Received: 26 September 2024 | Revised: 18 November 2024 | Accepted: 29 November 2024 | Published online: 17 December 2024

RESEARCH ARTICLE

Medinformatics
2024, Vol. 00(00) 1–9
DOI: 10.47852/bonviewMEDIN42024441

BON VIEW PUBLISHING

Towards Predicting Recurrence Risk of Differentiated Thyroid Cancer with a Hybrid Machine Learning Model

Subhagata Chattopadhyay[1,*]

[1] *Independent Researcher, India*

Abstract: Thyroid cancers (TC) often recur. This paper tests a novel unsupervised and supervised hybrid machine learning model to predict the recurrence risk (RR) and its score (RRS) in a population of “differentiated thyroid cancer” (DTC) cases as the prognostic measure. The DTC data (383 × 13) are collected from the UCI library. The population is grouped into “high risk of recurrence” (HRR) and “no high risk of recurrence” (NHRR) using the agglomerative clustering algorithm (ACA). Prior, the dataset is log-normalized [0,1], column-wise as a preprocessing step. Log-normalized values of the predictors, their corresponding coefficients, and the constant/intercept are used to construct a multiple linear regression to compute the RRS. Further, RRS values are normalized [0,1] using a log-sigmoidal function and termed “RRS_norm”. RRS_norms closer to the average RRS_norms of HRR and NHRR determine the predicted group. The model’s performance is measured with a confusion matrix, and RRS_norm results are matched with the RR labeled within the dataset. The result shows that ACA can correctly cluster the dataset into HRR and NHRR by 63.4%. Based on the coefficient values, predictors such as “Age”, “Gender”, “Smoking”, “History of smoking”, “History of Radiotherapy”, “Adenopathy”, and “Tumor staging” which comprise 53.84% of the total number of predictors show a positive correlation with “recurrence”. However, while matching the RRS_norms with the actual RRs, a 21.68% mismatch is observed, which mandates investigations with other DTC datasets.

Keywords: differentiate thyroid cancer, agglomerative clustering algorithm, multiple linear regressions, recurrence predictions, hybrid machine learning model

1. Introduction

Differentiated thyroid cancer (DTC) comprises more than 90% of thyroid cancers (TC) and includes papillary and follicular types [1]. There has been an increasing trend of DTC in the last three decades [2]. However, the overall mortality rate remains manageably low [3]. The five-year survival rate of stage-I and II DTCs is as high as 100% while it is 98% in those suffering from higher stages [4]. Although the overall clinical profile of DTC is good, studies have observed a recurrence rate of 12–20% [5], and post-operative recurrence is one of the most common causes of death [6]. Factors that increase the chance of recurrence are (i) large-sized tumors, (ii) extra-thyroid extensions such as lymph nodal metastases, (iii) multifocality, and (iv) BRAF gene mutation [7]. It has been therefore observed that clinical prediction of recurrences of DTC is highly complex [6].

Computer sciences play an important role in initiating and establishing the mission of “health for all”, helping create an electronic medical record to store all necessary health data of patients who avail services of an establishment. This vast domain brings automation to manual tasks yielding speed and thus saving cost in terms of time and money. Its algorithms help clean noisy data and provide analytics in the form of descriptive statistics to an advanced level of knowledge engineering through exploratory data analytics and insight mining. Algorithms also create several types of predictive models, which are data-driven, comprehensive, and explainable. There are special algorithms, e.g., supervised learning, unsupervised learning, and reinforcement learning based on the labeled vs. unlabeled datasets vs. agent-action-state-environment of any given problem that gives a desired optimal result through reward and punishments, respectively, that machines learn through training, testing, and validations iteratively on several types of datasets of various sources [8]. These are called machine learning (ML) algorithms. In 21st-century healthcare, a lot of emphasis has been put on adopting and using ML models for various clinical predictions where the scenario is utterly challenging to uphold value-based care [9] through precision medicine [10] practices. The algorithms, thus, can better be termed assistive intelligence than artificial intelligence in healthcare as it considers all “organic” variables for decision-making similar to human doctors and nothing is there as “artificial”. Attempts have also been made to enhance these algorithms to such an extent that they behave like human doctors [11] and highlight and explain why specific types of ML algorithms work the way human doctors arrive at a diagnosis [12].

The key objectives of using ML algorithms to model clinical predictions are to assist medical doctors in arriving at comprehensive data-driven decisions without forgetting any pertinent variable within the rule base and to explain the reasons behind making such decisions by highlighting the key driving variables. The algorithm learns to identify the key drivers iteratively

*__Corresponding author:__ Subhagata Chattopadhyay, Independent Researcher, India. Email: subhagata.chattopadhyay2017@gmail.com

to arrive at a decision and minimize decisional error to learn patterns effectively. Hence, it works as an assistive intelligence tool. As mentioned above, predicting DTC is a complex task. It needs continuous follow-ups and despite all, relapses occur due to new cancerous cell generations and/or undetected cancerous residues which eventually proliferate [13].

