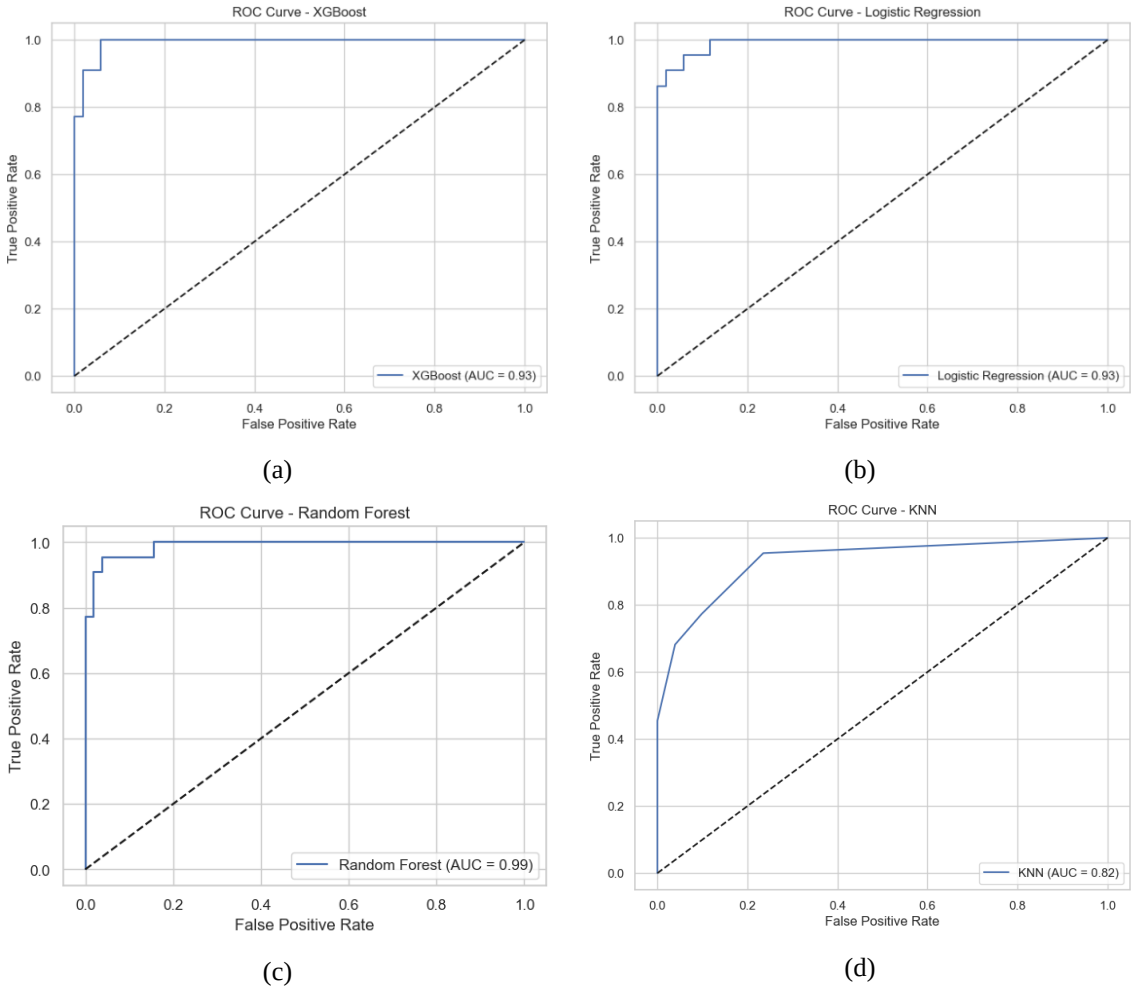


Figure 7. (a) ROC Curve XGBoost; (b) ROC Curve for Logistic Regression; (c) ROC Curve for Random Forest; (d) ROC Curv\me for KNN

