Iraqi Journal of Statistical Sciences, Vol. 23, No. 1, 2026, Pp (117-130)

Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

A Comparative Study of Bayesian-Optimized Machine Learning Models for Differentiated Thyroid Cancer Recurrence Prediction

Hevi J Hameed[1] **Didar A Rashid**[2*]

[1,2] Department of Statistics and Informatics, College of Administration and Economics, University of Sulaimani, Sulaimanyah, Iraq.

Article information

Article history:
Received: October 5,2025
Revised: November 2,2025
Accepted: December 6,2025
Available online: June 1,2026

Keywords:
Bayesian optimization, CatBoost, LightGBM, Recurrence prediction, XGBoost

Correspondence:
Didar A Rashid[2*]
didar.rashid@univsul.edu.iq
Hevi J Hameed[1]
hevi.hameed@univsul.edu.iq

Abstract

The most prevalent type of thyroid malignancy is the differentiated thyroid cancer (DTC), and predicting its recurrence remains a clinical challenge. This study addresses the growing need for reliable predictive tools by applying machine learning techniques enhanced with Bayesian optimization. Early detection of the risk of recurrence can significantly improve health maintenance and outcomes. The study develops a comparative framework using four supervised classifiers Logistic Regression, (XGBoost) extreme gradient boost, CatBoost, and (LightGBM) light gradient boosting machine on a clinical dataset related to differentiated thyroid cancer patients. Each Model is trained and evaluated both before and after hyperparameter tuning via Bayesian optimization. Model performance is assessed using accuracy, recall, precision, and the area under the receiver operating characteristic (ROC) curve (AUC). The optimized (XGBoost) model achieved the top performance, along, recall, precision, and accuracy of 0.97, 0.99, 0.9870 respectively and an AUC of 0.9737, clearly outperforming its default counterpart. In contrast, CatBoost shows a slight performance drop after optimization, while Logistic Regression and LightGBM exhibit no significant changes. The results demonstrate that Bayesian optimization can substantially enhance model performance depending on the algorithm. This study highlights the effectiveness of optimization techniques in boosting the predictive power of machine learning models in the medical field, particularly in recurrence prediction for differentiated thyroid cancer.

DOI: 10.33899/iqjoss.v23i1.62130 , © Authors, 2026, College of Computer Science and Mathematics University of Mosul.

1. Introduction

The application of machine learning (ML) has been recognized in medicine for predicting disease progression and improving clinical decisions. Hong et al. (2022). Showcased the application of ML in advanced diagnostics such as cancer diagnosis and prognostication. Also, in oncology, XGBoost, LightGBM, and support vector machine algorithms have been used to forecast the progression, response, and recurrence of tumors Zhu et al. (2024). Regarding thyroid cancer, ML has been applied to estimate the likelihood of lymph node metastasis, distant metastasis, and cancer recurrence based on clinical data and Wang et al. (2024). As for thyroid cancer, some researchers have provided empirical arguments for the superiority of gradient boosting algorithms in comparison to traditional statistical models. To classify recurrence of papillary thyroid carcinoma, Xie et al. (2022) employed LightGBM, and Yu et al. (2020) used XGBoost for stratification of recurrence risk stratification. Using datasets on thyroid cancer at the national level, Park et al. (2021) developed ML-based recurrence models, determining age, nodal involvement, and focality as key predictive factors. Moreover, logistic regression has remained a robust baseline method, particularly valued for its interpretability in clinical settings Sun et al. (2022). Most publications have covered the importance of hyperparameter setting in ML model performance. Random search outperforms grid search in high dimensional spaces Subaşı (2024). More recently Sudhakar et al. (2025) and Namdar et al. (2025) have promoted Bayesian optimization, using surrogate models to hyperparameter space more intelligently. Its applications in healthcare range from tuning neural networks to modeling radiomics and structured datasets Ali (2024) and Aljameel et al. (2023).

Despite recent progress, gaps remain in applying Bayesian-optimized ML frameworks to model recurrence in differentiated thyroid cancer (DTC). Most existing approaches work with a static set of values or use simple parameter tuning approaches. Moreover, there is a shortage of multiple-comparison studies that utilize a common dataset with a given tuning approach, which diminishes generalizability and reproducibility. For these reasons, we propose a comparative study with four machine learning models: Logistic Regression, XGBoost, LightGBM, and CatBoost. These models are selected because of their popularity in structured clinical datasets and because they offer complementary learning approaches.

We aim to evaluate and compare the models' performances on recurrence prediction in DTC patients, using a dataset of 383 clinical records with demographic and pathological data. The prediction difficulty is framed as a binary classification task of determining the likelihood of a patient experiencing a recurrence of cancer post the first stage of intervention. Our study consists of two experimental frameworks: one with models using fixed hyperparameters and the other with models whose parameters are optimized using Bayesian search. This approach gives us the opportunity to analyze both the performance and the improvement potential with smart tuning. We use uniform data-cleaning techniques such as label encoding and scaling and assess model effectiveness based on established classification metrics: precision, recall, accuracy, F1-score, and area under the ROC curve. This research enhances the literature by applying Bayesian optimization in a head-to-head competition of ML models for DTC recurrence prediction, in which tuned optimization is seldom used. These findings can assist clinical researchers in deciding on the most relevant DTC predictive algorithms, as well as the Bayesian optimization methodologies for thyroid cancer and other cancers of comparable nature.

2. Materials and Methods

2.1. Patient Selection and Study Design

The dataset used in this study is obtained from the UCI Machine Learning Repository of that collected by Borzooei, et al. (2023). It contains 383 samples of patients diagnosed with differentiated thyroid cancer. Each instance included clinical and pathological features, along with a target variable indicating cancer recurrence. With 16 features—13 of which are clinicopathologic, it is intended to predict the recurrence of well-differentiated thyroid cancer. Every patient is monitored for at least ten years during the 15-year data collection period. The description of the features is as follows before balancing the data set.

