engineering proceedings

MDPI

Proceeding Paper

Machine Learning-Based Prognostic Modeling of Thyroid Cancer Recurrence [†]

Duppala Rohan [1], Kasaraneni Purna Prakash [2], Yellapragada Venkata Pavan Kumar [3,*], Gogulamudi Pradeep Reddy [4,*], Maddikera Kalyan Chakravarthi [5] and Pradeep Reddy Challa [6]

[1] School of Computer Science and Engineering, VIT-AP University, Amaravati 522241, Andhra Pradesh, India; rohan.21bce9757@vitapstudent.ac.in
[2] Department of Computer Science and Engineering, Siddhartha Academy of Higher Education, Deemed to Be University, Vijayawāda 520007, Andhra Pradesh, India; kpurnaprakash@vrsiddhartha.ac.in
[3] School of Electronics Engineering, VIT-AP University, Amaravati 522241, Andhra Pradesh, India
[4] Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, Karnataka, India
[5] Engineering Department, College of Engineering and Technology, University of Technology and Applied Sciences, Muscat 133, Oman; kalyan.maddikera@utas.edu.om
[6] School of Engineering and Technology, CGC University, Mohali 140307, Punjab, India; pradeep1417@gmail.com
[*] Correspondence: pavankumar.yv@vitap.ac.in (Y.V.P.K.); pradeep.reddy@manipal.edu (G.P.R.)

[†] Presented at the 6th International Electronic Conference on Applied Sciences, 9–11 December 2025; Available online: https://sciforum.net/event/ASEC2025.

Abstract

Thyroid cancer is the most common type of endocrine cancer. Most cases are called differentiated thyroid cancer (DTC), which includes papillary, follicular, and hurthle cell types. DTC usually grows slowly and has a good prognosis, especially when found early and treated with surgery, radioactive iodine, and thyroid hormone therapy. However, cancer can come back sometimes even years after treatment. This recurrence can appear as abnormal blood tests or as lumps in the neck or other parts of the body. Being able to predict and detect these recurrences early is important for improving patient care and planning follow-up treatment. In this view, this research explores different machine learning algorithms and neural networks to effectively predict DTC recurrence. A total of 17 classifiers were utilized for the experiment, namely, logistic regression, random forest, k-nearest neighbours, Gaussian naïve Bayes, multi-layered perceptron, extreme gradient boosting, adaptive boosting, gradient boosting classifier, extra tree classifier (ETC), light gradient boosting machine, categorical boosting, Bernoulli naïve Bayes, complement naïve Bayes, multinomial naïve Bayes, histogram-based gradient boosting, and nearest centroid, followed by building an artificial neural network. Among the classifiers, ETC performed best with 95.3% accuracy, 95.1% precision, 87.92% recall, 98.18% specificity, 91.21% F1-score, 98.84% AUROC and 97.66% AUPRC on the first dataset, and 99.47% accuracy, 94.83% precision, 98.62% sensitivity, 99.54% specificity, 96.65% F1-score, 99.95% AUROC, and 99.37% AUPRC on the second dataset. To improve model interpretability, Shapley Additive Explanations (SHAP) was also used to explain the contribution of each clinical feature to the model's predictions, allowing for transparent, patient-specific insights into which factors were most important for predicting recurrence, thereby supporting the proposed model's clinical relevance.

Keywords: differentiated thyroid cancer; artificial intelligence; extra tree classifier; machine learning; medical prognosis

Check for updates

Academic Editor: Cosimo Trono

Published: 3 February 2026

https://doi.org/10.3390/engproc2026124013

1. Introduction

Differentiated Thyroid Cancer (DTC) includes papillary, follicular, and hurthle cell variants and accounts for the majority of cases. Although DTC is an indolent disease with favourable survival rates, recurrence remains a significant clinical problem. Despite progress in surgical approaches and radioactive iodine therapy (RAI), DTC recurrence continues to pose clinical challenges. Recurrence usually results from microscopic residual disease present at initial treatment that evades detection, biologically aggressive tumour subtypes, and tumour cells with variable iodine avidity. Extrathyroidal extension, lymph node metastases, less-than-adequate surgical resection, and multifocal tumours increase the risk of disease recurrence years or even decades later. Additionally, it is difficult to predict long-term outcomes based on routine clinical assessment only, as the responses of individual patients to RAI vary according to tumour heterogeneity and genetic alterations and differences in thyroid-stimulating hormone (TSH) suppression. Since such small risk factors accumulate over time, recurrence may occur frequently and insidiously without early clinical manifestations. This indicates the need for improved strategies that might identify individuals who are at high risk before the clinical appearance of recurrence.

The early detection of recurrence provides the opportunity to control risk in time, but there is an emerging need for a dependable system that would assist doctors in predicting recurrence trajectories based on already-available clinical parameters. Recent research has investigated the association of DTC with underlying endocrine or autoimmune diseases. Observations of the prognostic differences in patients with concomitant Graves' disease have pointed out the subtleties of recurrence patterns and the need for improved methods of risk stratification. Again, the meta-analysis of results points to how differences in outcome remain poorly understood due to diverse clinical manifestations and poor predictive models [1]. Although this research showed just how difficult the prediction of recurrence is, many current approaches depend upon routine clinical tests, which might not fully expose nonlinear associations or mild interactions between patient variables.

Artificial Intelligence (AI) has indeed been progressively transforming healthcare. Explainable AI shows the necessity of being transparent for medical decision-support tools. Instead of having clinicians rely on "black-box" models, they need systems that can interpret, understand, and be trusted [2]. AI-based diagnostic systems have shown great potential in many aspects, such as in predicting heart disease [3], for detecting chronic kidney disease [4], and in supporting AI-enabled healthcare infrastructures [5]. Collectively, these studies showcased how Machine Learning (ML) has helped to enhance diagnostic accuracy and smooth clinical workflows. The generalizability of findings across different datasets is an area where most studies have continuously failed. Other deficits include the limited emphasis placed on the interpretability of predictions and a focused lack on the prediction of disease recurrence in cancer-related studies. Although these findings support the potential of ML for medical diagnosis, they highlight certain consistent gaps, viz., a minimal focus on long-term disease recurrence and a lack of emphasis on personalized reasons of predictions, which are considered important in developing clinically deployable tools.

The rest of the paper is structured as follows: Section 2 summarizes the existing literature on DTC recurrence. Section 3 covers data pre-processing and describes the experimental setup; Section 4 reports the results of the experiment, both with regard to the performance of the different models and the SHAP analysis. Conclusions and future research directions are outlined in Section 5.

2. Literature Review

Recent progress in ML and DL has spurred efforts to improve both the diagnosis and recurrence prediction of thyroid cancer. The studies summarized in Table 1 highlight

https://doi.org/10.3390/engproc2026124013

how researchers have shifted toward interpretable and clinically aligned modelling tactics, focusing on feature selection, multimodal integration, and explainability.

Table 1. Summary of the recent literature on DTC.

Year	Objective	Approach	Relevance
2025 [6]	▪ Enhance the model's performance using hyperparameter tuning to accurately predict DTC.	▪ Integrated the Whale Optimization Algorithm (WOA) and a modified version of XGBoost to optimize feature selection and hyperparameters.	▪ Demonstrates the potential of combining meta-heuristic optimization and ML to enhance the early detection of DTC recurrence.
2025 [7]	▪ Investigate the convergence of AI and ML, with a focus on their potential to enhance diagnosis.	▪ Conducted a review of clinical trials and existing works, highlighting the applications of AI in diagnosing cancer and patient care.	▪ Highlights how AI can be applied in pathology, genomics, and radiology, improving biomarker development and treatment selection.
2025 [8]	▪ Improve DTC recurrence prediction by developing an interpretable ML framework incorporating clinical and pathological features.	▪ Employed an XGBoost model and utilized SHAP to identify the contribution of each feature influencing the recurrence risk.	▪ Provides clinicians with feature insights, supporting more informed and patient-centred management of DTC.
2025 [9]	▪ Advance recurrence and risk prediction in DTC by evaluating different ML models and feature selection techniques.	▪ Benchmarked multiple ML classifiers combined with wrapper-based feature selection techniques to identify the most influential features.	▪ Demonstrates how ML classifiers can produce better results using only a few key features.
2025 [10]	▪ Improve the recurrence prediction in DTC by evaluating linear and nonlinear dimensionality reduction techniques.	▪ Employed PCA, t-SVD, UMAP, and t-SNE with ML classifiers, evaluated through cross-validation and bootstrapping.	▪ Shows how optimized pipelines improve accuracy, supporting the integration of ML into clinical decision-support systems.
2025 [11]	▪ Enhance preoperative prediction of lateral lymph node metastases in thyroid carcinoma by developing an explainable DL framework.	▪ Built a 2-way attention multimodal model that incorporates images from ultrasound exams, radiologist assessments, pathology findings, and demographic data from over 39,451 patients across seven centres.	▪ Demonstrates that multimodal DL can outperform human experts, supporting more accurate surgical planning.

