Leveraging Machine Learning to Predict Treatment Success in patients with thyroid cancer

Mohammad Shihabul Islam
Department of Management Information Systems
University of Dhaka
shihabulislam2018@gmail.com

***Abstract*—The variability in treatment outcomes for thyroid cancer patients highlights the need for predictive models to assist clinicians in making informed decisions. This study employs multiple machine learning algorithms with hyperparameter tuning to identify the best-performing model. The XGBoost model emerged as the most effective, achieving a 98% accuracy and an Macro Avg F1-score of 98%. Feature importance analysis revealed that the most critical predictor was the `FirstRecurrenceType`. By leveraging machine learning and optimizing model performance, this study demonstrates the potential to enhance clinical decision-making and enable personalized treatment strategies for thyroid cancer patients, marking a significant step toward advancing predictive analytics in healthcare.**

***Index Terms*—XGBoost, Hyperparameter Tuning, SHAP, ROC, DTC**

I. INTRODUCTION

Thyroid cancer, particularly differentiated thyroid cancer (DTC), has emerged as the most prevalent form of endocrine malignancy. Despite its generally favourable prognosis, recurrence remains a significant concern, affecting approximately 10-30% of the patients. Accurate prediction of recurrence is crucial to improve patient outcomes and tailor personalized treatment strategies. Recent advances in machine learning (ML) and artificial intelligence (AI) have shown promising results in addressing this challenge.

Outcome prediction of various treatment interventions like surgical, radiation and hormonal therapy are essential to address from the tumour to devise individualized treatment plans. Recent developments in the field of ML give a chance to look at patient's data and reveal peculiarities that are difficult to point out by common statistical analysis. This study applies ML towards making the evaluation of patient treatment for thyroid cancer so that clinicians can make the right decision that would offer the patients the best results.

A. Problem Statement

The variability in treatment outcomes for thyroid cancer patients highlights the need for predictive models that can assist clinicians in making more informed decisions. Current clinical decision-making often relies on general guidelines and the clinician's experience, which may not account for the individual variability among patients. Predictive models using ML can analyze large datasets to identify patterns and factors that influence treatment success, providing personalized predictions for patients.

B. Objective

The primary objective of this research is to develop and evaluate an ML models that are used in predicting outcome of given treatment options in patients diagnosed with thyroid cancer. Considering a large set of features including patient characteristics, as well as diagnosis and treatment information, this work will open the possibility to develop models to predict the treatment efficacy by aiming at the survival of the patient and the cancer recurrence. The ultimate aim is to develop a tool that will help clinicians to come up with the right treatment plan with an aim to enhance the quality of the life of the patients as well as to manage the scarce health resources well.

C. Literature Review

Thyroid cancer, particularly DTC, has been the focus of numerous studies exploring the application of ML and AI techniques to predict recurrence and improve treatment strategies. Elizabeth et al. (2024) investigated the potential of six ML models, including Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), to predict DTC recurrence. The study used the Random Forest model that achieved the highest precision, demonstrating a balanced performance in precision, recall, and F1 score [1].

Manzoor et al. (2024) extended this work by incorporating explainable AI techniques into their predictive framework. Using a data set of 383 patients, they developed a deep learning model that achieved a sensitivity and specificity of 97.75% and 100%, respectively. Interpretability techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and Morris Sensitivity Analysis were used to elucidate the model's predictions, enhancing its trustworthiness among clinicians [2].

Similarly, Emmanuel et al. (2024) emphasized the role of unsupervised feature engineering techniques, including Principal Component Analysis (PCA) and Truncated Singular Value Decomposition (TSVD), in improving ML model performance. The study demonstrated that integrating these techniques into Logistic Regression and Random Forest pipelines significantly enhanced prediction accuracy. The findings suggest that robust feature engineering is essential for reliable recurrence prediction [3].

Md Shah Alam et al. (2024) compared ML models, such as Logistic Regression, SVM, and Random Forest, using the Thyroid Gland Dataset from Kaggle. Logistic Regression

slightly outperformed other models, achieving an accuracy of 92.32%. However, the study highlighted challenges with class imbalance and model interpretability, advocating for integrating ML into clinical workflows to enhance personalized treatment strategies and resource management [4].

Another study by Elizabeth et al. (2024) focused on developing predictive models for DTC recurrence, reaffirming the effectiveness of Random Forest in achieving superior accuracy. The study emphasized the utility of SMOTE and hyperparameter tuning in addressing data imbalance and optimizing model performance, further validating the potential of ML in oncology [5].