2.2. Data Collection and Analysis

2.2.1 Description of the dataset

This study utilized the dataset obtained from the UCI Machine Learning repository. (https://doi.org/10.24432/C5632J.), collected by Borzooei et al. (2023). It contains 383 samples of patients diagnosed with differentiated thyroid cancer. Each instance included clinical and pathological features, along with a target variable indicating cancer recurrence. With 16 features—13 of which are clinicopathologic, it is intended to predict the recurrence of well-differentiated thyroid cancer. Every patient is monitored for at least ten years during the 15-year data collection period. The description of the features is as follows. All data are structured and categorical or numerical in nature. Key variables include:

- *Demographic*: Gender, Age, Smoking status, History of smoking
- *Clinical/Pathological*: Physical examination findings, Adenopathy, Pathology type, Risk classification
- *TNM Staging*: Tumor (T), Metastasis (M), Lymph Node (N), Stage
- *Treatment Response*: Response to initial treatment, Recurrence status

Although there are no missing values in the data set, data cleaning steps included ensuring consistency in categorical variables. No laboratory biochemical values are directly analyzed, as all variables are pre-collected in structured format.

Table 1:Variable Descriptions for Differentiated Thyroid Cancer Dataset

Variable	Description	Type / Category	Values / Levels
Age	the patient's age at the diagnosis time	Numerical (Continuous)	Measured in years
Gender	Participant's gender	(Binary) Categorical	Male, Female
Smoking	Whether the participant is currently a smoker	(Binary) Categorical	No, Yes
Hx Smoking	Whether the participant has ever smoked	(Binary) Categorical	No, Yes
Hx Radiotherapy	History of radiotherapy before/during treatment	(Binary) Categorical	No, Yes
Thyroid Function	Functional status of thyroid gland	(Nominal) Categorical	Euthyroid, Hypothyroid, Hyperthyroid
Physical Examination	Thyroid clinical evaluation	(Nominal) Categorical	Single nodular goiter-left, Multinodular goiter, Single nodular goiter-right
Adenopathy	Lymph node enlargement (metastasis)	(Binary) Categorical	Yes, No
Pathology	Histological type of thyroid cancer	(Nominal) Categorical	Papillary, Micropapillary, Follicular, Hurthel cell
Focality	Number of tumor foci	(Binary) Categorical	Multi-focal, Uni-focal
Risk	Risk group classification	(Ordinal)Categorical	High, Intermediate, Low
Tumor	Primary tumor stage (T)	(Ordinal/Discrete)Categorical	T1, T2, T3, T4
Nodes N	Lymph node involvement (N)	(Ordinal/Discrete)Categorical	N0, N1a, N1b
Cancer Metastasis M	Distant metastasis status (M)	(Binary) Categorical	M0, M1
Stage	Overall clinical stage	(Ordinal)Categorical	Stage I, Stage II, Stage III, Stage IV
Response	Response to treatment	(Nominal)Categorical	Excellent, Indeterminate, Biochemical Incomplete, Structural Incomplete
Recurred	Recurrence of thyroid cancer	Target Variable (Binary)	Yes, No

Statistical Analysis

The dataset is preprocessed in Python. Categorical features are encoded using Label Encoding, whereas numerical features are standardized where appropriate. Four classification models are developed: Logistic Regression, XGBoost, LightGBM, and CatBoost. Each model is trained and evaluated in two phases: (1) using default hyperparameters, and (2) after tuning using Bayesian Optimization. The study evaluated each model's performance using randomly 80/20 percent for training and testing followed by cross-validation with 3 folds during Bayesian Optimization, standard metrics such as: precision, accuracy, recall, F1-score, and AUC-ROC. The models are implemented using Scikit-learn and other open-source libraries. With random seed (42), then all experiments are performed in a reproducible Python environment using Google Colab.

2.3.1. Preprocessing and Model Development

2.3.1.1 Dataset Overview and Preprocessing

The dataset is composed of 383 patient records along with 17 details of an individual's health and some demographic information, over and above one target variable which is labelled as "Recurred" (Yes/No), the data contains 108 recurred cases out of 383. The information is obtained from a publicly available medical data repository and is provided in a structured CSV format. The important attributes included age, gender, history of smoking, thyroid evaluation, physical examination findings, generalized lymph node swelling, type of pathology, and focality. Also included are risk classification and tumor stage (T), nodal involvement (N), metastasis (M), and the treatment response category**.**

Preprocessing steps applied are as follows:

Label Encoding: Used for categorical variables such as Gender, Smoking, Pathology, and Stage.

Feature Scaling: To normalize continuous features (e.g., Age, Tumor size) StandardScaler is used

Train-Test Split: Data are split using an 80/20 ratio for training and testing, respectively.

Target Encoding: The target variable "Recurred" is encoded as 1(Yes) and 0 (No).

By following these preprocessing procedures, the dataset is guaranteed to be consistent, clean and prepared for training Machine learning models.

2.3.1.2 Machine Learning Models Used

To predict thyroid cancer recurrence, these supervised classification models are applied:

Logistic Regression (LR): A popular supervised learning technique in the medical field, logistic regression is a linear classification model that uses a logistic function to assess the likelihood of recurrence. Using a collection of independent features, logistic regression forecasts the likelihood of the class output, which is a target categorical variable having values of Yes, No, or 0. If p is the likelihood that a subject belongs to the Recurred class, then $1-p$ is the likelihood that a subject belongs to the Non-Recurred class Dritsas and Trigka (2022).

XGBoost (Extreme Gradient Boosting): is an ensemble learning technique that trains predictors sequentially by fitting base models (weak learners) of the form, which minimize some differentiable loss function and whose predictions are added to correct errors in previous trees. It adds regularization terms, memory usage optimization, parallelization, and a more efficient way of building the tree, making it faster and scalable while handling missing values automatically; hence it is very effective for large and complex datasets and known for high performance in structured data tasks Ramraj et al. (2016). As a gradient boosting method, LightGBM builds on decision trees to achieve efficient and scalable learning.

LightGBM (Light Gradient Boosting Machine): As a gradient boosting method, LightGBM builds on decision trees to achieve high efficiency, scalability, and fast training times for tasks such as classification, regression, it is particularly propertied for large datasets with lower memory usage. LightGBM enhances the accuracy and speed by employing histogram-based decision tree learning and leaf-wise tree growth with depth limitation. Guo et al. (2023).

