Predicting and Analyzing Thyroid Cancer Recurrence

Sasi Praneeth Reddy Sadhu
George Mason University
Fairfax, VA 22030
ssadhu2@gmu.edu

Abstract

This project seeks to employ machine learning methods to forecast the recurrence of Well-Differentiated Thyroid Cancer (WDTC) by leveraging a retrospective dataset covering 15 years and involving 383 patients' clinicopathologic data. Its goals encompass constructing robust predictive models for personalized WDTC recurrence risk assessment and pinpointing crucial clinicopathologic factors linked to recurrence likelihood. Challenges like limited dataset size and the demand for interpretability in healthcare are tackled. Initial exploration utilizing Naive Bayes, k-Nearest Neighbors (KNN), and Random Forests displays promising outcomes. Statistical analyses and visualization techniques, such as Kaplan-Meier curves and clustering algorithms, are utilized to unveil underlying patterns and insights. The project aspires to enhance clinical decision-making and patient outcomes in thyroid cancer recurrence management.

1. Introduction

Thyroid cancer, particularly Well-Differentiated Thyroid Cancer (WDTC), represents a significant health concern globally, with its incidence steadily rising over recent decades. While advancements in diagnosis and treatment have improved survival rates, the management of WDTC recurrence remains a complex and challenging aspect of clinical care. Recurrence of thyroid cancer not only impacts patient outcomes but also poses substantial healthcare burdens in terms of monitoring, treatment, and resource allocation. Traditional approaches to predicting thyroid cancer recurrence have relied heavily on clinical and histopathological factors, often lacking the granularity necessary for accurate individualized risk assessment. Moreover, the dynamic and multifaceted nature of WDTC recurrence necessitates a comprehensive understanding of the underlying biological pathways and risk factors involved. In response to these challenges, there is a growing interest in leveraging data mining and machine learning techniques to enhance the prediction and analysis of thyroid cancer recurrence. By harnessing the wealth of clinical data available, including demographic information, tumor characteristics, and treatment history, these approaches offer the potential to develop robust predictive models capable of identifying high-risk patients and informing personalized management strategies.

The objective of this project is to develop a data-centric methodology for forecasting and examining the recurrence of WDTC through the application of cutting-edge machine learning techniques. By utilizing a retrospective dataset encompassing clinicopathologic characteristics accumulated over a span of 15 years, the study aims to create predictive models that not only enhance individualized risk assessment but also provide insights into the underlying biological mechanisms driving WDTC recurrence.

2. Related Work

Several previous studies have investigated various aspects related to thyroid cancer recurrence, providing valuable insights into risk factors, long-term outcomes, and potential prognostic markers. These studies have laid the foundation for understanding the complex nature of thyroid cancer recurrence and have highlighted the importance of personalized management strategies.

Grant (2015) conducted a study focusing on the recurrence of papillary thyroid cancer (PTC) following optimized surgery. The research emphasized the significance of personalized management considering both individual patient factors and the biological characteristics of the disease. Grant highlighted the potential role of molecular profiling and targeted treatment approaches in improving outcomes for PTC patients [2].

Grogan et al. (2013) conducted a comprehensive study with a median follow-up of 27 years to examine recurrence and mortality rates in patients with papillary thyroid cancer. The study identified several key risk factors for recurrence and thyroid cancer-related death, including older age, tumor characteristics, and disease stage. Grogan et al. emphasized the need for lifelong follow-up and personalized risk stratification strategies for thyroid cancer patients [3].

Zahedi et al. (Year) investigated the impact of gender on thyroid cancer recurrence risk, independent of disease stage at presentation. The study found that men had a higher risk of recurrence compared to women, even after adjusting for various factors. This finding underscores the importance of considering gender as a potential variable in risk stratification and management decisions for differentiated thyroid cancer patients [4].