https://doi.org/10.3390/engproc2026124013

Table 1. *Cont.*

Year	Objective	Approach	Relevance
2024 [12]	▪ Improve thyroid disease detection by identifying the most informative features and combining multiple classifiers through a stacking ensemble framework.	▪ Applied an information gain feature selection technique and built a stacking ensemble combining logistic regression, random forest, and SVM.	▪ Integrates filter-based feature selection with ensemble methods, improving diagnostic reliability while reducing noise.
2024 [13]	▪ Benchmark different ML models to identify the most reliable and efficient classifier for the accurate detection of thyroid cancer.	▪ Evaluates algorithms such as random forest, decision tree, SVM, KNN, naïve Bayes, and gradient boosting, followed by a detailed comparison.	▪ Provides foundational guidance on selecting effective ML algorithms for diagnosing thyroid cancer, highlighting which methods handle medical data efficiently.
2024 [14]	▪ Develop a highly reliable survival prediction ML model for thyroid cancer patients on a large-scale epidemiological dataset.	▪ Applied chi-square test and select-k-best feature selection techniques, and imbalance handling techniques like random oversampling and SMOTE, and evaluated classifiers which include SVM, random forest, AdaBoost, XGBoost, and MLP using accuracy, F1-score, and AUROC.	▪ Demonstrates that ML models supported by rigorous preprocessing can enhance risk stratification and healthcare management.
2024 [15]	▪ Optimize malignancy classification of thyroid tumours using an ML-based technique that is explainable and well optimized with a bio-inspired optimization policy.	▪ Applied a two-level Naked Mole-Rat Algorithm to identify optimal hyperparameters of classifiers, focusing on key features using SHAP values for model interpretability.	▪ Improves the accuracy of diagnosis and resolves the lack of explainability and optimization in research studies carried out for thyroid cancer using ML.
2024 [16]	▪ Investigate the relationship between papillary thyroid cancer and clinical, biochemical, and thyroid function-related indicators to identify significant predictive risk factors.	▪ Conducted binary logistic and LASSO regression for feature selection, Spearman's correlation analysis, ROC curve evaluation, and extensive statistical testing of thyroid function markers, urine iodine concentration, BMI, and serum biochemicals to compare PTC patients and controls.	▪ Highlights urinary iodine concentration and TSH-related patterns as clinically meaningful predictors, offering evidence that strengthens early risk identification.

https://doi.org/10.3390/engproc2026124013

Table 1. *Cont.*

Year	Objective	Approach	Relevance
2024 [17]	▪ Develop an accurate DL model capable of predicting mortality in DTC patients using demographic, histologic, and TNM staging features.	▪ Built Thy-DAMP, optimized through SGD with Nesterov momentum, cross-validation for hyperparameter tuning, and external validation using a 15-year cohort dataset.	▪ Utilizes DL to create a clinically significant mortality prediction tool that performs better than conventional rule-based staging systems.
2024 [18]	▪ Develop a model for predicting cervical lymph node metastasis in papillary thyroid cancer by integrating radiomics features from ultrasound images.	▪ Extracted high-dimensional features from ultrasound images, performed feature reduction using LASSO and Spearman's correlation analysis, and compared ML models' diagnostic efficacy using AUROC, calibration, and decision-curve analysis.	▪ Demonstrates that integrating radiomics with clinical data enhances the prediction of lymph node metastasis, supporting more precise surgical planning.
2024 [19]	▪ Accurately predict the occurrence of thyroid cancer recurrences while capturing complex clinicopathological relations in the data.	▪ Various ML classifiers were trained on clinical characteristics of patients, with performance comparison to determine the best method for predicting recurrence, with SHAP analysis used to interpret the contributions of the features.	▪ Emphasizes the utility of ML solutions in predicting the recurrence of thyroid cancer while addressing difficulties in model formulation for long-term follow-up.
2024 [20]	▪ Develop an interpretable ML framework that accurately classifies thyroid disease types using clinical and biochemical features while ensuring model transparency through XAI.	▪ Evaluated random forest, SVM, XGBoost, KNN, and decision tree, along with cross-validation, to benchmark performance and applied SHAP to interpret feature contributions.	▪ Addresses the black-box issues that prevent clinical use of ML systems by giving physicians clear, trustworthy decision-support tools for early thyroid disease identification.
2022 [21]	▪ Systematically evaluate the role of AI technology in the fight against the COVID-19 pandemic.	▪ Applied the methodological frameworks of PRISMA and PICO to screen in excess of 21,000 published works down to 41 of high quality using bibliometric searches.	▪ Highlights AI's critical role in pandemic response efforts and illustrates the knowledge of collaboration networks among global partners and challenges such as data quality and privacy.

Table 1. *Cont.*

Year	Objective	Approach	Relevance
2022 [22]	▪ Develop different machine learning models for the accurate prediction of cardiovascular diseases to support clinical decision-making.	▪ Used preprocessing with standard scaling, after which the KNN and ANN models were trained on Kaggle datasets with up to 96% accuracy by using ANN.	▪ Shows the potential of AI-driven prediction systems in helping cardiologists to detect these issues in their early stages, thus reducing mortality rates and improving patient outcomes.

3. Methodology

3.1. Data Preprocessing

Two datasets were used for the experimentation. The first dataset [23] consists of 383 rows and 16 columns, out of which 14 features are categorical, and 1 is of integer type. The second dataset [24] was pre-cleaned, in which redundant attributes and missing laboratory measurements had already been handled, and consists of 26 features. Categorical values in the first dataset were handled via one-hot encoding, and SMOTE–Tomek [25] was used to address data imbalance, as shown in Figure 1. SMOTE reduces class imbalances by randomly choosing a sample from the minority class, finding the near neighbours of the sample, and creating new instances on the line segment connecting the sample with its neighbours. Tomek connections are pairs of samples from distinct classes that are the closest neighbours. If two samples create a Tomek link, at least one of them is considered borderline or noisy. By deleting the majority class samples from Tomek connections, the dataset becomes better, and the class boundary becomes more distinct. SMOTE–Tomek can be defined mathematically as Equation (1). SMOTE alone may introduce overlap between classes, leading to noisy regions. Tomek links alone only reduce overlapping but do not handle imbalance strongly. SMOTE–Tomek enhances minority representation and removes noisy samples near the decision boundary.

$$D_{SMOTE-Tomek} = (D \cup D_{SMOTE}) - D_{Tomek} \tag{1}$$

where D = original dataset, D_{SMOTE} = set of synthetic minority samples generated using SMOTE, and D_{Tomek} = set of majority class samples identified as Tomek links.

3.2. ANN Architecture and Training Configuration

The artificial neural network was built using a sequential approach that included two fully connected hidden layers. The hidden layers included 64 units with ReLU activation functions, followed by a dropout layer with a dropout rate of 0.3. This was followed by another hidden layer that included 32 units with ReLU activation functions, followed by another dropout layer with the same dropout rate of 0.3. Finally, the model ends with an output layer that includes one neuron with a sigmoid activation function for binary classification, as shown in Figure 2.

https://doi.org/10.3390/engproc2026124013

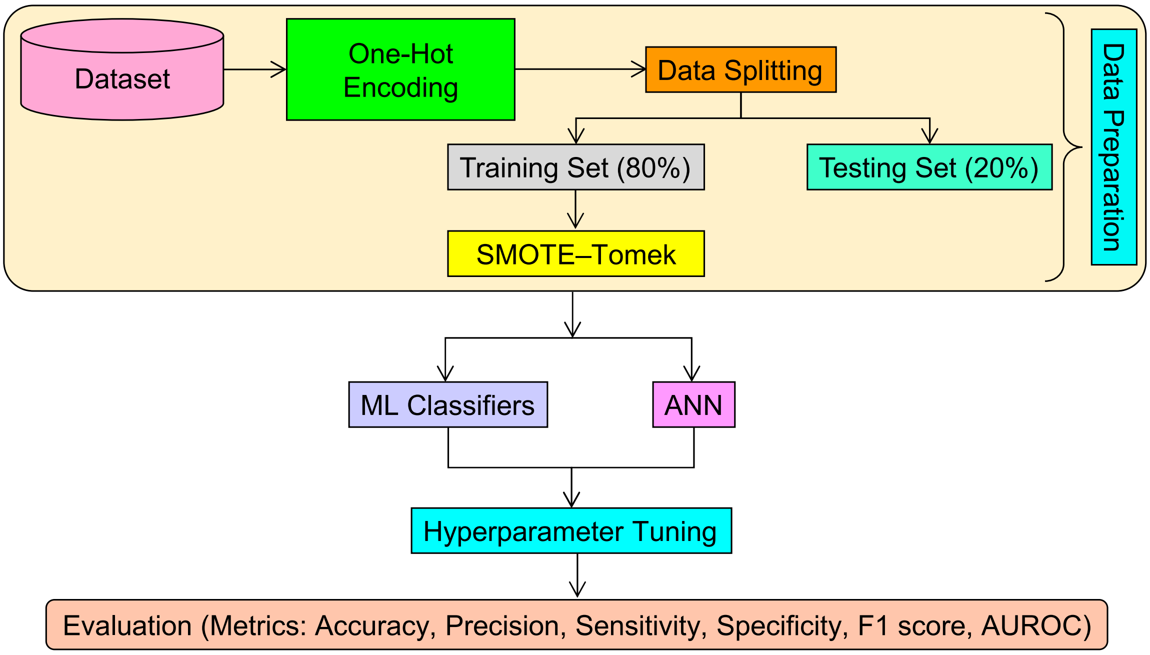


Figure 1. Workflow of the experiment.

