Ensemble Learning for Improved Prediction of Thyroid Cancer Recurrence Risk Using Machine Learning and Deep Learning Algorithms

Ziyad Hossam, Mohamed Elsayed, Amgad Atef, Abdelrahman Shawky

Department of Systems and Biomedical Engineering, Faculty of Engineering, Cairo University

K E Y W O R D

Mahine Learning
Artificial intelligence
Deep learning
Thyroid cancer
Recurrence
SVM
KNN
ANN
Decision Tree
Random Forest
Ensemble Learning

A B S T R A C T

The objective of this study was to train machine learning models for predicting the likelihood of recurrence in patients diagnosed with well-differentiated thyroid cancer. While thyroid cancer mortality remains low, the risk of recurrence is a significant concern, making the identification of individual patient recurrence risk crucial for guiding subsequent management and follow-ups. In this prospective study, a cohort of 383 patients was observed for a minimum duration of 10 years within a 15-year timeframe, assessing thirteen clinicopathologic features to predict recurrence potential. Motivated by the limitations of traditional predictive models and the increasing incidence of thyroid cancer, we sought to develop a more accurate and personalized risk prediction tool. Our approach involved re-implementing various machine learning (ML) models, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and an Artificial Neural Network (ANN), and combining them into an ensemble to leverage their strengths. The ensemble model achieved accuracy comparable to the best-performing individual model (SVM), demonstrating its potential as a reliable tool for predicting recurrence risk and aiding in personalized treatment planning. Our findings highlight the value of ensemble learning in enhancing predictive performance and provide insights into the potential and limitations of these techniques in this context, ultimately guiding future advancements in predictive modeling for thyroid cancer recurrence.

1. Introduction

The incidence of thyroid cancer has risen significantly, largely due to advancements in diagnostic techniques that enable the detection of smaller and less aggressive tumors. Although the mortality rate remains low, the risk of recurrence is a significant concern for patients, underscoring the need for precise risk prediction to tailor individual management plans. While the American Thyroid Association (ATA) guidelines provide a foundational risk stratification framework, emerging research indicates that machine learning (ML) models can offer superior predictive capabilities, potentially improving patient outcomes.

This increasing incidence of thyroid cancer, combined with the potential for recurrence, highlights the necessity for more reliable risk prediction models. Traditional methods, such as the ATA guidelines, offer a structured approach but may not capture the complex, multifactorial nature of thyroid cancer recurrence risk as effectively as modern ML techniques. Accurate risk prediction is crucial for determining the appropriate intensity of monitoring and intervention for each patient, thereby preventing recurrence and improving quality of life.

Our motivation for studying this problem stems from the desire to leverage technological advancements to enhance patient care. By applying ML techniques such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), Decision Trees (DT), Random Forests (RF), voting classifiers, and stacking classifiers, we can uncover complex patterns and interactions within the data that traditional methods might miss. These advanced ML models hold promise for providing more accurate and personalized risk assessments, ultimately improving the management and outcomes of thyroid cancer patients.

2. Data Description and Preprocessing Summary

The dataset includes clinical and pathological features of 383 patients diagnosed with well-differentiated thyroid cancer over a 15-year period, with a minimum follow-up of 10 years. Key features include:

- **Age**: Patient's age at diagnosis, **Gender**: Biological sex (Female, Male), **Smoking and Hx Smoking**: Current and past smoking status, **Hx Radiotherapy**: History of head and neck radiation, **Thyroid Function**: Thyroid status (e.g., euthyroid,

hyperthyroidism), **Physical Examination**: Goiter and adenopathy presence, **Pathology**: Cancer subtype (e.g., papillary, follicular), **Focality**: Tumor focality (unifocal, multifocal), **Risk**: ATA risk assessment (low, intermediate, high), **TNM Staging**: Tumor size (T), lymph node involvement (N), and metastasis (M), **Stage**: Combined TNM stage, **Response**: Initial treatment response, **Recurred**: Recurrence status.