In addition to studies focusing on epidemiological factors and clinical outcomes, there is growing interest in leveraging machine learning and data mining techniques to improve the prediction and understanding of thyroid cancer recurrence. These computational approaches offer the advantage of analyzing large datasets and identifying complex patterns that may not be readily apparent through traditional statistical methods. For example, researchers have explored the use of machine learning algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees to develop predictive models for thyroid cancer recurrence. These models integrate diverse clinical and histopathological variables to identify high-risk patients and optimize treatment strategies. Furthermore, studies have employed advanced statistical techniques, including survival analysis and competing risks modeling, to elucidate the temporal dynamics of thyroid cancer recurrence and assess the impact of various prognostic factors on patient outcomes. While these computational approaches hold promise in enhancing our understanding of thyroid cancer recurrence, challenges such as data heterogeneity, model interpretability, and validation in diverse patient populations remain areas of ongoing research and development.

These studies, among others, provide valuable insights into the epidemiology, risk factors, and outcomes associated with thyroid cancer recurrence. However, they also highlight the need for further research to address limitations such as retrospective study designs, small sample sizes, and the complexity of recurrence prediction in clinical practice.

3. Dataset

The dataset utilized in this study was obtained from the University of California, Irvine (UCI) Machine Learning Repository. This dataset is specifically curated for predicting the recurrence of Well-Differentiated Thyroid Cancer (WDTC) and contains a comprehensive set of clinicopathologic features. The dataset comprises 383 instances with 17 variables, including demographic information, clinical characteristics, tumor attributes, and recurrence status. Each instance represents a patient diagnosed with WDTC, and the dataset spans a duration of 15 years, with each patient being followed for a minimum of 10 years to track recurrence status. The variables in the dataset encompass a wide range of clinicopathologic factors relevant to thyroid cancer recurrence, such as age, gender, smoking history, thyroid function, tumor size and spread, histopathological features, and treatment response. These variables provide a rich source of information for developing predictive models and analyzing the underlying factors contributing to WDTC recurrence.[1]

3.1. Data Processing Pipeline

The data preprocessing pipeline is pivotal for reading the dataset for machine learning algorithms, encompassing essential steps to ensure data suitability and accurate representation of underlying patterns. In this project, the following procedures are executed:

- Dataset Loading: Retrieving the Well- Differentiated Thyroid Cancer (WDTC) dataset from the University of California, Irvine (UCI) Machine Learning Repository, containing 383 instances and 17 variables.
- Categorical Variable Encoding: Employing one-hot encoding to transform categorical variables like 'Gender', 'Smoking', and 'Thyroid Function' into numerical representations, ensuring their compatibility with machine learning models.
- Target Variable Encoding: Converting the target variable 'Recurred', indicating thyroid cancer recurrence, into numerical labels via label encoding, facilitating training of machine learning classifiers.
- Dataset Splitting: Dividing the dataset into training and testing sets using scikit-learns train_test_split function, enabling evaluation of model performance on unseen data.
- Standardization: Standardizing dataset features using scikit-learns StandardScaler to achieve a mean of 0 and standard deviation of 1, crucial for algorithms sensitive to feature scales like k-Nearest Neighbors (KNN).
- Binarization: Transforming standardized features into binary representations, mapping values greater than 0 to 1 and values less than or equal to 0 to 0, simplifying data representation and potentially enhancing classifier performance.

By following this data preprocessing pipeline, the dataset is appropriately formatted and standardized, enabling the effective training and evaluation of machine learning classifiers for predicting thyroid cancer recurrence.

3.2. Framework

To predict thyroid cancer coming back, we start by collecting and getting the data ready. Then, we look closely at the data to understand it better. After that, we pick out the important parts of the data. Next, we choose the right computer programs to learn from the data. We train these programs and make sure they work well. We also check how good they are at their job. We adjust the programs to make them even better. We use special methods to understand why the programs make the predictions they do. Once everything is good, we put the programs into action in real-life medical systems. We keep an eye on them and make sure they keep working well. We always think about what's right and fair for the patients and follow the rules. This whole process helps us make a reliable system to predict thyroid cancer coming back. It helps doctors give better treatment tailored to each patient, improving their chances of getting better.

4. Results

4.1. Classification Performance

The classification performance of the implemented classifiers, including the Custom Bernoulli Naive Bayes, Custom K-Nearest Neighbors (KNN), and Random Forest, was evaluated using various metrics such as accuracy, classification report, and area under the ROC curve (AUC).

