Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 1s (2025)

Stacking Classifier for Enhanced Detection of Thyroid Cancer Recurrence- A Novel Approach

[1]Avijit Kumar Chaudhuri, [2]Pranab Gharai, [3]Mithun Biswas, [4]Sulekha Das, 5Ranjan Banerjee, [6]Amartya Ghosh,[7]Payel Sengupta

[1]Professor Computer Science and Engineering

Brainware University

c.avijit@gmail.com

[2]Assistant Professor Computer Science and Engineering

Brainware University

pranab.g10@gmail.com

[3]Assistant Professor Computer Science and Engineering

Brainware University

mithunbiswas0707@gmail.com

[4]Research Scholar Information Technology, GCECT

shu7773sea@gmail.com

[5]Assistant Professor Computer Science and Engineering

Brainware University

rnb.cse@brainwareuniversity.ac.in

[6]Assistant Professor Computer Science and Engineering

Brainware University

com.amartya@gmail.com

[7]Assistant Professor Computer Science and Engineering

Brainware University

Payel9433@gmail.com

Article History:

Received: *02-12-2024*

Revised: *19 -01- 2025*

Accepted: *29-01-2025*

Abstract - The performance of different machine learning models for predicting well-differentiated thyroid cancer recurrence is compared in this study using several accuracy metrics such as accuracy, sensitivity, precision, F1 score, specificity, the area under the curve (ROC), and Kappa statistics. The models that the paper considered for ranking are Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), and the proposed Stacked model. The results suggest that the use of ensemble learning methods, especially the proposed Stacked model, results in a generalized improvement over individual classifiers in terms of most of the measures. From Stacked models, there was a boosted level of sensitivity, precision, and F1-score, and the AUC in the higher train-test split (such as 80-20%) and 30-fold cross-validation where the accuracy was at par 100% and consistent. Random Forest also showed good accuracy of results and increased their speed when working with large data sets. The best outcomes were achieved using Decision Trees depending on the 80-20 split and 30-fold cross-validation. However, in Naive Bayes, which was used as a baseline, all the metrics were the lowest, indicating its inapplicability

to this data set. Among the ensemble models, the newly designed Stacked model is the best for prediction accuracy of thyroid cancer recurrence; Random Forest is preferred for volume datasets. The results imply that using ensemble methods of constructing classifiers and selecting training data splits are indicative of operationalizing better models in intricate classification problems.

Keywords: Differentiated Thyroid Cancer, Machine-Learning Classifiers, Stacking Classifier

Introduction

Thyroid cancer starts in the thyroid gland, a small gland shaped like a butterfly and situated at the base of the neck just below the larynx. It is an endocrine gland that secretes hormones, and its most important hormones are thyroxine (T4) and triiodothyronine ((T3) hormones, which control metabolism and heart rate and supply the body with energy. Thyroid cancer, as mentioned, is slow to develop and may not be seen for several years, provided there are no signs in the initial stage[1].

Symptoms and Detection

Signs of throat cancer usually do not manifest or manifest in very mild ways. A thyroid nodule or a lump in the neck may often be felt or noticed. Other symptoms may include dysphagia, persistent hoarseness or voice changes, breathing problems, and, in some instances, neck or earache. These symptoms require additional assessment, such as physical Examination, blood tests, thyroid function tests, and ultrasonography.

These developments in diagnostic imaging skills have improved thyroid cancer visualization. The increased number of observed thyroid nodules can be explained by the availability and increased usage of high-resolution ultrasonography and ultrasound (US)-guided fine needle aspiration (FNA) biopsies. According to the GLOBOCAN 2023 report, thyroid cancer is estimated to be the seventh most common cancer globally, and it comprises roughly 1% of all kinds of cancer[2,3].

Gender Disparity and Prevalence

Contrary to this case, it may be regarded as a gender-neutral disease; the occurrence of thyroid cancer is higher among females. Thyroid cancer affects females more than males, with at least 75% of the patients being females, and it is the seventh most common cancer among women. Other researchers believe that hormonal factors, especially estrogen and progesterone, are some of the reasons for such a gender disparity. Other predisposing factors include autoimmune thyroid diseases like Hashimoto Thyroiditis and iodine levels as well as other environmental influences[4,5].

Types of Thyroid Cancer

There are different types of thyroid cancer, and the most common is the well-differentiated thyroid cancer (WDTC), which is usually the slowest-growing type. WDTC includes two main subtypes:

1. Papillary Thyroid Cancer (PTC): This is the most frequent type, making up between 80% and 85% of all cases of thyroid cancer. It is usually slow-growing and commonly extends to nearby lymph nodes, even though it rarely metastasizes to other organs.

2. Follicular Thyroid Cancer (FTC): Still less frequent than PTC, FTC constitutes 10-15 % of all cases. It has been known that FTC can metastasize to any organs at distant sites, including the lungs or the bones, but generally, fatality rate of FTCs are less than many other types of cancer.

Both types of WDTC originate from the follicular cells of the thyroid gland that are responsible for synthesizing and secreting thyroid hormones. These cancers look a lot like normal thyroid tissue and, in most cases if caught early, are not life-threatening.

Prognosis and Treatment

The overall survival rate of WDTC is usually excellent. WDTC has an excellent prognosis when diagnosed at an early stage, with five-year survival rates frequently in access of 98%. Thyroidectomy, the removal of a portion of or the entire thyroid gland, is the general treatment provided. In addition, based on the size and degree of disbursement of the tumor, the surgeon may also remove specific lymph nodes to curtail metastasis.