Additional research by Manzoor et al. (2024) explored ML models for staging thyroid cancer, demonstrating the applicability of algorithms like CatBoost, Logistic Regression, and XGBoost in predicting cancer stages. Accurate staging is vital for formulating effective treatment plans and reducing mortality risks. This study reinforces the importance of data-driven approaches in enhancing cancer management [**?**].

Collectively, these studies illustrate the transformative role of ML and AI in thyroid cancer research. They demonstrate that integrating advanced algorithms with feature engineering and interpretability techniques can significantly improve recurrence and staging predictions. However, challenges such as class imbalance, data privacy, and the need for high-quality datasets persist. Future research should focus on validating these models across diverse patient populations and incorporating real-time patient data to enhance their clinical applicability and generalizability.

II. Research Gap and Rationale of the Study

Although important progress is made in the utilization of the ML and AI technologies to forecast thyroid cancer prognosis, several limitations exist. The latest work has mainly concentrated on the use of ML algorithms for the recurrence and staging of well differentiated thyroid cancer (DTC). Thus, there is a limited body of work addressing various combinations of surgical, radiation and hormonal therapies, and their effects on patient prognosis. Moreover, the issues of class imbalance and model interpretability are partially discussed while the problems with data privacy, the demands on the high-quality datasets, and model portability to different patient populations have not been solved yet. Secondly, patient data are often incorporated into the models in non-real-time and clinical use of such models is therefore restricted.

To close these gaps, the present research endeavor to design and undertake the assessment of various ML models that can predict treatment outcomes among thyroid cancer patients. With wealth of basic information about patients, their diagnosis, the prescription plan and results of treatment in hand, the aim of this research is to build refined models for gauging the probability of treatment effectiveness, particularly for patients' survival rate and the likelihood of cancer recurrence. The developed rationale for this study is to give a tool that will help the clinician to make a reasonant decision hence enhancing the patient's well-being and the efficient usage of the available resources in the healthcare sector. In the end, the goal of this study is to make a contribution to the body of knowledge that should assist in the development of individualised treatment plans for those with thyroid cancer.

III. Method

A. Data Collection

The dataset used in this study was obtained from secondary sources, specifically the Penn Medicine cancer registries, and is available at [9]. It includes data from 1362 patients with thyroid cancer, collected over two years from January 2, 2013, to December 31, 2015. The data set was selected based on specific criteria, including PrimarySiteCode C73.9, PrimarySiteCategory Endocrine, and PrimarySiteDesc thyroid gland. The data has been anonymized to ensure privacy, with identifiers and dates modified, and hospital names replaced with generic labels. Additionally, the hospital ID, index and Patient ID columns were excluded.

B. Dataset Description

The data set used in this study comprises 40 columns, among which the most significant columns are presented under some key variables that have been identified to predict the success of treatment in patients with thyroid cancer. Table I provides a comprehensive description of these key variables-Clinical Data, Demographic Data, Treatment Details, and Stages. The target variable, *Success*, is defined as a binary outcome indicating recurrence or survival.

This detailed feature set enables clinicians to input the necessary variables into the model, ensuring accurate predictions of treatment outcomes. The dataset includes additional features that were also considered during model training to enhance performance and generalizability.

C. Data Preprocessing

The data preprocessing phase ensures the dataset is clean, consistent, and suitable for machine learning models. The following steps were conducted as part of this process:

1) **Pipeline Creation:** A preprocessing pipeline was created to streamline and automate the steps for data preparation. This ensured reproducibility and minimized errors in data handling.
2) **Handling Missing Values:** Missing values in numerical features were imputed using the **mean** value, while categorical features were imputed using the **mode**. This approach ensured the completeness of the dataset without introducing significant biases.
3) **Data Cleaning:** Redundant and irrelevant features that did not contribute to the classification task were removed. Duplicate entries were also identified and eliminated.
4) **Feature Encoding:** Categorical features were transformed into numerical format using one-hot encoding and label encoding techniques to make them suitable for machine learning algorithms.