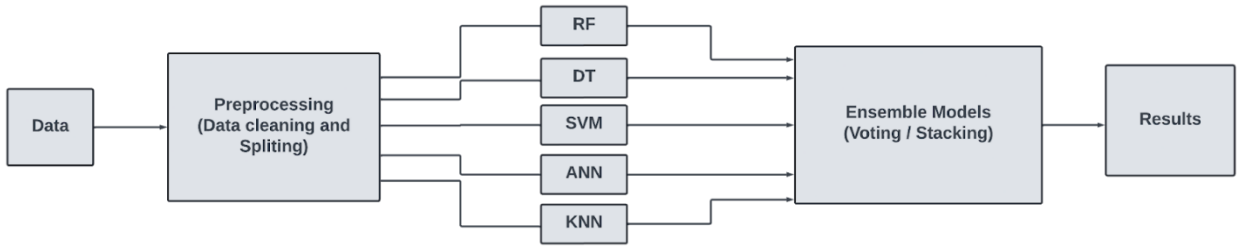
Data preprocessing involved several steps. First, 19 duplicate records were identified and removed to ensure data quality. There were no missing values or outliers detected in the dataset. Label encoding was used for all categorical variables, transforming them into numerical values suitable for machine learning algorithms. The encoded categorical variables include: Gender, Smoking, Hx Smoking, Hx Radiotherapy, Thyroid Function, Physical Examination, Pathology, Focality, Risk, T, N, M, Stage, Response, and Recurred.

Feature selection was conducted using a correlation matrix, identifying features highly correlated with the target variable (recurrence status). Features with correlation values greater than 0.3 or less than -0.3 were selected for model training. This step ensured that only the most relevant features were used, enhancing the model's predictive performance.

These preprocessing steps ensured the dataset was clean and consistent, making it ready for training machine learning models to accurately predict thyroid cancer recurrence.

3. Methods

Workflow Diagram



4 Model Implementation

We re-implemented the following ML algorithms:

- **Support Vector Machine** (SVM), **Decision Tree** (DT), **Random Forest** (RF), **K-Nearest Neighbors** (KNN), **Artificial Neural Network** (ANN)

4.1 Decision Tree Model

Decision trees divide the feature space into smaller regions based on feature values, selecting splits that maximize purity (Gini impurity). They're used for classification and regression tasks.

4.1.1 Model Training and Hyperparameter Tuning

We employed the Decision Tree Classifier from the scikit-learn library. To find the optimal hyperparameters, we used Grid Search Cross-Validation (GridSearchCV with 10 folds), which searches over specified parameter distributions.

4.1.2 The loss function

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{C} (p_i)^2$$

4.1.3 Best Parameters

The Grid Search Cross-Validation identified the following best parameters for our Decision Tree model:

- **criterion**: gini, **splitter**: best, **min_samples_split**: 5, **min_samples_leaf**: 2

4.2 Random Forest Model

Random Forest is an ensemble learning method that builds multiple decision trees during training. Each tree is trained on a random subset of the data and considers a random subset of features for splitting. During prediction, individual tree predictions are combined through averaging (for regression) or voting (for classification). This approach helps reduce overfitting and improves generalization performance, making Random Forest effective for classification and regression tasks, particularly with high-dimensional or noisy data.

4.2.1 Model Training and Hyperparameter Tuning

We employed the Random Forest classifier from the scikit-learn library. To find the optimal hyperparameters, we used Randomized Search Cross-Validation (GridSearchCV with 3 folds), which searches over specified parameter distributions.

4.2.2 The loss function

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{C} (p_i)^2$$

4.2.3 Best Parameters

The Randomized Search Cross-Validation identified the following best parameters for our Random Forest model:

- **n_estimators**: 273, **max_features**: 'sqrt', **max_depth**: 30, **min_samples_split**: 3, **min_samples_leaf**: 2, **bootstrap**: True

4.3 KNN Model

KNN classifies data points based on the majority class of their nearest neighbors, using a similarity measure like Euclidean distance. It's effective for classification and regression tasks

4.3.1 Model Training and Hyperparameter Tuning

We employed the K-Nearest Neighbors (KNN) classifier from the scikit-learn library. To find the optimal hyperparameters, we used Grid Search Cross-Validation (GridSearchCV with 5 folds), which exhaustively searches over specified parameter grids.

4.3.2 Best Parameters

The Grid Search Cross-Validation identified the following best parameters for our KNN model:

- **n_neighbors**: 5, **weights**: 'distance', **metric**: 'manhattan',

4.3.4 Distance metric (Manhattan)

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

4.4 SVM Model

SVM finds the hyperplane that best separates classes in feature space, maximizing the margin between classes. It's versatile for linear and non-linear classification and regression tasks using different kernel functions.