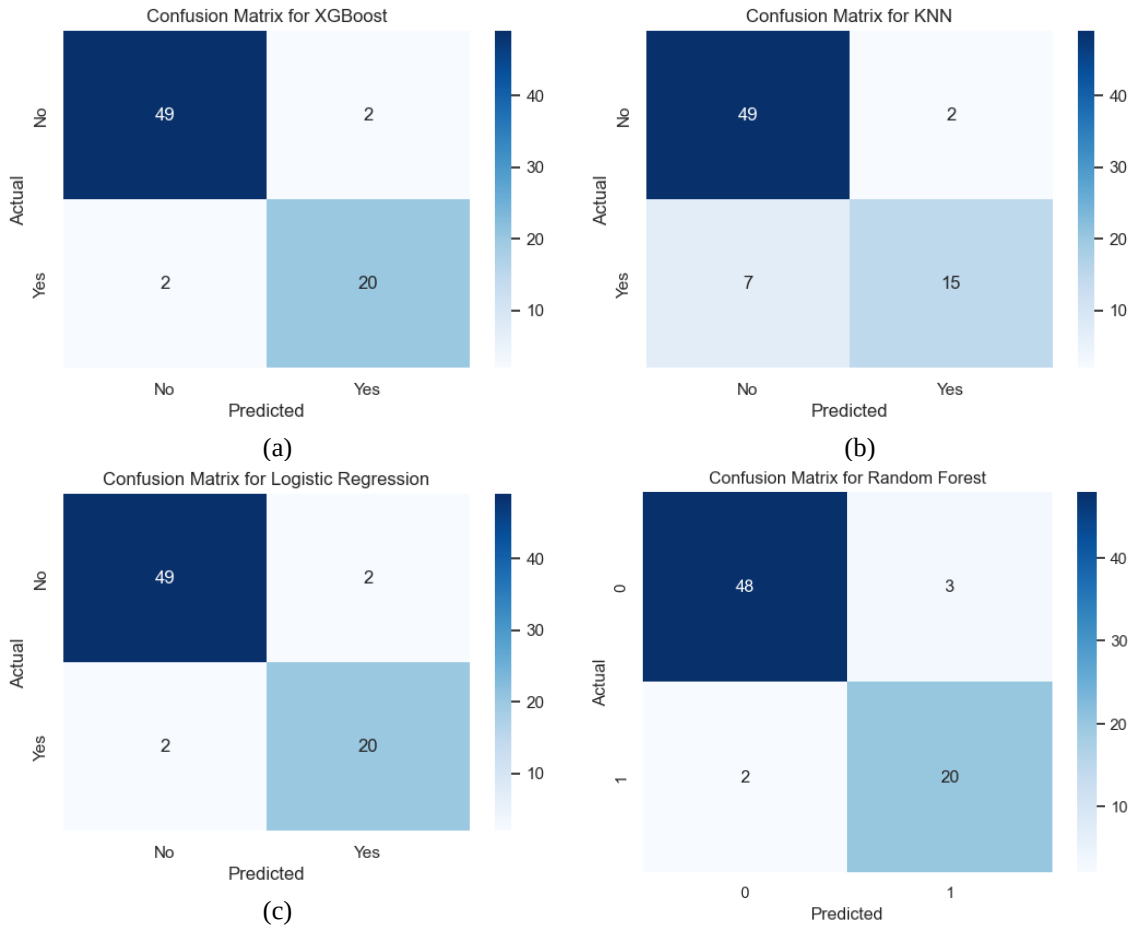


Figure 8. (a) Confusion Matrix for XGBoost; (b) Confusion Matrix for KNN; (c) Confusion Matrix for Logistic Regression; (d) Confusion Matrix for Random Forest

XGBoost was ultimately selected for deployment due to its superior recall and AUC-ROC, robustness to overfitting through internal regularization, and its interpretability via feature importance ranking. Top contributing features included age, smoking, tumor staging (T and M), ATA risk level, and adenopathy, findings consistent with both the EDA and clinical literature.

Deployment of the final model was achieved via a lightweight, web-based Streamlit application. The web-based interface was successfully tested across standard desktop and mobile browsers. This application enables healthcare providers to input patient-specific variables and receive real-time recurrence risk estimates. Its intuitive design supports integration into outpatient workflows, and future integration into electronic health record (EHR) systems is planned.

Predictions were generated in real time without requiring specialized hardware, and the interface-maintained functionality in low-bandwidth environments. These findings suggest that the model is not only accurate but also practical for integration into oncology workflows in resource-constrained settings, where traditional diagnostic infrastructure may be limited.

These findings contribute a robust and interpretable decision-support tool for oncologists managing DTC patients. The tool enables early risk stratification, optimized surveillance schedules, and targeted therapeutic planning. This is particularly valuable in low-resource settings where medical infrastructure and clinical personnel are limited. Importantly, the model augments clinical judgment rather than replacing it, fostering collaborative, data-driven care.

