SETSCI Conference Proceedings
21, 33-37, 2024

International Trend of Tech Symposium
7-8 December 2024, İstanbul - Türkiye

https://doi.org/10.36287/setsci.21.6.033

Prediction of Recurrence of Differentiated Thyroid Cancer with Hybrid SMOTE-Stacking Model

Erkan AKKUR*[1], Serkan Cizmecioglulları[2] and Ahmet Cankat OZTURK [3]

[1]*Turkish Medicine and Medical Devices Agency, Ankara, Türkiye*
[2]*Vocational School of Technician Sciences Electronics and Automation, Kırsehir Ahi Evran University, Kırsehir, Turkey*
[3] *Presidency of The Republic of Turkey Secretariat of Defence Industries, Ankara, Türkiye*
(eakkur@gmail.com)

Abstract – Differentiated thyroid cancer (DTC) is the most frequent form of thyroid cancer. Although this type of cancer shows a favourable prognosis, the risk of recurrence remains a critical concern. Early and accurate prediction of the risk of recurrence is essential to improve patient outcomes and minimize this risk. This study proposes a hybrid model combining SMOTE (Synthetic Minority Oversampling Technique) and a stacking ensemble approach to predict DTC relapse. First, the dataset is balanced using the SMOTE technique, ensuring equal representation across classes. Then, the overall accuracy of the model is improved by the stacking method, which combines the predictions of multiple classifiers. This model has been tested on a publicly available dataset, with impressive results such as an accuracy of 99.09% and an AUC-ROC score of 0.998.

Keywords – Differentiated thyroid cancer, machine learning, ensemble learning, stacking, SMOTE

I. Introduction

Thyroid cancer is a form of cancer that occurs when normal thyroid cells in the thyroid gland grow uncontrollably into abnormal cells [1]. Differentiated thyroid cancer (DTC) is the most frequent form, constituting approximately 95% of thyroid cancers [2]. Although the overall survival rate for this cancer is relatively high, the risk of recurrence after treatment remains a major clinical challenge. Studies in the literature have reported that the risk of recurrence ranges from 5% to 30%, making disease management complicated [3]. Recurrence varies based on multiple factors, including the stage at diagnosis, treatment modalities and histopathologic characteristics of the disease [4]. Accurate prediction of the risk of recurrence is therefore crucial for the regular follow-up of patients. However, the current prognostic method may not always yield the desired results. Currently, artificial intelligence technology has become increasingly recognized as an important tool in the prediction and treatment of diseases [5].

Machine learning (ML), one of the current artificial intelligence methods, plays an extremely critical position in the prediction of cancer prognoses. Therefore, as in other cancer types, the potential use of ML algorithms in a clinically significant situation such as DTC recurrence may help identify high-risk patients and improve treatment strategies [5-7]. However, class imbalance in datasets is one of the challenges that adversely impair the prediction performance of machine learning algorithms. The SMOTE (Synthetic Minority Oversampling Technique) is a method that aims to balance the dataset by augmenting the samples of the minority class, is a common method used to balance the data. [8].

Ensemble learning is another method that aims to improve the prediction of ML algorithms. It is a ML meta-approach that strives to optimize prediction performance by combining predictions from many models. The stacking approach, one of the well-known ensemble learning algorithms, aims to achieve higher performance than models used individually by combining the predictions of multiple models. This model can serve an important role in early diagnosis of the disease and effective treatment management by providing a more reliable prediction of rare events such as DTC recurrence [9].

In this study, a hybrid model combining SMOTE and Stacking methods is proposed to predict DTC recurrence risk more accurately. This hybrid SMOTE-Stacking approach aims to contribute to clinical decision-making processes by predicting DTC recurrence early and developing more efficient patient management and treatment planning.

II. Materials and Method

This study aims to introduces a hybrid model combining SMOTE-Stacking for DTC recurrence prediction. Fig. 1 illustrates the workflow of the proposed model.

A. Dataset

The dataset examined in this study was retrieved from published research by Borzooei et al. [10]. It is named Differentiated Thyroid Cancer Recurrence and is a publicly available dataset accessible through the Machine Learning Repository at UC Irvine [11]. The dataset comprises 383 patients and 17 features. The last feature represents whether the individual has a recurrence or not. The age of patients with DTC recurrence was 47.11±18.27 years and 38±12.95 years for those without recurrence. Table 1 lists the features and descriptions of the data set.