TABLE I: Key Features and Their Descriptions

Variables	Feature	Description
Clinical Data	*PrimarySequence*	Indicates the sequence of the primary tumor (e.g., first, second).
	DateFirstContact.x	The date when the patient first contacted the reporting hospital.
	DateDx.x	The date when the patient was diagnosed with thyroid cancer.
	FirstRecurrenceType	The type of recurrence (e.g., local, regional, distant).
	MstDefRtDate	The date of the most definitive radiation therapy the patient received.
	Grade	The grade of the tumor, indicating how much the tumor cells differ from normal cells (e.g., well-differentiated, poorly differentiated).
Demographic Data	*AgeAtDx.x*	The patient's age at the time of diagnosis.
	Sex	The patient's sex (e.g., male, female).
	Race	The patient's race (e.g., White, Black or African American).
Treatment Details	*SurgeryDesc*	Description of the surgery performed (e.g., total thyroidectomy).
	RtDesc	A description of the radiation therapy performed.
	ChemoDesc	A description of the chemotherapy administered.
	HormoneDesc	A description of the hormone therapy administered.
Stages	*ClinStage*	The overall clinical stage of the cancer, combining T, N, and M stages.
	PathStage	The overall pathological stage of the cancer, combining PathT, PathN, and PathM stages.
Target Variable (Success)	*recurred*	Indicates whether the cancer has recurred (TRUE or FALSE).
	death	Indicates whether the patient has died (TRUE or FALSE).

5) **Feature Scaling:** To ensure all features contributed equally during model training, numerical features were standardized using **z-score normalization**, transforming them to have a mean of 0 and a standard deviation of 1.
6) **Combining Pre-processing Steps:** The pre-processing steps were combined using a **ColumnTransformer**, which applies the appropriate transformations to numerical and categorical features. This ensures that all features are pre-processed correctly and consistently
7) **Train-Test Split:** The dataset was split into training and testing sets using an 80-20 ratio. **Stratified sampling** was employed to preserve the original class distribution in both sets.

By leveraging the preprocessing pipeline, the dataset was effectively prepared for model building, ensuring consistency and reproducibility across the experiments.

IV. Data Analysis Procedure

The data analysis procedure encompasses the systematic steps undertaken to evaluate the performance of various machine learning models on the preprocessed dataset. This section outlines the models applied, the metrics used for evaluation, and the insights derived from the comparative analysis.

A. Model Selection and Training

For this study, several machine learning models were selected and trained, including:

- **Random Forest:** A robust ensemble-based algorithm used to handle classification tasks effectively.
- **XGBoost:** An advanced gradient boosting technique known for its efficiency and scalability.
- **Neural Network:** A deep learning model capable of capturing complex patterns in the data.
- **Support Vector Machine (SVM):** A powerful classifier that constructs hyperplanes for optimal separation of classes.

Each model was evaluated with and without hyperparameter tuning to determine its optimal performance. Grid search and random search techniques were employed to identify the best set of hyperparameters.

B. Evaluation Metrics

The following metrics were utilized to assess the models' performance:

- **Accuracy:** The proportion of correctly classified samples.
- **Macro Average F1-Score:** A harmonic mean of precision and recall, averaged across all classes to account for imbalanced data.
- **Confusion Matrix:** A visual representation of the model's predictions, highlighting true positives, false positives, true negatives, and false negatives.

C. Comparative Results

Table II summarizes the performance of each model in terms of accuracy and macro average F1-score. From the comparison, it is evident that **XGBoost with hyperparameter tuning** outperformed other models, achieving the highest accuracy and F1-score.

TABLE II: Model Evaluation Summary

Model	Tuning	Accuracy	Macro Avg F1-Score
Random Forest	No	0.95	0.93
Random Forest (Best)	Yes	0.96	0.94
XGBoost	No	0.96	0.97
XGBoost (Best)	Yes	**0.98**	**0.98**
Neural Network	No	0.96	0.94
Neural Network (Best)	Yes	0.97	0.96
SVM	No	0.96	0.94
SVM (Best)	Yes	0.97	0.96

D. Insights from the Analysis

The analysis revealed that hyperparameter tuning significantly improved the performance of all models. Among them, XGBoost demonstrated superior capability, achieving the best results in terms of both accuracy and macro average F1-score.

These findings suggest that XGBoost is particularly effective in handling the complexities of this dataset.

Figure 1 presents the confusion matrix of the best-performing XGBoost model, illustrating its ability to correctly classify the majority of samples.

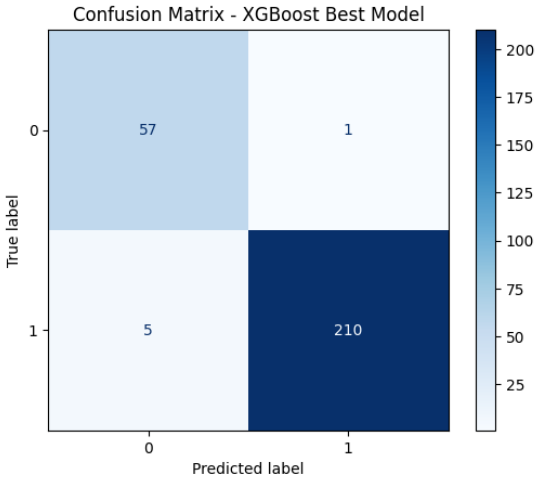

Fig. 1: Confusion Matrix of the XGBoost Model.

