THYROID DISEASE DETECTION USING MACHINE LEARNING

Ruchi singh, Raushan kumar mishra, Roshan kumar, Ronith samhal·

School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073 .

Abstract

Thyroid disease has been on the rise during the past few years. Owing to its importance in metabolism, early detection of thyroiddiseaseisataskofcriticalimportance.Despiteseveralexistingworksonthyroiddiseasedetection,theproblemofclass imbalance is not investigated very well. In addition, existing studies predominantly focus on the binary-class problem. This study aims to solve these issues by the proposed approach where ten types of thyroid diseases are considered. The proposed approach uses a differential evolution (DE)-based optimization algorithm to fine-tune the parameters of machine learning models. Moreover, conditional generative adversarial networks are used for data augmentation. Several sets of experiments are carried out to analyze the performance of the proposed approach with and without model optimization. Results suggest that a 0.998 accuracy score can be obtained using AdaBoost with DE optimization which is better than existing state-of-the-art models.

1 Introduction

The thyroid is a small, but very important gland in the neck that allows the human body to maintain digestion and heart rate .The thyroid organ releases the hormones that control metabolisms such as body temperature and heart rate. It produces two important hormones, T4 and T3. For several metabolic activities, these hormones are responsible such as heart rate and body weight. The thyroid gland produces thyroid hormones that travel in the blood to help control several organs. When the function of the thyroid gland is affected, it leads to inappropriate production of the thyroid hormone. The symptoms of thyroid disease may involve high cholesterol, an unusual pulse rate, and high blood pressure. There are five common types of thyroid disease including hypothyroidism, structural abnormalities, hyperthyroidism, tumors, and subclinical hyperthyroidism or subclinical hypothyroidism. To diagnose hypothyroidism, tle hair, increased perspiration, hand tremors, heart racing, anxiety, skin thinning, and muscular weakness. Hyperthyroidism is very common after the age of 60 years. There are three major treatment methods for hyperthyroidism disease such as medications, radioiodine therapy, and thyroid surgery.

The thyroid gland creates hormones to carry out several important functions in the body, and lack of thyroid balance (creating too much or too little) leads to thyroid disease. There are various diseases related to the thyroid which lead to imbalance and malfunctioning of various other organs] If not properly treated, it can lead to complications such as goiter, heart disease, pregnancy problems, and more dangerous myxoedema coma .According to thyroid disease affects 200 million people worldwide and an estimated 40% of people are at risk of iodine deficiency which helps to produce thyroid hormone. Several kinds of thyroid diseases may

the blood sample is tested in a laboratory where medical specialists are needed to analyze the test reports for the hormones and other parameters to diagnose thyroid disease. In hypothyroidism, the thyroid gland does not produce sufficient thyroid hormone. Hypothyroidism may involve several symptoms such as the feeling of tiredness, poor capability to tolerate colds, constipation, depression, slow heart rate, and weight gain. Doctors adjust the medicine dose according to the patient's conditions to normalize TSH levels and thyroxine. Excessive thyroid hormones are produced in hyperthyroidism disease. Symptoms of hyperthyroidism include troubled sleeping, irritability, nervousness, fine brit-

Table 1 Thyroid test features

Feature	Description
TSH Thyroid-stimulating hormone	This is produced by pituitary glands to manage the thyroid hormones (TSH level in the blood from laboratory work)
T3 Triiodothyronine tests	100–200 ng/dL
Free T3 (FT3)	2.3–4.1 pg/mL
Test T4 (TT4)	
Thyroxine tests (11>T4>5) (T4)	Low T4 refers to hypothyroidism; high refers to hyperthyroidism
Free T4 or free thyroxine (FTI)	
	0.9–1.7 ng/dL

occurandeachhasitsownimpactonhumanhealth,forexample, hypothyroidism and hyperthyroidism and the two main types of thyroid disease that cause thyroid imbalance. To avoid such complications, early prediction of the correct thyroid disease type is important so that treatment can be done according to thyroid type.

In the past, various tests have been carried out that record different symptoms for thyroid diseases. These symptoms can be used to diagnose a specific thyroid disease For example, these tests and symptoms can be used to predict and diagnose thyroiditis/Graves', disease/Hashimoto's disease/goiter/thyroid, nodule/thyroid cancer, etc. Thyroid disease can be categorized based on various symptoms and features. Existing studies provide many features relevant to thyroiddiseases.Table1providesthenamesanddescriptions of a few features that can be used to predict thyroid disease; for example, lithium, goiter, hypopituitary, Psych, TSH, T3, TT4, T4U, FTI, and TBG can be used in thyroid detection.

Previously, several studies investigated thyroid diseases and their symptoms Some approaches focus on data analytics, while others carry out the recording of tests and symptoms. Based on the technique used for disease detection like statistical analysis, machine learning model, or deep learning model, the accuracy and robustness of such approaches vary. Predominantly, existing approaches make use of publicly available datasets and suffer from model overfitting. Most of the available datasets suffer from an imbalanced class problem where the number of samples for the positive (disease) class is substantially small. When such a dataset is used with a machine or deep learning model, the model overfits the majority class and produces false predictions for the minority class. Another limitation of existing studies is that only a few thyroid diseases are used for classification; for the most part, existing studies focus on the binary-class problem which makes those approaches unsuitable for real-world disease detection. This study focuses on resolving this issue by increasing the number of samples of the minority class using synthetic data samples.