However, when the disease extends to glands other than the thyroid, Radioactive Iodine(RAI) therapy may be administered after surgery. RAI is helpful in the ablation of any extra-thyroidal disease or the elimination of micro-invasive cancer cells that may accompany the thyroid lesion after surgery. Furthermore, the patients may have to undergo a lifetime of thyroid hormone supplementation since the absence of the thyroid gland means the body lacks thyroid in its metabolic processes[6,7].

Follow-up and Recurrence Risk

However, adequate follow-up is essential because of the typical presentation of well-differentiated thyroid cancer, whose prognosis is otherwise reasonably good. They also found that even in patients with NOV, which is no visible sign of the malignant cells in their body, there is still the possibility of relapse, even if a small one. Recurrence can be years or sometimes decades after treatment, affecting the lymph nodes or distant sites like the lungs or bones. In most cases, the patient is followed up using imaging, blood tests for thyroglobulin, a tumor marker, and a physical examination.

Thyroid cancer, although not frequent, has increased trends in recent years owing to improved diagnostic tools. It is commonly diagnosed in women, and the two most frequent subtypes, papillary and follicular thyroid carcinomas, are generally indolent with a favourable prognosis. Huge impacts on early diagnosis by state-of-the-art imaging and early operative intervention with RAI therapy where required have increased the patient survival rate. Nevertheless, they need long-term follow-up to address relapse and overall patient health.

Classification of diseases has been a critical area where ML algorithms are widely applied due to their capacity to capture patient information about diseases and place them into requisite classes. These algorithms involve themselves in understanding the complicated trends and dependencies of the data; therefore, they play a critical role in making diagnoses

of the medical inputs, which may include symptom assessment, lab results, and imaging. Consequently, by analyzing this data, the ML models can guess which disease a specific patient might suffer from or their probability of contracting one or more illnesses.

This is why the evaluative aspect of machine learning depends on the model's ability to identify patterns associated with particular diseases. For instance, from a set of patient's symptoms and corresponding diagnoses, an algorithm can develop a pattern to classify future patients into different diagnosis categories based on the new data set. It is often employed in circumstances that require an early and accurate diagnosis in clinical practice whenever a condition is minimally manifest or represented in a complicated form. Apart from simply categorizing knowledge, the machine learning models also help in diagnosis by pointing out the diseases that one is likely to develop and prognosis, including predicting a recurrence of the disease[8,9].

In clinical practice, machine learning (ML) models, especially models based on deep learning (DL), have shown high prediction accuracy in interpreting medical images. This is highly successful in caring for diseases such as cancer through procedures that use MRI and CT scans. They can precisely identify irregularities within the images as effectively as or even better than the doctors who specialize in treating specific diseases[10].

Relevant Literature

Thyroid cancer is one of the uncommon types of cancer, but if not diagnosed in its early stage, it can have severe implications for the health of the affected individual. It develops on the throat and controls the energy levels, pulse rates, and body temperatures through the hormones it produces; this gland may not show symptoms in its preliminary stage and thus cannot be easily diagnosed. Some symptoms are associated with a swelling in the thyroid gland these are neck lumps and difficulties swallowing. The last decade witnessed a significant rise in the development of ML, which improved diagnostic and prognostic indicators for thyroid cancer. Most recently, deep learning methods have often been applied to image data, including US and CT images, in the diagnosis of thyroid nodules and differentiation between benign and malignant ones with high diagnostic performance. In addition, the prognosis of the result and the adaptation of a treatment plan for the patient have also been aided by ML. Research that has been conducted has shown that artificial intelligence (AI) holds great promise by assisting in the early diagnosis of disease and refusal-making regarding the identification of the right treatment plans for the diseases[11,12]. These technologies are helping in improved survival of different types of thyroid cancer patients, thus signifying that the patient management outcomes are competent.

[13]Cao et al., 2021 studied machine learning for detecting differentiated thyroid cancer (DTC). Radionics, a quantitative image extraction technique, was used to identify DTC in medical images. Their research highlighted radionics as a significant advancement in medical image analysis, as it enables the efficient extraction of quantitative features that can be used as machine learning inputs to predict DTC presence.

[14]Zhu et al., 2022 also investigated machine learning for understanding DTC tumor behavior and outcomes, employing an unsupervised clustering approach. They introduced the

Ensemble Algorithm for Clustering Cancer Data (EACCD) to develop predictive algorithms for well-differentiated thyroid cancer. This method involves three key steps: establishing initial dissimilarities with the Gehan-Wilcoxon test, learning dissimilarities, and performing hierarchical clustering. They used data on well-differentiated thyroid cancer cases from 2004 to 2021 from the National Cancer Institute's SEER database.

Using the SEER database, [15]Liu et al., 2022 also focused on supervised machine-learning approaches to predict lung metastasis in thyroid cancer. They built six models, including SVM, logistic regression, XGBoost, decision trees, random forest, and k-nearest neighbors, with random forest proving the most effective.

In another study, [16]Shin et al., 2020 applied machine learning to differentiate follicular adenomas of the thyroid, employing supervised learning models like artificial neural networks (ANN) and SVM on patient data from two hospitals in South Korea from 2012 to 2015. Similarly, [17]Masuda et al. 2021 used an SVM-based model to identify lymph node metastasis in thyroid cancer. They trained it on data from 117 patients to classify metastatic versus benign lymph nodes, achieving an area under the curve (AUC) of 0.64.