The paper aims to predict the chances of recurrences of DTC in a given population as the prognostic measure. The objective is to build and implement a hybrid machine learning model (HMLM) using the agglomerative clustering algorithm (ACA, which is an unsupervised learning algorithm) and the multiple linear regressions (MLR, which is a supervised learning algorithm) to predict the recurrence risk score (RRS) of DTC patients. This work also tests the credibility of ACA in learning clinicopathological information in the domain of oncology and grouping the population based on the risks of cancer recurrences by recognizing case patterns.

2. Material and Methods

This section showcases the 1. Dataset and its features, 2. Analytics performed, 3. ACA model construction and implementation, 4. Observed outcomes and discussions, 5. Construction of MLR to compute the RRS_norm as the metric of chances of recurrences of DTC in the patient population, and finally, 6. Validating RRS-norm-based recurrences (HMLM-based) to the observed recurrences (real-world) mentioned within the dataset, and 7. Hand calculations for one case to understand the RRS_norm calculations and the corresponding interpretations. The last two steps are shown in the following section. The rationale behind combining ACA and MLR to develop an HMLM is that ACA can group the population based on the recurrence risks (RRs) by learning the case patterns, i.e., clinicodemographic data while MLR can log-linearly translate it further to compute the risk score, making more objective decision-making than a subjective interpretation, done by the clinicians and population health specialists.

2.1. Dataset and features

The DTC dataset is available in the UCI library (public domain) [14]. The data were collected for 15 years and followed for at least 10 years [15]. The original shape of the data is 383 rows (cases) and 17 "predictors" out of which 4 are "target or dependent" variables and the remaining are the "independent" variables. In this work, all target variables, e.g., "Stage", "Risk", "Treatment response", and "whether recurred or not" are excluded to convert it as an unlabeled dataset, which is fit for clustering. However, later, the last target variable, e.g., "recurred or not" is used for model validations when the predicted risk and actual risk are compared case-wise.

For ease of computation, all categorical values are converted into numerical values, and Supplementary Table 1 presents it. The values are then log-normalized column-wise to convert the skewed distribution due to high variance within the features/predictors to a normal distribution (see Equation (1)). Normalization is a preprocessing step in data mining, which enhances data quality by reducing redundancies, and noise within the data, such as duplicate and inconsistent data.

$$X_{log} = \log(X) \tag{1}$$

In this equation, X represents all the predictors (x).

The data type is verified with Python IDLE before proceeding to exploratory data analytics, described below.

2.2. Exploratory data analytics

Here the predictors are divided into three constructs – A. Demographic data, B. Social determinant of health, and C. Clinical data. For each construct, a histogram of one of the variables is plotted (refer to Supplementary Figure 1) to give a glimpse of the data distribution within the given variable. The plots can be seen in the following section.

2.3. ACA model construction and implementation

The ACA is a popular hierarchical clustering method (unsupervised learning) that starts clustering data points based on dissimilarities. Its working principle is described below:

Step 1: Compute the proximity matrix using distance metrics, e.g., the Euclidean (see Equation (2)) distance

$$d(p,q) = \sum\nolimits_{i=1}^{N} (p_i - q_i)^2 \tag{2}$$

In this equation, "p" and "q" are two points in the Euclidean "N" space and are the Euclidean vectors, starting from the origin of the space or initial point; and "N" refers to the N-space.

Step 2: Use the Linkage function (here, the average linkage function is applied, see Equation (3)) to group objects to form a hierarchical cluster tree with the help of Step 1

$$D(r,s) = \frac{T_{r,s}}{(N_r N_s)} \tag{3}$$

Here, $D(r,s)$, $T_{r,s}$, r, s, N_r, and N_s refers to the average linkage function, the sum of the pairwise distances between cluster "r" and "s", and respective cluster sizes.

Step 3: The data points with closer proximities are merged to form the clusters

Step 4: Repeat steps 2–3 till all data points are grouped into clusters or a single cluster remains as either the outlier or a well-defined cluster.

It is important to state although there are several other popular clustering techniques available; however, in this paper, the author has arbitrarily chosen ACA to note how it behaves in this type of multimodal dataset. To minimize the increased computational complexity due to data dimensionality, principle component analysis (PCA) has been applied. By transforming the original variables into a new set of non-correlated variables, i.e., the principle components (PCs), PCA simplifies data for clustering. It helps to focus on the significant features instead of all the features. In this way, it brings more efficiency and interpretability to the clustering process. Also, based on the PCs, the DTC dataset can be clustered into HHR and NHHR sub-groups instead of linkage-based clusters, which is the key objective of using PCA to get crisp clusters through ACA.

The steps of PCA are described below:

Step 1: Data normalization

Step 2: Computation of covariance matrix, which is a symmetric/ square matrix

Step 3: Selection of Eigenvectors (perpendicular direction) and their respective Eigenvalues (amount of variance for the given direction, expressed as scalar values) in linear algebra to explain linear transformations.

Step 4: Selection of the PC by computing Eigenvectors and their respective Eigenvalues. The highest Eigenvalue decides the first PC, the immediate next lower value refers to the second PC, and so on.

Step 5: Data transformation into a new dimension space, defined by the PCs. This transformation/reorientation is done by multiplying the original data by the previously computed Eigenvectors.