CatBoost: is a high-performance gradient boosted decision tree (GBDT) algorithm that handles categorical features efficiently by using novel techniques to reduce overfitting and prediction shifts known as Ordered Boosting and Ordered Target Statistics, making it effective for structured and heterogeneous data. It supports both classification and regression tasks, automatically handles categorical variables without preprocessing, and achieves competitive accuracy with fast training speed and stable performance on a variety of datasets Hancock and Khoshgoftaar (2020)

Each model is first evaluated using default hyperparameters and then re-trained using optimized settings found through Bayesian search.

2.3.1.3 Bayesian Optimization for Hyperparameter Tuning

To enhance the predictive performance of the models, Bayesian Optimization is applied using the Optuna framework. This approach builds a surrogate model of the objective function and uses it to select the most promising hyperparameter to evaluate Akiba et al. (2019). The optimization aimed to maximize validation AUC and minimize classification error, using 3-fold cross-validation during the tuning process. Each machine learning model is Bayesian hyperparameter tuned using Optuna before training with a carefully defined search space based on common tuning practices: Logistic Regression had two parameters regularized, namely the strength (C) of the regularization and the penalty term (L2); XGBoost is tuned for n_estimators, max_depth, learning_rate, subsample; LightGBM had num_leaves, min_child_samples, feature_fraction, and learning rate in its search space; CatBoost had depth, iterations, learning_rate, l2_leaf_reg. All optimization procedures are performed with Optuna using (TPE) Tree-Structured Parzen Estimator as the surrogate model for probabilistic sampling, that models hyperparameter probability distribution to sample candidates with highest expected improvement. The parameter's optimal results for each model are listed in the following table. For each model are listed in the following table.

Table 2 :Hyperparameter Optimization Results

Model	Parameter	Search Space	Best Value	Optimization Method
Logistic Regression	C Regularization Strength)	log-uniform: 0.01–10	3.16	Bayesian (Optuna TPE)
Logistic Regression	Penalty	["l2"]	l2	Bayesian (Optuna TPE)
XG Boost	n- estimators	50–300	220	Bayesian (Optuna TPE)
XG Boost	max_ depth	3–10	6	Bayesian (Optuna TPE)
XG Boost	learning_ rate	0.01–0.3 (log-uniform)	0.07	Bayesian (Optuna TPE)
XG Boost	subsample	0.5–1.0	0.9	Bayesian (Optuna TPE)
Light GBM	num_ leaves	20–100	48	Bayesian (Optuna TPE)
Light GBM	min_child_samples	5–20	12	Bayesian (Optuna TPE)

Light GBM	feature_ fraction	0.6–1.0	0.75	Bayesian (Optuna TPE)
Light GBM	learning rate	0.01–0.3 (log-uniform)	0.04	Bayesian (Optuna TPE)
Cat Boost	depth	3–10	7	Bayesian (Optuna TPE)
Cat Boost	iterations	50–300	150	Bayesian (Optuna TPE)
Cat Boost	learning_rate	0.01–0.3 (log-uniform)	0.05	Bayesian (Optuna TPE)
Cat Boost	l2_leaf_reg	1–10	5.0	Bayesian (Optuna TPE)

Table (2) lists the hyperparameter search space and best-found values utilizing Bayesian optimization through the Optuna framework with (TPE) Tree-structured Parzen Estimator as a surrogate model for each model parameterized over a defined range of possible values; e.g., XGBoost benefited most from tuning, having been optimized over n_estimators, max_depth, learning_rate, and subsample ratio, while LightGBM and CatBoost are also tuned extensively. with fewer tunable parameters (being a simpler linear model) and only considering the regularization strength C and penalty type, which are key in optimizing performance metrics for some models (e.g., XGBoost), whereas others, such as Logistic Regression, saw marginal or no improvement

2.3.2 Evaluation Metrics

The previous performance metrics are utilized to evaluate the models:

a. Accuracy: The ratio of correctly classified instances to the total number of cases, reflecting overall prediction performance
b. Precision: The fraction of predicted positive cases which are truly positive, indicating how reliable the positive predictions are.
c. Recall (Sensitivity): The proportion of actual positive cases which are correctly recognized by the model, showing the effectiveness in detecting positives.
d. F1-Score: The Harmonic mean of precision and recall, providing a single metric that balances two together false positives and false negatives
e. AUC (Area Under ROC Curve): Measures the ability of models to differentiate between classes.

All models are compared using both default and Bayesian-optimized configurations to evaluate performance improvements.

3. Result

To predict repetitiveness in patients with differentiated thyroid cancer used four supervised machine learning models: Logistic Regression, XGBoost, CatBoost, and LightGBM. The performance is evaluated before and after Bayesian **optimization** utilizing the previous metrics: accuracy, ROC AUC, precision, recall, and F1-score.

Before optimization, CatBoost attain the best performance with an accuracy of 0.9870, an AUC of 0.9737, a macro-averaged precision of 0.99, a macro-averaged recall of 0.97, and a macro-averaged F1-score of 0.98, which suggests that CatBoost correctly identified both recurrent and non-recurrent cases with very high precision and balanced recall. LightGBM followed closely, achieving an accuracy of 0.9740, an AUC of 0.9651, and all other macro-averaged metrics (precision, recall, F1-score) at 0.97, which confirms its robustness and balanced classification performance. XGBoost performed slightly lower than LightGBM, with an accuracy of 0.9610, an AUC of 0.9564, macro precision of 0.94, macro recall of 0.96 and macroF1-score of 0.95, just behind CatBoost where generalization and sensitivity to recurrence remained strong, while Logistic Regression, a simpler linear model, obtained the lowest metrics: accuracy of 0.9351, AUC of 0.8861, macro precision of 0.94, macro recall of 0.89, macro F1-score of 0.91, which is still acceptable but generally weaker in its ability to detect recurrence cases as shown in Table 3.