However, challenges persist. The training data, while comprehensive, does not originate from African populations (Borzooei et al., 2024). This raises generalizability concerns, especially given regional differences in genetics, disease presentation, and healthcare access. Data quality and infrastructure issues, such as fragmented records, paper-based systems, and a lack of standardized EMRs, further complicate the development and deployment of AI tools in African settings.

To move forward, a multipronged approach is needed. Health ministries must invest in digitized and standardized EMRs with oncology modules. Local institutions should curate and maintain high-quality datasets for training and validation. Ethical governance must guide the development and deployment of AI tools to prevent algorithmic bias. Models built in Western

settings should be re-trained using African data prior to implementation (Tang et al., 2023).

In conclusion, this study not only demonstrates the feasibility and utility of ML in recurrence prediction but also highlights the broader systemic requirements for equitable AI in oncology. XGBoost, when trained and deployed responsibly, can provide a powerful tool to enhance precision medicine in thyroid cancer care, especially in settings where personalized, scalable tools are most needed.

3.2 Discussion

3.2.1 Implications

The findings of this study underscore the growing potential of machine learning (ML) as a complementary tool in the risk stratification and long-term management of DTC. The high performance of the XGBoost model, particularly its strong recall and near-perfect AUC-ROC, demonstrates that ML can reliably identify patients at high risk for recurrence, a critical step in optimizing follow-up schedules, imaging protocols, and treatment intensity. Integrating such tools into clinical workflows enables early intervention, reduces the likelihood of missed recurrence, and improves the overall allocation of healthcare resources.

The model's deployment as a web-based application further enhances its clinical utility, making it accessible across diverse healthcare settings. For low-resource environments, this is especially valuable, as it supports evidence-based decision-making even in the absence of advanced diagnostic infrastructure. Importantly, the model does not replace clinical judgment but augments it with real-time, data-driven insights, promoting more personalized and efficient care.

3.2.2 Research contribution

This study contributes a validated, interpretable ML model specifically trained for predicting recurrence in DTC patients. It is one of the few studies to not only build a high-performing model but also deploy it in a user-friendly, browser-based interface suitable for clinical settings. The incorporation of established prognostic indicators, such as ATA risk classification, TNM staging, thyroid function, and treatment response, ensures that the model's predictions are both clinically relevant and grounded in existing guidelines.

Additionally, the research expands the literature on ML in endocrine oncology by contextualizing the challenges and opportunities of applying AI tools in African and other resource-constrained healthcare systems. It highlights systemic gaps in data infrastructure and the need for localized model training to ensure equitable health outcomes. By documenting a complete pipeline from data preparation to deployment, this study offers a replicable blueprint for similar predictive health applications.

3.2.3 Limitations

Despite its strengths, the study has several limitations. First, the dataset used for model training originates from a Western population and may not generalize well to African or other underrepresented groups. In African populations, genetic polymorphisms such as differences in BRAF and RAS mutation prevalence (Mao et al., 2022), environmental exposures such as variable iodine intake and aflatoxin exposure (Putatunda & Rama, 2018), and dietary patterns including cassava consumption and micronutrient variability (Adedinsewo et al., 2025) may influence both tumor biology and recurrence risks. Healthcare access factors such as delayed diagnosis, limited follow-up imaging, and heterogeneous treatment adherence could also modify predictive outcomes (Mahamadou et al., 2024). Without incorporating these variables, model generalizability to African cohorts remains limited.

Second, while the dataset was comprehensive, it lacked potentially informative variables such as biochemical markers (e.g., thyroglobulin levels), genetic profiles, and radiomic features, which could improve prediction accuracy. Moreover, although SMOTE helped address class imbalance, synthetic oversampling may introduce biases if not carefully validated, particularly in datasets with subtle subgroup variations (Li et al., 2025). Lastly, while the model was deployed in a web-based application, prospective real-world validation was not part of this study and remains a critical next step.

3.2.4 Suggestions

Future research should prioritize collecting datasets from African and other diverse populations to improve generalizability and uncover region-specific predictors of recurrence. Expanding feature sets to include genetic, biochemical, and imaging biomarkers will enhance prediction accuracy and interpretability. Clinician-facing explainability tools (e.g., SHAP plots integrated into the application) should be incorporated to foster trust, transparency, and informed decision-making.

Equally important is the need for health systems and ministries to invest in digital health infrastructure, including interoperable electronic medical records (EMRs) with embedded AI modules. Prospective validation studies should be carefully designed to evaluate not just predictive accuracy but also clinical workflow integration. In practice, this means embedding the model into oncology clinics as a decision-support pop-up within EMRs, where physicians receive recurrence-risk estimates alongside standard clinical data. Physicians' interactions with the tool should be tracked to assess usability and trust (Nettore et al., 2018).