This study aims at mitigating the influence of class imbalance and increasing disease detection performance. In summary, this study makes the following contributions:

- An efficient machine learning approach is designed to predict thyroid disease with high accuracy. The performance of the models is optimized using a differential evolution (DE)-based optimization algorithm.
- DE interacts with hyperparameters of various machine learning models to identify the best hyperparameters. The DE algorithm is used to find the optimal parameters for machine learning models to obtain significant improvement in the accuracy of models for thyroid disease detection.
- This study specifically deals with the class imbalance problem. The impact of class imbalance is mitigated using conditional generative adversarial networks (CTGAN) to perform data augmentation.
- A higher number of classes are considered compared to existing studies. Thyroid disease detection is performed using ten classes with several machine learning models including random forest (RF), gradient boosting (GB), AdaBoost, logistic regression (LR), and support vector machine (SVM). In addition, long short-term memory (LSTM), convolutional neural network (CNN), and

recurrent neural network (RNN) are also employed. • Extensive experiments are performed regarding the performance of models with and without DE optimization. The performance is further validated using k-fold crossvalidation.

2 Related Work

For thyroid disease prediction, deep learning and machine learning methods have been applied in various existing research works. Prediction of thyroid disease at its early stages and categorization into cancer, hyperthyroidism, or hypothyroidismisveryhelpfulfortreatingandrecoveringthe maximum number of patients. To identify the recent research studies in the presented work, various thyroid disease classification and detection methods have been described here.

In ,the authors present a machine learning approach for B-Raf proto-oncogene, serine/threonine kinase (BRAF) mutation presence in cancer thyroid nodules. The study also presented ultrasonic images of 96 thyroid nodules. Machine learningmodelssuchasRF,LR,andSVMareusedfordetecting the presence of BRAF mutation. Using these models, a classification accuracy of higher than 60% is reported. In anothersimilarstudy,fine-needleaspiration(FNA)andultrasonic features were used to reduce the false-negative rate for thyroid cancer. The RF model reported better results than other methods such as gradient descent and decision tree (DT). The authors applied LR and the least absolute shrinkageandselectionoperator(LASSO)modelsintochoose the malignant thyroid nodule associated with the ultrasonic features. To classify the malignant thyroid nodules, the RF model is used along with a scoring system. The logistic lasso regression(LLR)withRFattainedhigherthan80%accuracy. In , the authors analyzed the data by applying different machine learning algorithms. The results are compared with ten different classifiers. An 84% accuracy was achieved by using an extra tree classifier.

The study used SVM for detecting thyroid disease. The reported accuracy is 83.37%. Additionally, the model correctly distinguishes between four thyroid states. In , theauthorsperformedexperimentationtopredictandclassify thyroid disease using the DT model. In addition, researchers introduced a machine learning-based tool, a machine learning tool for thyroid disease diagnosis (MLTDD) to predict intelligently thyroid diseases. MLTDD shows an accuracy of 98.7%. Machine learning algorithms, including LR, RF, SVM, GBM, and DNN, are used to predict the highest probable molecules that initiate the homeostasis thyroid hormone in . The authors investigated feature engineering using deep learning and machine learning methods in . Backward feature elimination, forward feature selection, machine learning-based feature selection, and bidirectional feature elimination with an extra tree classifier were adopted. The proposed approach can predict non-thyroidal syndrome, Hashimoto's thyroiditis, autoimmune thyroiditis, and binding protein. Results indicate an improved accuracy of 99% using extra tree classifier-based selected features and with the RF classifier.

A multi-kernel SVM is presented in to predict cancer and thyroid diseases. The gray-wolf optimization is applied for feature selection and improves performance. The study reports a 97.49% accuracy using the multi-kernel SVM. In , the authors applied image processing methods and feature selection

techniques to select the important features from the database and obtain improved performance for thyroid diseases. Using machine learning and selective features techniques, performed multi-class hypothyroidism. Hypothyroidism is categorized into four groups. Experimental results demonstrate that RF achieved 99.81% accuracy compared with KNN, DT, and SVM algorithms. The study investigated three feature selection approaches using DT, SVM, RF, naive Bayes (NB), and LR for the prediction of hypothyroidism. Principal component analysis, univariate feature selection, and recursive feature selection were used for feature selection. Recursive feature selection combined with machine learning algorithms reported improved performance than other techniques. DT, SVM, RF, naive Bayes (NB), and LR algorithms are combined with RFE and achieve 99.35% accuracy.

The study introduced a multiple multi-layer perceptron (MMLP) model for thyroid disease classification. The MMLP is reported to achieve 99% accuracy for large-scale datasets. Another study presented the XGBoost technique to predict and classify thyroid disease. The XGBoost algorithm's efficiency is compared with DT, LR, and KNN approaches. The XGBoost algorithm improves the accuracy by 0.02% than the KNN algorithm. In , a comparative analysis for machine learning-based techniques RF, DT, artificial neural network (ANN), and KNN is presented. Experiments are carried out on a large-scale dataset. Moreover, both original and sampled data are considered for experiments. RF attained improved performance with 94.8% accuracy.