2.4. Constructing an HMLM using the MLR equation

Focus has been made on each case of high risk of recurrence (HRR) and no high risk of recurrence (NHRR) having predictors (x) with its average log-normal (LN) values and the respective outcome (y) as 1. In this respect, it is important to note that any LN distribution is a rightward skewed continuous probability distribution and is popularly used for modeling various natural phenomena where negative values do not make sense, e.g., recurrence of DTC.

While constructing the MLR equation, at the initial step, regression coefficients are calculated to establish the relationships between the predictors and the outcome. Below are the steps through Equations (4)–(7):

Step 1: Computing the average of all log-normalized predictors

Step 2: Computing the coefficients (a) of each predictor using Equation (4)

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \tag{4}$$

Step 3: Computing the constant (b) using Equation (5)

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x^2)} \tag{5}$$

In both equations, "n" determines the number of predictors.

Step 4: RRS calculations for all the cases using Equation (6).

$$RRS_j = \sum a_i x_i + b \tag{6}$$

where "i" varies from 1 to "n" and "j" varies from 1 to "N" (total number of cases).

Step 5: RRS values are normalized [0,1] using log-sigmoidal transfer function using Equation (7).

$$RRS_{norm} = \frac{1}{1 + e^{-RRS}} \tag{7}$$

where "e" denotes the Euler number having a value of 2.718.

Validation results and hand calculations are showcased below.

It is worth noting that the paper aims to conceive, develop, and implement an HMLM using one unsupervised (ACA) and one supervised (MLR) ML algorithm to predict the RR of DTC for a given set of cases and calculate their respective RRS. No attempt has been made here to compare the developed model with other ML methods.

3. Results and Discussions

In this section, the outcomes of the experiment have been presented sequentially. Please refer to the Supplementary file to see Supplementary Table 1 and Supplementary Figures 1–2, which supports the results. Primary corroborators, such as Figure 1 (ACA-based cluster plots) and Table 1 (computation of RRS), are retained in the main text.

3.1. Exploratory data analytics

3.1.1. Demographic data

1) ***Gender distribution:*** Male 71 (18.53%) and Female 312 (81.47%).
2) ***Age distribution:*** Male and Female (15–82 years) with an average age of 47 and 41 years, respectively with a standard deviation of 15. Supplementary Figure 1(A) displays the histogram plot of "Age".

3.1.2. Social determinant of health

1) ***History of Smoking:*** 17% and 5% of Males and Females, with average ages of 55.83% and 42.31%, respectively. There are 0.73% of people with a History of Smoking in the population.
2) ***Smokers:*** 56% of Males with an average age of 52.77 years and 2% of Female populations with an average age of 54.44 years. There are 12.8% smokers in the population. Supplementary Figure 1(B) shows the histogram of Smokers.

3.1.3. Clinical data

1) ***History of Radiotherapy:*** In the population, 0.18% has a history of Radiotherapy. Out of which 8% of Males and 0.3% of Females with an average age of 60 years each received it. Radiotherapy, especially radioactive iodine significantly reduces the risk of recurrence [16]. The data show that 100% of Males are Smokers and have a History of Smoking in the past. Interestingly, cigarette smoking decreases the overall chance of TC [17].
2) ***Pathological state of the Thyroid gland:*** In the population, 13.57% of cases have pathological states of the Thyroid gland, elaborated below.
   - **"Subclinical hypothyroidism":** 0.7% Males of average age of 52.4 years have it while 2.88% of Females are recorded to have this condition with an average age of 46.55 years. Out of it, 40% of Males have a History of Smoking and also currently are Smokers. None has any history of Radiotherapy. In the case of Females, each of the 11% are Smokers and have a History of Smoking. None has any history of Radiotherapy.
   - **"Subclinical hyperthyroidism":** It is observed in a relatively younger age (29.6 years). It is observed that 0.14% of Males and 0.12% of Females are suffering from it and there is no relationship with smoking and Radiotherapy.
   - **"Clinical hypothyroidism":** The average age of this condition is 39 years in the population affecting 0.14% of Males and 0.35% of Females and there is no relationship with Smoking habits and Radiotherapy.
   - **"Clinical hyperthyroidism":** The average age of this group is 42.4 Years with the involvement of 0.7% Males and 0.48% Females without any relationships with Radiotherapy and Smoking habits.
3) ***Physical examination of the Thyroid gland:*** Only 0.18% of the total population have normal Thyroid gland and 100% of this population are Females, while the remaining have goiters (99.82%), e.g., (a) Single nodular goiter right (36.55% in the total population, 12.14% of Males with the average age group of 41.82 years and 87.85% of Females with 38.6 years are suffering from it); (b) Single nodular goiter left (23.23% in the total population, affected 25.55% Males with 47.63 years of average age and 74.44% of Females with 39.4 years of average age are

affected). It is important to note that in this population of Single nodular goiters, there are 12.66% Smokers, but none received any Radiotherapy. A size of 36.55% of the population has (c) Multinodular goiter where 23.57% are Males (60% are smokers) and the rest are Females (10% are smokers).