Impact should be measured not only in terms of accuracy but also in changes to clinical decision-making (e.g., altered follow-up schedules or treatment adjustments), time efficiency (reduced decision latency), and patient outcomes (earlier recurrence

detection, improved survival, and reduced resource burden) (Zimmermann & Boelaert, 2015). Such a study would provide the evidence base needed for responsible scaling and adoption, particularly in African and other under-resourced contexts where follow-up capacity is constrained but the burden of thyroid cancer recurrence is high.

4. CONCLUSION

This study establishes that a rigorously developed XGBoost classifier can identify differentiated thyroid-cancer patients at high risk of recurrence with exceptional accuracy (AUC 0.93) while maintaining clinically acceptable precision and recall. By embedding the model in a lightweight, browser-based application, we show that sophisticated machine learning can be translated into real-time decision support without imposing additional technical burden on busy clinics. Crucially, the model's most influential features, risk classification, thyroid function, age, smoking status, tumour stage, and treatment response- mirror factors long recognised in endocrine-oncology guidelines, reinforcing the biological plausibility of its predictions and mitigating concerns about "black-box" outputs.

Why does this matter? Current follow-up strategies still rely heavily on broad-brush risk categories or single-marker rules that can miss early recurrence or prompt unnecessary interventions. Our multi-feature, probability-based approach gives clinicians a clearer, patient-specific signal, enabling tighter surveillance of truly vulnerable patients while sparing low-risk individuals the costs and anxiety of over-monitoring. When integrated into electronic medical-record systems, such stratification can optimise use of imaging, laboratory testing, and specialist time, resources that remain scarce even in well-resourced centres.

The findings also extend the growing body of evidence that ensemble learning methods outperform traditional linear and nearest-neighbour techniques in oncology prognostication. Nevertheless, responsible adoption demands two next steps. First, external validation in geographically and ethnically diverse cohorts, particularly from under-represented African populations, will confirm generalisability and surface latent biases. Second, coupling the model with explainability frameworks (e.g., SHAP) will foster clinician trust, support regulatory review, and pave the way for shared decision-making at the bedside. If these conditions are met, the tool holds genuine promise for democratising precision follow-up in thyroid cancer, delivering concrete patient benefits while advancing the field toward transparent, AI-augmented care.

5. ACKNOWLEDGEMENT

The authors wish to acknowledge the support and collaborative contributions of their respective institutions throughout this research. We extend our sincere gratitude to the University of California, Irvine's Machine Learning Repository for providing the dataset used for the project. Our respective institutions, Lakeshore Cancer Center, Federal University Lokoja, Tennessee Technological University, Bells University of Technology, Lagos State University, and the Federal Medical Center Ebute-Metta, and Afe Babalola University, College of Medicine and Health Sciences for providing the academic and infrastructural environments that enabled this work. We also want to thank the team a Datalab for providing the technical support needed in training, building, and deploying the machine learning model. We also appreciate the collective efforts of our first author, who spearheaded and led the research, Dr Joy Aifuobhokhan, and co-authors: Ahmad Khalid Hussain, Chijioke Cyriacus Ekechi, Aisha Olasunbo Olanrewaju, Emmanuel Afuadajo, Deborah Adetola Bowale and Oluwadare Marvellous Inioluwa whose interdisciplinary input was essential to the successful execution of this study.

6. AUTHOR CONTRIBUTION STATEMENT

JA led the research from conception to deployment, including the development, training, and deployment of the machine learning model, as well as the writing, editing, and submission of the manuscript. AK supported the model development and machine learning implementation. CE contributed to the literature review and publication sourcing. AO supported the research design and methodological framework. EA, DB, and OM contributed to the analysis and interpretation of results and the discussion of findings. All authors reviewed and approved the final manuscript.