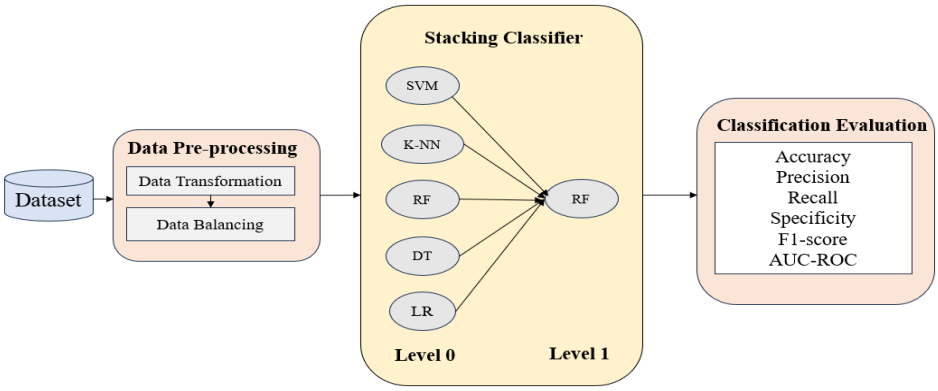

Fig. 1: General framework of the proposed model

Table 1: The features of the dataset

No	Feature	Description
1	Age	Represents the ages of individuals in the dataset.
2	Gender	Indicates the gender of individuals
3	Smoking	Possibly an attribute related to smoking behaviour.
4	Hx Smoking	Indicates whether individuals have a history of smoking
5	Hx Radiotherapy	Indicates whether individuals have a history of radiotherapy treatment
6	Thyroid Function	Possibly indicates the status or function of the thyroid gland.
7	Physical Examination	Describes the results of a physical examination
8	Adenopathy	Indicates the presence and location of adenopathy
9	Pathology	Describes the types of thyroid cancer based on pathology examinations
10	Focality	Indicates whether the thyroid cancer is unifocal or multifocal.
11	Risk	Represents the risk category associated with thyroid cancer.
12	Tumor (T)	Represents the T (Tumor) stage of thyroid cancer.
13	Lymph Nodes	Represents the N (Node) stage of thyroid cancer.
14	Metastasis	Represents the M (Metastasis) stage of thyroid cancer.
15	Stage	Represents the overall stage of thyroid cancer based on the combination of T, N, and M stages.
16	Treatment Response	Describes the response to treatment
17	Recurred	Indicates whether thyroid cancer has recurred

B. Data pre-processing

Data pre-processing is a critical step in the modelling process for organising the raw data and improving model performance. In the first stage of this process, the data set was checked for missing data and no missing data was detected [12]. In the next step, the categorical data in the data set were converted into numerical values by the label coding method. It allows ML algorithms to analyze data more effectively and speeds up the training of models [13].

When the dataset used in the study was analyzed, it was observed that the class distribution was unbalanced. This poses the risk of the model not learning the minority class sufficiently. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the class distribution. SMOTE is a resampling technique that aims to reduce data imbalance. However, the difference of SMOTE is that instead of directly replicating existing minority class samples, it creates new and synthetic data points in the gaps between minority class samples [8]. The working principle of SMOTE is as follows:

i. Determination of Minority Class Instances: Data samples in the minority class are selected and their nearest neighbours are determined.
ii. Random Selection of Points: Random points are selected among these neighbours.
iii. Synthetic Data Generation: New synthetic data samples are generated between the selected points and added to the dataset.
iv. Creating a Balanced Dataset: This process makes the data set more balanced.
v. More Accurate Learning: As the imbalance decreases, the model learns the minority class better.
vi. Improved Performance: As a result, the performance of the model and its ability to predict the minority class increases.

The status of the dataset before and after data balancing is shown in Table 2.

Table 2: Class distribution in the data before and after data balancing

Sample	Before data balancing	After data balancing
Recurrence	108	275
No recurrence	275	275

C. Stacking ensemble method

The stacked is an ensemble approach that combines various machine learning algorithms to produce a more robust and overall performing model. This method involves the creation of base models and a meta-model from these base models to ensure final predictions. Since the base learners and the meta-model could be trained with the same training set, a standard stacking model may experience overfitting. Therefore,

stacking and cross-validation (CV) are commonly used jointly to avoid overfitting. The CV technique starts with an initial division of the data into k folds. In each k iteration, k-1 folds are used to place the base classifiers. In each round, base classifiers are installed for the remaining subset that does not have model fitting. The 5-fold CV technique is adopted for this process. In the next stage, the resulting predictions are summed and served as input data for the meta-model. The resulting predictions are then stacked and supplied as input data to the second-level classifier [14-15]. Fig. 2 indicates the framework of the stacking approach.