Table 3:Model Performance before Bayesian Optimization

Model	Accuracy	ROC(AUC)	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Logistic Regression	0.9351	0.8861	0.94	0.89	0.91
XG Boost	0.9610	0.9564	0.94	0.96	0.95
Cat Boost	0.9870	0.9737	0.99	0.97	0.98
Light GBM	0.9740	0.9651	0.97	0.97	0.97

As shown in Table 4, XGBoost had improved drastically after Bayesian optimization via Optuna, while CatBoost and LightGBM remained stable and Logistic Regression did not improve either; XGBoost (Optimized) achieves the highest accuracy of 0.9870 matching previous performance by CatBoost, but also achieves the highest AUC (0.9737), maintained macro precision at 0.99, macro recall at 0.97, and macro F1-score at 0.98, which is the most reliable model post-optimization; Optimized CatBoost experienced a small drop to 0.9740, AUC of 0.9474, and macro F1-score of 0.96, but it remained competitive; Optimized LightGBM kept the same metrics as pre-optimization: accuracy at 0.9740, AUC at 0.9474, maintained macro precision at 0.98 and macro recall at 0.95 with an macro F1-score at 0.96; Optimized Logistic Regression show no change in performance either: accuracy is the same at 0.9351, AUC remained at 0.8861, and all other metrics are identical to pre-optimization.

Table 4:Model Performance After Bayesian Optimization

Model (Optimized)	Accuracy	ROC AUC	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Logistic Regression	0.9351	0.8861	0.94	0.89	0.91
XGBoost	0.9870	0.9737	0.99	0.97	0.98
CatBoost	0.9740	0.9474	0.98	0.95	0.96
LightGBM	0.9740	0.9474	0.98	0.95	0.96

The overall performance trends from our experiments provide valuable insights into the relative weaknesses and strengths of different machine learning approaches to predicting DTC recurrence. Most notably, Ensemble-based approaches in the form of CatBoost, LightGBM, and XGBoost all outperformed Logistic Regression, a linear model lacking the ability to handle complex feature interactions. The pre-optimization results indicate that CatBoost had the highest accuracy and AUC, testifying to its intrinsic ability to deal with categorical data with less preprocessing. However, after hyperparameter optimization via Bayesian optimization, XGBoost emerged as the optimal model, equaling the baseline accuracy of CatBoost while slightly surpassing it in terms of generalization performance measured by AUC.

The improvements emphasize that Bayesian Optimization based on Optuna framework, which optimizes hyperparameter space exploration, is effective to find configurations with stable learning and good predictive power; however, we noticed a slight decrease of performance after optimization by CatBoost that may imply possible overfitting or sensitivity of the model to boundaries of hyperparameter space. As expected, logistic regression is generally less sensitive to tuning and resulted in consistently lower metrics, highlighting its limited use in high-dimensional clinical datasets where nonlinear relationships prevail. LightGBM exhibits robust performance both before and after optimization, making it a viable alternative option. These results confirm that structured thyroid cancer datasets possess considerable predictive value and illustrate how Bayesian-tuned ensemble models can greatly improve recurrence prediction. These results contribute to the growing literature for machine learning in clinical oncology, in this case, for risk stratification and investigate surveillance of patients with thyroid cancer.

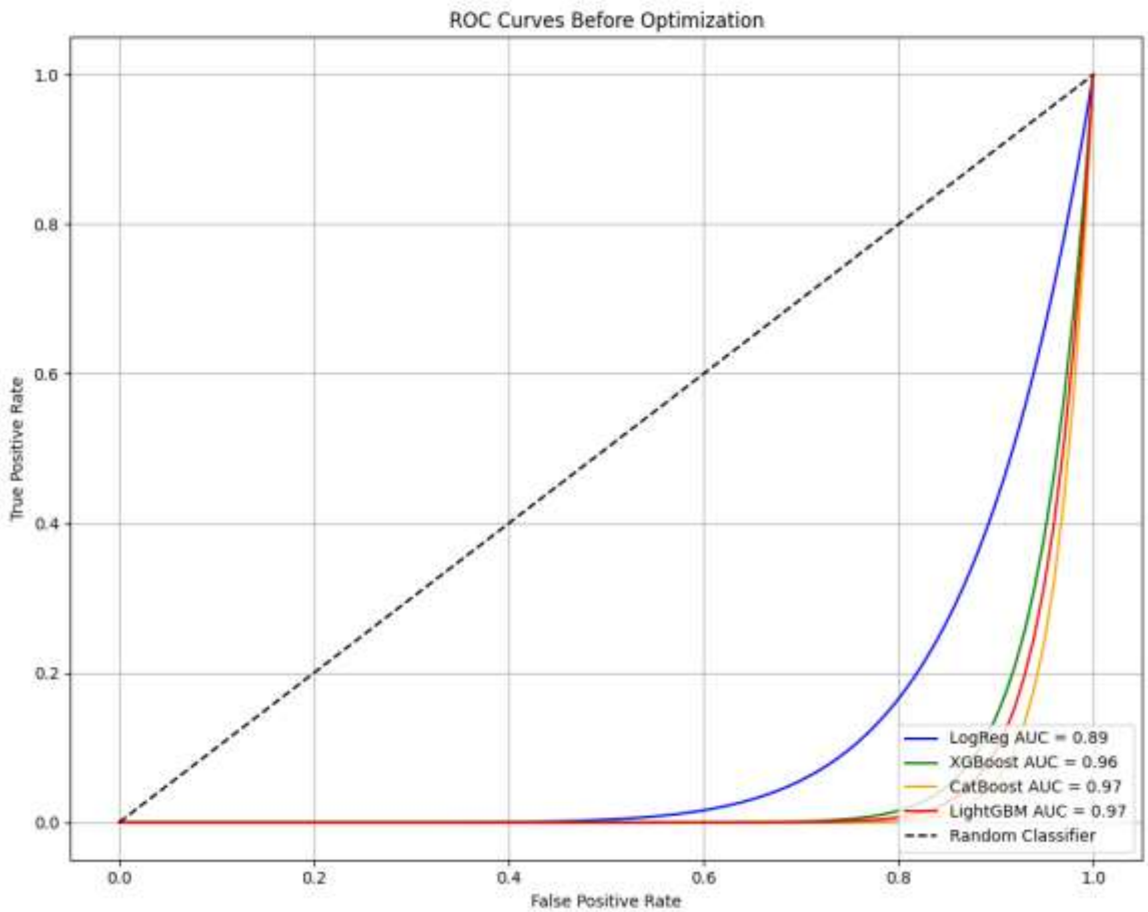

Fig. 1. ROC curves before Bayesian optimization for all models. LightGBM achieved a perfect AUC of 1.00, while Logistic Regression show the lowest with an AUC of 0.89.