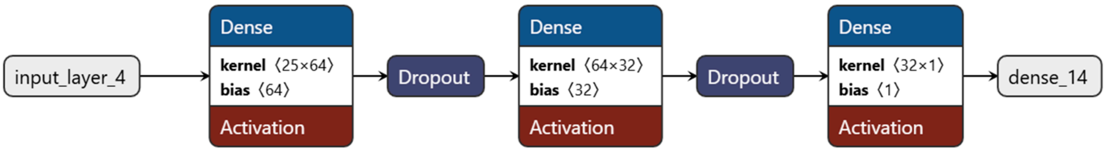


Figure 2. Model architecture of ANN.

The model was optimized using the "Adam" optimizer and the "binary cross-entropy" loss function, while accuracy was used to monitor the training. The maximum number of epochs used during training was 300, and the batch size used was 32, while 20% of the training data was used for validation. During training, "early stopping" was used to validate the loss on the validation set with a patience of 10 epochs. If there is no improvement in validation loss for five consecutive epochs, the learning rate is reduced by 0.5 times, with the minimum limit of the learning rate being set to 10^{-6}. Hyperparameter tuning of the classifiers, including the ETC, was performed manually. For ETC, parameters such as the number of estimators (100–500), maximum depth (5–20), and minimum samples per leaf (1–5) were tuned to identify the optimal configuration.

3.3. SHAP

In this study, SHAP [26] was used to interpret the predictions of ML models on the dataset. Shapley values indicate the average marginal contribution of a feature when added to all possible coalitions of features. SHAP achieves this by constructing a set of feature subsets by systematically adding and removing features from the model and tracking changes in the predicted output. It is assumed that the prediction can be divided into the summation of the contributions of individual features, as formulated in Equation (2).

$$f(x) = \varnothing_0 + \sum_{j=1}^{M} \varnothing_j \tag{2}$$

where $f(x)$ = model's prediction, $\varnothing_0$ = base value (Expected prediction across the dataset), $\varnothing_j$ = Shapley value representing the contribution of each feature, and M = total number of features.

https://doi.org/10.3390/engproc2026124013

By adding together SHAP values throughout the whole dataset, interpretability can be obtained. This makes it clear how the model makes its decisions.

4. Results and Discussion

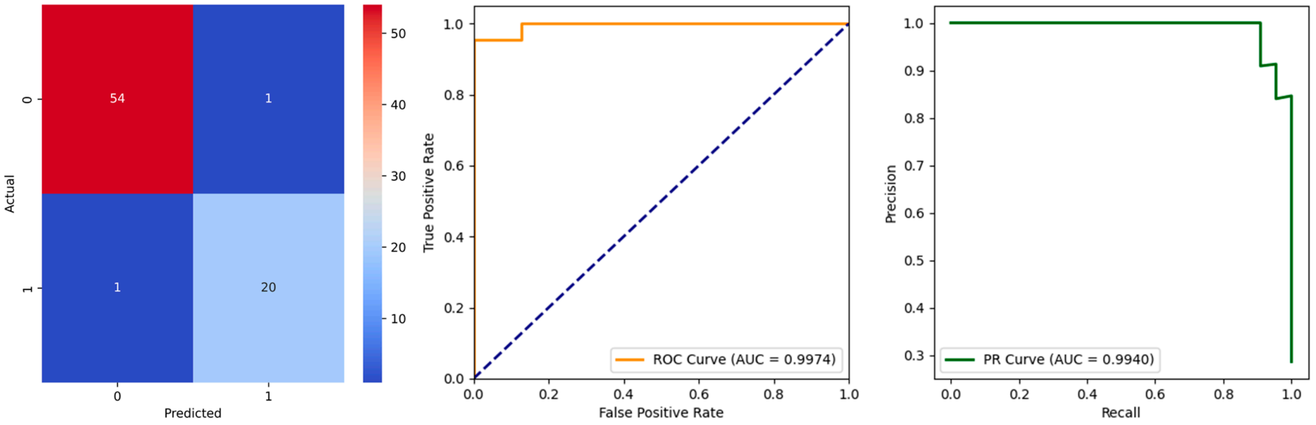
4.1. Results of Classifiers on the First Dataset

Table 2 compares the performance of 17 classifiers that have been evaluated on the first dataset by five-fold cross-validation. Overall, ensemble-based models achieved consistently better results compared with linear and probabilistic classifiers, which underlines their increased capability for modelling complex, nonlinear relationships present in clinical data. Traditional classifiers that included LR, KNN, and variants of naïve Bayes reported overall lower sensitivity and F1-scores, while neural and tree-based ensembles had more balanced and robust predictive behaviour. Out of all the models tested, ETC performed best in overall results with the highest testing accuracy (95.30 ± 2.00%) and precision (95.10 ± 3.26%), and highest specificity (98.18 ± 1.29%), besides having one of the best F1-scores (91.21 ± 4.21%). In terms of model discriminability, ETC also had the highest AUROC (98.84 ± 0.71%) and AUPRC (97.66 ± 1.45%) scores with a minor edge over the best competitors in the area of ensemble models like XGB (AUROC: 98.64 ± 1.04%, AUPRC: 97.17 ± 2.26%), CB (AUROC: 98.61 ± 0.84%, AUPRC: 97.29 ± 1.68%), and GBC (AUROC: 98.42 ± 1.05%, AUPRC: 96.97 ± 1.86%). Although the other models were comparable to ETC in terms of accuracy and F1-scores, ETC was better in terms of model discriminability.

Table 2. Results of various classifiers on the first dataset (mean ± standard deviation).

S. No.	Model	Testing Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUROC (%)	AUPRC (%)
1	LR	94.26 ± 1.18	89.49 ± 5.84	90.74 ± 3.21	95.64 ± 2.87	89.96 ± 1.57	96.83 ± 0.18	94.14 ± 4.36
2	RF	94.51 ± 3.41	91.4 ± 5.1	88.79 ± 9.95	96.73 ± 1.99	89.89 ± 6.93	98.79 ± 0.85	97.57 ± 1.66
3	KNN	91.9 ± 1.74	86.59 ± 7.12	85.15 ± 3.99	94.55 ± 3.4	85.62 ± 2.62	95.96 ± 1.04	89.89 ± 2.92
4	GNB	91.12 ± 2.14	94.67 ± 5.06	73.12 ± 9.99	98.18 ± 1.82	82 ± 5.59	97.51 ± 1.7	93.20 ± 5.23
5	MLP	93.98 ± 2.59	89.67 ± 4.11	88.79 ± 7.33	96 ± 1.52	89.14 ± 5.2	97.4 ± 1.1	95.75 ± 2.04
6	XGB	94.77 ± 1.86	90.78 ± 3.02	90.69 ± 5.88	96.36 ± 1.29	90.65 ± 3.66	98.64 ± 1.04	97.17 ± 2.26
7	ADB	94.79 ± 3.31	90.48 ± 7.91	91.69 ± 3.78	96 ± 3.5	90.99 ± 5.43	97.98 ± 0.8	96.49 ± 1.54
8	GBC	95.04 ± 1.44	90.88 ± 2.97	91.65 ± 3.99	96.36 ± 1.29	91.21 ± 2.65	98.42 ± 1.05	96.97 ± 1.86
9	ETC	95.3 ± 2	95.1 ± 3.26	87.92 ± 7.27	98.18 ± 1.29	91.21 ± 4.21	98.84 ± 0.71	97.66 ± 1.45
10	LGBM	94.52 ± 2.34	90.05 ± 4.32	90.74 ± 6.74	96 ± 1.99	90.27 ± 4.29	98.33 ± 0.79	96.55 ± 1.5
11	CB	95.3 ± 2	91.68 ± 3.93	91.65 ± 3.99	96.73 ± 1.52	91.64 ± 3.65	98.61 ± 0.84	97.29 ± 1.68
12	BNB	90.33 ± 2.43	80.94 ± 8.03	86.97 ± 6.16	91.64 ± 3.77	83.54 ± 4.11	97.68 ± 0.56	95.08 ± 1.29
13	CNB	91.64 ± 2.02	84.24 ± 9.12	87.92 ± 5.31	93.09 ± 4.15	85.64 ± 3.12	98.06 ± 0.58	95.95 ± 1.12
14	MNB	91.64 ± 2.02	84.24 ± 9.12	87.92 ± 5.31	93.09 ± 4.15	85.64 ± 3.12	98.06 ± 0.58	95.95 ± 1.12
15	HGB	93.99 ± 1.76	89.11 ± 3.62	89.78 ± 6.18	95.64 ± 1.63	89.32 ± 3.45	98.35 ± 0.79	96.87 ± 1.37
16	NC	89.02 ± 2.26	78.18 ± 5.12	85.11 ± 7.69	90.55 ± 2.7	81.28 ± 4.24	97.6 ± 0.52	94.46 ± 1.97
17	ANN	95.29 ± 2.58	95.03 ± 3.29	87.88 ± 9.39	98.18 ± 1.29	91.08 ± 5.67	98.28 ± 0.41	96.99 ± 0.88

The confusion matrix for the best-performing fold of ETC has also been shown in Figure 3a. ETC has managed to classify 54 out of 55 non-recurrence samples and 20 out of 23 recurrence samples in a correct manner while also obtaining only one false positive instance and one false negative instance. This clearly shows that ETC has achieved a superior trade-off between its sensitivity and specificity compared with the other top-performing models that have had a greater variability in results among the folds. The ROC curve and precision–recall (PR) curve for ETC, shown in Figure 3b,c, clearly show that it has a superb discriminatory power compared with the other models since it has obtained an AUROC value of 99.74% and an AUPRC of 99.4%, which is higher compared with that of the other ensemble models. Although ensemble classifiers performed well in predicting the first dataset, ETC stood out in terms of stability, discriminability, and class balance characteristics and thus proved to be the best classifier.