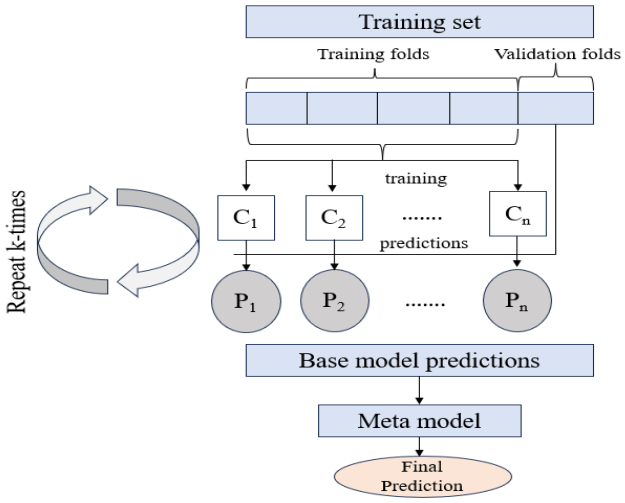

Fig. 2: The framework of the suggested stacking model

In this study, the value of k was set to 5 in the CV technique. Random forest (RF), k-nearest neighbour (K-NN), AdaBoost, support vector machine (SVM), logistic regression (LR) and decision tree (DT) algorithms were utilized as base-level classifiers and random forest algorithm was used as a meta-model.

III. RESULTS

Performance criteria such as accuracy, sensitivity, precision, sensitivity, F1-score and specificity were used to evaluate the prediction performance of the proposed model. In this context, the proposed SMOTE-Stacking ensemble (SEM) model is compared with Random Forest (RF), K-Nearest Neighbour (K-NN), Logistic Regression (LR), Support Vector Machine (SVM) and Decision Trees (DT) algorithms. The experiments were performed using Jupyter Notebook 3.8.16 on Python. The models were fed with training and test data, and the data was split as 80% training and 20% test.

Table 2 shows the prediction performances of the base classifiers and the SEM algorithm. The proposed SEM model achieved 99.09% of accuracy, 98.21% of precision, 100% of sensitivity, 98.18% of specificity and 99.1% of F1-score. Fig. 3 compares the performance of the proposed SEM model and the base classifiers in terms of accuracy. The base classifiers RF, K-NN, SVM, DT and LR performed with 98.18%, 87.27%, 87.27%, 87.27%, 96.36% and 86.36% accuracy respectively. On the other hand, the SEM model outperformed all base classifiers with an accuracy of 99.09%.

Table 3: The comparison of classification techniques for DTC recurrence prediction

Model	Accuracy %	Precision %	Sensitivity %	Specificity %	F1-Score %
SEM	**99.09**	**98.21**	**100**	**98.18**	**99.09**
RF	98.18	96.49	100	96.39	98.18
DF	96.36	96.36	96.36	96.36	96.36
SVM	87.27	90.2	83.64	90.91	86.79
K-NN	87.27	87.27	87.27	87.27	87.27
LR	86.36	90	81.82	90.91	85.71

IV. DISCUSSION

Accurate prediction of DTC recurrence is crucial for timely intervention and improved patient outcomes. This study investigated the effectiveness of ML algorithms in predicting DTC recurrence. Borzooei et al. [10] show that based on the dataset used in this study, the SVM model achieved remarkable results with a sensitivity of 0.99, specificity of 0.97 and AUC of 0.99. In Yaşar [16] study, the effectiveness of RF and AdaBoost algorithms in predicting DTC recurrence was investigated and an accuracy rate of 95.7% was obtained with the RF algorithm. However, these studies ignored the class imbalance in the dataset.

Class imbalance in datasets is one of the important problems affecting the prediction performance of ML algorithms. In this study, this problem is solved by increasing the number of samples of the minority class with the SMOTE technique. In addition, a stacking approach is adopted to improve the prediction performance by combining the predictions of different classifiers to obtain a final prediction result. The accuracy rates and ROC values of this hybrid model and the individual classifiers are presented in Figures 3 and 4. The proposed model outperforms the individual classifiers with an accuracy rate of 99.09% and an AUC value of 0.998.