4.4.1 Model Training and Hyperparameter Tuning

We employed the Support Vector Machine (SVM) classifier from the scikit-learn library. To find the optimal hyperparameters, we used Grid Search Cross-Validation (GridSearchCV with 5 folds), which exhaustively searches over specified parameter grids.

4.4.2 Best Parameters

The Grid Search Cross-Validation identified the following best parameters for our SVM model:

- **n_neighbors**: 5, **weights**: 'distance', **metric**: 'manhattan',

4.5 Artificial Neural Network Model

ANNs consist of interconnected nodes organized into layers, learning complex patterns through weighted connections and activation functions. They excel in classification, regression, and feature learning tasks.

4.5.1 Model Training and Hyperparameter Tuning

We constructed a neural network model using the TensorFlow and scikit-learn libraries. The model architecture was defined using the Sequential API from the tensorflow.keras.models module, allowing for a sequential arrangement of layers. To identify the optimal hyperparameters, we employed Grid Search Cross-Validation (GridSearchCV with 3 folds), which exhaustively searched over a predefined grid of hyperparameters.

4.5.3 Best Parameters

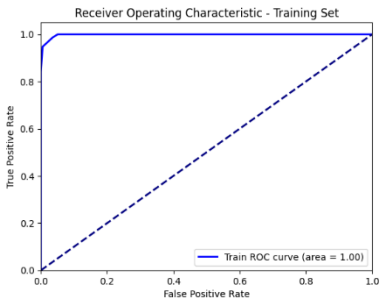
The grid search identified the following optimal hyperparameters:

- **Optimizers**: Adam, **Weight Initialization Methods:** Glorot Uniform, **Batch Size**: 10, **Number of Epochs**: 150

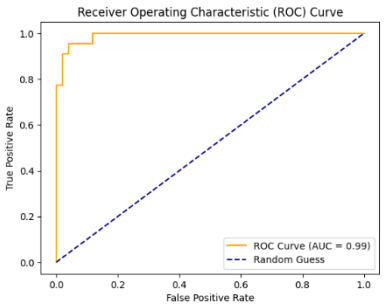
4.6 Models Evaluation

	DT	RF	KNN	SVM	ANN
TRAIN ACCURACY	96.7	97.9	98.17	96.06	97.24
TEST ACCURACY	95.6	95.8	96.7	97.27	97.27
TEST SENSITIVITY	90.62	90.9	96.87	94.73	97.36
TEST SPECIFITIY	98.31	98.04	96.64	98.61	96.22
TEST AUC	98.52	98.9	96.58	99.12	97.29
TEST SIZE	25%	20%	25%	30%	30%

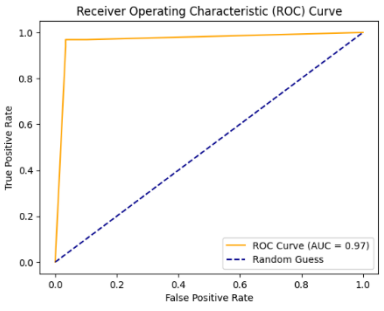
4.7 ROC Curves



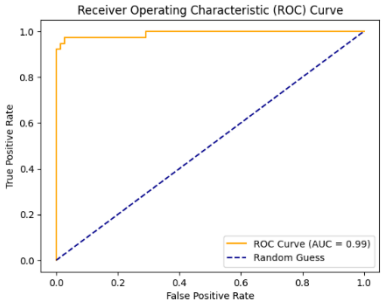
Decision Tree



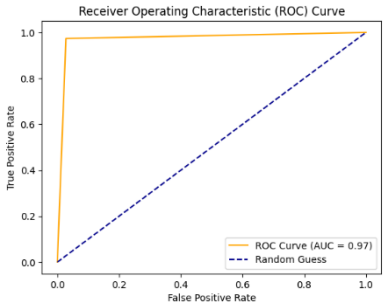
Random Forest



KNN



SVM



ANN

4.7 Ensemble Model

Ensemble methods combine predictions from multiple models to improve performance by leveraging the strengths of each model. In our project, we used stacking and voting classifiers.