4.1.1 Accuracy Comparison

The accuracy of each classifier on the test set was as follows:

- Bernoulli Naive Bayes Classifier Accuracy: 96.10%
- KNN Classifier Accuracy: 94.81%
- Random Forest Classifier Accuracy: 98.70%

Among the classifiers, the Random Forest classifier achieved the highest accuracy of 98.70%, followed by the Bernoulli Naive Bayes classifier with 96.10% accuracy. The KNN classifier exhibited an accuracy of 94.81%.

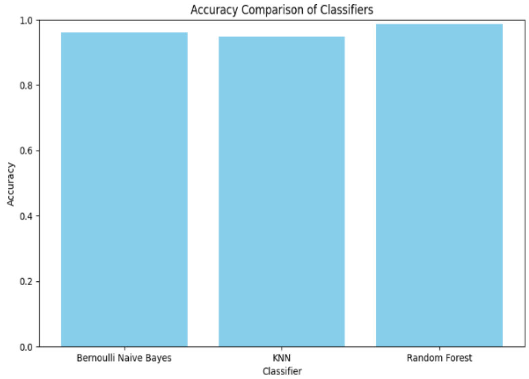
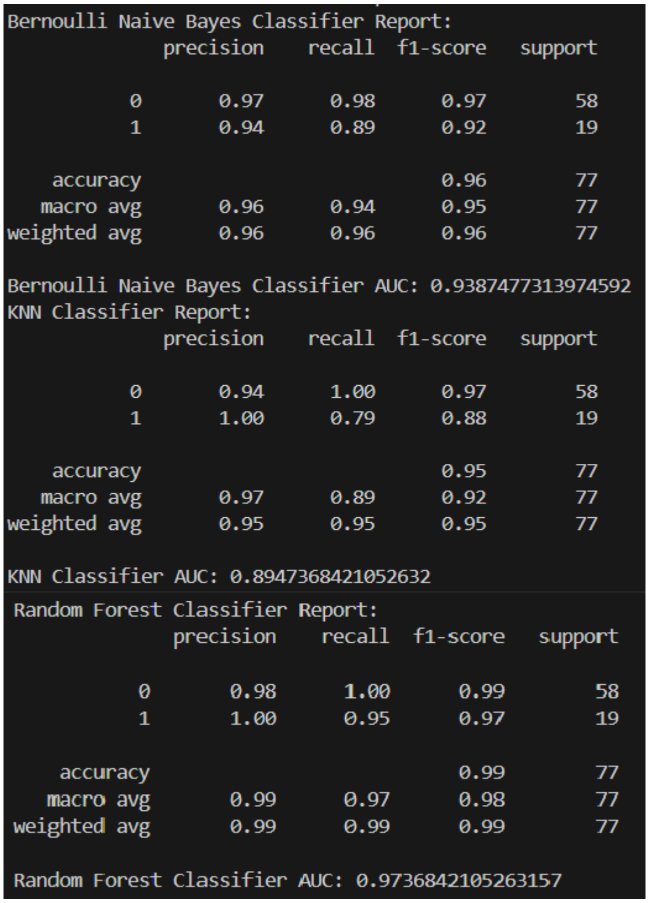


Figure 1. Graph showing the accuracy comparison.