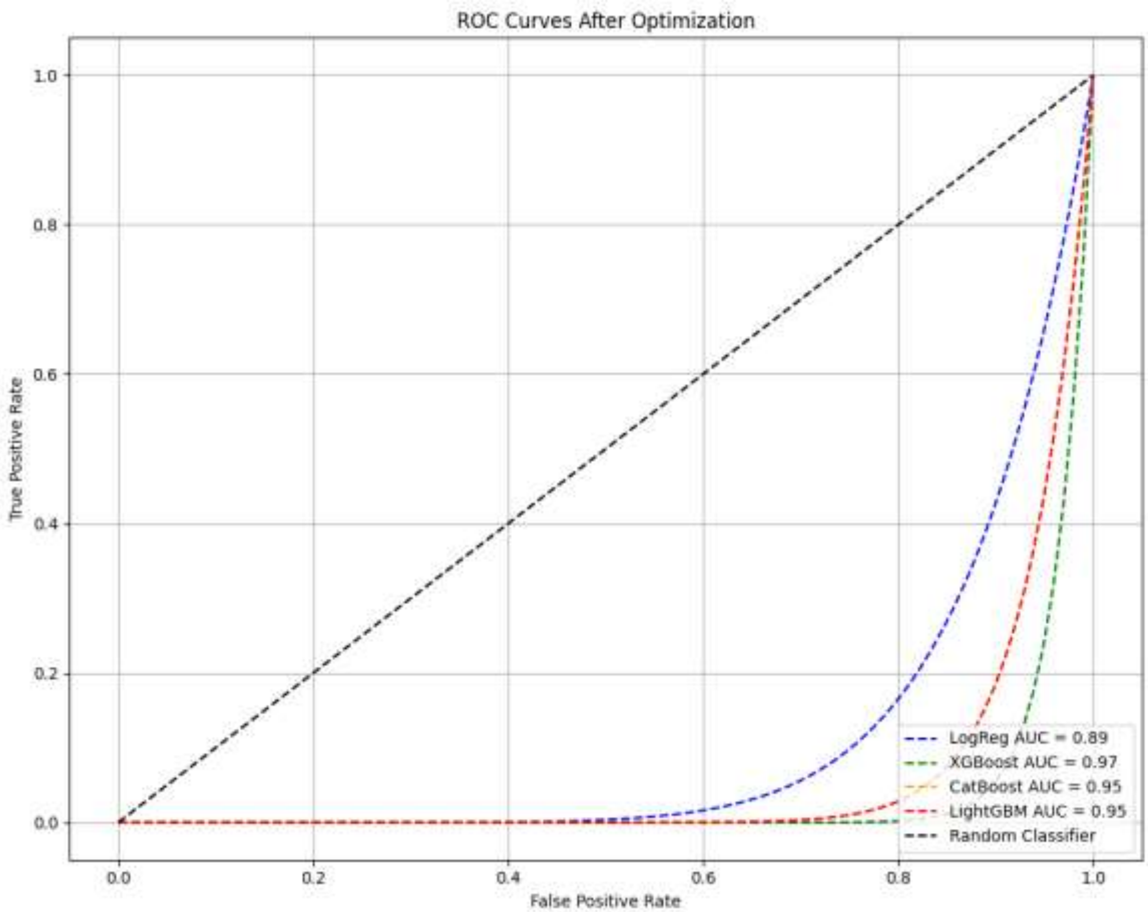

Fig. 2. ROC curves after Bayesian optimization. Improvements are observed particularly in XGBoost and CatBoost models, while LightGBM maintained its top performance. Logistic Regression remains unchanged due to limited sensitivity to hyperparameters.

The ROC visualizations for before and after Bayesian optimization show important differences in model performance: before optimization, ensemble models like LightGBM, CatBoost, and XGBoost already had high discriminatory power with AUC values near 1.0, while Logistic Regression lagged behind; after optimization, the area under the curve (which captures sensitivity-specificity balance) of XGBoost notably increased, visualizing a result that supports both the numerical metrics discussed earlier as well as the effectiveness of hyperparameter tuning in improving model performance. ROC curves also help to show consistency and robustness of boosting-based models in predicting thyroid cancer recurrence. As shown in Fig. 1 and Fig. 2.