4) ***Adenopathy:*** It is observed that 72.32% do not have any adenopathy (swollen and or painful lymph nodes in the body). In the remaining group having adenopathy, 16.03%, 45.28%, 30.18%, 0.19%, and 6.6% have "left", "right", "bilateral", "posterior", and "extensive" adenopathy. In the population, 33% Males and 67% Females have adenopathy. Almost 25.5% are Smokers and 3.77% have a history of Radiotherapy. Supplementary Figure 1(C) refers to the histogram plot of Adenopathy.
5) ***Pathology:*** There are four subtypes, e.g., (a) Follicular (7.3%) with Female predominance (78.57%) having average age of 43.31 years; (b) Hurthel cell (5.22%) with 68.18% Female dominance having average age of 43.28 years; (c) Micropapillary (12.53%) comprising of 89.58% Females with average age of 43.81 years; and (d) Papillary (74.93%) mostly found in 81.25% of Females with average age of 38.14%. Most of the patients in these pathology groups are Euthyroid and do not have any mentionable relationships with smoking, history of smoking, or radiotherapy.
6) ***Focality:*** It has been observed that 47.25% are (a) Unifocal mostly affecting Females (85.71%) with an average age of 35 years and the remaining are (b) Multifocal affecting 72.64% of Females in the population. Similar revelations can be noted concerning smoking, history of smoking, radiotherapy, thyroid state (Euthyroid), and Goiters.

Figure 1 shows clear demarcations of HRR and NHRR subpopulations

Python 3.11.1 preloaded with pandas, matplotlib, scipy, numpy, math, and sklearn on 64-bit Windows OS × 64-based Processor Intel(R) Core TM @ 2.80 GHz has been used to cluster the DTC dataset into two groups – (a) having and (b) not having a chance of recurrence. Case-wise cluster labels are "0" which refers to the HRR and "1" indicates the NHRR group (see Supplementary Figure 1) and the corresponding plots in Supplementary Figure 2 where the blue and red dots refer to HRR and NHRR, respectively. As ACA is sensitive to outliers, Supplementary Figure 1 shows all 13 variables are devoid of outliers.

3.2. ACA model outcome

ACA clustered the thyroid dataset ($N = 383$) into two groups – (a) HRR ($N = 180$) and (b) NHRR ($N = 203$), while in the original data, there are 275 NHRR and 108 HRR cases. HRR are the "positive" cases of recurrences, otherwise, these are NHRR. Thus, ACA has misclassified HRR and NHRR by 37.6%.

3.3. Validation

Predictors such as "Age", "Gender", "Smoking", "History of smoking", "History of Radiotherapy", "Adenopathy", and "Tumor staging" comprises 53.84% of the total number of predictors ($n = 13$) and each shows a positive correlation with recurrence while remaining predictors have negative relationships based on the positive and negative coefficient values, respectively, calculated using Equations (4) and (5). A systematic literature search corroborates this finding [18–21]. Hence, it can be said that these are well-acclaimed cofounding predictors against the backdrop of the DTC recurrence.

HRR cases show high RRS_norms, which are greater than its average (0.4664) and the value range (±0.0373), i.e., 0.43 to 0.5073. Thus, values greater than or equal to or even close to 0.5073 are considered as the high possibility of getting HRR cases. It is further validated against the original label to note the mismatches, if any. In this experiment, a 21.68% mismatch is observed as the HMLM correctly predicts HRR cases by 78.32%.

The author, however, believed that to bring more generalizability to the developed HMLM, it should be tested on other DTC datasets, labeled with "recurrence".

3.4. Hand calculations

To understand the steps of computing RRS_norm, below is the hand calculation for one case:

Independent variables: (1) Age (27 yrs.); (2) Gender (F); (3) Smoker (No); (4) Smoking history (No); (5) Radiation history (No); (6) Thyroid function (Euthyroid); (7) Physical examination (Single nodular goiter left); (8) Adenopathy (No); (9) Pathology (micropapillary); (10) Focality (unifocal); (11) T (1a); (12) N (0); and (13) M (0).

Dependent variable: RS (No).