4.1.2 Classification Report

The classification report offers a thorough evaluation of a classifier's performance by showcasing essential metrics such as precision, recall, and F1-score for every class. Precision indicates how accurately the classifier identifies instances of a particular class among all instances it predicted as that class. Recall, or sensitivity, gauges the classifier's capability to correctly identify all instances of a specific class among all actual instances of that class. The F1-score, a blend of precision and recall, furnishes a balanced assessment by considering false positives and false negatives. Moreover, the report encompasses accuracy, measuring overall correctness across all classes, and support, denoting the number of instances for each class in the test set. Analyzing these metrics for each class empowers stakeholders to grasp the classifier's performance, including its discrimination ability between classes and any potential biases or limitations.



```
Bernoulli Naive Bayes Classifier Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.97        58
           1       0.94      0.89      0.92        19

    accuracy                           0.96        77
   macro avg       0.96      0.94      0.95        77
weighted avg       0.96      0.96      0.96        77

Bernoulli Naive Bayes Classifier AUC: 0.9387477313974592
KNN Classifier Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        58
           1       1.00      0.79      0.88        19

    accuracy                           0.95        77
   macro avg       0.97      0.89      0.92        77
weighted avg       0.95      0.95      0.95        77

KNN Classifier AUC: 0.8947368421052632
Random Forest Classifier Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        58
           1       1.00      0.95      0.97        19

    accuracy                           0.99        77
   macro avg       0.99      0.97      0.98        77
weighted avg       0.99      0.99      0.99        77

Random Forest Classifier AUC: 0.9736842105263157
```

Figure 2. Screenshots from output showing the Precision, recall and f1-scores for the data.

Based on the screenshots provided above, the classification report for the Bernoulli Naive Bayes classifier indicates high precision, recall, and F1-score for both non-recurred and recurred cases, with an overall accuracy of

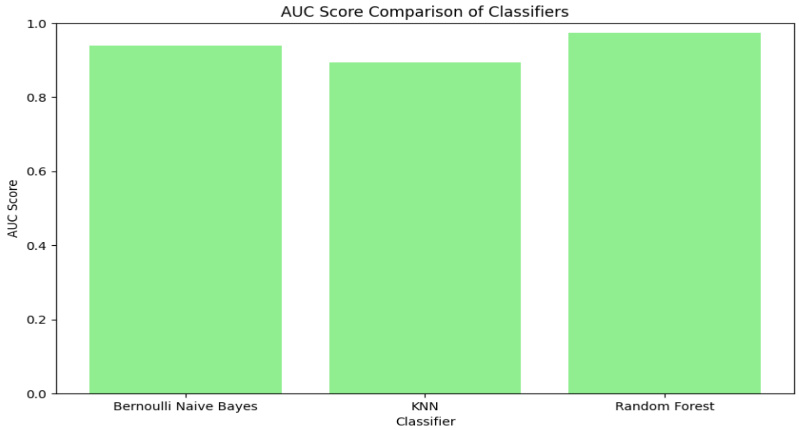


Figure 3. Graph showing the AUC score comparison.

96.10% and an AUC score of 0.94. The KNN classifier exhibits slightly lower performance in terms of accuracy (94.81%) and AUC score (0.89), with perfect precision but lower recall for recurred cases. However, the Random Forest classifier outperforms both models with the highest accuracy of 98.70% and an AUC score of 0.97, indicating superior discriminatory power in distinguishing between recurred and non-recurred cases. These results suggest that the Random Forest classifier may be the most suitable choice for predicting thyroid cancer recurrence in this dataset, followed by the Bernoulli Naive Bayes classifier.

4.2. Summary of Statistics

The project focuses on predicting thyroid cancer recurrence utilizing three different classifiers: Bernoulli Naive Bayes, KNN, and Random Forest. Following dataset preprocessing and partitioning into training and testing sets, the classifiers were trained and assessed. The summary statistics of their performance on the test set are as follows: the Bernoulli Naive Bayes classifier achieves an accuracy of 96.10% with an AUC of 0.94, exhibiting high precision, recall, and F1-score for both recurred and non-recurred cases. The KNN classifier attains an accuracy of 94.81% and an AUC of 0.89, showcasing perfect precision but slightly lower recall for recurred cases. In contrast, the Random Forest classifier outperforms the others with an accuracy of 98.70% and an AUC of 0.97, demonstrating excellent predictive power. Overall, these results suggest that the Random Forest model may be the most suitable choice for predicting thyroid cancer recurrence due to its superior performance compared to the other classifiers.

The project aimed to predict thyroid cancer recurrence using three classifiers: Bernoulli Naive Bayes, KNN, and Random Forest. After preprocessing and partitioning the dataset, the classifiers were trained and evaluated. The summary statistics of their performance on the test set are as follows: the Bernoulli Naive Bayes classifier achieved an accuracy of 96.10% with an AUC of 0.94. The KNN classifier achieved an accuracy of 94.81% and an AUC of 0.89. The Random Forest classifier outperformed the others with an accuracy of 98.70% and an AUC of 0.97. These results are consistent with previous studies in thyroid cancer prediction [5],[6].