Besides using machine learning models, some studies specificallyfocusonusingdeeplearningmodelsfordetecting thyroid disease. For example, a deep neural network (DNN) is used to predict and classify thyroid disease in . DNN is reported to obtain an accuracy of 99.95%. The authors compared several machine learning algorithms such as extra

trees, CatBoost, LightGBM, ANN, KNN, SVM, RF, DT, XGBoost, and GaussianNB, to improve the thyroid prediction accuracy in. The accuracy, recall, precision, and F1 scoreareexaminedtoevaluatetheperformance.Theresearch reported 96% accuracy using an ANN classifier. The study utilized a CNN-based ResNet architecture to detect thyroid from an image dataset. The research focused on five types of thyroid conditions and achieved a 94% accuracy rateusingResNetmodelswiththestochasticgradientdescent (SGD) optimizer. The study, introduces a novel transfer learning approach, the distant domain high-level feature fusion (DHFF) model. DHFF aims to narrow the distribution gap between the source and target domains while preserving their unique characteristics. This approach prevents the overblending of features while facilitating a more effective transfer of knowledge acquisition. The proposed approach achieved 88.92% accuracy when applied to thyroid ultrasound auxiliary source domains.

The above-discussed studies investigate various machine learning and deep learning approaches for thyroid disease detection and report different results regarding the accuracy, precision, F1 score, etc. However, these studies have several limitations; in particular, we identified four major gaps concerning thyroid disease prediction. First, several studies perform experiments with smaller datasets and their results cannot be generalized. Secondly, the majority of the studies use thyroid detection, and the type of disease is not investigated. Thirdly, predominantly, the existing studies utilized imbalanced

datasets. Although high accuracy is reported in these studies, the models lack generalizability. The models may experience bias and overfitting, leading to wrong predictions for the minority classes. The model overfitting can cause a higher number of false positives for the minor class. Lastly, for the most part, the optimization is carried out concerning the feature engineering and there is a lack of emphasis on model tuning. The model's fine-tuning holds a prominent significance, especially with respect to dataset characteristics. Traditional tuning methods have proven to be inefficient. Therefore, this study focuses on hyperparameter optimization using the DE algorithm. Materials and Methods

This section describes the dataset used for experiments, the use of CTGAN for data balancing, the working of the DE algorithm inthe context of hyperparameter optimization, and a brief overview of machine learning models used in this study.

This study designs a machine learning approach for thyroid disease detection. the architecture of the proposed methodology. First, we acquire the dataset from the Kaggle The dataset contains 25 target classes, of which the top 10 target classes are selected for experiments. These classes are selected based on the high number of samples. The rest of the classes have very few samples, so they are not included in this study. The selected targeted dataset is imbalanced, so to make the dataset balanced we used the CTGAN augmentation technique. This technique generates samples for the minority class. Data splitting is done to divide the dataset into training and testing sets with a 0.8 to 0.2 ratio, where 80% is used for training and 20% is used for testing. Machine learning models work on numeric data, so we use a Label-encodertoconvertdataintonumericformbeforepassing it to machine learning models. We train machine learning models with a training set and perform hyperparameter optimization using DE optimizer which helps to select the best hyperparameter setting for models. In the end, we evaluate models in terms of accuracy, precision, recall, F1 score, and confusion matrix.

2.1 Thyroid Disease Dataset

The datasets used in this study are taken from the Kaggle repository. The thyroid disease dataset comprises 9172 samples and every sample has 31 features. The dataset consists of various records for different thyroid diseases and the target classes. The target classes include health condition state and diagnosis classes. The importance of features should be evaluated to select the optimal number for features of thyroid disease classification. In addition, these features belong to different types including float, boolean, string, and int. The proposed approach works with ten classes, namely '-', 'K', 'G', 'I', 'F', 'R', 'A', 'L', 'M', and 'N'

The class counts indicate that the dataset is imbalanced. For example, many samples in the dataset are not used in any particular class. However, data pre-processing is performed to attain the standard dataset to evaluate the performance. The classification count is used to classify as "no condition". Additionally, categorization is not performed on data samples for any other classes such as hypothyroid, hyperthyroid, general health, binding proteins, anti-thyroid treatment, replacement therapy, and many more. The patients who do not have thyroid