The hybrid model proposed in this study shows high performance in predicting DTC recurrence and it is envisioned that this model can be an alternative approach for healthcare professionals in patient follow-up and formulation of treatment strategies. Furthermore, it is expected that this model can serve as a framework for other cancer prediction approaches. In upcoming studies, the most significant clinical parameters affecting recurrence prediction can be identified by incorporating feature selection into the model.

The presented hybrid model has limitations such as being tested on publicly available data. It is important to validate the model with independent cohorts to ensure its reliability and to fully represent different patient populations, assisting the model to be more widely accepted in clinical applications.

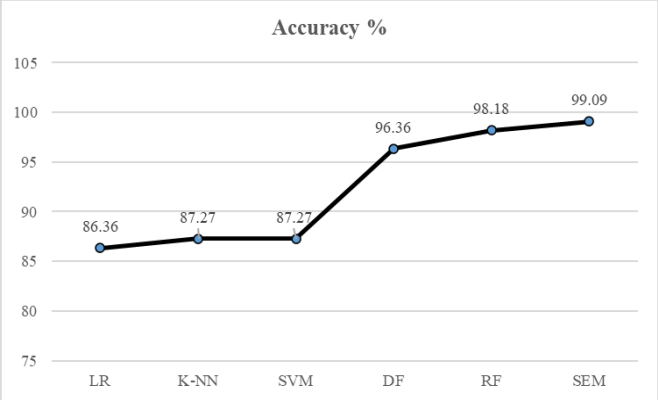

Fig. 3: Accuracy graphs of classifiers for DTC recurrence prediction.

V. CONCLUSION

Early prediction of DTC recurrence risk may be useful for appropriate treatment strategies and regular patient follow-up. This study introduces a hybrid model integrating SMOTE and Stacking for effective prediction of DTC recurrence. This proposed model utilizes SMOTE to solve the class imbalance problem in the dataset and Stacking techniques to improve the prediction performance of classifiers and achieves an impressive prediction performance with 99.09% accuracy and 0.998 AUC. This study provides an innovative and valuable approach for both the prediction of DTC recurrence risk and the prediction of other cancer types. This approach is expected to support the elaboration of more personalized and effective strategies for cancer treatment.