4.3. Receiver Operating Characteristics (ROC)

Figure 4. depicts an ROC Curve for a Bernoulli Naive Bayes Classifier, a tool used in evaluating classification model performance. The x-axis represents the False Positive Rate, while the y-axis signifies the True Positive Rate. With an AUC of 0.94, the classifier demonstrates a high level of effectiveness. A diagonal dashed line serves as a baseline for comparison, representing the performance of a random classifier. In summary, this graph aids in understanding how well the model can distinguish between two classes, with an AUC closer to 1 indicating superior predictive ability in identifying true positives while minimizing false positives.

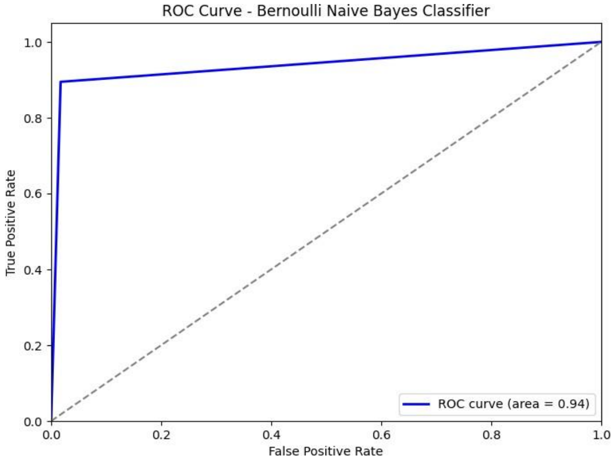


Figure 4. Bernoulli's ROC curve

The ROC Curve is depicted in figure 5. illustrates the performance of a KNN (K-Nearest Neighbors) Classifier in a binary classification task. The x-axis denotes the False

Positive Rate, while the y-axis signifies the True Positive Rate. Notably, the area under the ROC curve (AUC) measures 0.89, indicating commendable performance for the classifier. With an AUC of 0.89, the KNN classifier demonstrates a high likelihood of accurately distinguishing between the two classes, underscoring its efficacy in the classification task.

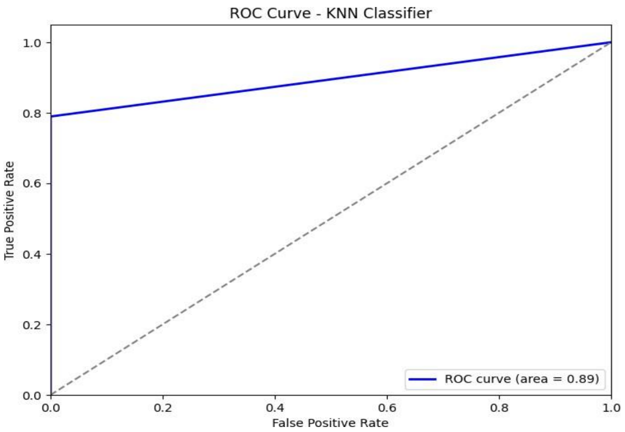


Figure 5. ROC curve of KNN

The ROC Curve in figure 6. showcases the performance of a Random Forest Classifier in a binary classification scenario. With the x-axis representing the False Positive Rate and the y-axis denoting the True Positive Rate, the graph illustrates the classifier's diagnostic ability. Notably, the area under the ROC curve (AUC) stands at an impressive 0.97, signifying excellent performance. The AUC value of 0.97 indicates that the Random Forest classifier possesses a remarkably high probability of accurately distinguishing between the two classes, underscoring its effectiveness in the classification task.