(**a**) Confusion Matrix (**b**) AUROC (**c**) AUPRC

Figure 3. Results of ETC on the first dataset.

4.2. Results of Classifiers on the Second Dataset

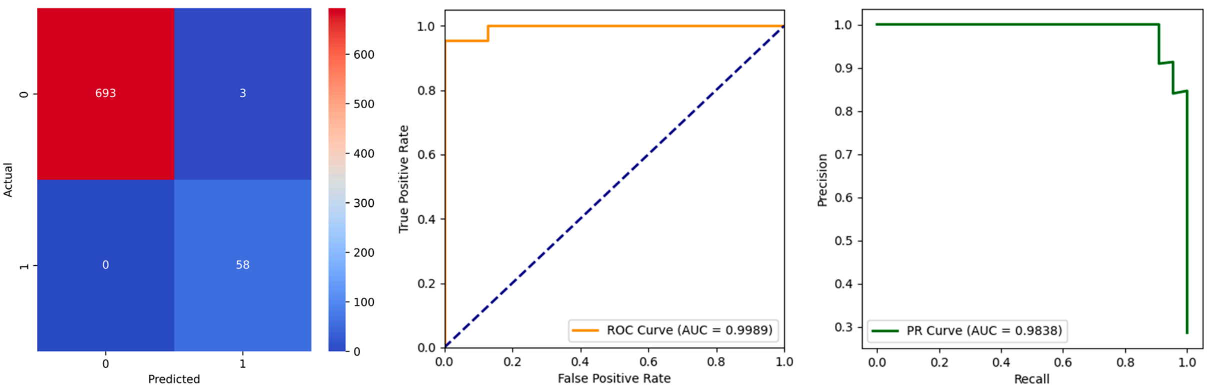
Table 3 summarizes the results of the comparative analysis of 17 classifiers applied to the second dataset in a five-fold cross-validation setting. Ensemble machine learning algorithms demonstrated a significant advantage over linear predictors, probabilistic predictors, and distance-based predictors due to their ability to learn complex patterns in the imbalanced health datasets. On the other hand, poor results in precision, F1-score, and AUPRC metrics demonstrated that the results of the KNN and naïve Bayes machine learning algorithm variants are not very reliable in predicting recurrence. Among all the investigated models, ETC yielded the highest testing accuracy, precision, specificity, and F1-score—99.47 ± 0.16%, 94.83 ± 2.91%, 99.54 ± 0.28%, and 96.65 ± 0.96%, respectively. In terms of discrimination capability, ETC also attained the highest AUROC and AUPRC, 99.95 ± 0.04% and 99.37 ± 0.59%, respectively, which is very close to those of other powerful ensemble classifiers, including RF, XGB, CB, and ADB: AUROC of 99.87 ± 0.13%, 99.73 ± 0.22%, 99.84 ± 0.13%, and 99.72 ± 0.37%; and AUPRC of 98.29 ± 1.92%, 95.33 ± 4.89%, 97.89 ± 2.20%, and 97.45 ± 2.49%, respectively. Although these models have a similar accuracy and sensitivity, ETC consistently exhibited superior precision with better class-wise balance, leading to a more stable and clinically reliable performance.

To further demonstrate the efficacy of ETC, Figure 4a presents a confusion matrix corresponding to its best-performing fold: the model appropriately classified 693 non-recurrence and 58 recurrence cases, with only three false positives and no false negatives, which is indicative of perfect sensitivity and near-perfect specificity. The strong classification performance is further reflected by the ROC curve presented in Figure 4b, wherein ETC has obtained an AUROC of 99.89%, thus confirming an excellent discriminability of recurrent versus non-recurrent cases. The PR curve in Figure 4c reports an AUPRC of 98.38%, again illustrating the robustness of ETC for identifying recurrence cases even under severe conditions of class imbalance. In conclusion, while some of the ensemble techniques showed excellent prediction capabilities on the second dataset, ETC was clearly better than the rest of the classifiers in terms of accuracy, discriminative power, and the stability of results across cross-validation runs. This further proves that ETC has consistently performed across both datasets, emerging as the proposed method of this research. The hyperparameters of the proposed ETC model are tabulated in Table 4.

https://doi.org/10.3390/engproc2026124013

Table 3. Results of various classifiers on the second dataset (mean ± standard deviation).

S. No.	Model	Testing Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUROC (%)	AUPRC (%)
1	LR	96.98 ± 0.45	77.91 ± 2.39	84.85 ± 4.83	97.99 ± 0.23	81.2 ± 3.1	96.33 ± 1.72	85.87 ± 6.43
2	RF	99.23 ± 0.37	91.51 ± 3.11	99.31 ± 1.54	99.22 ± 0.3	95.24 ± 2.25	99.87 ± 0.13	98.29 ± 1.92
3	KNN	85.76 ± 1.29	26.91 ± 4.03	49.11 ± 7.67	88.82 ± 1.23	34.72 ± 5.06	76.91 ± 4.96	31.44 ± 6.87
4	GNB	42.85 ± 2.24	10.38 ± 0.8	83.82 ± 5.58	39.43 ± 2.33	18.47 ± 1.39	62.56 ± 3.66	10.32 ± 0.98
5	MLP	97.72 ± 0.41	88.74 ± 5.79	81.43 ± 7.38	99.08 ± 0.63	84.57 ± 3.14	96.16 ± 2.11	88.5 ± 5.18
6	XGB	99.18 ± 0.29	92.34 ± 3.66	97.59 ± 1.54	99.31 ± 0.34	94.85 ± 1.75	99.73 ± 0.22	95.33 ± 4.89
7	ADB	99.26 ± 0.47	92.4 ± 4.29	98.62 ± 2.25	99.31 ± 0.4	95.38 ± 2.87	99.72 ± 0.37	97.45 ± 2.49
8	GBC	99.2 ± 0.57	91.68 ± 5.49	98.97 ± 2.31	99.22 ± 0.57	95.11 ± 3.38	99.38 ± 0.98	93.75 ± 5.81
9	ETC	99.47 ± 0.16	94.83 ± 2.91	98.62 ± 1.89	99.54 ± 0.28	96.65 ± 0.96	99.95 ± 0.04	99.37 ± 0.59
10	LGBM	99.02 ± 0.38	90.98 ± 2.65	96.9 ± 2.25	99.2 ± 0.24	93.85 ± 2.4	99.35 ± 0.74	93.38 ± 2.89
11	CB	99.1 ± 0.34	90.63 ± 3.54	98.62 ± 1.44	99.14 ± 0.34	94.43 ± 2.09	99.84 ± 0.13	97.89 ± 2.2
12	BNB	79.47 ± 1.21	22.32 ± 3.16	67.31 ± 11.34	80.49 ± 0.94	33.52 ± 4.96	82.41 ± 5.12	53.3 ± 10.39
13	CNB	44.68 ± 0.83	9.72 ± 0.96	74.53 ± 8.65	42.18 ± 1.24	17.19 ± 1.73	72.9 ± 6.74	50.3 ± 10.3
14	MNB	44.68 ± 0.83	9.72 ± 0.96	74.53 ± 8.65	42.18 ± 1.24	17.19 ± 1.73	72.9 ± 6.74	50.3 ± 10.3
15	HGB	89.4 ± 22.29	78.18 ± 36.71	96.24 ± 3.69	88.82 ± 23.95	81.18 ± 33.06	93.25 ± 13.52	79.35 ± 37.61
16	NC	78.89 ± 0.79	21.43 ± 1.87	65.27 ± 7.74	80.03 ± 0.91	32.25 ± 3.04	83.17 ± 3.33	50.99 ± 7.44
17	ANN	98.33 ± 0.63	94.54 ± 2.76	83.13 ± 7.59	99.6 ± 0.21	88.33 ± 4.79	97.45 ± 2.14	92.96 ± 4.46



(**a**) Confusion Matrix (**b**) AUROC (**c**) AUPRC

Figure 4. Results of ETC on the second dataset.

Table 4. Hyperparameters of ETC.