disease are considered in "no condition". Therefore,concurrentnon-

Fig. 1 Architecture of the proposed methodology
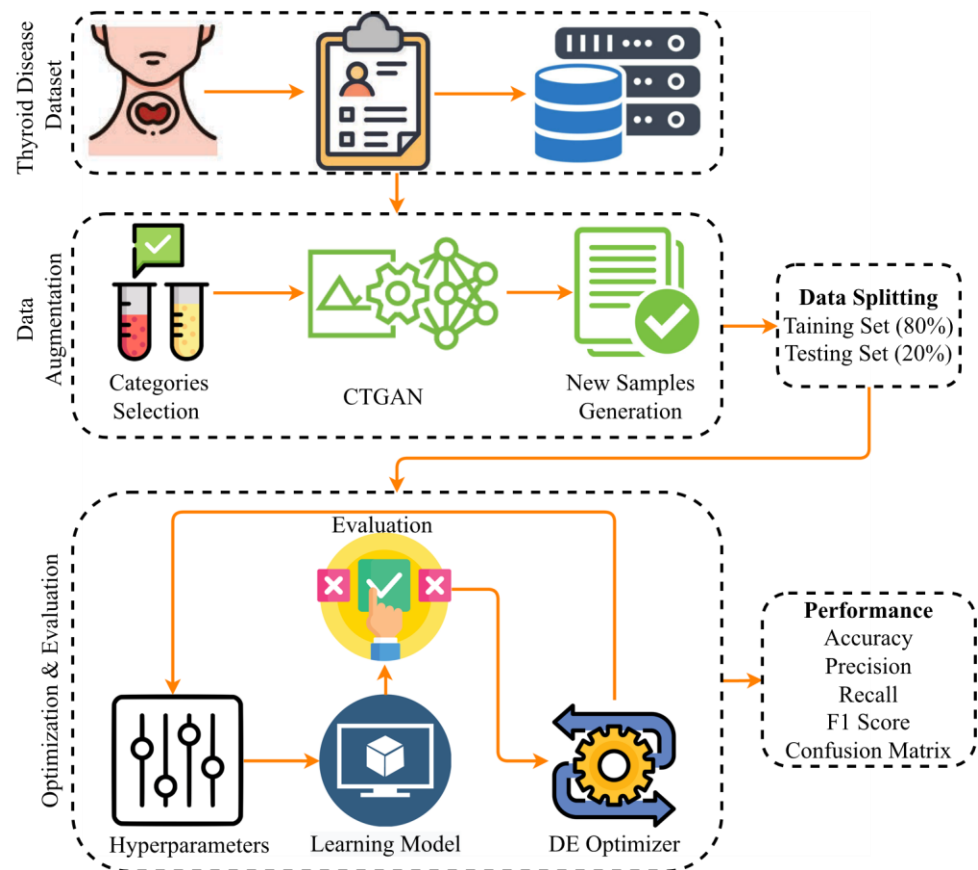
A total of 800 samples were randomly selected from the normal class. The primary hypothyroid increased binding protein, hypothyroid, hyperthyroid, consistent with replacement therapy, over-replaced, discordant assay results, and under-replaced and concurrent non-thyroid disease counts were not changed.

2.2 Data Balancing Using CTGAN

In this study, data balancing is carried out using CTGAN to generate samples. Inthedataset used inthisstudy, thenormal class has the highest number of samples, while some other classes have a very small number of samples. To balance the dataset, 400–500 samples are generated using CTGAN. Table 5 shows the number of samples before and after data augmentation using CTGAN.

The purpose of dataset balancing is to avoid model overfitting which happens when a model is trained on a highly unbalanced dataset. For the unbalanced dataset, the feature distributionisskewedconcerningthemajorityclassasshown in Fig.2b. Dataset balancing helps to normalize the feature distribution and reduces the probability of model overfitting. Feature distribution of the balanced dataset is shown in Fig.2 ifferential Evolution (DE)

ThisstudyusesaDE-basedoptimizationalgorithmtofindthe set of optimal hyperparameters for machine learning models such as RF, LR, SVM, AdaBoost, GB, etc . The hyperparameter optimization improves the performance of the models. The DE optimizer is divided into five phases: (1) Initialization, (2) fitness evaluation, (3) mutation, (4) crossover, and (5) stopping condition.

1. **Initialization:** In this phase, n^{th} random solutions are generated for a given problem. For this study, a random combination of hyperparameters (max_depth and a number of iterations (n_ite)) is generated. Each randomly generated solution istreated as a chromosome. This phase keepsarecordofinitializingallthebasicparameterslisted below:

   - **Number of the population (n):** 50.

   - **Number of iteration (I):** 100.

   - **Weighting factor(wf):** 0.9 (0 <wf<50).

   - **Crossover probability (CP):** 0.5 (0<CP<1), where

   $50 \leq Max\_Depth \leq 300, 50 \leq n\_ite \leq 300.$

2. **Fitness Evaluation:** This phase evaluates the fitness of each chromosome/solution. For our problem, fitness