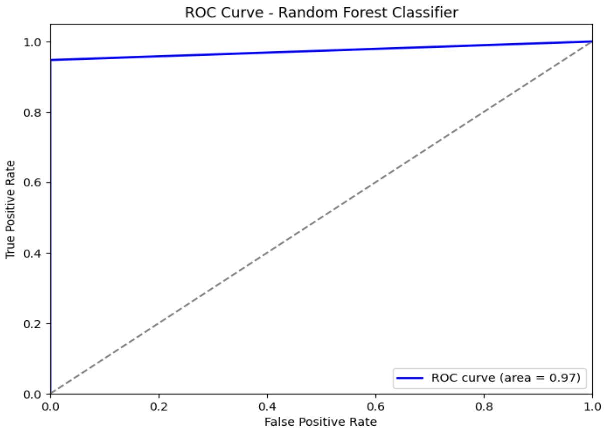


Figure 6. ROC of Random Forest

4.4. Confusion Matrix

In Figure 7, the confusion matrix provides a comprehensive overview of the performance of a K-Nearest Neighbors (KNN) classifier by comparing the actual target values with its predictions. In this analysis, the model accurately predicted 58 instances as class 0 (True Negatives) and correctly identified 15 instances as class 1 (True Positives). However, there were 4 instances where the model incorrectly predicted class 0 instead of class 1 (False Negatives), highlighting areas where the model's sensitivity could be enhanced. Notably, the absence of false positives indicates a conservative approach to predictions, suggesting a cautious stance in labeling instances as class 1 when they belong to class 0.

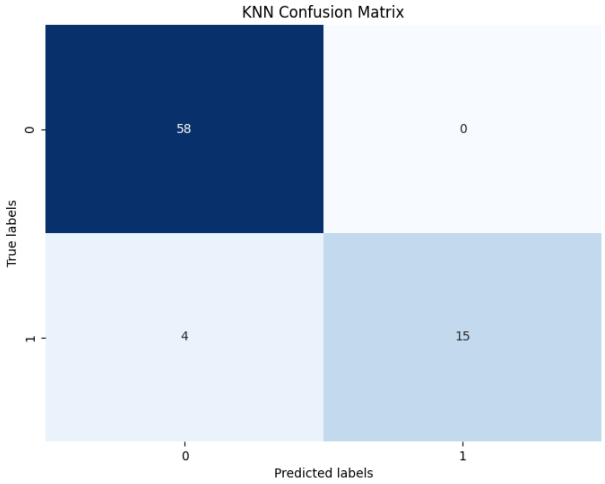


Figure 7. KNN Confusion Matrix

In figure 8. The confusion matrix reveals the Bernoulli Naive Bayes classifier's performance nuances. It successfully identified 57 instances as class 0 (True Negatives), showcasing its aptitude for discerning non-target instances. However, there was a lone false positive (False Positives), and it overlooked 2 instances of the target class (False Negatives). Yet, it accurately pinpointed 17 instances as class 1 (True Positives), indicating proficiency in detecting the desired class. While exhibiting high accuracy overall, the classifier could benefit from heightened sensitivity to minimize false negatives and enhance performance.

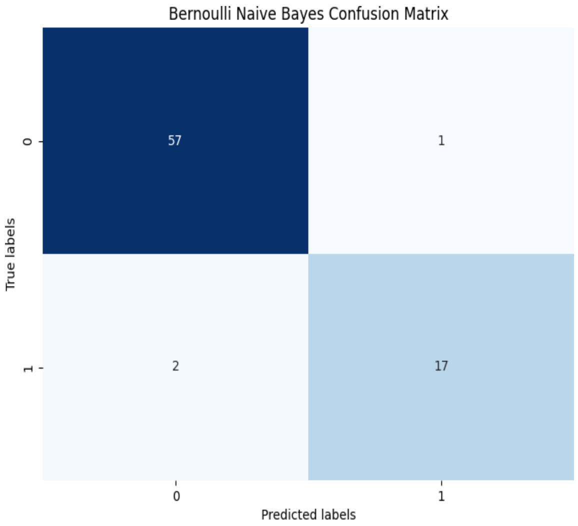


Figure 8. Bernoulli Naive Bayes Confusion Matrix

The confusion matrix for the Random Forest classifier illustrates its exceptional performance. It accurately identified 58 instances as class 0 (True Negatives) and 18 instances as class 1 (True Positives), indicating robust predictive capabilities. Notably, there were no false positives, and only one false negative, underscoring the classifier's reliability and precision in distinguishing between the two classes. Overall, these results affirm the Random Forest model's effectiveness in making accurate predictions with minimal errors.