S. No.	Hyperparameter	Value
1	n_estimators	500
2	max_depth	None
3	criterion	"entropy"
4	min_samples_split	2
5	min_samples_leaf	1
6	max_features	None
7	class_weight	"balanced"
8	random_state	42
9	n_jobs	−1

4.3. Discussion

The results presented in Tables 2 and 3 clearly reveal that model performance is rather sensitive to both the learning paradigm and underlying data characteristics, given that the ensemble-based classifiers systematically outperformed classifiers from linear, probabilistic, and distance-based learning for most of the metrics used. For the first dataset (Table 2), the performance of classifiers was relatively smooth, with testing accuracies mostly distributed between 90% and 95%. Ensemble models, including GBC, ETC, CB, XGB, and ADB, performed the best and most balanced based on accuracy, F1-score, AUROC, and AUPRC. Among them, ETC and CB obtained the highest testing accuracy of 95.30 ± 2.00% and 95.30 ± 2.00%, respectively. Meanwhile, ETC demonstrated the best precision of 95.10 ± 3.26% and specificity of 98.18 ± 1.29%. Although reasonable performances in accuracy were seen for LR and KNN, their relatively lower values in AUROC and AUPRC did manifest an increasingly weaker discrimination capability in classes with significant imbalance. Then, variants of naïve Bayes and the NC further manifested lower F1-scores and higher variability, indicating weaker robustness for the prediction of recurrence.

In contrast, more diverse performance differences among the classifiers can be seen within the second dataset, as presented in Table 3. The noticeable class imbalance, probably along with more complex interactions between features, induced such differences. Although many of the ensemble models reached a very high testing accuracy—a value of over 99% in some cases—accuracy was not enough to reflect clinical reliability in every case. The probabilistic and distance-based models, including KNN, GNB, BNB, CNB, and MNB, exhibited noticeably poor precision, F1-scores, and AUPRC values, although sometimes yielding high sensitivity; this would mean that they had a strong bias toward the majority class and a higher rate of false positives. ANN and MLP showed competitive performances but lower sensitivity and AUPRC values compared with the ensemble methods, therefore showing reduced stability in the detection of the minority class.

Although accuracy and AUROC are useful indicators of the overall performance of the classification task, they might not address the issue of class imbalance, which is very common in recurrence forecasting problems. Henceforth, there was greater importance attached to F1-score and PR curve, which are generally better indicators to address the precision–sensitivity trade-off. As made evident in Tables 2 and 3, several ensemble models like ETC, GBC, and CB achieved high F1-scores in both datasets, thus concluding an effective trade-off between predicting the recurrence and avoiding false positives. For the first dataset (Table 2), the precision for the ETC model came out to be 95.10% with slightly reduced sensitivity at 87.92%, and the F1-score and AUPRC were calculated to be 91.21% and 97.66%, respectively. Likewise, for the same dataset, the precision for the ANN model stood at 95.03%, with a slightly reduced sensitivity of 87.88%, and the F1-score and AUPRC were calculated to be 91.08% and 97.67%, respectively. This highlights the appropriateness and efficiency of these models in terms of correctly identifying the actual cases, although a few cases may have gone undetected.

In the second dataset (Table 3), the ETC model outperformed others with an F1-score of 96.65%, an AUROC of 99.95%, and an AUPRC of 99.37%, and this showcases the robustness of ETC irrespective of higher class imbalance. Other ensemble models, such as RF, XGB, and ADB, also had F1-scores of more than 94%, which again ensures reliable performances for making predictions. From a clinical perspective, these results are significant in relation to the trade-off in precision and recall observed in the models. A false negative might indicate a failure in a recurrence case that could have delayed follow-up or treatment. On the other hand, a false positive could result in unnecessary follow-up, an increase in health expenses, and patient anxiety. Models such as ETC and ANN follow a conservative predictive approach that favours high precision in order to minimize the occurrence of false positives at the expense of a slight decrease in recall. The addition of the F1 measure and AUPRC ensures the clinical significance of the comparison in accordance with distinct clinical considerations.

4.4. Comparison with State-of-the-Art Models

The performance of the proposed ETC model on various state-of-the-art models has been presented in Table 5. The models include XGBoost [8,19], KNN [9], LR [10,16], LightGBM [15], Thy-DAMP [17], and RF [18]. Though the existing models performed well in achieving moderate to high accuracy on their respective data sets, they did not perform well in achieving a combined result of accuracy, sensitivity, specificity, F1-score, AUROC, and AUPRC. XGBoost [19] was able to attain a high specificity of 89.6%, but it attained a low sensitivity of 60.1%, implying it was not effective in identifying recurrence well. The same applies to Thy-DAMP models that attained 93.3% accuracy and 83.24% sensitivity. By contrast, the proposed ETC model outperformed all existing methods in two independent datasets.

Table 5. Comparison of ETC vs. state-of-the-art models.

Ref.	Proposed Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUROC (%)	AUPRC (%)
[8]	XGBoost	-	-	79	78	-	84	-
[9]	KNN	95.9	-	93.7	88.9	-	93.7	-
[10]	LR	95	-	94	-	-	99	-
[15]	LightGBM	81.82	84.97	88.4		86.62	86	-
[16]	LR	-	-	70.1	71.4	-	67.3	-
[17]	Thy-DAMP	93.3	-	83.24	93.53	-	95	-
[18]	RF	77.5	-	67.6	78.4	33.1	76.6	-
[19]	XGBoost	72.4	74.9	60.1	89.6	66.7	85.7	-
Proposed Model	ETC (On first dataset)	95.3	95.1	87.92	98.18	91.21	98.84	97.66
	ETC (On second dataset)	99.47	94.83	98.62	99.54	96.65	99.95	99.37

With the first dataset, ETC yielded 95.3% accuracy, 95.1% precision, 87.92% sensitivity, 98.18% specificity, 91.21% F1-score, 98.84% AUROC, and 97.66% AUPRC. Notice that this result shows an effective balance between true positives and false positives regarding the identification of recurrence cases. This model further improved its scores on the second dataset: 99.47% accuracy, 94.83% precision, 98.62% sensitivity, 99.54% specificity, 96.65% F1-score, 99.95% AUROC, and 99.37% AUPRC. This certainly proves that even with a higher-class imbalance, the proposed model is robust. These findings clearly illustrate that the ETC not only outperforms conventional models in standard evaluation metrics but also offers a good trade-off between precision and recall rates, which is essential in order to avoid false negatives (missed recurrence) and false positives (unwanted monitoring). The proposed ETC certainly establishes a new benchmark in recurrence risk prediction models.

4.5. SHAP Analysis

The SHAP values for each feature group after one-hot encoding indicate their impact on recurrence prediction (Table 6). The Response group made the largest contribution (0.31 ± 0.07), suggesting that the model's judgments are heavily influenced by treatment response. The Risk group came next (0.12 ± 0.02), while Focality and Adenopathy both had a significant impact. While clinical feature groups like Physical Examination and Pathology contributed less, TNM-related features (N and T) contributed moderately. Age, Gender, and Smoking were among the lifestyle and demographic characteristics that had very little effect, and variables like M, Hx Smoking, and Hx Radiotherapy made very little difference. Overall, the SHAP analysis verifies that response-, risk-, and tumour-related factors are the main factors influencing recurrence prediction, as depicted in Figure 5. A beeswarm plot of the feature groups is shown in Figure 6.

Table 6. The contribution of each feature of the first dataset to recurrence prediction.

S. No.	Feature Group	Mean ± Standard Deviation (SHAP Value)
1	Response	0.31 ± 0.07
2	Risk	0.12 ± 0.02
3	Adenopathy	0.05 ± 0.01
4	Focality	0.03 ± 0.01
5	T	0.03 ± 0.01
6	Physical Examination	0.02 ± 0.01
7	N	0.01 ± 0.007
8	Pathology	0.01 ± 0.01
9	Age	0.01 ± 0.009
10	Thyroid Function	0.01 ± 0.01
11	Stage	0.006 ± 0.01
12	Gender	0.006 ± 0.006
13	M	0.002 ± 0.003
14	Smoking	0.001 ± 0.002
15	Hx Smoking	0.0007 ± 0.001
16	Hx Radiotherapy	0.00008 ± 0.0001

SHAP value analysis shows the relative importance of various clinical features to the predictions made by the model (Table 7). Among all the features, TSH has been shown to have the highest impact on the model predictions, with a mean SHAP value of 0.33 ± 0.09, showing the predominant role of TSH in the model's decision-making process. This is followed by FTI (0.05 ± 0.02) and TT4 (0.04 ± 0.02), emphasizing the role of biochemical indices for evaluating thyroid function. The treatment feature "on thyroxine" (0.04 ± 0.08) and the test feature "TSH measured" (0.03 ± 0.06) have also been shown to have a significant effect on the predictions. While demographic and auxiliary clinical features are lower in impact, including sex (0.02 ± 0.03) and age (0.002 ± 0.003), several flags of laboratory measurements contribute marginally, such as FTI measured, T3 measured, and T4U measured. Variables describing comorbid conditions and clinical history, such as psych, sick, tumour, and thyroid surgery, have low impacts, whereas the SHAP values for very rare conditions or treatments, like hypopituitary, lithium, I131 treatment, and pregnant, become negligible. By and large, biochemical thyroid markers and features related to treatment are confirmed as the major drivers of the model by the SHAP results, while demographic factors and less frequent clinical conditions contribute minimally, as illustrated in Figure 7.