In deep learning for thyroid cancer, Zhao et al., 2021 explored CNN models, including Xception, SE-ResNeXt50, DenseNet169, DenseNet121, and ResNet50, to differentiate malignant from benign thyroid nodules. They trained these models on Northern Jiangsu People's Hospital CT images involving 880 patients from 2017 to 2019.

[18]Chan et al. investigated DTC diagnosis using CNNs like VGG19, ResNet101, and InceptionV3 on Chang Gung Memorial Hospital ultrasound images. InceptionV3 achieved the best accuracy rate at 83.7%, with ResNet101 and VGG19 at 72.5% and 66.2%, respectively.

Materials and Methods

Data Source

The Differentiated Thyroid Cancer Recurrence dataset, sourced from the University of California at Irvine Machine Learning Repository, was utilized in this study[19]. It includes retrospective clinical data from 383 patients diagnosed with Differentiated Thyroid Cancer (DTC), all of whom were followed for at least 10 years. The dataset contains 16 clinical features: age at diagnosis, gender, current smoking status, prior smoking history, history of head and neck radiation, thyroid function, presence of goitre, presence of adenopathy on physical examination, cancer pathological subtype, focality, ATA risk assessment, TNM staging, initial treatment response, and recurrence status. Among the patients, 312 (81%) were female and 71 (19%) were male, with an average age of diagnosis of 41 years. The pathological subtypes included 287 Papillary (75%), 48 Micropapillary (13%), 28 Follicular (7%), and 20 Hurthel Cell (5%) cases. Based on the ATA risk classification, 249 patients (65%) were categorized as low risk, 102 (27%) as intermediate risk, and 32 (8%) as high risk. Most cases (333, or 87%) were classified as Stage 1. In terms of initial treatment response, 208 patients (54%) had an excellent response, while 91 (24%) had structural incomplete, 61

(16%) had indeterminate, and 23 (6%) had biochemical incomplete reactions—a total of 108 patients (28%) experienced recurrence. The dataset does not contain any missing values.

The dataset is available for access at

https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence. Descriptions of the 16 features of the dataset are provided in Table 1 below.

Table 1. Explanations about 16 different features used in the study

Features	**Features Type**
Age	Refers to the age of individuals in the dataset
Gender	Specifies the gender of individuals (e.g., Male or Female)
Smoking	Related to smoking behavior, with further investigation needed to clarify the specific values or categories
Smoking History	Identifies whether individuals have a history of smoking
Radiotherapy History	Identifies whether individuals have received radiotherapy treatment.
Thyroid Function	Potentially refers to the condition or function of the thyroid gland
Physical Examination	Provides details from a physical examination, likely focused on the thyroid
Adenopathy	Identifies the presence and location of adenopathy (enlarged lymph nodes)
Types of Thyroid Cancer (Pathology)	Categorizes the types of thyroid cancer based on pathology, including specific subtypes like "Micropapillary Papillary", "Follicular" and "Hürthle cell."
Focality	Specifies whether thyroid cancer is unifocal or multifocal.
Risk	Denotes the risk classification related to thyroid cancer
Tumor	Describes the T (Tumor) stage of thyroid cancer, reflecting the size and spread of the primary tumor.
Lymph Nodes	Determines the N (Node) stage of thyroid cancer; relates to involvement of nearby lymph nodes.
Cancer Metastasis	Represents M (Metastasis) stage of thyroid cancer, which demonstrate if cancer has affected distant organs.
Stage	Represents the general state of thyroid cancer, which results from T, N, and M stages.
Treatment Response	Describes the response to the treatment and has the options of “Indeterminate,” “Excellent” “Structural Incomplete,” and “Biochemical Incomplete.”
Recurred	Indicates the presence or absence of recurrence of thyroid cancer

Methods

In this study, the stacking model works at the stacking level to improve the predictive performance by systematically combining several classifiers. Here is a detailed breakdown of the approach:

Data Preparation: The dataset is preprocessed by a train and test split using a 10-fold cross-validation to enhance the validation. StandardScaler is also used before feature rescaling to adjust the range of features to 0-1, which is beneficial when all inputs are approximately on the same scale, which is expected in many models of this study.

Base Learners: Two base classifiers, Random Forest and Gradient Boosting, are chosen. The classifiers are tuned through the GridSearchCV, which allows us to find the best versions of the model in a grid by adjusting the parameters.

Stacking Classifier: The optimized results of the base models are then connected through a Stacking Classifier. In the case of training, each base model produces predictions on the training data and subsequently generates a dataset from these outputs. This new dataset then makes the basis for the other stages in the stacking process.

Meta-Learner: The final estimator we use is the MLP Classifier, and a new dataset that consists of the base learners' outputs is used to feed this final estimator. The meta-learner learns the combination of the base models, which means the meta-learner learns when and how to use the classifiers to minimize the total classification error.

Prediction and Evaluation: The fitted stacking model is trained with a test set, and evaluation parameters include accuracy, confusion matrix, ROC AUC, and Cohen's Kappa. These allow one to estimate the model's overall effectiveness.

This hierarchical notion uses many learning algorithms that enhance precision classification, making stacking a significant sign in machine learning. The flowchart of the experimental design and model building and the flowchart are depicted in the Figure 1 and Figure 2.