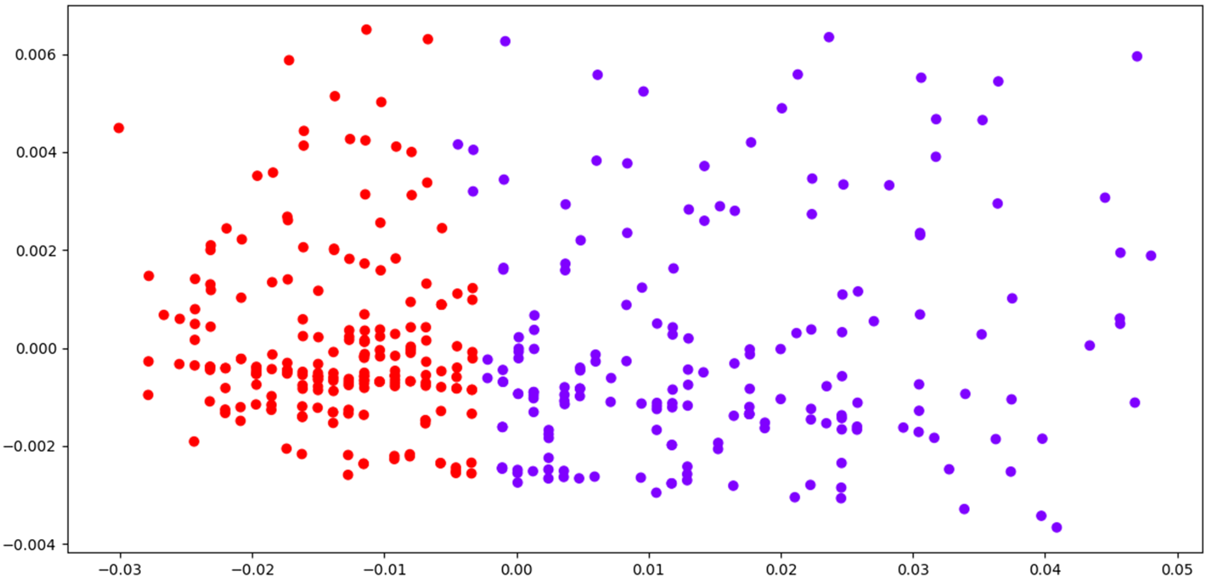


Figure 1. Agglomerative cluster plots

Table 1. Log-normalized 13 variables (V1-13) and their respective coefficients (Coeff 1-13) and the constant to compute the RRS and the RRS_norm as the final score of one sample case

V1	**Coeff1**	**V2**	**Coeff2**	**V3**	**Coeff3**	**V4**	**Coeff4**	**V5**	**Coeff5**
0.0314	0.002	0	1.44	0	27.54	0	48.86	0	112.9
V6	**Coeff6**	**V7**	**Coeff7**	**V8**	**Coeff8**	**V9**	**Coeff9**	**V10**	**Coeff10**
0	−74.87	0.0011	12.86	0	−4.2	0.0011	76.94	0	−43.95
V11	**Coeff11**	**V12**	**Coeff12**	**V13**	**Coeff13**				
0	−22.95	0	−98.05	0	80.54				

Constant: −0.10409.

RRS: −0.0293 (using Equation (6))

RRS_norm: using log-sigmoid function as the final output: 0.4926 (using Equation (7))

Mean of HRR: 0.44

Mean of NHRR: 0.48

Log-normalized values of the Independent variables and their corresponding coefficients to construct the MLR model are displayed in Table 1.

HMLM interprets based on the RRS-norm's distance from the mean RRS_norms of HRR and NHRR subpopulations. In case 1, the computed RRS_norm is closer to NHRR and hence the interpretation of RRS is "Nil", which matches with the actual occurrence in case 1. Similarly, the model interprets the remaining 382 cases with 78.32% accuracy. Case 1 features and their corresponding coefficients can be seen in Table 1. From this table, it can be noted that "Age", "Physical examination finding", and "Pathological finding" plays a crucial role in the RRS prediction. Here, the patient is a 27-year-old female having single nodular goiter on the left side and micropapillary type of tumor cells, which seldom show a high degree of recurrences. Similarly, "Age", "Gender", "Smoking", "History of smoking", "History of Radiotherapy", "Adenopathy", and "Tumor staging" comprises 53.9% of the significant predictors for recurrences.

3.5. Discussions

Previous research such as Survival prediction of DTC using artificial neural network (ANN) which falls under soft computing technique and logistic regressions (LR) which falls under traditional statistical technique has been carried out by a group of researchers on Surveillance, Epidemiology, and End Result 1 (SEER1) dataset. The study observed that ANN and LR both gave high average accuracy (>80%) in predicting 1, 3, and 5-year survival [22].

In another study on the de-identified SEER dataset, retrospective information has been used to recommend treatment choices in DTC cases using feature selection algorithms, e.g., Fisher's discriminant ratio, Kruskal-Wallis' analysis, and Relief-F. The patients were labeled as those who did not survive more than 5 years and who survived for ten years or more post their diagnoses. ML algorithms (optimized ANNs) are developed with 34 unique features taking records between 6,756 and 20,344 for different models giving 94.5% accuracy [23].

A comprehensive review by Cao et al. [24] used throughput features, radionics, and a quantitative extraction technique from the pictures to detect DTC within these pictures. The authors emphasized radiomics as one of the best measures for automatic mining of the quantitative characteristics which can then be fed into the ML algorithms for predictions. On the other hand, to understand the tumor behavior, unsupervised learning methods (clustering) have been applied by Yang et al. [25]. Authors used the ensemble algorithm for clustering cancer data to develop a prognostic algorithm for the DTC cases. The algorithm has three steps: 1. Application of Geha-Wilcoxon test statistics to explore the initial dissimilarities, 2. Obtaining learned dissimilarities, and finally, 3. Performing a hierarchical clustering technique for DTC detection and prognostic measures.