In addition to feature-level behaviour analysis through feature-level plots, beeswarm plots for the SHAP values of each feature have also been created for both datasets. Unlike the aggregated SHAP bar charts, the feature-level behaviour of the beeswarm charts provides insights into the spread of feature behaviour across patients. In the first dataset (Figure 6), the features Response, Risk, T, N, and Adenopathy have a positive SHAP value with less variability. Adenopathy (0.05 ± 0.01) is one of the most influential features; most data points have a positive effect. This implies that the involvement of the lymph nodes increases the chance of recurrence. This result is clinically well supported because cervical lymph node involvement has long been a well-accepted prognostic indicator in thyroid carcinomas and has a strong positive correlation with persistence and recurrence. Several studies support the importance of lateral lymph node involvement in specifically impairing recurrence-free survival rates and are therefore ranked so highly in SHAP. In contrast, demographic and lifestyle variables like age, gender, smoking, and radiotherapy have SHAP values huddled around zero, indicating that these variables have little and unstable effects on the model predictions.

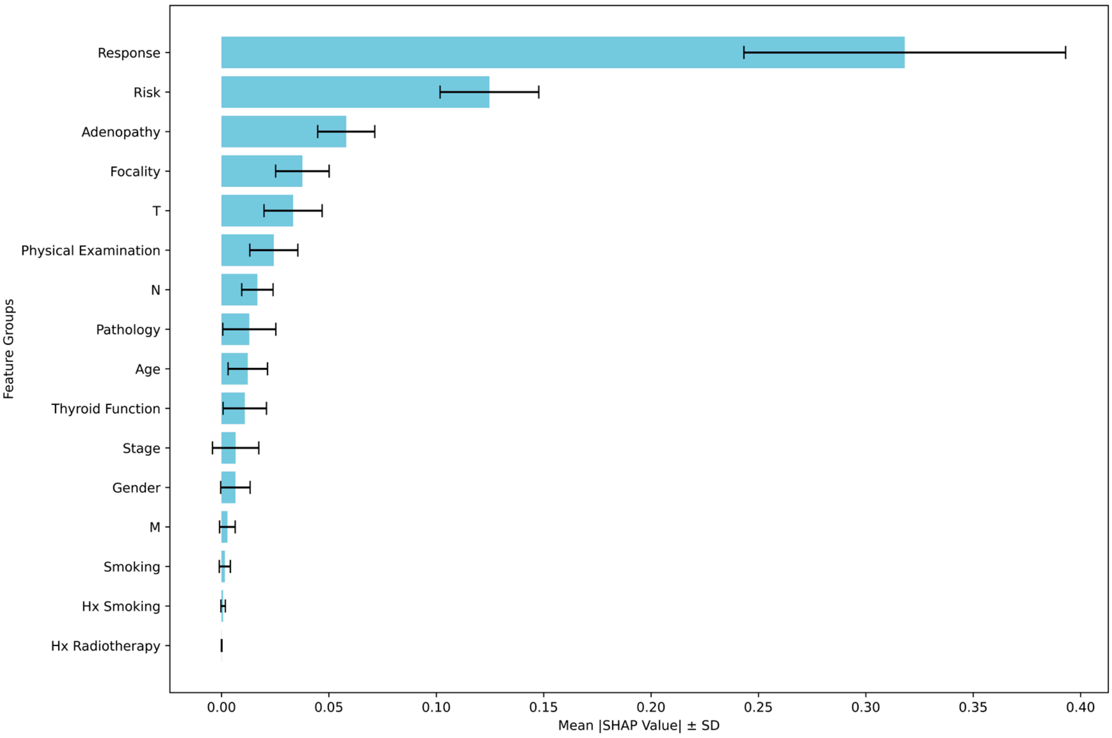


Figure 5. Mean SHAP values of feature groups of the first dataset.

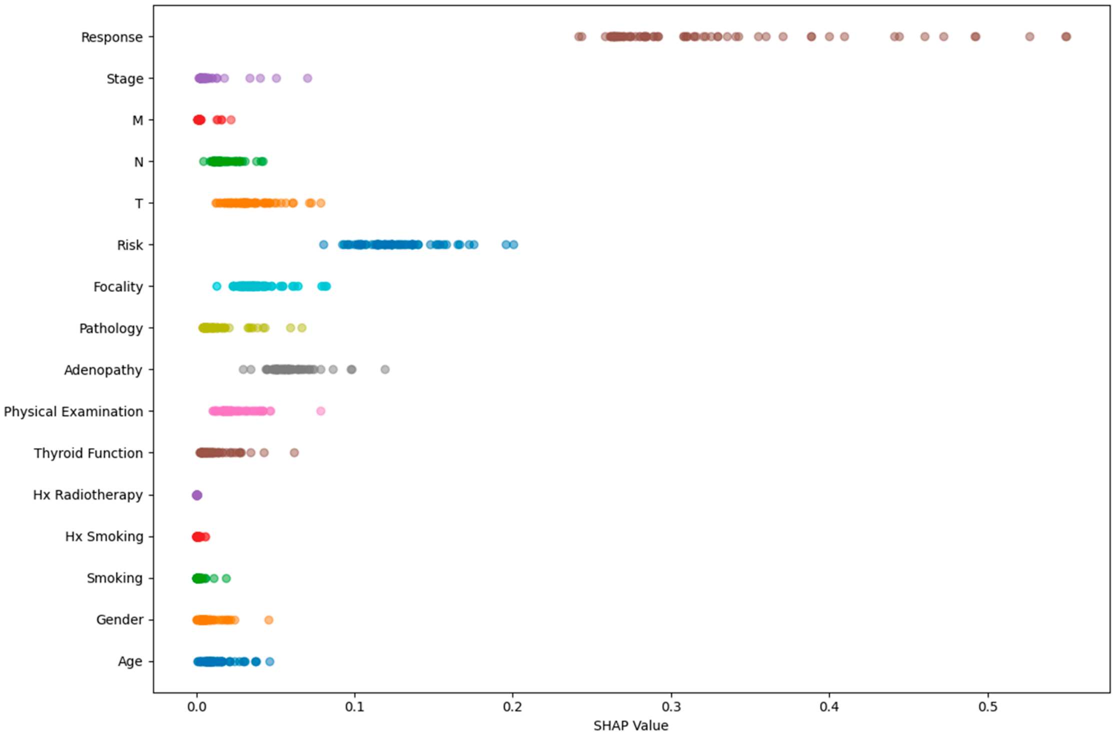


Figure 6. Beeswarm plot of feature groups of the first dataset.

Table 7. The contribution of each feature of the second dataset to recurrence prediction.

S. No.	Feature	Mean ± Standard Deviation (SHAP Value)
1	TSH	0.33 ± 0.09
2	FTI	0.05 ± 0.02
3	TT4	0.04 ± 0.02
4	on thyroxine	0.04 ± 0.08
5	TSH measured	0.03 ± 0.06
6	sex	0.02 ± 0.03
7	FTI measured	0.02 ± 0.01
8	T3 measured	0.01 ± 0.008
9	T4U measured	0.004 ± 0.005
10	thyroid surgery	0.003 ± 0.03
11	age	0.002 ± 0.003
12	query hyperthyroid	0.002 ± 0.006
13	T4U	0.002 ± 0.004
14	psych	0.001 ± 0.005
15	query hypothyroid	0.001 ± 0.003
16	TT4 measured	0.0009 ± 0.002
17	sick	0.0009 ± 0.003
18	tumour	0.0004 ± 0.003
19	I131 treatment	0.0002 ± 0.002
20	on antithyroid medication	0.0001 ± 0.0009
21	pregnant	0.0001 ± 0.0005
22	query on thyroxine	0.0001 ± 0.0007
23	goitre	0.0001 ± 0.0007
24	lithium	0.00008 ± 0.0005
25	hypopituitary	0.0000008 ± 0.000002