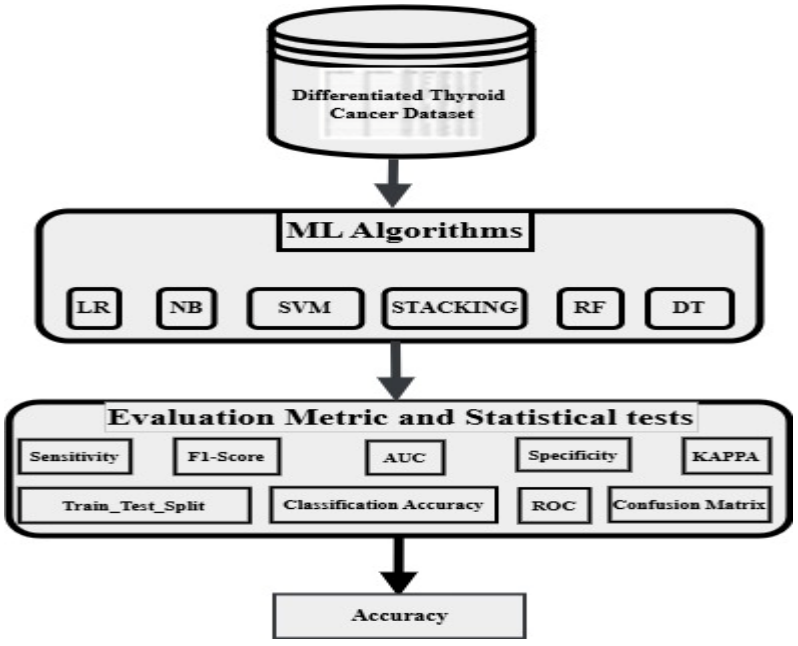

Figure 1. WDTC prediction model

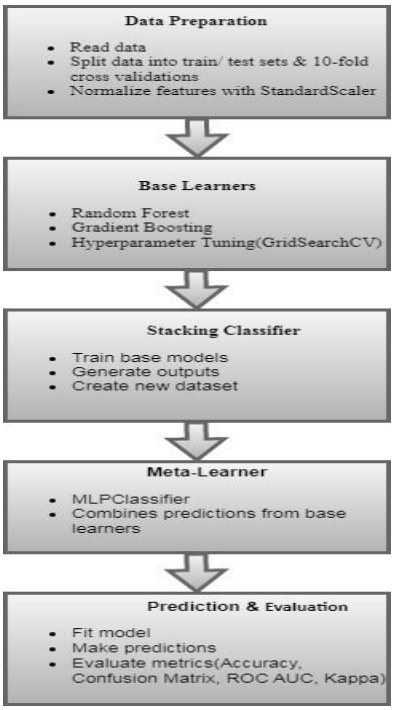

Figure 2. Flow Chart WDTC prediction model

Algorithm

Stacked Ensemble Model with Pre-processing and Cross-validation

Step 1. Preprocessing of the Dataset

- **Train-Test Split**:

  $The\ dataset\ D = \{X, y\} is\ first\ divided\ into\ a\ training\ set\ D_{train} = \{X_{train}, y_{train}\} and\ a\ test\ set\ D_{test} = \{X_{test}, y_{test}\}$

  $The\ training\ set\ is\ partitioned\ into\ thirty\ subsets, denoted\ as$

  $D_{train}^{(1)}, D_{train}^{(2)}. \ldots, D_{train}^{(30)}$

  $For\ each\ fold\ i, the\ model\ is\ trained\ on\ D_{train}^{(-i)} = D_{train} \setminus D_{train}^{(i)}\ and\ validated\ on\ D_{train}^{(i)}$

  Feature Scaling:

- $Standard\ Scaler\ is\ applied\ to\ the\ features\ X\ to\ ensure\ that\ each\ feature\ is\ centred$
- $(mean = 0) and\ scaled\ (variance = 1): X_{scaled} = \frac{X-\mu x}{X-\sigma x}\ represent\ the\ mean\ and\ standard\ deviation\ of\ X, respectively.$
- $This\ operation\ ensures\ all\ features\ are\ on\ the\ same\ scale, which\ is\ important$
- $for\ many\ models, including\ neural\ networks\ and\ distance-based\ algorithms.$

  Step 2. Base Classifiers

- **Random Forest Classifier**:

- Random Forest (RF) is an ensemble of decision trees. The final random forest prediction is obtained by aggregating the predictions of individual trees (via majority voting for classification problems).
- $The\ mathematical\ formulation\ for\ a\ single\ tree's\ prediction\ is\ f_{tree}(x) = sign(\sum_{i=1}^{N} w_i\ .\ h_i(x))where\ h_i(x)is\ the\ prediction\ of\ tree\ i, and\ w_i\ is\ the\ weight\ of\ the\ i-th\ tree$
- **Gradient Boosting Classifier**:
- $Gradient\ Boosting\ builds\ a\ sequence\ of\ weak\ learners\ (typically\ decision\ trees)\ where\ each\ subsequent\ tree\ corrects\ the\ errors\ made\ by\ the\ previous\ one.$

$The\ model\ can\ be\ formulated\ as\ F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)\ where\ h_m(x)\ is\ the\ prediction\ of\ the\ m - th\ weak\ learner\ and\ \gamma_m\ is\ the\ corresponding\ weight.$

Step 3. Hyperparameter Tuning with GridSearchCV

- **GridSearchCV** is used to search over a specified parameter grid to find the best combination of parameters for each base model:
- Define a grid of hyperparameters $P = \{p_1, p_2, \ldots, p_k\}$ for each classifier.
- Perform cross-validation on each combination of hyperparameters $\mathrm{p_i}$ to identify the optimal configuration:

$$\hat{\mathrm{P}} = \arg\min_{\mathrm{p_i}} \mathrm{CV\ Loss}\left(\mathrm{p_i}\right)$$

Where $\mathrm{CV\ Loss}\left(\mathrm{p_i}\right)$ is the average cross-validation error for a particular setting of hyperparameters.