4. Conclusion

The paper demonstrates the development of an HMLM for the prediction of recurrence in DTC cases. The model correctly predicts HRR cases by 78.32% when matched with real cases. The proposed model is a hybrid of unsupervised (ACA) and supervised learning algorithms (MLR) where ACA groups the cases into HRR and NHRR and MLR equates the RR as the possibility in the patients with a "confidence" value. Recurrences in cancers are a common occurrence. This simple method of computing the RRS can be adapted for several other cancers. It also assists doctors in taking a closer look at each of the mismatched cases to curb human and/or algorithm bias in computer-driven clinical decision-making. Put together, it helps obtain a positive patient outcome.

The limitation of this work is the lack of generalizability of the proposed model, which can be tested on several other datasets. The contribution, on the other hand, is to conceptualize, develop, and successfully implement the HMLM in the oncology domain to predict RRs.

The developed model can be tested with other ML methods as the extension of this work in the future.

Recommendations

The finding reveals that the proposed HMLM can efficiently predict the recurrence of DTC in a given population with its respective RRS to prioritize monitoring and preventive efforts.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by the author.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence, reference number [14].

Author Contribution Statement

Subhagata Chattopadhyay: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Sherma, S. I. (2003). Thyroid carcinoma. *The Lancet*, *361*(9356), 501–511. https://doi.org/10.1016/S0140-6736(03)12488-9

[2] Pellegriti, G., Frasca, F., Regalbuto, C., Squatrito, S., & Vigneri, R. (2013). Worldwide increasing incidence of thyroid cancer: Update on epidemiology and risk factors. *Journal of Cancer Epidemiology*, *2013*(1), 965212. https://doi.org/10.1155/2013/965212

[3] Ahn, H. S., Kim, H. J., & Welch, H. G. (2014). Korea's thyroid-cancer "epidemic"—Screening and overdiagnosis. *New England Journal of Medicine*, *371*(19), 1765–1767. https://doi.org/10.1056/NEJMp1409841

[4] Tsang, R. W., Brierley, J. D., Simpson, W. J., Panzarella, T., Gospodarowicz, M. K., & Sutcliffe, S. B. (1998). The effects of surgery, radioiodine, and external radiation therapy on the clinical outcome of patients with differentiated thyroid carcinoma. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *82*(2), 375–388. https://doi.org/10.1002/(SICI)1097-0142(19980115)82:2%3C389::AID-CNCR19%3E3.0.CO;2-V

[5] Luo, X., Chen, A., Zhou, Y., Jiang, Y., Zhang, B., & Wu, J. (2019). Analysis of risk factors for postoperative recurrence of thyroid cancer. *Journal of BUON*, *24*(2), 813–818.

[6] Kim, S. Y., Kim, Y. I., Kim, H. J., Chang, H., Kim, S. M., Lee, Y. S., . . . , & Park, C. S. (2021). New approach of prediction of recurrence in thyroid cancer patients using machine learning. *Medicine*, *100*(42), e27493. https://doi.org/10.1097/MD.0000000000027493

[7] Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E., . . . , & Wartofsky, L. (2016). 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid*, *26*(1), 1–133. https://doi.org/10.1089/thy.2015.0020

[8] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

[9] Crowson, M. G., & Chan, T. C. (2020). Machine learning as a catalyst for value-based health care. *Journal of Medical Systems*, *44*(9), 139. https://doi.org/10.1007/s10916-020-01607-5

[10] Nilius, H., Tsouka, S., Nagler, M., & Masoodi, M. (2024). Machine learning applications in precision medicine: Overcoming challenges and unlocking potential. *TrAC Trends in Analytical Chemistry*, *179*, 117872. https://doi.org/10.1016/j.trac.2024.117872

[11] Chattopadhyay, S. (2013). Mathematical modelling of doctors' perceptions in the diagnosis of depression: A novel approach. *International Journal of Biomedical Engineering and Technology*, *11*(1), 1–17. https://doi.org/10.1504/IJBET.2013.053702

[12] Chattopadhyay, S. (2024). Decoding medical diagnosis with machine learning classifiers. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN42022583

[13] Shaha, A. R. (2012). Recurrent differentiated thyroid cancer. *Endocrine Practice*, *18*(4), 600–603. https://doi.org/10.4158/EP12047.CO

[14] Arvidsson, J. (2024). Differentiated thyroid cancer recurrence. *Kaggle*. Retrieved from: https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence

[15] Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhian, A. (2024). Machine learning for risk stratification of thyroid cancer patients: A 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*, *281*(4), 2095–2104. https://doi.org/10.1007/s00405-023-08299-w

[16] Toraih, E., Webster, A., Pineda, E., Pinion, D., Baer, L., Persons, E., . . . , & Kandil, E. (2024). Radioactive iodine ablation therapy reduces the risk of recurrent disease in pediatric differentiated thyroid carcinoma. *Surgical Oncology*, *56*, 102120. https://doi.org/10.1016/j.suronc.2024.102120

[17] Jiang, H., Li, Y., Shen, J., Lin, H., Fan, S., Qiu, R., . . . , & Chen, L. (2023). Cigarette smoking and thyroid cancer risk: A Mendelian randomization study. *Cancer Medicine*, *12*(19), 19866–19873. https://doi.org/10.1002/cam4.6570