AUTHOR INFORMATION

Corresponding Authors

Joy Aifuobhokhan, Digital Health and Research – Lakeshore Cancer Center, Nigeria
https://orcid.org/0009-0007-1747-931X
Email: joyaifuobhokhan@gmail.com

Authors

Ahmad Khalid Hussain, Computer science - Federal University Lokoja, Nigeria
https://orcid.org/0009-0009-6067-8079
Email: ahmadkhalidhussain408@gmail.com

Chijioke Cyriacus Ekechi, Engineering - Tennessee Technological University, USA
https://orcid.org/0009-0006-8920-6719
Email: chijiokekechi@gmail.com

Aisha Olasunbo Olanrewaju, Biomedical engineering - Bells University of Technology, Nigeria
https://orcid.org/0009-0002-7338-5725
Email: Ishaolan1@gmail.com

Emmanuel Afuadajo, Electronic and Computer Engineering - Lagos State University, Nigeria
https://orcid.org/0009-0005-1408-5409
Email: emmanuel.afuadajo160211024@st.lasu.edu.ng

Deborah Adetola Bowale, Federal Medical Center Ebute-Metta, Nigeria
https://orcid.org/0009-0002-8400-4633
Email: bowaledeborah@gmail.com

Oluwadare Marvellous Inioluwa, College of Medicine and Health Sciences - Afe Babalola University, Nigeria
https://orcid.org/0009-0002-7841-7773
Email: Marvellous.ini.oluwadare@gmail.com

REFFERENCE

Adedinsewo, D. A., Onietan, D., Morales-Lara, A. C., Moideen Sheriff, S., Afolabi, B. B., Kushimo, O. A., Mbakwem, A. C., Ibiyemi, K. F., Ogunmodede, J. A., Raji, H. O., Ringim, S. H., Habib, A. A., Hamza, S. M., Ogah, O. S., Obajimi, G., Saanu, O. O., Aborisade, S., Jagun, O. E., Inofomoh, F. O., … Carter, R. E. (2025). Contextual challenges in implementing artificial intelligence for healthcare in low-resource environments: insights from the SPEC-AI Nigeria trial. *Frontiers in Cardiovascular Medicine*, *12*(March), 1–9. https://doi.org/10.3389/fcvm.2025.1516088

Ahmad, M. A. S., & Haddad, J. (2024). An Explainable AI Model for Predicting the Recurrence of Differentiated Thyroid Cancer. *Second Jordanian International Biomedical Engineering Conference (JIBEC)*, 84–89. https://doi.org/10.1109/JIBEC63210.2024.10932125

Alawiyah, T., Wibisono, T., & Mulyani, Y. S. (2024). Journal of Computer Networks , Architecture and High Performance Computing The Prediction of Thyroid Cancer Recurrence with the XGBoost Method : The Clinicopathological Feature-Based Approach Journal of Computer Networks , Architecture and High Performanc. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(3), 1035–1045. https://doi.org/10.47709/cnahpc.v6i3.4101

Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhin, A. (2024). Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. *European Archives of Oto Rhino Laryngology*, *280*, 2095–2104. https://doi.org/10.1007/s00405-023-08299-w

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chu, C. S., Lee, N. P., Adeoye, J., Thomson, P., & Choi, S. W. (2020). Machine learning and treatment outcome prediction for oral cancer. *Journal of Oral Pathology and Medicine*, *49*(10), 977–985. https://doi.org/10.1111/jop.13089

Gomes Mantovani, R., Horváth, T., Rossi, A. L. D., Cerri, R., Barbon Junior, S., Vanschoren, J., & Carvalho, A. C. P. L. F. d. (2024). Better trees: an empirical study on hyperparameter tuning of classification decision tree induction algorithms. *Data Mining and Knowledge Discovery*, *38*(3), 1364–1416. https://doi.org/10.1007/s10618-024-01002-5

Gordon, A. J., Dublin, J. C., Patel, E., Papazian, M., Chow, M. S., Persky, M. J., Jacobson, A. S., Patel, K. N., Suh, I., Morris, L. G. T., & Givi, B. (2022). American Thyroid Association Guidelines and National Trends in Management of Papillary Thyroid Carcinoma. *JAMA Otolaryngology - Head and Neck Surgery*, *148*(12), 1156–1163. https://doi.org/10.1001/jamaoto.2022.3360

Habchi, Y., Himeur, Y., Kheddar, H., Boukabou, A., Atalla, S., Chouchane, A., Ouamane, A., & Mansoor, W. (2023). AI in Thyroid Cancer Diagnosis: Techniques, Trends, and Future Directions. *Systems*, *11*(10), 1–33. https://doi.org/10.3390/systems11100519

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00973-y

Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E., Pacini, F., Randolph, G. W., Sawka, A. M., Schlumberger, M., Schuff, K. G., Sherman, S. I., Sosa, J. A., Steward, D. L., Tuttle, R. M., & Wartofsky, L. (2016). 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid*, *26*(1), 1–133. https://doi.org/10.1089/thy.2015.0020