V. Results

The performance evaluation of the models highlighted XGBoost with hyperparameter tuning as the best-performing algorithm. This section presents the ROC curve and SHAP waterfall visualization for the XGBoost model, illustrating its effectiveness and providing insights into the factors influencing its predictions.

A. ROC Curve Analysis for XGBoost Model

The ROC curve in Fig. 2 evaluates the performance of the best XGBoost model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The area under the curve (AUC) is impressive 1.00, indicating excellent predictive performance.

This result demonstrates the model's capability to distinguish between classes effectively. The proximity of the curve to the top-left corner signifies a high level of accuracy, reinforcing the XGBoost model as the most reliable predictor in this study.

B. Feature Importance Analysis using SHAP

The SHAP waterfall visualization in Fig. 3 illustrates the contributions of individual features to the XGBoost model's prediction for a specific instance. Each bar represents the impact of a feature on the model's output, with red bars indicating positive contributions and blue bars indicating negative contributions.

The main contributing features, such as `DateFirstContact.y` and `FirstRecurrenceType`, exhibit the most significant influence on the prediction. Specifically, `DateFirstContact.y` reduces the prediction

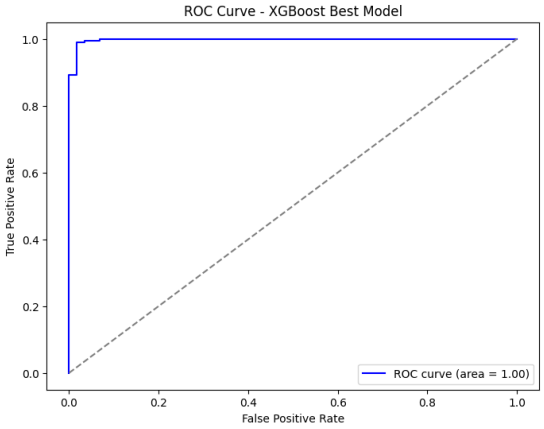

Fig. 2: ROC Curve of the XGBoost Model.

score by -1.31, while `FirstRecurrenceType` increases it by $+0.94$. Other features, including `DeathDate` and `DateDx.x`, also demonstrate substantial contributions, either amplifying or mitigating the prediction score.

The high number of features observed in the SHAP visualization is due to the one-hot encoding of categorical variables, which expands each categorical column into multiple binary columns representing unique categories, thereby increasing the total feature count used by the model.

This visualization effectively highlights the importance of these features in driving the model's decision-making process, providing valuable insights into the interpretability of the model.

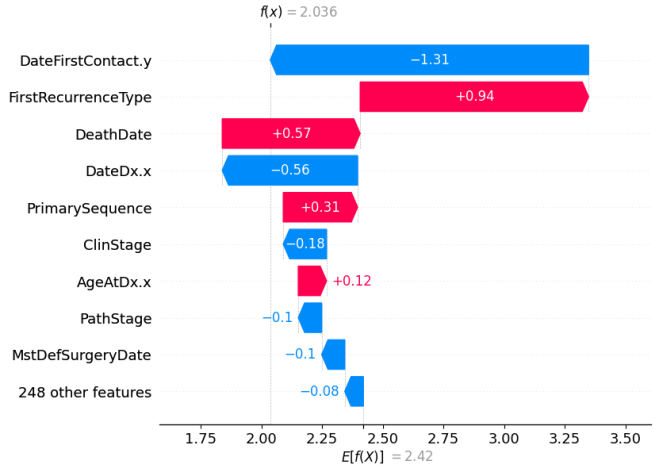

Fig. 3: SHAP Waterfall Visualization for the XGBoost Model.

C. Comparison of Model Performance

The XGBoost model demonstrated superior performance compared to other machine learning models used in this study. With an AUC of 1.00 and an accurate visualization of the importance of features provided by SHAP, the model outperformed others in both predictive capability and interpretability. This high level of performance underlines its ability to provide

reliable and actionable insights into treatment success for thyroid cancer patients. These results validate the robustness of XGBoost as a predictive tool for complex data sets in healthcare.

VI. Discussion

The results of this study highlight the potential of machine learning, particularly XGBoost, to predict the success of treatment for patients with thyroid cancer. This section discusses the implications, limitations, and future directions of this research.