Table 3 Description of classes in the dataset

Class	Description	Detail
K	Concurrent non-thyroid illness	Non-thyroidal disease is normally used to define the changes in hormones related to the thyroid that may arise in serum or chronic disease that is not produced with an intrinsic irregularity in the thyroid function
G	Compensated hypothyroid	Compensated hypothyroidism is also known as subclinical hypothyroidism. It is a condition that is associated with a high serum concentration for TSH, but a normal serum-free thyroxine (FT4)
I	Increased binding protein	A binding protein is some protein that acts as an agent to combine two or many molecules
F	Primary hypothyroid	Primary hypothyroidism is described as low levels of blood thyroid hormone because it damages the thyroid gland. This type of destruction is typically due to auto-immunity, including surgery, radiation, and radio-iodine
R	Discordant assay results	Assay interfering may be a reason for abnormal thyroidal function tests. Recognition at an early stage prevents inappropriate patient management
N	Over-replaced	In the use of a high free T4 along with TSH to describe the over-replaced group, which is combined by use of a normal free T4 with low TSH to define the group not likely to be over-replaced, will reduce the possibility of error in allocating patients to the affected and control groups
A	Hyperthyroid	
L	Consistent with replacement therapy	Excessive thyroid hormones are produced in hyperthyroidism disease. Symptoms of hyperthyroidism include trouble sleeping, irritability, nervousness, fine, brittle hair, increased perspiration, hand tremors, heart racing, anxiety, thinning of the skin, and muscular weakness
M	Under-replaced	Hypothyroidism is generally treated by taking hormone replacement tablets daily, known as levothyroxine. It replaces thyroxine hormone, which is not enough in the patient's body]. It can rarely comprise: its use to control the growth of thyroid goiter
-	No condition	No thyroid disease in patient / Normal report

Class	# of samples
Normal	800
Over-replaced	436
Concurrent non-thyroidal illness	359
Compensated hypothyroid	346
Increased binding protein	233
Primary hypothyroid	196
Discordant assay results	147
Hyperthyroid	115
Consistent with replacement therapy	111
Under-replaced	110

Table 5 Details of the dataset before and after data augmentation

Class	Count	New sample count
Normal	800	800
Over-replaced	436	836
Concurrent non-thyroidal illness	359	859
Compensated hypothyroid	346	846
Increased binding protein	233	733
Primary hypothyroid	196	596
Discordant assay results	147	647
Hyperthyroid	115	615
Consistent with replacement therapy	111	611
Under-replaced	110	610

refers to the accuracy score of the model using the given set of hyperparameters.

3. **Mutation:** In this phase, a new offspring is generated from the existing solution in search of a new better solution. This involves the random selection of two chromosomes. Then using the weighting factor, one target chromosome is selected with maximum fitness. Figure 3 shows the mutation process. A new chromosome can be defined as

$$New_{chromosome} = target_{vector} + wf \cdot (random\_selected_1 - random\_selected_1).$$

(1)

4. **Crossover:** In this process, some of the values of the new solution are interchanged with the existing solution. The crossover probability defines the probability of swapping the values. For this study, it is 0.5 which means 50% of the chromosomes are updated. Figure4 shows the process carried out in the crossover.

5. **Stopping Condition:** This step defines when to stop the process of searching for new solutions. The stopping condition is the number of iterations to find the solution. In
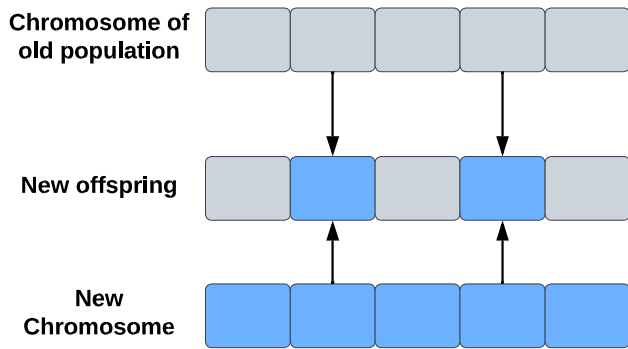

Fig. 4 Processofcrossover

the proposed approach, if the condition for the number of iterations is met, the process stops, and it returns the best-explored solution along with the best accuracy score achievedusingthatglobalsolution.Figure5illustratesthe flowchart of the DE and the interaction between DE and ML as proposed in this study.

Fig. 5 DE optimizer flow diagram

2.3 Machine Learning Models

The presented approach employs various machine learning models to detect thyroid disease. RF, SVM, LR, AdaBoost, GBM, CNN, RNN, and LSTM are used in the presented study. These classifiers are tuned for performance Random Forest

RF is employed for regression and classification problems. RF is an ensemble classifier that uses a tree-based classification

models aretrainedonboot-strapped samples that areusedfor classification in the bagging approach. RF can produce more consistent ensemble forecasts than a DT. The test statistic of a single function is computed using the feature selection method in Eq. 2

Table 6 Hyperparameter setting and optimization range for machine learning models

Model	Hyperparameters	Hyperparameters range
RF	n_estimators=282,max_depth=15	n_estimators = {2 to 300}, max_depth = {2 to 300}
GB	n_estimators = 148, max_depth = 107, learning_rate = 0.5	n_estimators = {2 to 300}, max_depth = {2 to 300}, learning_rate = {0.1 to 0.9}
AdaBoost	n_estimators = 286, learning_rate = 0.6	n_estimators = 2 to 300, learning_rate = {0.1 to 0.9}
LR	Solver = 'saga', multi_class = 'multinomial', C = 4	Solver = {'newton-cg', 'lbfgs', 'sag', 'saga' }, multi_class = 'multinomial', C = { 1 to 10}
SVM	kernel= 'linear', C=5.0	kernel = {'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'}, C = { 1.0 to 10.0}