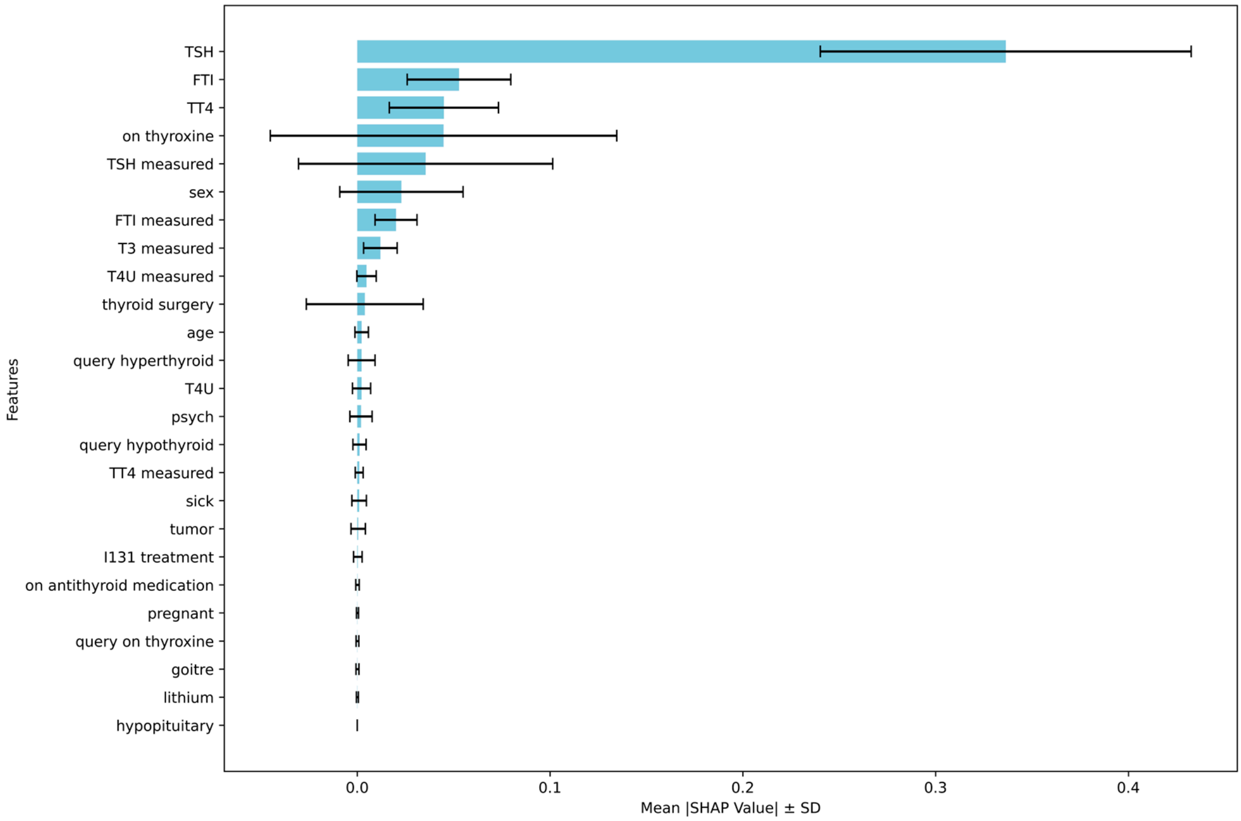


Figure 7. Mean SHAP values of features of the second dataset.

Likewise, for the second dataset (Figure 8), biochemical variables like TSH, FTI, TT4, and treatment variables (on thyroxine, thyroid surgery) strongly influence the SHAP values. These variables have large SHAP ranges, suggesting variability in the treatment responses, especially for those with irregular levels of thyroid hormones and treated cases. Features like TSH are consistent and large in size, re-emphasizing its significance as an endpoint for biochemical analysis for assessing the thyroid condition. Overall, the beeswarm plots revealed that the predictions made by the model are influenced largely by tumour-related factors, lymph node status, treatment response, and markers of thyroid function, and only marginally by other less important factors. This level of explainability at the instance level helps to make the proposed framework more credible and transparent. Although SHAP is a model-agnostic explainability framework, in this study, it was applied specifically to the ETC, as it demonstrated the best overall predictive performance among all evaluated models. The purpose of the SHAP analysis was to enhance the clinical interpretability of the final selected model rather than to conduct a comparative interpretability assessment across classifiers. Since other high-performing models, such as RF, are also tree-based ensemble methods, similar feature contribution patterns are expected. Therefore, SHAP-based interpretation was limited to the ETC model.

The findings of this study indicate that ETC is a reliable model for predicting the recurrence of ETC. With only one recurrence and non-recurrence case incorrectly classified in the first dataset and three false positives and no false negatives in the second dataset, the confusion matrices demonstrate the model's steady behaviour. When compared with existing works, ETC demonstrated a better performance across all metrics. XGBoost, KNN, LR, LightGBM, RF, and Thy-DAMP often reported lower sensitivity and uneven performances across the metrics. The usage of SHAP added further explainability to the model's decision-making process. Features related to post-operative response, such as clinical risk categories, lymph node involvement, and tumour focality, contributed the most to the recurrence prediction, whereas demographic features showed only a marginal influence on the model's decision-making process.

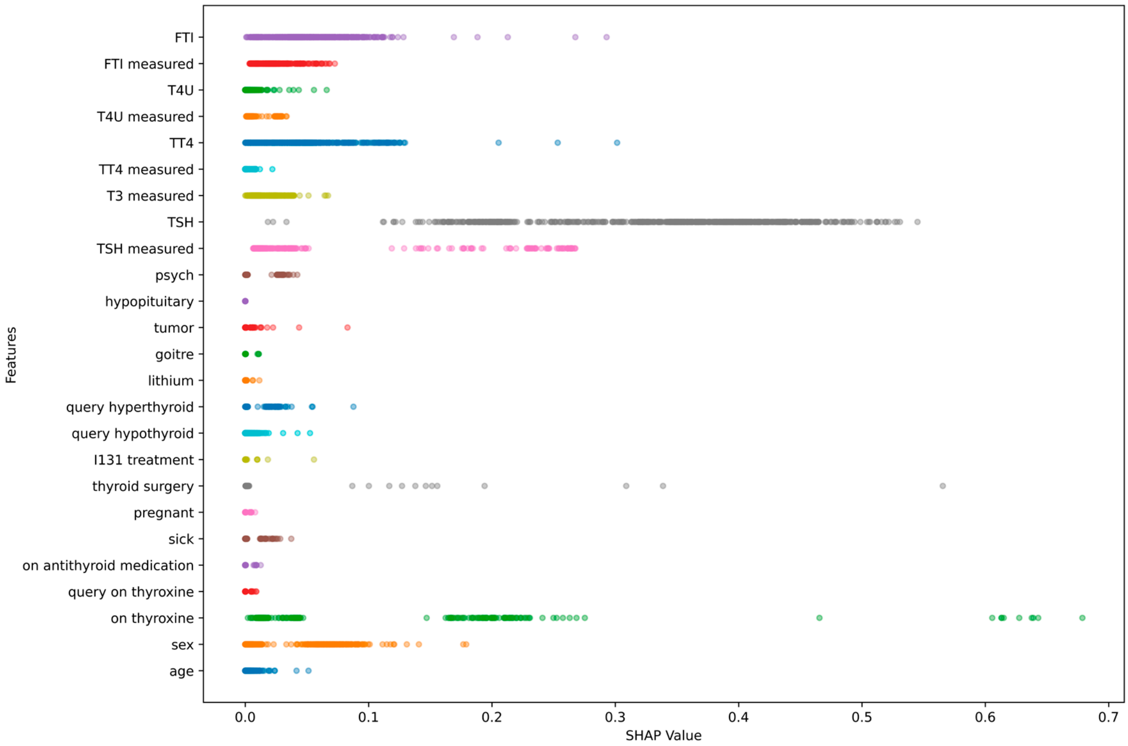


Figure 8. Beeswarm plot of features of the second dataset.

5. Conclusions

The early identification of thyroid cancer recurrence plays a crucial role in preventing disease progression and guiding personalized medical management. A comprehensive experimental framework was developed to evaluate the recurrence risk using clinical information. After applying one-hot encoding, SMOTE–Tomek balancing, and performing hyperparameter tuning, ETC outperformed the other classifiers with testing accuracies of 95.3% and 99.47%, precision values of 95.1% and 94.83%, sensitivities of 87.92% and 98.62%, specificities of 98.18% and 99.54%, F1-scores of 91.21% and 96.65%, AU-ROC values of 98.84% and 99.95%, and AUPRC values of 97.66% and 99.37% on the two datasets respectively, demonstrating its predictive power. SHAP feature group analysis was conducted to enhance interpretability, revealing that Response (0.31 ± 0.07), Risk (0.12 ± 0.02), and Adenopathy (0.05 ± 0.01) in the first dataset, and TSH (0.33 ± 0.09), FTI (0.05 ± 0.02), and TT4 (0.04 ± 0.02) in the second dataset emerged as the influential predictors, whereas, features such as Smoking (0.001 ± 0.002), Hx Smoking (0.0007 ± 0.001), and Hx Radiotherapy (0.00008 ± 0.0001) in the first dataset, and goitre (0.0001 ± 0.0007), lithium (0.00008 ± 0.0005), and hypopituitary (0.0000008 ± 0.000002) in the second dataset had minimal impact on the prediction from both datasets respectively. These findings highlight the importance of the post-treatment behaviour of the patient and tumour-related characteristics, which drive the recurrence rather than baseline patient features. This study demonstrates that the integration of balanced pre-processing, optimized classifiers, and XAI techniques enabled reliable and interpretable thyroid cancer recurrence. The proposed experimental framework can be made accessible, practical, and well suited for integration into routine follow-up workflows using standard clinical data. While the performances are all very strong, the lack of diversity in demographics may make the generalization of our findings to populations different from those in the dataset challenging. Future work will include prospective validation studies on patient cohorts over time, integration with electronic health record systems for real-time risk assessment, inter-institutional validation on a wide range of patient demographics for further improved clinical validity and generalizability in the context of personalized medicine practices, and extending the evaluation to larger and more diverse cohorts with a view to further validating the proposed approach.