Step 4. Stacking Classifier

- **Training the Base Learners**:
- Once the base classifiers (Random Forest and Gradient Boosting) are tuned, they are trained on the full training data $\mathrm{X_{train}}$. Each base classifier produces predictions $\hat{\mathrm{y}}_{\mathrm{RF}}(\mathrm{X})$ and $\hat{\mathrm{y}}_{GB}(\mathrm{X})$.
- These predictions form a new dataset $\mathrm{X_{stack}} = [\hat{\mathrm{y}}_{\mathrm{RF}}(\mathrm{X}), \hat{\mathrm{y}}_{GB}(\mathrm{X})]$.

Meta-Learning (Final Estimator):

- A **Meta-Learner** (Multi-Layer Perceptron Classifier-MLPClassifier) is trained on the new dataset $\mathrm{X_{stack}}$.

- The goal of the meta-learner is to learn the best combination of the base classifiers' predictions to minimize the total classification error.
- The output of the final meta-learner is:

$$\hat{y}_{stack} = \text{MLPClassifier}\left(X_{stack}\right)$$

The described approach combines multiple machine-learning techniques into a cohesive pipeline to optimize classification performance. The working of the proposed method is depicted in Figure 3

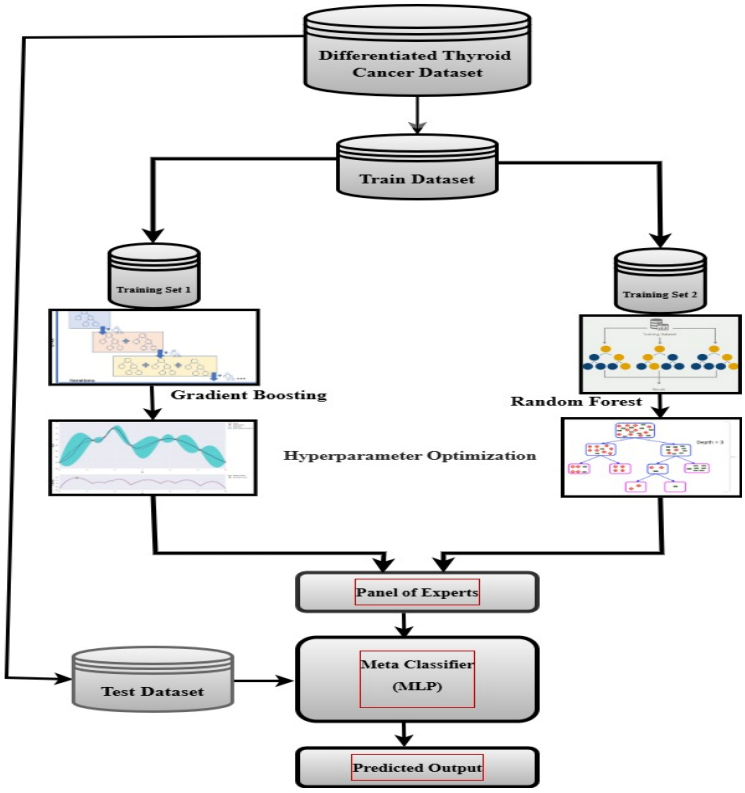

Figure 3. Working of the proposed classifier with n=2

Random Forest (RF): In this research authors evaluate machine learning approaches for thyroid cancer prognosis. It has been shown from the previous studies that Random Forest (RF) classifier rank among the most accurate tools for reoccurrence prediction, specifically in patients with DTC. The strength of the RF models is their ability to deal with the high dimensionality of the data and feature selection, which may help temper the overfitting problem a little when developing cancer prognosis models. They are instrumental in being highly specific for DTC, with accuracy and specificity as high as 94% in the recent comparative studies where RF surpassed other models, including SVM and logistic regression models, regarding sensitivity and consistency[20].

Subsequent research explores the combination of RF models with clinical features, including patient age, metastasis-affected lymph node ratio, and tumor morphology, to create more accurate prediction models. These models are especially helpful in identifying patients at

higher risk of recurrence and helping clinicians decide on appropriate treatments where certain cases may need more attention[21].

Gradient Boosting(GB): Gradient boosting is one of the most popular ensemble learning methods to construct complex predictive models because it can successfully work with nonlinear relationships between factors. An XGBoost or LightGBM system can be used as a gradient-boosting framework for evaluating the clinical, pathological, and molecular risk factors for recurrent thyroid cancer. The procedure includes data pre-processing that provides for the treatment of missing values, management of categorical features through encoding, and feature-scaling. The feature selection techniques such as recursive features elimination or SHAP(Shapley Additive exPlanations) values may be employed to select important features. The model is constructed on the framework based on labelled data in a supervised learning algorithm, structured in sequential decision trees is iteratively built to solve classification or regression by solving the loss functions for log loss or mean squared error. Hyperparameter tuning using the grid search method or Bayesian optimization guarantees the model result. The model's performance uses the area under the receiver operating characteristic curve (AUC-ROC) and cross-validation to reduce overfitting. The reasons for using this approach lie in the high performance in cancer recurrence prediction and the ability to quickly and easily interpret the results using feature importance analysis[22,23].