[18] Kaur, J., Nadarajan, A., Janardhan, D., George, N. A., Thomas, S., Varghese, B. T., & Krishna, J. (2023). Predictive factors for nodal recurrence in differentiated thyroid cancers. *Cancer Treatment and Research Communications*, *36*, 100728. https://doi.org/10.1016/j.ctarc.2023.100728

[19] 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. https://doi.org/10.3389/fendo.2023.1268282

[20] Lee, J. H., Chai, Y. J., & Yi, K. H. (2021). Effect of cigarette smoking on thyroid cancer: Meta-analysis. *Endocrinology and Metabolism*, *36*(3), 590–598. https://doi.org/10.3803/EnM.2021.954

[21] Shokoohi, A., Berthelet, E., Gill, S., Prisman, E., Sexsmith, G., Tran, E., . . . , & Ho, C. (2020). Treatment for recurrent differentiated thyroid cancer: A Canadian population based experience. *Cureus*, *12*(2), e7122. https://doi.org/10.7759/cureus.7122

[22] Jajroudi, M., Baniasadi, T., Kamkar, L., Arbabi, F., Sanei, M., & Ahmadzade, M. (2014). Prediction of survival in thyroid cancer using data mining technique. *Technology in Cancer Research & Treatment*, *13*(4), 353–359. https://doi.org/10.7785/tcrt.2012.500384

[23] Mourad, M., Moubayed, S., Dezube, A., Mourad, Y., Park, K., Torreblanca-Zanca, A., . . . , & Wang, J. (2020). Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis. *Scientific Reports*, *10*(1), 5176. https://doi.org/10.1038/s41598-020-62023-w

[24] 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. https://doi.org/10.3390/cancers13102436

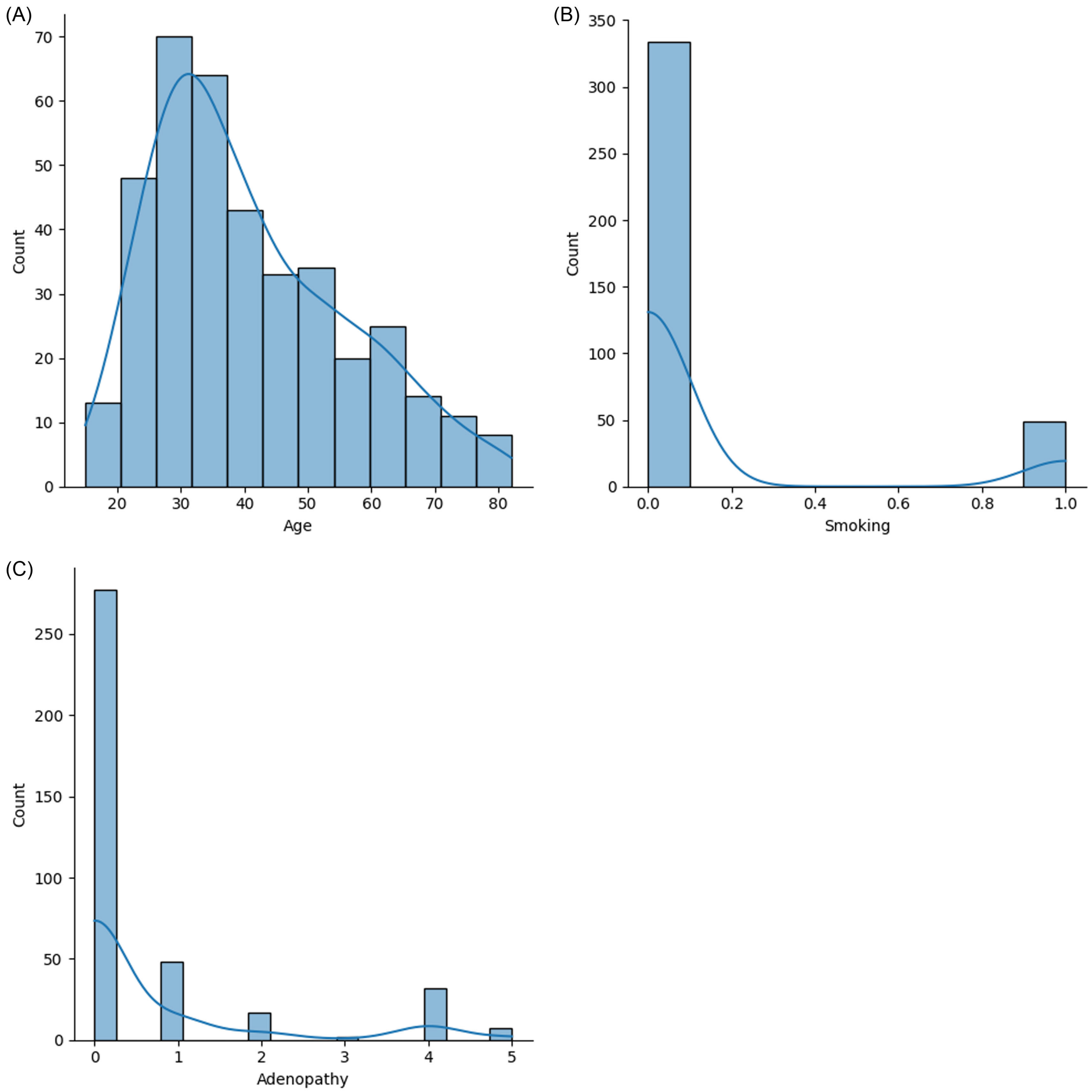
[25] Yang, C. Q., Gardiner, L., Wang, H., Hueman, M. T., & Chen, D. (2019). Creating prognostic systems for well-differentiated thyroid cancer using machine learning. *Frontiers in Endocrinology*, *10*, 288. https://doi.org/10.3389/fendo.2019.00288