Author Contributions: Conceptualization, D.R.; methodology, G.P.R.; formal analysis, Y.V.P.K.; software, K.P.P.; validation, M.K.C.; data curation, P.R.C.; resources, Y.V.P.K.; investigation, D.R.; writing—original draft preparation, D.R. and G.P.R.; writing—review and editing, M.K.C. and K.P.P.; visualization, K.P.P.; supervision, G.P.R. and Y.V.P.K.; project administration, P.R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article. The source code used to implement the proposed methodology is available at: https://github.com/rohandgkp/DTC-Codes (25 January 2026).

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Zhang, D.; Zhang, G.; Li, Y.; Liu, H.; Shan, Z.; Teng, W. Prognosis of Differentiated Thyroid Cancer in Patients with Graves' Disease: A Meta-Analysis. *Thyroid Res.* **2025**, *18*, 51. [CrossRef] [PubMed]
2. Reddy, G.P.; Kumar, Y.V.P. Explainable AI (XAI): Explained. In *Proceedings of the 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*; IEEE: Vilnius, Lithuania, 2023; pp. 1–6. [CrossRef]

3. Rohan, D.; Reddy, G.P.; Kumar, Y.V.P.; Prakash, K.P.; Reddy, C.P. An Extensive Experimental Analysis for Heart Disease Prediction Using Artificial Intelligence Techniques. *Sci. Rep.* **2025**, *15*, 6132. [CrossRef] [PubMed]
4. Pradeep Reddy, G.; Rohan, D.; Venkata Pavan Kumar, Y.; Purna Prakash, K.; Kalyan Chakravarthi, M. Artificial Intelligence-Based Approach for Chronic Kidney Disease Detection. *ASEAN Sci. Technol. Rep.* **2025**, *28*, e258012. [CrossRef]
5. Sanjeev, S.; Sai Ponnekanti, G.; Pradeep Reddy, G. Advanced Healthcare System Using Artificial Intelligence. In *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 28–29 January 2021*; IEEE: Noida, India, 2021; pp. 76–81. [CrossRef]
6. Shrestha, K.; Rifat, H.M.J.O.; Biswas, U.; Tiang, J.-J.; Nahid, A.-A. Predicting the Recurrence of Differentiated Thyroid Cancer Using Whale Optimization-Based XGBoost Algorithm. *Diagnostics* **2025**, *15*, 1684. [CrossRef]
7. Fountzilas, E.; Pearce, T.; Baysal, M.A.; Chakraborty, A.; Tsimberidou, A.M. Convergence of Evolving Artificial Intelligence and Machine Learning Techniques in Precision Oncology. *NPJ Digit. Med.* **2025**, *8*, 75. [CrossRef]
8. Schindele, A.; Krebold, A.; Heiß, U.; Nimptsch, K.; Pfaehler, E.; Berr, C.; Bundschuh, R.A.; Wendler, T.; Kertels, O.; Tran-Gia, J.; et al. Interpretable Machine Learning for Thyroid Cancer Recurrence Predicton: Leveraging XGBoost and SHAP Analysis. *Eur. J. Radiol.* **2025**, *186*, 112049. [CrossRef]
9. Penner, M.A.; Berger, D.; Guo, X.; Levman, J. Machine Learning in Differentiated Thyroid Cancer Recurrence and Risk Prediction. *Appl. Sci.* **2025**, *15*, 9397. [CrossRef]
10. Onah, E.; Eze, U.J.; Abdulraheem, A.S.; Ezigbo, U.G.; Amorha, K.C.; Ntie-Kang, F. Optimizing Unsupervised Feature Engineering and Classification Pipelines for Differentiated Thyroid Cancer Recurrence Prediction. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 182. [CrossRef]
11. Shen, P.; Yang, Z.; Sun, J.; Wang, Y.; Qiu, C.; Wang, Y.; Ren, Y.; Liu, S.; Cai, W.; Lu, H.; et al. Explainable Multimodal Deep Learning for Predicting Thyroid Cancer Lateral Lymph Node Metastasis Using Ultrasound Imaging. *Nat. Commun.* **2025**, *16*, 7052. [CrossRef]
12. Obaido, G.; Achilonu, O.; Ogbuokiri, B.; Amadi, C.S.; Habeebullahi, L.; Ohalloran, T.; Chukwu, C.W.; Mienye, E.D.; Aliyu, M.; Fasawe, O.; et al. An Improved Framework for Detecting Thyroid Disease Using Filter-Based Feature Selection and Stacking Ensemble. *IEEE Access* **2024**, *12*, 89098–89112. [CrossRef]
13. Clark, E.; Price, S.; Lucena, T.; Haberlein, B.; Wahbeh, A.; Seetan, R. Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach. *Knowledge* **2024**, *4*, 557–570. [CrossRef]
14. Alhashmi, S.M.; Polash, M.S.I.; Haque, A.; Rabbe, F.; Hossen, S.; Faruqui, N.; Abaker Targio Hashem, I.; Fathima Abubacker, N. Survival Analysis of Thyroid Cancer Patients Using Machine Learning Algorithms. *IEEE Access* **2024**, *12*, 61978–61990. [CrossRef]
15. Książek, W. Explainable Thyroid Cancer Diagnosis Through Two-Level Machine Learning Optimization with an Improved Naked Mole-Rat Algorithm. *Cancers* **2024**, *16*, 4128. [CrossRef] [PubMed]
16. Liu, J.; Feng, Z.; Gao, R.; Liu, P.; Meng, F.; Fan, L.; Liu, L.; Du, Y. Analysis of Risk Factors for Papillary Thyroid Carcinoma and the Association with Thyroid Function Indicators. *Front. Endocrinol.* **2024**, *15*, 1429932. [CrossRef] [PubMed]
17. Barfejani, A.H.; Rahimi, M.; Safdari, H.; Gholizadeh, S.; Borzooei, S.; Roshanaei, G.; Golparian, M.; Tarokhian, A. Thy-DAMP: Deep Artificial Neural Network Model for Prediction of Thyroid Cancer Mortality. *Eur. Arch. Otorhinolaryngol.* **2025**, *282*, 1577–1583. [CrossRef] [PubMed]
18. Wang, H.; Zhang, C.; Li, Q.; Tian, T.; Huang, R.; Qiu, J.; Tian, R. Development and Validation of Prediction Models for Papillary Thyroid Cancer Structural Recurrence Using Machine Learning Approaches. *BMC Cancer* **2024**, *24*, 427. [CrossRef]
19. Chun, L.; Wang, D.; He, L.; Li, D.; Fu, Z.; Xue, S.; Su, X.; Zhou, J. Explainable Machine Learning Model for Predicting Paratracheal Lymph Node Metastasis in cN0 Papillary Thyroid Cancer. *Sci. Rep.* **2024**, *14*, 22361. [CrossRef]
20. Kumari, P.; Kaur, B.; Rakhra, M.; Deka, A.; Byeon, H.; Asenso, E.; Rawat, A.K. Explainable Artificial Intelligence and Machine Learning Algorithms for Classification of Thyroid Disease. *Discov. Appl. Sci.* **2024**, *6*, 360. [CrossRef]
21. Isiaka, R.M.; Babatunde, R.S.; Ajao, J.F.; Yusuff, S.R.; Popoola, D.D.; Arowolo, M.O.; Adewole, K.S. A Study of Impacts of Artificial Intelligence on COVID-19 Prediction, Diagnosis, Treatment, and Prognosis. *J. Adv. Math. Comput. Sci.* **2022**, *10*, 11–62. [CrossRef]
22. Abdulsalam, S.O.; Arowolo, M.O.; Udofot, E.C.; Sanni, A.M.; Popoola, D.D.; Adebiyi, M.O. A KNN and ANN Model for Predicting Heart Diseases. In *Explainable Artificial Intelligence in Medical Decision Support Systems*; Imoize, A.L., Hemanth, J., Do, D.-T., Sur, S.N., Eds.; Institution of Engineering and Technology: Stevenage, UK, 2022; pp. 335–356, ISBN 978-1-83953-620-5. [CrossRef]
23. Borzooei, S.; Briganti, G.; Golparian, M.; Lechien, J.R.; Tarokhian, A. Machine Learning for Risk Stratification of Thyroid Cancer Patients: A 15-Year Cohort Study. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 2095–2104. [CrossRef]
24. Thyroid Disease Data Set. Available online: https://www.kaggle.com/datasets/yasserhessein/thyroid-disease-data-set/data (accessed on 18 January 2026).

25. Sasada, T.; Liu, Z.; Baba, T.; Hatano, K.; Kimura, Y. A Resampling Method for Imbalanced Datasets Considering Noise and Overlap. *Procedia Comput. Sci.* **2020**, *176*, 420–429. [CrossRef]
26. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/engproc2026124013