Results and Discussions

The machine-learning approaches yielded different levels of accuracy as shown in Table 2.

Table 2: Confusion matrix and performance evaluation metrics and statistical tests

S/N	Metrics	Formula/Description			
1	Confusion Matrix			Actual	
				Without CVD	With CVD
		Predicted	Without CVD	*True Positive (TP)*	*False Positive (FP)*
			With CVD	*False Negative (FN)*	*True Negative (TN)*
2	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}*100$			
3	Sensitivity	$\frac{TP}{TP+FN}*100$			
4	Specificity	$\frac{TN}{TN+FP}*100$			
5	Kappa statistics	$\frac{pa-pac}{(1-pac)}$,'pa' represents total agreement probability and 'pac' represents probability 'by chance'. Its range is (-1,1).			
6	Area under the curve (AUC)	Receiver operating characteristic (ROC) is plotted between Sensitivity and (1-Specificity). The area under the curve (AUC) measures the degree to which the curve is up in the northwest			

corner.

As shown in Table 3, Table 4, Table 5, Table 6, Table 7, Table 8, and Table 9 the machine-learning approaches yielded different levels of accuracies, sensitivity, precision, f score, specificity, AUC-ROC, and kappa statistics respectively.

Table 3. Comparison of Accuracies

Train-Test Split	**LR**	**NB**	**SVM**			**STACKED**
50-50	91	89	88	91	95	93
66-34	89	86	88	94	90	97
80-20	92	88	90	96	95	95
30 fold c.v	87	87	86	95	92	96.05

Comparison of Accuracies-Key Insights

Train/ Test split: Models obtain superior performance with a more extensive training set, as revealed in the 80-20 split and the 30-fold cv.

Ensemble Classifiers: This means that the Stacked model is superior to individual models, showing the advantage of assembling several algorithms.

Selection of Classifiers: Naive Bayes can be eliminated from being a good choice based on the evaluation outcome. At the same time, Random Forest and the Stacked model are excellent choices for this data set.

Random Forest and Stacked outperform other models and have the highest levels of accuracy for significantly larger configurations within the training data. Like other models, decision trees also depict the model's reliable performance. Naive Bayes, as simple as it is, disappoints in this case as well. To improve performance in future evaluations, the analysis could be further expanded, for instance, through hyperparameter optimization, more models, or constructive feature engineering methods.

Table 4. Comparison of Sensitivity(recall)

Train-Test Split	**LR**	**NB**	**SVM**			**STACKED**
50-50	91	89	88	91	95	99
66-34	89	86	88	94	90	99

80-20	92	88	90	96	95	95
30 fold c.v	87	87	86	95	92	96.05

Comparison of Sensitivity-Key Insights

Ensemble Classifiers: The Stacked model is more accurate than the individual models at the 50/50 and 66/34 splits, proving the use of ensembles.

Train/ Test split: As in other analyses, larger training sets (80-20) provide better performance, yet the Stacked model remains insensitive to the choice of training set size.

Variability in Performance: LR and SVM exhibit some fluctuation, and the stability of these two algorithms decreases with the reduced training sample size. This means it may not be as accurate and will need more tuning or a different set of features selected to improve its accuracy.

Table 5. Comparison of Precision

Train-Test Split	**LR**	**NB**	**SVM**	**DT**		**STACKED**
50-50	91	88	88	93	95	93
66-34	89	86	88	95	90	97
80-20	92	90	89	96	95	95
30 fold c.v	91	86	88	96	94	96.90

Comparison of Precision-Key Insights

Best Performing Models:

For the 66-34 split, the Stacked Model has the highest measured precision (97%); for all the other splits, it delivers a perfect score or, at worst, a few percentage points lower than the best algorithm in the stack. RF has a 95% and above accuracy in most and has shown better performance and reliability. The DT has high results in the 80-20 split and 30-fold cross-validation at 96%. One benefit involves interpretability significantly, as it could quickly fall prey to the issue of overfitting. LR has constant accuracy from 89% to 92 %, which the researchers used as a base model. Motion may not cause, but it offers steady and stable performance. NB performs worst always, with precision values slightly higher than 90% for most cases. This indicates that it could be folly to base feature selection on the assumption that all features are independent.

Train/ Test split:

The performance of models differs much depending on the train-test split. The Stacked model wins big with the 66-34 ratio, and Decision Trees and Random Forests win with the 80-20 split.

Cross-Validation Robustness:

The performance of the models tends to vary significantly during a one-fold cv, but it can be considered stable enough in terms of 30-fold cross-validation; in this case, the Stacked model yields an accuracy of 96,90%, and Random Forest – 94%.

Table 6. Comparison of F-1 Score

Train-Test Split	LR	NB	SVM	DT		STACKED
50-50	91	88	88	91	95	93
66-34	89	85	87	94	90	97
80-20	92	87	89	96	95	95
30 fold c.v	89	86	87	96	93	97

Comparison of F-1 Score-Key Insights

Train/ Test split:

Consistency: The performance of Random Forest and Stacked models stays high regardless of the choice of train-test splits. These statistics prove their stability and good performance when tested on unseen data.