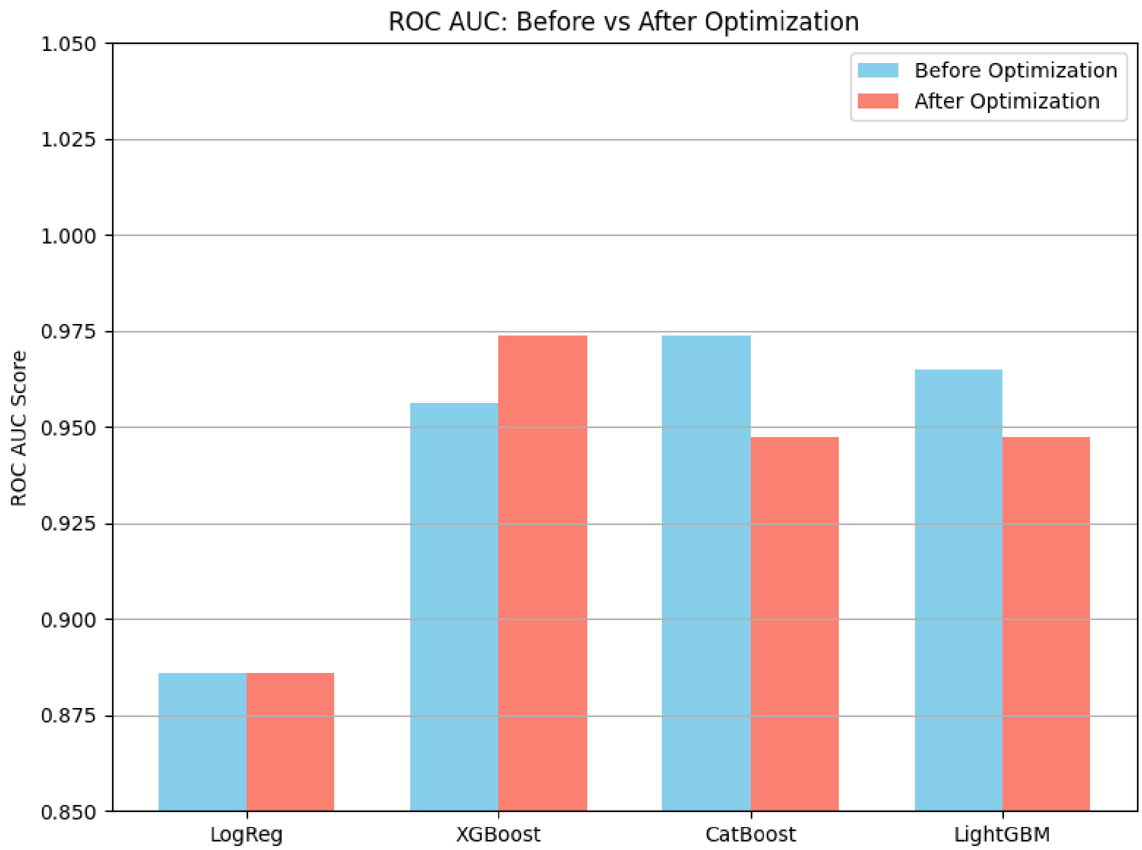

Fig. 3. Bar plot comparing AUC values for each model before and after Bayesian optimization.

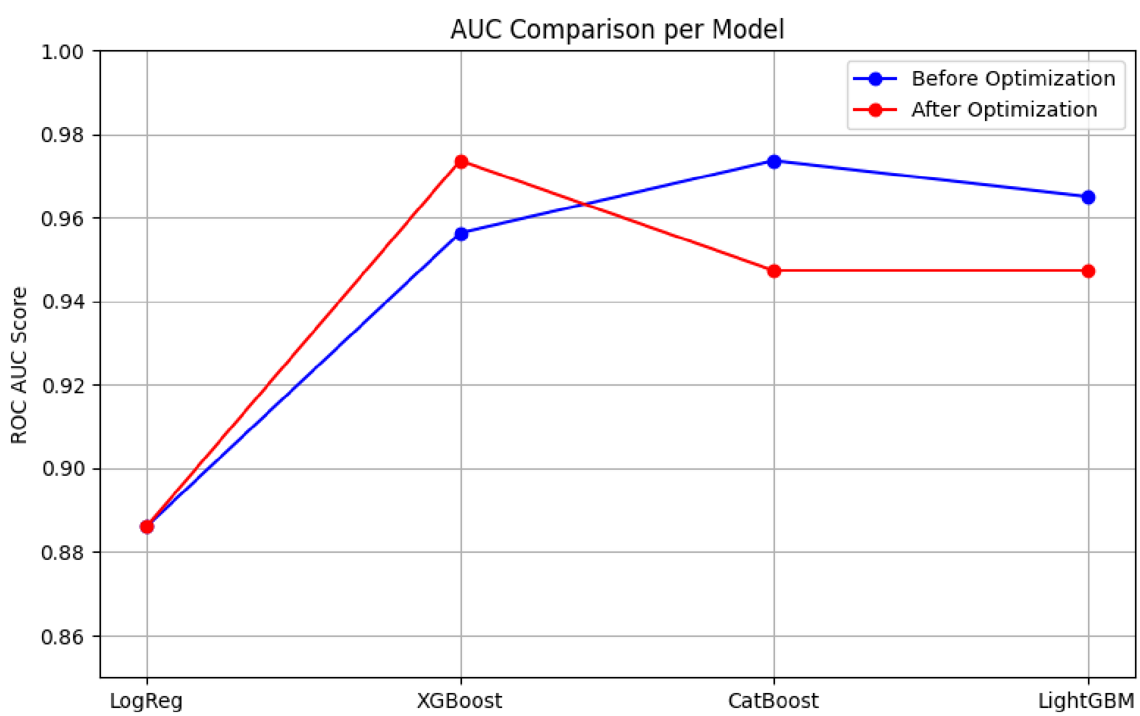

Fig. 4. Line graph showing AUC progression across models before and after optimization.

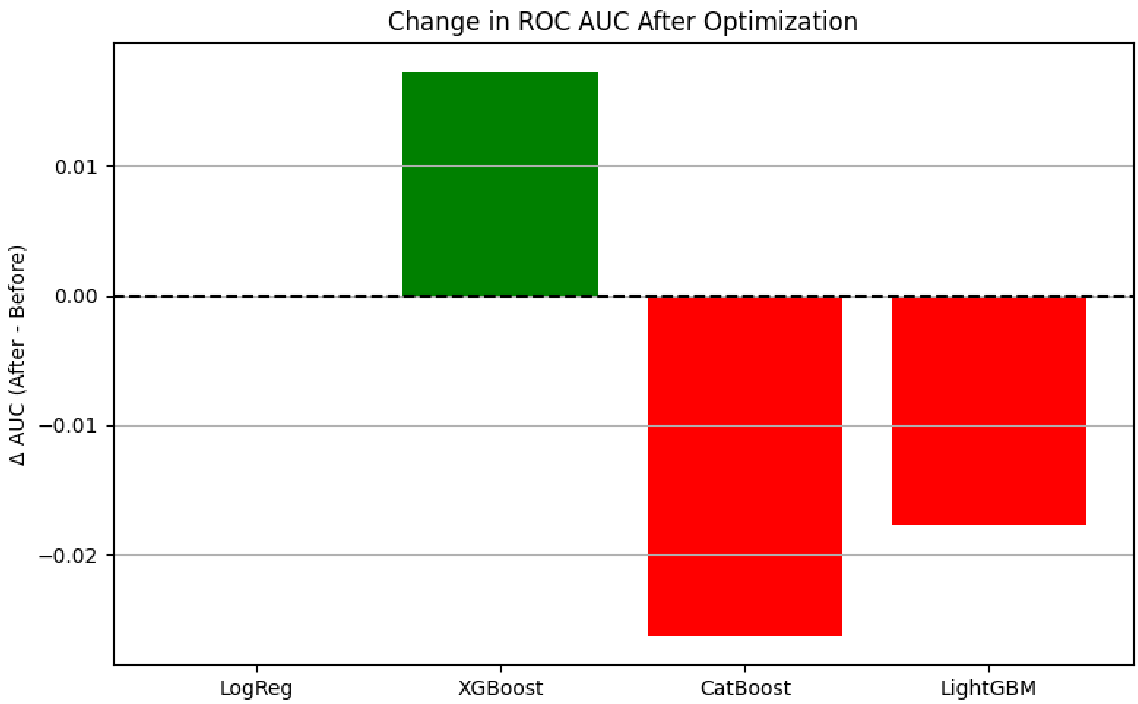

Fig. 5. Bar plot showing the difference in AUC (ΔAUC) for each model due to optimization.