technique. Additionally, RF is applied to reduce the overfitting problems using a bootstrap approach for sampling. It defines the best prediction by the voting process. Additionally, it detects the significant elements within a dataset and reports a simple indicator for feature significance. Feature selection is applied in classification research to reconstruct the data and also improve the accuracy. Many
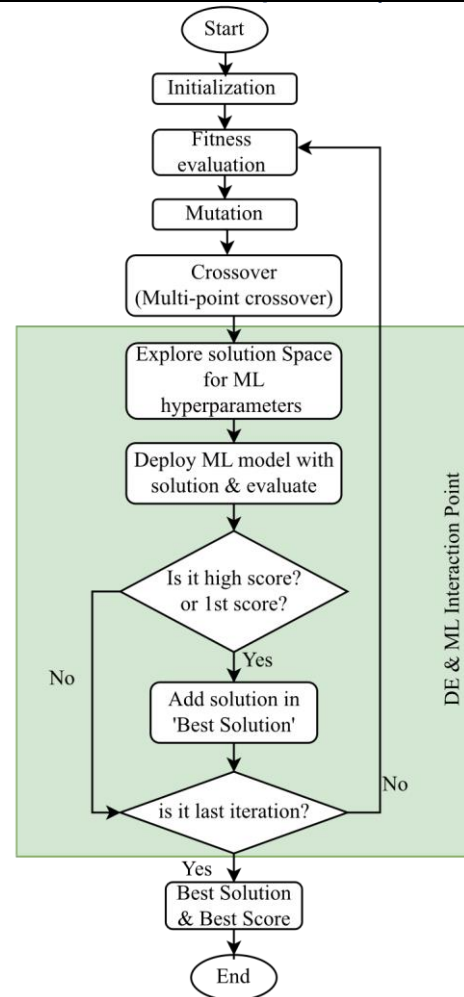
Table 7 RF performance comparison with and without DE optimizer

Class	Without optimization			Models	With optimization		
	Precision	Recall	F1 score		Precision	Recall	F1 score
-	1.00	1.00	1.00	–	0.99	1.00	1.00
A	0.96	0.99	0.97	A	1.00	1.00	1.00
F	1.00	1.00	1.00	F	0.99	0.99	0.99
G	0.98	1.00	0.99	G	0.99	1.00	1.00
I	0.98	0.98	0.98	I	0.99	0.99	0.99
K	0.99	1.00	1.00	K	1.00	0.99	1.00
L	0.99	0.97	0.98	L	1.00	1.00	1.00
M	1.00	0.99	1.00	M	0.98	1.00	0.99
N	1.00	1.00	1.00	N	1.00	1.00	1.00
R	0.98	0.94	0.96	R	1.00	0.97	0.98
Average	0.99	0.99	0.99	Average	1.00	0.99	1.00

where F_b and F_0 are functions for the state transition and output,respectively,intheequation.Everyfunctionisparameterized using a set of parameters, ϕb and $\phi 0$.

Suppose there is a set of M training sequences $A = ((\alpha_1^{(m)}, c_1^{(m)}), ..., ((\alpha_{Dm}^{(m)}, c_{Dm}^{(m)})_M$ estimated to minimize cost function as described in Eq. m=1. RNN's parameters are16 [44].

$$V(\phi) \quad \sum_{m=1}^{M} \sum_{d=1}^{Dm} j(c_d^{(m)}, F_o(b_d^{(m)})), \quad \frac{1}{M} = \quad (16) \quad —$$

where $= F_b(\alpha) =_d^{(m)}, c_{d-1}^{(m)}$ $b_d^{(m)}$ and $b_d^{(m)} = 0$, $j(x, y)$ is a predefined divergence value between x and y, as crossentropy or Euclidean distance.

3 Results and Discussions

This section contains the results of machine learning models to analyze the performance of models with and without the DE optimizer. In addition, k-fol cross-validation and performance comparison with existing state-of-the-art models is also carried out.

3.1 Experimental Setup

The machine learning models are investigated in terms of accuracy, precision, recall, and F1 score. Moreover, each model is evaluated using confusion matrix values. This study performs experiments using an Intel Core i7 11th generation system with 16 GB RAM, 1TB SSD, and Windows 10.0 operating system. We used Jupyter Notebook and Python language toimplementtheproposedapproach. Theproposed approachusesthesci-kitlearnlibraryandTensorFlowframework.

3.2 Results of Machine Learning Models

optimizer. Although GBM shows better performance due to its boosting operations, DE optimization further improves its performance.

performance of the AdaBoost model. AdaBoost is similar to GBM, but it identifies the shortcomingsoftheexistingweaklearnersbyhigh-weightdatapoints, while GBM uses gradients. Results show that Adaboost outperforms in all evaluation parameters, as it achieved a 1.00 score each for precision, recall, and F1 score when used with

Table 10 GBM confusion metrics with and without the DE optimizer

Withoutoptimizer

205	0	0	0	0	0	0	0	0	0
3	133	0	0	0	0	0	0	0	0
2	0	155	0	0	0	0	0	0	0
0	0	0	178	0	0	0	0	0	0
2	0	0	2	165	0	0	0	0	0
0	0	0	0	0	149	0	0	0	0
1	0	0	0	2	0	113	0	0	0
3	0	0	0	0	0	0	115	0	0
0	0	0	0	0	0	0	0	121	0
0	0	0	0	1	0	0	0	0	121