Sensitivity: Other models, such as Logistic Regression and SVM, present different results depending on the different train/ test data split, probably because these methods might be more sensitive to an amount of training data and other dataset features.

Ensemble Classifiers:

The Stacked model in the experiments performed better than when any model was used singly in some splits, underlining the gains made by forming ensembles out of models. This could use the advantages of the multiple models, thereby producing better forecasts and better addressing different data distributions.

Table 7. Comparison of Specificity

Train-Test Split	LR	NB	SVM	DT		STACKED
50-50	94	89	93	99	95	80
66-34	90	86	89	99	91	92.3

80-20	93	87	92	98	95	94.11
30 fold c.v	78	85	77	92	88	95.15

Comparison of Specificity-Key Insights

The present analysis of specificity in various machine learning models and train-test splits presents the variable level of performance. Table 7 shows high specificity across all methods, with an LR achieving 94% specificity at a 50-50 split, and Naive Bayes achieves lower specificity at all splits. SVM emerged as more vital, at 93%, in efficiency settings with a distribution of 50-50, and it fell to 77% with 30-fold cross-validation. Decision Trees (DT) specifically detected 99% in the 50-50 and 66-34 splits. The Random Forest (RF) shows pretty good accuracy, with the maximum accuracy being 95% for 50-50 and 80-20 splits. Compared with Stacked models, variability was found, with 30-fold cross-validation giving 95.15%. Thus, the highest increase in accuracy is observed for DT and RF, especially with more significant numbers of training data; at the same time, NB demonstrates low accuracy across all settings.

Table 8. Comparison of AUC (ROC area)

Train-Test Split	**LR**	**NB**	**SVM**	**DT**		**STACKED**
50-50	89	81	85	93	92	94
66-34	84	78	83	95	86	97
80-20	88	76	84	96	91	91.24
30 fold c.v	98	99	96	94	98	100

Comparison of AUC (ROC area)-Key Insights

Ensemble Classifier:

The Stacked model shows high stability across cross-validation but is outstanding in the 30-fold cross-validation, where the AUC is 100. This also shows that building on the idea of ensemble methods, combining several algorithms results in a marked improvement in the model's predictive ability.

Cross-Validation Importance:

The degree of performance variation across the splits emphasizes the need for more sophisticated validation methods, such as 30-fold cross-validation.

Model Interpretability vs. Performance:

Methods such as LR and DT are easy to interpret as opposed to some of the ensemble methods that can be slightly less accurate than Stacked model.

Table 9. Comparison of Kappa Statistics

Train-Test Split	LR	NB	SVM	DT	RF	STACKED
50-50	76	68	69	79	86	82.22
66-34	72	62	68	86	75	93
80-20	78	63	71	90	86	86
30 fold c.v	70	62	63	88	81	90.1

Comparison of Kappa Statistics -Key Insights

For random forest (RF), Kappa statistics comparison with decision trees was high and most promising in 66-34 and 80-20, subjected to Kappa scores of 86 and 90, respectively, for the 80-20 split. The Stacked classifier that combined two models is better than the single classifiers, reaching 93% in the models with 66% in the training set and 34% in the test set. The results revealed that both Logistic Regression (LR) and Support Vector Machine (SVM) had a moderate performance; however, Naive Bayes (NB) was universally low across all splits and, therefore, potentially not suitable for this dataset. In conclusion, RF and Stacked models were the most accurate for the classification task.

Conclusion and directions for future research

In conclusion, the presented results imply that the well-differentiated thyroid cancer recurrence might be associated with the recurrence of different factors like age, microscopic remnants after thyroid cancer surgery, lymph node conditions, tumor aggressiveness, size, and iodine therapy response. In the present research, several machine-learning algorithms were used to evaluate and classify the risk of recurrence in well-differentiated thyroid cancer. As this study illustrates, ensemble models, especially the Stacked classification model, dominate all classifiers in several accuracy matrices such as sensitivity, specificity, precision, F-1 score, kappa score, and AUC/ROC. Random Forest also demonstrated high accuracy and was suitable for large training sets. Nevertheless, the Naive Bayes classifier represented the lowest performance compared to the other approaches, which disclosed its inefficiency in such a scenario. The results reveal that the train/ test split ratio directly influences accuracy where more extensive test data are suggested, such as 80-20%, on decision trees and random forests. Additionally, applying other validation techniques highlights that the 30-fold CV gave our model the most reliable performance. Integrating many algorithms using the ensemble method enhances stability and accuracy compared to single algorithms, indicating that applying such strategies for intricate classification problems is extremely useful.

Also, as several studies have been reviewed in order to identify approaches for DTC recurrence prediction within the paper, the researchers expand the overall understanding of the model's performance and applicability. Such knowledge can be a valuable asset to healthcare workers who are seeking reliable means through which diseases can be diagnosed,

modelled and managed with improved patient outcomes. Therefore, employing a single model to detect multiple diseases demonstrates the idea of model generality and the possibility of leveraging the result to improve the development of healthcare analytics, paving the way for further breakthroughs in disease study and clinical application.