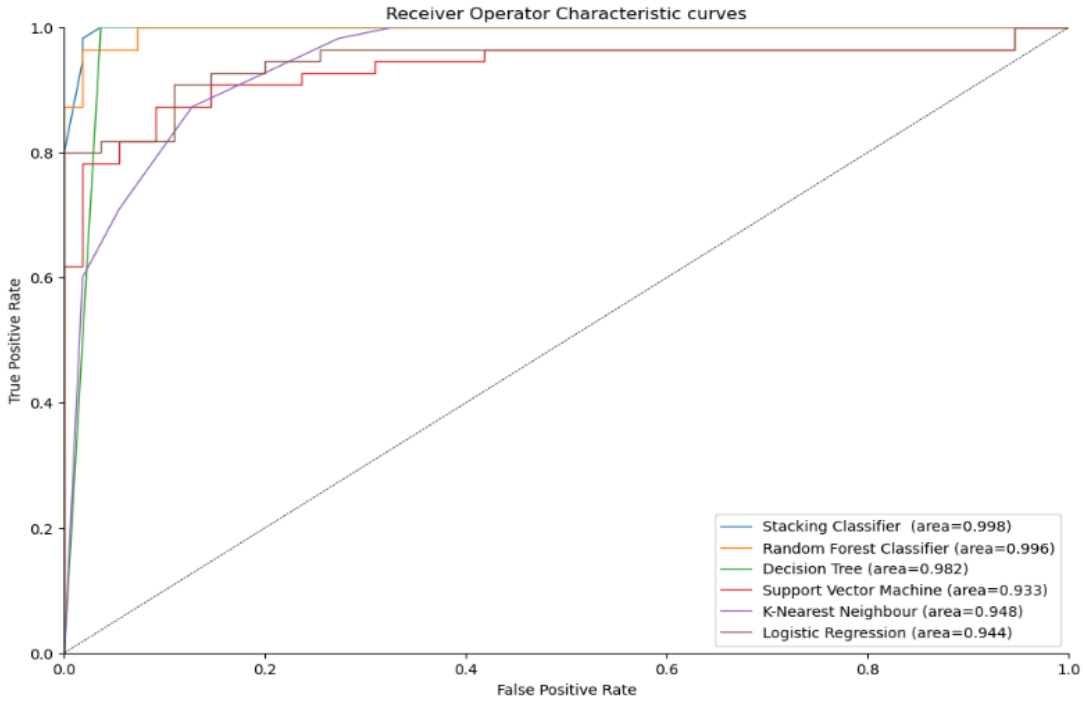

Fig. 4: ROC curve of ML algorithms

REFERENCES

[1] Cabanillas, M. E., McFadden, D. G., & Durante, C. "Thyroid cancer", *The Lancet*, 388(10061), 388(10061), 2783-2795, 2016.

[2] Hoff, A. O., Chaves, A. L. F., de Oliveira, T. B., Ramos, H. E., Penna, G. C., Santos, L. V. D., ... & Vizzotto, F. P. "Differentiated thyroid carcinoma: what the nonspecialists needs to know," *Archives of Endocrinology and Metabolism*, 68, e230375, 2024.

[3] Pałyga, I., Rumian, M., Kosel, A., Albrzykowski, M., Krawczyk, P., Kalwat, A., ... & Kowalska, A. (2024). "The frequency of differentiated thyroid cancer recurrence in 2302 patients with excellent response to primary therapy", *The Journal of Clinical Endocrinology & Metabolism*, 109(2), e569-e578, 2024.

[4] Yu L, Hong H, Han J, Leng SX, Zhang H, Yan X. "Comparison of Survival and Risk Factors of Differentiated Thyroid Cancer in the Geriatric Population", *Front Oncol*. 2020 Feb 3; 10:42, 2020.

[5] Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., ... & Albekairy, A. M. "Revolutionizing healthcare: the role of artificial intelligence in clinical practice", *BMC Med Educ*, 23(1), 689, 2023.

[6] Maurya, S., Tiwari, S., Mothukuri, M. C., Tangeda, C. M., Nandigam, R. N. S., & Addagiri, D. C. "A review on recent developments in cancer detection using Machine Learning and Deep Learning models". *Biomedical Signal Processing and Control*, 2023, 80, 104398.

[7] Mooijman, P., Catal, C., Tekinerdogan, B., Lommen, A., & Blokland, M. (2023). "The effects of data balancing approaches: A case study". *Applied Soft Computing*, 2023, 132, 109853.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. "SMOTE: synthetic minority over-sampling technique". *Journal of artificial intelligence research*, 2002, 16, 321-357.

[9] Mahajan P, Uddin S, Hajati F, Moni MA. "Ensemble Learning for Disease Prediction: A Review", *Healthcare*. 11(12):1808, 2023.

[10] Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhian, A. "Machine learning for risk stratification

of thyroid cancer patients: a 15-year cohort study", *European Archives of Oto-Rhino-Laryngology*, 281(4), 2095-2104, 2024.

[11] UCI Machine Learning Repository: Differentiated Thyroid Cancer Recurrence Available online: https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence (accessed on 1 August 2024).

[12] Alasadi, S. A., & Bhaya, W. S. "Review of data preprocessing techniques in data mining". *Journal of Engineering and Applied Sciences*, 2017, *12*(16), 4102-4107

[13] Jadhav, A., Dhaulakhandi, D., Shandilya, S. K., Malviya, L., & Mewada, A. "Data transformation: A preprocessing stage in machine learning regression problems". *In Artificial Intelligence Techniques in Power Systems Operations and Analysis*, 2023, (pp. 183-194). Auerbach Publications.

[14] Sagan, A., & Łapczyński, M. "SEM-Tree hybrid models in the preferences analysis of the members of Polish households", *Advances in Data Analysis and Classification*, 14, 855-869, 2020.

[15] Perlich, C., & Świrszcz, G. "On cross-validation and stacking: "Building seemingly predictive models on random data", *ACM SIGKDD Explorations Newsletter*, 12(2), 11-15, 2011.

[16] Yaşar, Ş. "Determination of Possible Biomarkers for Predicting Well-Differentiated Thyroid Cancer Recurrence by Different Ensemble Machine Learning Methods*." Middle Black Sea Journal of Health Science*, 2024, 10(3), 255-265.