The stacking classifier uses a meta-model to combine predictions from base models, enhancing overall accuracy. The voting classifier uses majority voting to aggregate predictions from multiple models, increasing robustness.

Both stacking and voting classifiers achieved good accuracy, demonstrating the effectiveness of ensemble methods in predicting thyroid cancer recurrence.

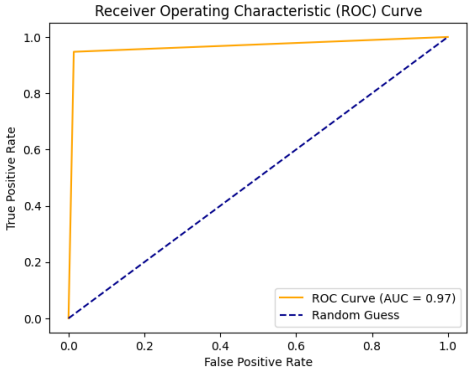
4.7.1 Evaluation

The ensemble model exhibited superior performance compared to the individual models. Key findings are summarized as follows:

Voting Classifier:
- **Sensitivity**: 94.73%, **Specificity**: 98.61%, **AUC**: 0.97, **Accuracy**: 97.27%

Stacking Classifier:
- **Sensitivity**: 94.73%, **Specificity**: 97.22%, **AUC**: 0.97, **Accuracy**: 96.36%

The ensemble models achieved the highest accuracy among all models, thereby effectively enhancing risk stratification for recurrence prediction



5. Discussion

The integration of multiple ML algorithms into an ensemble model enhances the robustness and reliability of recurrence risk predictions. This model's superior performance underscores the potential of ensemble learning in clinical decision-making, providing a more accurate tool for guiding treatment and follow-up strategies.

6. Conclusion

In conclusion, our study highlights the potential of machine learning and deep learning algorithms in predicting thyroid cancer recurrence risk. By implementing ensemble methods, we achieved higher accuracies compared to individual models, demonstrating the effectiveness of our approach. While our ensemble approach yields promising results, there remains ample room for future work. Exploring more sophisticated ensemble techniques, such as stacking or boosting, could further enhance predictive accuracy. Additionally, incorporating novel data sources or biomarkers, such as genetic markers or molecular signatures, may offer deeper insights into recurrence risk factors and improve the precision of our models. Continued research in this area holds the potential to refine risk prediction models and ultimately improve outcomes for patients with thyroid cancer.

Contributors:

Name	Code	Paper
Abdulrahman Shawky	KNN and SVM	Abstract & Methods
Amgad Atef	Decision Tree	Introduction & Methods
Muhammad El-Sayed	Random Forest	Data Description & Preprocessing
Ziyad Hossam	ANN and Ensemble	Data Description, Preprocessing & Conclusion

References:

[1] Shiva Borzooei, G. Briganti, Mitra Golparian, J. R. Lechien, and Aidin Tarokhian, "Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study," *European Archives of Oto-Rhino-Laryngology*, Oct. 2023, doi: https://doi.org/10.1007/s00405-023-08299-w.

[2] J. Wu *et al.*, "The Prospective Implementation of the 2015 ATA Guidelines and Modified ATA Recurrence Risk Stratification System for Treatment of Differentiated Thyroid Cancer in a Canadian Tertiary Care Referral Setting," *Thyroid*, vol. 32, no. 12, pp. 1509–1518, Dec. 2022, doi: https://doi.org/10.1089/thy.2022.0055.

[3] Hasna El Haji *et al.*, "Evolution of Breast Cancer Recurrence Risk Prediction: A Systematic Review of Statistical and Machine Learning–Based Models," *PubMed*, no. 7, Aug. 2023, doi: https://doi.org/10.1200/cci.23.00049.

[4] "ATA Professional Guidelines | American Thyroid Association," *American Thyroid Association*, 2016. https://www.thyroid.org/professionals/ata-professional-guidelines/

The used packages:
- Pandas: For data manipulation and analysis.
- NumPy: For numerical computations.
- Matplotlib: For creating static, animated, and interactive visualizations.
- scikit-learn: For machine learning algorithms and data preprocessing.
- scikeras: For integrating scikit-learn with Keras models.
- TensorFlow Keras: For building and training deep learning models.
- Pickle: For serializing and de-serializing Python object structures.