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information: Not Applicable

Author contribution: Sulekha Das: Conceptualization,Writing- Original draft, Software, Investigation, Data Curation, Visualization. **Avijit Kumar Chaudhuri:** Methodology, Validation, Formal Analysis, Supervision, Writing- Review & Editing, Methodology, Validation, Formal Analysis. **Nobhonil Roy Choudhury:** Review & Editing**. Partha Ghosh:** Supervision

Acknowledgments: Not Applicable

Research involving human participants and/or animals: Not Applicable

Informed consent: Not Applicable

Data availability: https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence

References

1. Yaşar, Ş. (2024). Determination of Possible Biomarkers for Predicting Well-Differentiated Thyroid Cancer Recurrence by Different Ensemble Machine Learning Methods. Middle Black Sea Journal of Health Science, 10(3), 255-265.
2. Chen, D., Lang, B., McLeod, D., & Newbold, K. M. (2023). Haymart. Thyroid cancer. Lancet (London England), 401, 1531-44.
3. Chiang, H. T., Fu, S. W., Wang, H. M., Tsao, Y., & Hansen, J. H. (2024). Multi-objective non-intrusive hearing-aid speech assessment model. The Journal of the Acoustical Society of America, 156(5), 3574-3587.
4. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: a cancer journal for clinicians, 73(1), 17-48.
5. Vaish, R., Mahajan, A., Sable, N., Dusane, R., Deshmukh, A., Bal, M., & D'cruz, A. K. (2023). Role of computed tomography in the evaluation of regional metastasis in well-differentiated thyroid cancer. Frontiers in Radiology, 3, 1243000.
6. Chaudhuri, A. K., & Das, S. (2024). The Performance of Feature Selection Approaches on Boosted Random Forest Algorithms for Predicting Cardiovascular Disease. In Computer Vision and AI-Integrated IoT Technologies in the Medical Ecosystem (pp. 288-310). CRC Press.
7. Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A dataset centric feature selection and Stacked model to detect breast cancer. International Journal of Intelligent Systems and Applications, 13(4), 24.

8. Boina, R., Ganage, D., Chincholkar, Y. D., Chinthamu, N., & Shrivastava, A. (2023). Enhancing Intelligence Diagnostic Accuracy Based on Machine Learning Disease Classification. International Journal of Intelligent Systems and Applications in Engineering, 11, 765-774.
9. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. International Journal of Intelligent Systems and Applications, 10(5), 46.
10. Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, 20, 100402.
11. Zhao, H. B., Liu, C., Ye, J., Chang, L. F., Xu, Q., Shi, B. W., ... & Shi, B. B. (2021). A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images. Endokrynologia Polska, 72(3), 217-225.
12. Sun, Z., Wang, C., Zhao, Y., & Ling, Q. (2024). CAR-T cell therapy in advanced thyroid cancer: from basic to clinical. Frontiers in Immunology, 15, 1411300.
13. Cao, Y., Zhong, X., Diao, W., Mu, J., Cheng, Y., & Jia, Z. (2021). Radiomics in differentiated thyroid cancer and nodules: explorations, application, and limitations. Cancers, 13(10), 2436.
14. Zhu, Y., Yang, S., & He, X. (2022). Prognostic evaluation models for primary thyroid lymphoma, based on the SEER database and an external validation cohort. Journal of Endocrinological Investigation, 45(4), 815-824.
15. Shin, I., Kim, Y. J., Han, K., Lee, E., Kim, H. J., Shin, J. H., ... & Kwak, J. Y. (2020). Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland. Ultrasonography, 39(3), 257.
16. Shin, I., Kim, Y. J., Han, K., Lee, E., Kim, H. J., Shin, J. H., ... & Kwak, J. Y. (2020). Application of machine learning to ultrasound images to differentiate follicular neoplasms of the thyroid gland. Ultrasonography, 39(3), 257.
17. Masuda, T., Nakaura, T., Funama, Y., Sugino, K., Sato, T., Yoshiura, T., ... & Awai, K. (2021). Machine learning to identify lymph node metastasis from thyroid cancer in patients undergoing contrast-enhanced CT studies. *Radiography*, *27*(3), 920-926.
18. Chan, W. K., Sun, J. H., Liou, M. J., Li, Y. R., Chou, W. Y., Liu, F. H., ... & Peng, S. J. (2021). Using deep convolutional neural networks for enhanced ultrasonographic image diagnosis of differentiated thyroid cancer. Biomedicines, 9(12), 1771.
19. Mukhtar, H., Qaisar, S. M., & Zaguia, A. (2021). Deep convolutional neural network regularization for alcoholism detection using EEG signals. Sensors, 21(16), 5456.
20. Li, Y., Tian, J., Jiang, K., Wang, Z., Gao, S., Wei, K., ... & Li, Q. (2023). Risk factors and predictive model for recurrence in papillary thyroid carcinoma: a single-center retrospective cohort study based on 955 cases. Frontiers in Endocrinology, 14, 1268282.
21. Aggarwal, A., Kaur, E., & Lu, S. (2024). Comparative Analysis of Machine Learning Models for Thyroid Cancer Recurrence Prediction.

22. Habchi, Y., Himeur, Y., Kheddar, H., Boukabou, A., Atalla, S., Chouchane, A., ... & Mansoor, W. (2023). Ai in thyroid cancer diagnosis: Techniques, trends, and future directions. Systems, 11(10), 519.
23. Bellantuono, L., Tommasi, R., Pantaleo, E., Verri, M., Amoroso, N., Crucitti, P., ... & Bellotti, R. (2023). An eXplainable Artificial Intelligence analysis of Raman spectra for thyroid cancer diagnosis. Scientific Reports, 13(1), 16590.