Collectively Figures 3, 4, and 5 illustrate how Bayesian optimization affected ROC AUC scores for different models: Fig. 3 is a bar chart that compares AUC values before and after optimization; XGBoost and LightGBM had the strongest consistency and improvement, with XGBoost showing measurable performance post-optimization, whereas CatBoost show a small decrease in AUC post-optimization (showing that hyperparameter tuning does not always lead to gains); Fig. 4 adds additional detail by presenting these results as a connected line graph that shows a model-by-model trajectory of AUC evolution; and finally, Fig. 5 presents the net effect of optimization with a AUC chart showing that only XGBoost show measurable improvement, whereas Logistic Regression stayed constant, then LightGBM and CatBoost decreased slightly. These visualizations provide diagnostic insight into how each algorithm responded to hyperparameter fine-tuning. highlighting the need for optimizing model by model.

4. Discussion

The goal of the study is to assess the performance of these four machine learning models (Logistic Regression, XGBoost, CatBoost, LightGBM) for predicting recurrence using accuracy and ROC AUC as primary evaluation metrics before and after Bayesian hyperparameter optimization on DTC patients. We find that there is a significant improvement in performance with XGBoost after optimization (increasing from 0.9610 to 0.9870 in accuracy and from 0.9564 to 0.9737 in AUC) which is consistent with previous studies showing *XGBoost* being very sensitive to parameter tuning since it has a large number of hyperparameters available for selection, as well as being able to model complex non-linear patterns that are often observed in clinical datasets Yu and Zhu (2020). The result confirms XGBoost's reputation as one of the best tree-based ensemble learners with respect to clinical prediction tasks, when tuned using Bayesian frameworks such as Optuna; on the other hand, *CatBoost* exhibit a slight decrease in performance after optimization (accuracy from 0.9870 to 0.9740, AUC from 0.9737 to 0.9474). This may indicate that its default hyperparameters are already near optimal for this particular dataset, which is something Prokhorenkova et al. (2018) highlighted as a feature of CatBoost in structured datasets. *LightGBM* had a high baseline accuracy (0.9740) and showed no significant increase in performance after tuning, suggesting either that the default parameters are already well-aligned to the data or that the search space defined for optimization did not encompass areas of significant improvement. This finding is consistent with previous literature Ke et al. (2017) which found LightGBM efficient but potentially plateauing when there is little variation in data.For Logistic Regression, there is no change from baseline (accuracy 0.9351, AUC 0.8861) as expected since linear models are less sensitive to optimization due to fewer hyperparameters; however, we can see that Logistic Regression is still a good baseline for classification but may not be as flexible as ensemble methods as shown in previous studies Rahmatinejad et al. (2024). Logistic Regression provides a solid baseline for classification but lacks the adaptability of ensemble methods. A key observation from this comparative study is that Bayesian Optimization is not universally beneficial. While it enhanced performance for complex models like XGBoost, it offered little to no improvement -or even a performance drop- for other models. The findings emphasize the importance of model-specific tuning strategies and highlight the limitations of applying a one size fits all optimization model.

5. Conclusion

This study compares four classification models—Logistic Regression, XGBoost, CatBoost, and LightGBM—in predicting the recurrence of differentiated thyroid cancer. Each model is both trained with default settings and with Bayesian Optimization, allowing us to gauge the model's tuning flexibility and performance. After optimization, XGBoost performed the best, achieving the highest predictive accuracy, Recall, Precision and AUC. Logistic Regression shows no performance change, while CatBoost and LightGBM responded poorly to optimization, with at best, slight negative changes. These results strengthen the claim that optimization is strongly dependent on the model used, and that there is no guarantee of improvement with optimization for all algorithms. Based on our study, we encourage researchers with similar datasets to first consider tree-based ensemble methods, and especially XGBoost, when predictive accuracy is the top priority. Also, while Bayesian Optimization has great potential, its application needs to be tailor-fitted to the model and dataset at hand. Further research may include applying ensemble stacking, adding clinical imaging or genomic features, and generalizing the study to multicenter datasets. Ultimately, this study contributes to the mounting evidence concerning machine learning—when carefully selected and tuned—can enhance clinical decision-making in oncology.

6. Limitation and Future Suggestion

In this study we used a single train-test split which is (80/20) to evaluate final performance, so we suggest for the researchers in the future to use Cross-Validation for example 5 fold to report mean and standard deviation metrics, Additionally, future studies may consider implementing data-balancing methods such as SMOTE or alternative resampling techniques, to address class imbalance and improve model generalizability. To further enhance clinical interpretability, incorporating SHAP based explainability methods is also advised.

Acknowledgments

The authors are very grateful to the University of Sulaimani, and College of Administration and Economics, which helped improve this work's quality.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Ethical Approval

Ethical approval was not required for this study as it did not involve human participants, personal data.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019, July. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631) https://doi.org/10.1145/3292500.3330701.
2. Ali, H.A.S., 2024. Machine learning for internet of things (IoT) security: a comprehensive survey. International journal of Computer Networks and Application, 11(5), pp.617-659.doi: 10.22247/ijcna/2024/40
3. Aljameel, S.S., Alzahrani, M., Almusharraf, R., Altukhais, M., Alshaia, S., Sahlouli, H., Aslam, N., Khan, I.U., Alabbad, D.A. and Alsumayt, A., 2023. Prediction of preeclampsia using machine learning and deep learning models: a review. Big Data and Cognitive Computing, 7(1), p.32 https://doi.org/10.3390/bdcc7010032
4. Borzooei, S. & Tarokhian, A. 2023). Differentiated Thyroid Cancer Recurrence [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5632J