A. Clinical Implications

The XGBoost model achieved outstanding predictive performance, with an AUC of 1.00. The SHAP-based feature importance analysis provided valuable information on key factors that influence treatment success, such as `DateFirstContact.y` and `FirstRecurrenceType`. These findings underscore the potential for integrating machine learning models into clinical decision-making processes. By identifying critical variables, clinicians can tailor treatment plans to improve patient outcomes and optimize resource allocation.

B. Model Strengths and Interpretability

One of the notable strengths of XGBoost lies in its ability to handle complex data sets while maintaining high accuracy. Additionally, the integration of SHAP for feature importance analysis enhances model interpretability, making it easier for healthcare professionals to understand and trust the model's predictions. This interpretability is crucial for fostering adoption in clinical settings, where transparency and reliability are paramount.

C. Limitations of the Study

While the results are promising, there are some limitations to address:

- The dataset is specific to thyroid cancer patients from a particular region and time frame, which may limit the generalizability of the findings.
- External factors, such as treatment costs and patient preferences, were not considered, potentially influencing treatment decisions.
- Potential overfitting may occur due to the exceptionally high accuracy achieved by the best model.

This limitation underscores the importance of using broader datasets and incorporating additional variables in future research to enhance model robustness and applicability.

D. Future Directions

Future work should focus on validating the proposed approach across diverse datasets to ensure generalizability. Additionally, integrating other data sources, such as genomic data and real-time patient monitoring, could further enhance the model's predictive capabilities. Another avenue for exploration is the development of user-friendly interfaces that allow clinicians to interact with the model's predictions and visualizations seamlessly. Finally, efforts should be made to evaluate the integration of machine learning models into real-world clinical workflows to assess their practical utility and impact on patient outcomes.

VII. Conclusions

This study demonstrates the effectiveness of machine learning, particularly the XGBoost model, in predicting the success of treatment for patients with thyroid cancer. The model's high accuracy and interpretability, supported by SHAP-based feature importance analysis, highlight its potential for enhancing clinical decision-making. By identifying critical factors influencing treatment outcomes, the proposed approach can assist clinicians in optimizing treatment plans. However, further research with diverse datasets and real-world validation is essential to ensure generalizability and practical applicability. This work lays the foundation for integrating machine learning into personalized cancer care, ultimately improving patient outcomes and resource efficiency.

References

[1] Elizabeth, A., Clark, S. A., Price, T., Lucena, B., Haberlein, A., Wahbeh, R., & Seetan, R. (2024). Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach. *Knowledge*, 4(4), 557-570.

[2] Manzoor, A., & Haddad, J. J. (2024). An Explainable AI Model for Predicting the Recurrence of Differentiated Thyroid Cancer.

[3] Emmanuel, O., Eze, J. U., Abdulraheem, A. S., Ezigbo, G. U., & Amorha, K. C. (2024). Optimizing Unsupervised Feature Engineering and Predictive Models for Thyroid Cancer Recurrence Prediction.

[4] Md Shah Alam, H., Al Mukaddim, A., Rahman Anonna, F., Hossain, M. S., Rahman, M., & Nasiruddin, M. (2024). Machine Learning Models for Predicting Thyroid Cancer Recurrence: A Comparative Analysis. *Journal of Medical and Health Studies*, 5(4), 113-129.

[5] Elizabeth, A., Clark, S. A., Price, T., Lucena, B., Haberlein, A., Wahbeh, R., & Seetan, R. (2024). Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach.

[6] Shivam, K. B., Ghosh, D., Gourisaria, M. K., Jena, J. J., Pattnayak, P., & Patra, S. S. (2024). Beyond the Biopsy: A Comprehensive Machine Learning Based Approach to Thyroid Cancer Staging. *Proceedings of the International Conference on Computational Intelligence and Communication Networks*, 846-851.

[7] Hegde, S. K., Hegde, R., & Thangavel, M. (2024). Early Prediction of Thyroid Cancer using Hybrid Combination of Swarm Optimization and Meta Classifier based Machine Learning Algorithm. *Proceedings of the International Conference on Computational Intelligence and Communication Networks*, 1400-1406.

[8] Joshi, H., Vijayalakshmi, A., & George, S. M. (2024). Unraveling the Complexity of Thyroid Cancer Prediction. *Advances in Computational Intelligence and Robotics*, 367-388.

[9] Penn Medicine Cancer Registries. (2015). Thyroid Cancer Dataset. Mendeley Data. https://data.mendeley.com/datasets/57726vkm48/1