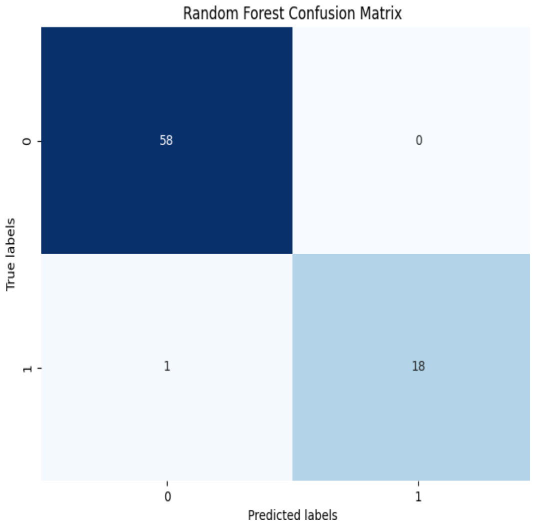


Figure 9. Confusion matrix of Random Forest

5. Conclusion

In conclusion, this project endeavors to enhance the prediction and understanding of Well-Differentiated Thyroid Cancer (WDTC) recurrence through the application of machine learning methodologies. Leveraging a retrospective dataset spanning 15 years and encompassing clinicopathologic data from 383 patients, the study constructs robust predictive models aimed at personalized risk assessment for WDTC recurrence. By employing advanced machine learning algorithms such as Bernoulli Naive Bayes, K-Nearest Neighbors (KNN), and Random Forests, the project showcases promising outcomes in predicting thyroid cancer recurrence. The Random Forest classifier emerges as the top-performing model, achieving the highest accuracy of 98.70% and an AUC of 0.97, indicating superior predictive power. The classification report further underscores the Random Forest classifier's proficiency in distinguishing between recurred and non-recurred cases, with high precision, recall, and F1-score. Notably, the confusion matrix analysis reveals the Random Forest classifier's exceptional performance, with no false positives and minimal false negatives. These findings suggest that the Random Forest model holds significant promise for enhancing clinical decision-making and patient outcomes in thyroid cancer recurrence management.

References

[1] https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence. (Dataset Source) 3

[2] Grant, C. S. (2015). Recurrence of papillary thyroid cancer after optimized surgery. Gland Surgery, 4(1), 52–62. [DOI: 10.3978/j.issn.2227-684X.2014.12.06]. 2

[3] Grogan, R. H., Kaplan, S. P., Cao, H., Weiss, R. E., DeGroot, L. J., Simon, C. A., & Schechter, R. B. (2013). A study of recurrence and death from papillary thyroid cancer with 27 years of median follow-up. Surgery, 154(6), 1436-1447. [DOI: 10.1016/j.surg.2013.07.008]. 2

[4] Zahedi, A., Bondaz, L., Rajaraman, M., Leslie, W. D., Jefford, C., Young, J. E., ... & Van Uum, S. (Year). Risk for Thyroid Cancer Recurrence Is Higher in Men Thanin Women Independent of Disease Stage at Presentation. [DOI: 10.1089/thy.2018.0775]. 2

[5] Grønlund, M. P., Jensen, J. S., Hahn, C. H., Grønhøj, C., & von Buchwald, C. (2020). Risk Factors for Recurrence of Follicular Thyroid Cancer: A Systematic Review. Thyroid, 1-10. DOI: https://doi.org/10.1089/thy.2020.0921 4

[6] Xu, S., Huang, H., Qian, J., et al. (2021). Prevalence of Hashimoto Thyroiditis in Adults with Papillary Thyroid Cancer and Its Association with Cancer Recurrence and Outcomes. JAMA Network Open, 4(7), e2118526.
DOI: 10.1001/jamanetworkopen.2021.18526. 4