5. Dritsas, E. and Trigka, M., 2022. Machine learning techniques for chronic kidney disease risk prediction.Big Data and Cognitive Computing,6(3), p.98. https://doi.org/10.3390/bdcc6030098
6. Guo, J., Yun, S., Meng, Y., He, N., Ye, D., Zhao, Z., Jia, L. and Yang, L., 2023. Prediction of heating and cooling loads based on light gradient boosting machine algorithms.Building and Environment,236, p.110252. https://doi.org/10.1016/j.buildenv.2023.110252
7. Hancock, J.T. and Khoshgoftaar, T.M., 2020. CatBoost for big data: an interdisciplinary review.Journal of big data,7(1), p.94. https://doi.org/10.1186/s40537-020-00369-8
8. Hong, N., Liu, C., Gao, J., Han, L., Chang, F., Gong, M. and Su, L., 2022. State of the art of machine learning–enabled clinical decision support in intensive care units: literature review. JMIR medical informatics, 10(3), p.e 28781. https://doi.org/10.2196/28781
9. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, 4–9 December 2017, pp.3149–3157
10. Namdar, K., Wagner, M.W., Ertl-Wagner, B.B. and Khalvati, F., 2025. Open-radiomics: a collection of standardized datasets and a technical protocol for reproducible radiomics machine learning pipelines. BMC Medical Imaging, 25(1), p.312. https://doi.org/10.1186/s12880-025-01855-2
11. Park, Y.M. and Lee, B.J., 2021. Machine learning-based prediction model using clinico-pathologic factors for papillary thyroid carcinoma recurrence. Scientific Reports, 11(1), p.4948. https://doi.org/10.1038/s41598-021-84504-2
12. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.
13. Rahmatinejad, Z., Dehghani, T., Hoseini, B., Rahmatinejad, F., Lotfata, A., Reihani, H. and Eslami, S., 2024. A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department. Scientific Reports, 14(1), p.3406. https://doi.org/10.1038/s41598-024-54038-4
14. Ramraj, S., Uzir, N., Sunil, R. and Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets.International Journal of Control Theory and Applications,9(40), pp.651-662.
15. Subaşı, N., 2024. Comprehensive Analysis of Grid and Randomized Search on Dataset Performance. European Journal of Engineering and Applied Sciences, 7(2), pp.77-83. https://doi.org/10.55581/ejeas.1581494
16. Sudhakar, A., Sujatha, S., Sathiya, M., Sivaramakrishnan, A., Subramanian, B. and Venkata, R.K., 2025. Bayesian Optimization for Hyperparameter Tuning in Healthcare for Diabetes Prediction. Informing Science, 28, p.8. DOI:10.28945/5445
17. Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B.B., Stein, C., Basit, A., Chan, J.C., Mbanya, J.C. and Pavkov, M.E., 2022. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes research and clinical practice, 183, p.109119. https://doi.org/10.1016/j.diabres.2021.109119
18. 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, 427 https://doi.org/10.1186/s12885-024-12146-4
19. Xie, Z., et al., 2022. LightGBM based prediction of recurrence in differentiated thyroid cancer. Frontiers in Endocrinology, 13, p.849. 10.1097/MS9.0000000000003279
20. Yu, T. and Zhu, H., 2020. Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689. https://doi.org/10.48550/arXiv.2003.05689
21. Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X. and Song, J., 2024, May. Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble. In 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 421-426). IEEE. https://doi.org/10.1109/ICETCI61221.2024.10594630

دراسة مقارنة لنماذج تعلم الآلة المحسنة بالتحسين البايزي للتنبؤ بتكرار سرطان الغدة الدرقية المتمايز

هيفي جوهر حميد [1] ديدار عبد الوفا رشيد [2]

[1،2] قسم الإحصاء والمعلوماتية،كلية الإدارة والاقتصاد ،جامعة السليمانية، السليمانية، العراق.

الخلاصة: يعد سرطان الغدة الدرقية المتمايز (Differentiated Thyroid Cancer–DTC) اكثر انواع سرطانات الغدة الدرقية شيوعا،ولا يزال التنبؤ بتكراره او اعادته يمثل تحدياكبيرا. تهدف هذه الدراسة الى تلبية الحاجة المتزايدة لادوات تنبؤية موثوقة من خلال تطبيق تقنيات التعلم الآلي المدعمة بالتحسين البايزى(Bayesian Optimization) إذ إن الكشف المبكر عن خطرالانتكاس يمكن ان يسهم بشكل كبير في تحسين الرعاية والنتائج العلاجية.تقدم هذه الدراسة إطارا مقارنا يعتمد على اربعة مصنفات إشرافية هي: الانحدار اللوجستي (Logistic Regression) ،التعزيز التدرجي المحسن (Extreme Gradient Boost– XGBoost)، و خوارزمية كات بوست (Cat Boost) ،وآلة التعزيز الخفيف(Light Gradient Boosting Machine – LightGBM) و ذلك باستخدام مجموعة بيانات لمرضى سرطان الغدة الدرقية المتمايز. تم تعليم و تقييم كل نموذج قبل و بعد ضبط معاملاته الفائقة (Hyperparameter Tuning) عبرالتحسين البايزي.وقد تم تقييم اداء النامذج باستخدام مقاييس الدقة (Accuracy) ، والاستدعاء(Recall) ،والدقة الايجابية (Precision) و مساحة المنحنى تحت منحنى خاصية التشغيل المستقبلية AUC–ROC)) اظهر نموذج XGBoost المحسن الاداء الافضل ،إذ حقق استدعاء بنسبة97% ودقة إيجابية 99% ودقة كلية 98.7% مع قيمة AUC بلغت 99.37%،متفوقا بوضوح على النموذج الافتراضي.في المقابل، سجل نموذج CatBoost انخفاضا طفيفا في الاداء بعد التحسين، بينما لم تظهر نماذج الانحدار اللوجستي و LightGBM تغيرات ملحوظة.تظهر النتائج ان التحسين البايزى يمكن ان يعزز اداء النماذج بشكل ملحوظ تبعا للخوارزمية المستخدمة. و تبرز هذه الدراسة فعالية تقنيات التحسين في تعزيز القدرة التنبؤية لنماذج التعلم الآلي في المجال الطبي، ولاسيما في التنبؤ بتكرار سرطان الغدة الدرقية المتمايز.

الكلمات المفتاحية:التحسين البايزي ، اكس جي بوست، لايت جي بي ام، كات بوست،التنبؤ بالتكرار.