Optimizer

198	0	0	0	0	0	0	0	0	0
0	115	0	0	0	0	0	0	0	0
0	0	135	0	0	0	0	0	0	0
0	0	0	188	0	0	0	0	0	0
0	0	0	0	177	0	0	0	0	0

Table 14 LR confusion metrics with and without the DE optimizer

Without Optimizer

206	0	0	0	2	0	0	1	13	1

3.3 Results of Deep Learning Models

In addition to machine learning models, this study deployed several deep learning models for comparison. DE optimization is not performed for deep learning models; these models are deployed on the data after applying CTGAN. We used fourdeeplearningmodelsincludingLSTM,CNN,RNN,and CNN–LSTM. We used all models with an embedding layer with a 5000 vocabulary size and 200 output size. After the embedding layer, the layer of each model is used. The ending layer of all

models consists of ten neurons and a softmax function. We used all models with categorical_crossentropy loss function, the Adam optimizer, and 100 epochs. Architectural details of all deep learning models are provided in

The resultsofdeeplearningmodelsforthyroid disease detection. Results suggest that the performance of deep learning models is not good. The LSTM achieved a 0.90accuracyscoreandCNNachieved0.86accuracy.LSTM is a recurrent architecture that has feedback connections, as it is capable of processing the entire sequence of data, apart from single data points, while CNN requires a large feature set to make correct predictions. We also used a combination of CNN and LSTM in which CNN extracts the features for LSTM. CNN does not show good performance. Since the feature set from CNN is not good, LSTM could not perform well.WeusedRNNwhichalsohasrecurrentarchitectureand can perform well on a small feature set.

3.4 Results ofT-test

We conducted a statistical analysis to compare the performance of this study with that of previous studies. We employed a statistical T-test and examined the results of all approaches. The T-test involves the formulation of two hypotheses.

- Null hypothesis (H_0): there is no significant difference in accuracy between this study and previous studies.

- Alternative hypothesis (H_a): there is a significant difference in accuracy between this study and previous studies.

T-test results are presented in Table 22. With a significance level (alpha) set at 0.05, the critical value is 2.35. When the absolute T-score exceeds the critical value, the T-test typically leads to the rejection of the H_0. In this case, the T-scores significantly exceed the critical value, indicating strong evidence of significant differences in all three comparisons. Notably, in the comparison between this study and] while the T-Score is lower compared to the other two comparisons, it is still sufficiently high to suggest a significant difference in the performance metrics.

References

1. Sharma, K.A., Arya, R., Mehta, R., Sharma, R., Sharma, K.A.: Hypothyroidism and cardiovascular disease: factors, mechanism and future perspectives. Curr. Med. Chem. **20**(35), 4411–4418 (2013)
2. healthline: The 6 Common Thyroid Problems & Diseases. https:// www.healthline.com/health/common-thyroid-disorders (2018)
3. UK, N.H.S.: Underactive thyroid (hypothyroidism). https://www. nhs.uk/conditions/underactive-thyroid-hypothyroidism/ (2021)
4. Diabetes, T.L., Endocrinology: The untapped potential of thethyroid axis. https://www.thelancet.com/journals/landia/article/ PIIS2213-8587(13)70166-9/fulltext#articleInformation (2013)
5. Zhang, B., Tian, J., Pei, S., Chen, Y., He, X., Dong, Y., Zhang, L.,Mo, X., Huang, W., Cong, S., et al.: Machine learning-assisted system for thyroid nodule diagnosis. Thyroid **29**(6), 858–867 (2019)
6. Idarraga, A.J., Luong, G., Hsiao, V., Schneider, D.F.: False negative rates in benign thyroid nodule diagnosis: Machine learning for detecting malignancy. J. Surg. Res. **268**, 562–569 (2021)
7. Razia, S., Rao, M.N.: Machine learning techniques for thyroid disease diagnosis-a review. Indian J. Sci. Technol. **9**(28), 1–9 (2016)
8. Aversano, L., Bernardi, M.L., Cimitile, M., Iammarino, M., Macchia, P.E., Nettore, I.C., Verdone, C.: Thyroid disease treatment prediction with machine learning approaches. Procedia Computer Science **192**, 1031–1040 (2021)
9. Kwon, M.-R., Shin, J., Park, H., Cho, H., Hahn, S., Park, K.: Radiomics study of thyroid ultrasound for predicting braf mutation in papillary thyroid carcinoma: preliminary results. Am. J. Neuroradiol. **41**(4), 700–705 (2020)
10. Chen, D., Hu, J., Zhu, M., Tang, N., Yang, Y., Feng, Y.: Diagnosisof thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest. BioData mining **13**(1), 1–21 (2020)
11. Razia, S., Siva Kumar, P., Rao, A.S.: Machine learning techniques for thyroid disease diagnosis: a systematic review. Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough, 203–212 (2020)
12. Lomana, M., Weber, A.G., Birk, B., Landsiedel, R., Achenbach,J., Schleifer, K.-J., Mathea, M., Kirchmair, J.: In silico models to predict the perturbation of molecular initiating events related to thyroidhormonehomeostasis.Chem.Res.Toxicol.**34**(2),396–411 (2020)
13. Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J.L.V.,Rodríguez,C.L.,Ashraf,I.:Thyroiddiseasepredictionu singselective features and machine learning techniques. Cancers **14**(16), 3914 (2022)
14. Shankar, K., Lakshmanaprabu, S., Gupta, D., Maseleno, A., DeAlbuquerque, V.H.C.: Optimal feature-based multi-kernel svm approach for thyroid disease classification. J. Supercomput. **76**(2), 1128–1143 (2020)
15. Das, R., Saraswat, S., Chandel, D., Karan, S., Kirar, J.S.: An aidriven approach for multiclass hypothyroidism classification. In: International Conference on Advanced Network Technologies and Intelligent Computing, pp. 319–327 (2021). Springer
16. Riajuliislam, M., Rahim, K.Z., Mahmud, A.: Prediction of thyroid disease (hypothyroid) in early stage using feature selection and classification techniques. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), pp. 60–64 (2021). IEEE
17. Hosseinzadeh, M., Ahmed, O.H., Ghafour, M.Y., Safara, F., Ali,S., Vo, B., Chiang, H.-S., et al.: A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. J. Supercomput. **77**(4), 3616–3637 (2021)
18. Sankar, S., Potti, A., Chandrika, G.N., Ramasubbareddy, S.: Thyroiddiseasepredictionusingxgboostalgorithms.Journalo fMobile Multimedia **18**(3), 1–18 (2022)
19. Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., Ahmad,A.: Empirical method for thyroid disease classification using a machine learning approach. BioMed Research International **2022** (2022)
20. Jha, R., Bhattacharjee, V., Mustafi, A.: Increasing the predictionaccuracy for thyroid disease: A step towards better health for society. Wireless Pers. Commun. **122**(2), 1921–1938 (2022)