Kim, S. Y., Kim, Y. Il, Kim, H. J., Chang, H., Kim, S. M., Lee, Y. S., Kwon, S. S., Shin, H., Chang, H. S., Park, C. S., & Moorthy, B. T. (2021). New approach of prediction of recurrence in thyroid cancer patients using machine learning. *Medicine (United States)*, *100*(42), E27493. https://doi.org/10.1097/MD.0000000000027493

Li, Z., Wang, N., Li, X., Xie, Y., Dou, Z., Xin, H., Lin, Y., Si, Y., Feng, T., & Wang, G. (2025). Thyroid cancer: From molecular insights to therapy (Review). *Oncology Letters*, *30*(5). https://doi.org/10.3892/ol.2025.15266

Lickert, H., Wewer, A., Dittmann, S., Bilge, P., & Dietrich, F. (2020). Selection of Suitable Machine Learning Algorithms for Classification Tasks in Reverse Logistics. *Procedia CIRP*, *96*(March), 272–277. https://doi.org/10.1016/j.procir.2021.01.086

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4766–4775. https://doi.org/10.48550/arXiv.1705.07874

Mahamadou, A. J. D., Ochasi, A., & Altman, R. B. (2024). Data Ethics in the Era of Healthcare Artificial Intelligence in Africa: An Ubuntu Philosophy Perspective. *ArXiv Preprint ArXiv:2406.10121*. https://doi.org/10.48550/arXiv.2406.10121

Mao, Y., Huang, Y., Xu, L., Liang, J., Lin, W., Huang, H., Li, L., Wen, J., & Chen, G. (2022). Surgical Methods and Social Factors Are Associated With Long-Term Survival in Follicular Thyroid Carcinoma: Construction and Validation of a Prognostic Model Based on Machine Learning Algorithms. *Frontiers in Oncology*, *12*(June), 1–17. https://doi.org/10.3389/fonc.2022.816427

Nettore, I. C., Colao, A., & Macchia, P. E. (2018). Nutritional and environmental factors in thyroid carcinogenesis. *International Journal of Environmental Research and Public Health*, *15*(8). https://doi.org/10.3390/ijerph15081735

Park, Y. M., & Lee, B. J. (2021). Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence. *Scientific Reports*, *11*(1), 1–7. https://doi.org/10.1038/s41598-021-84504-2

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), 1–19. https://doi.org/10.1002/widm.1301

Putatunda, S., & Rama, K. (2018). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. *ACM International Conference Proceeding Series*, 6–10. https://doi.org/10.1145/3297067.3297080

R, K., & E, I. (2021). Hyperparameter tuning of AdaBoost algorithm for social spammer identification. *International Journal of Pervasive Computing and Communications*, *5*(17), 462–482. https://doi.org/10.1108/IJPCC-09-2020-0130

Sankar, S., & Sathyalakshmi, S. (2024). A Study on the Explainability of Thyroid Cancer Prediction: SHAP Values and Association-Rule Based Feature Integration Framework. *Computers, Materials and Continua*, *79*(2), 3111–3138. https://doi.org/10.32604/cmc.2024.048408

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Tang, J., Zhanghuang, C., Yao, Z., Li, L., Xie, Y., Tang, H., Zhang, K., Wu, C., Yang, Z., & Yan, B. (2023). Development and validation of a nomogram to predict cancer-specific survival in middle-aged patients with papillary thyroid cancer: A SEER database study. *Heliyon*, *9*(2). https://doi.org/10.1016/j.heliyon.2023.e13665

Wang, H., Zhang, C., Li, Q., Tian, T., Huang, R., Qiu, J., & Tian, R. (2024). Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches. *BMC Cancer*, *24*(1), 1–12. https://doi.org/10.1186/s12885-024-12146-4

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yousefi, M., Maleki, S. F., Jafarizadeh, A., Youshanlui, M. A., Jafari, A., Pedrammehr, S., Alizadehsani, R., Tadeusiewicz, R., & Pławiak, P. (2024). Advancements in Radiomics and Artificial Intelligence for Thyroid Cancer Diagnosis. *ArXiv*, *1*(39). https://doi.org/10.48550/arXiv.2404.07239

Zimmermann, M. B., & Boelaert, K. (2015). Iodine deficiency and thyroid disorders. *The Lancet Diabetes and Endocrinology*, *3*(4), 286–295. https://doi.org/10.1016/S2213-8587(14)70225-6