21. Islam,S.S.,Haque,M.S.,Miah,M.S.U.,Sarwar,T.B.,Nugraha, R.: Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. PeerJ Computer Science **8**, 898 (2022)

22. Prathibha, S., Dahiya, D., Robin, C., Nishkala, C.V., Swedha, S.: Anovel technique for detecting various thyroid diseases using deep learning. Intell. Autom. Soft Comput **35**(1), 199–214 (2023)

23. Tang, F., Ding, J., Wang, L., Ning, C.: A novel distant domaintransfer learning framework for thyroid image classification. Neural Process. Lett. **55**(3), 2175–2191 (2023)

24. Sureshkumar, V., Balasubramaniam, S., Ravi, V., Arunachalam,A.: A hybrid optimization algorithm-based feature selection for thyroid disease classifier with rough type-2 fuzzy support vector machine. Expert. Syst. **39**(1), 12811 (2022)

25. Kaggle, E.F.W.: Thyroid diseases dataset. https://www.kaggle. com/datasets/emmanuelfwerr/thyroid-disease-data (2022)

26. Yasar, Z., Kirakli, C., Cimen, P., Ucar, Z.Z., Talay, F., Tibet, G.: Isnon-thyroidal illness syndrome a predictor for prolonged weaning in intubated chronic obstructive pulmonary disease patients? Int. J. Clin. Exp. Med. **8**(6), 10114 (2015)

27. Salerno, M., Improda, N., Capalbo, D.: Management of endocrinedisease subclinical hypothyroidism in children. Eur. J. Endocrinol. **183**(2), 13–28 (2020)

28. Bielli, P., Busà, R., Paronetto, M.P., Sette, C.: The rna-bindingprotein sam68 is a multifunctional player in human cancer. Endocr. Relat. Cancer **18**(4), 91–102 (2011)

29. Thyroiditis, C.A.: Chronic autoimmune thyroiditis. The Thyroidand Its Diseases: A Comprehensive Guide for the Clinician, 379 (2019)

30. Srichomkwun, P., Scherberg, N.H., Jakšic´, J., Refetoff, S.: Diagnostic dilemma in discordant thyroid function tests due to thyroid hormone autoantibodies. AACE clinical case reports **3**(1), 22–25 (2017)

31. Livingston, M., Birch, K., Guy, M., Kane, J., Heald, A.: No rolefor tri-iodothyronine (t3) testingin the assessment of levothyroxine (t4) over-replacement in hypothyroid patients. Br. J. Biomed. Sci. **72**(4), 160–163 (2015)

32. Sharma, A.: Thyroid: An updated study on the diagnosis and treatment of hyper and hypo thyroidism (2017)

33. Biondi, B., Cooper, D.S.: Thyroid hormone therapy for hypothyroidism. Endocrine **66**(1), 18–26 (2019)

34. Jonklaas, J., Bianco, A.C., Bauer, A.J., Burman, K.D., Cappola,A.R., Celi, F.S., Cooper, D.S., Kim, B.W., Peeters, R.P., Rosenthal, M.S., *et al.*: Guidelines for the treatment of hypothyroidism: prepared by the american thyroid association task force on thyroid hormone replacement. thyroid **24**(12), 1670–1751 (2014)