How to Cite: Chattopadhyay, S. (2024). Towards Predicting Recurrence Risk of Differentiated Thyroid Cancer with a Hybrid Machine Learning Model. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN42024441

Supplementary Files

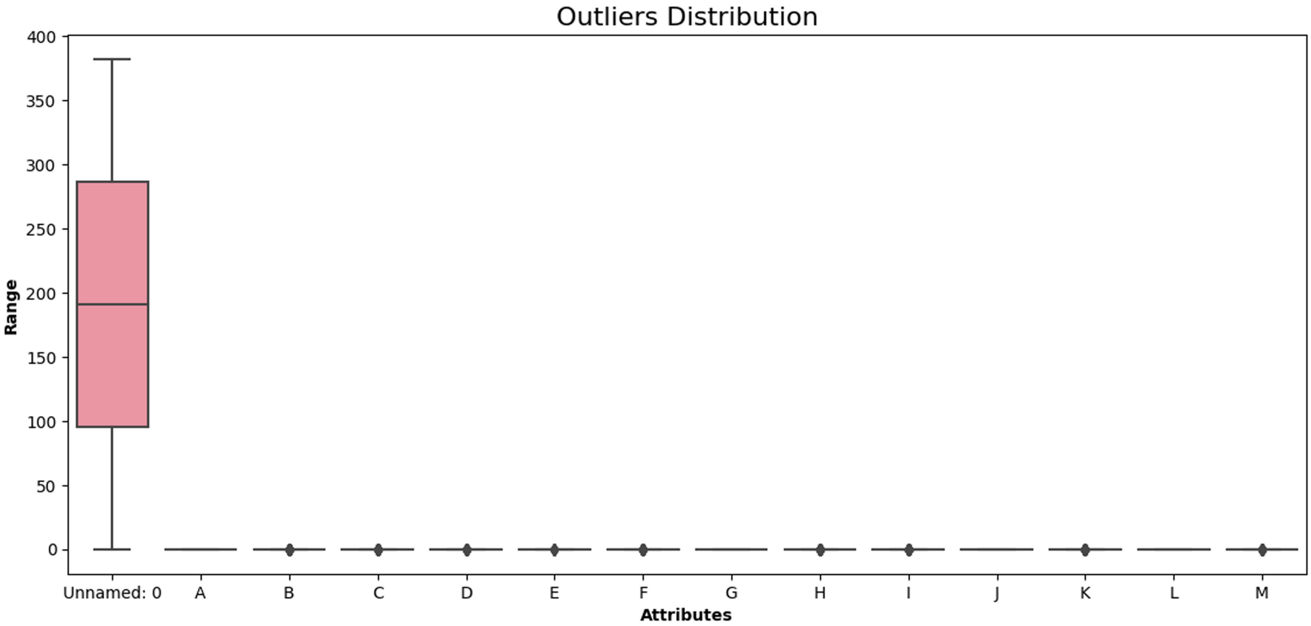
Supplementary Table 1. Conversion of categorical variables to numerical variables

Predictor	Categorical value	Numerical value
Gender	M	1
	F	0
Smoking	Yes	1
	No	0
History (Hx) of Smoking	Yes	1
	No	0
History (Hx) of Radiotherapy	Yes	1
	No	0
Thyroid Function	Euthyroid	0
	Subclinical hypothyroidism	1
	Subclinical hyperthyroidism	2
	Clinical hypothyroidism	3
	Clinical hyperthyroidism	4
Physical Examination	Normal	0
	Single nodular goiter left	1
	Single nodular goiter right	2
	Multinodular goiter	3
	Diffuse goiter	4
Adenopathy	No	0
	Right	1
	Left	2
	Posterior	3
	Bilateral	4
	Extensive	5
Pathology	Follicular	1
	Micropapillary	2
	Papillary	3
	Hurthel cell	4
Focality	Unifocal	1
	Multifocal	2
T (Tumor)	T1a	1
	T1b	2
	T2	3
	T3a	4
	T3b	5
	T4a	6
	T4b	7
N (Node)	N0	0
	N1a	1
	N1b	2
M (Metastasis)	M0	0
	M1	1



Supplementary Figure 1. (A) Histogram plot of Age. (B) Histogram plot of Smokers. (C) Histogram plot of Adenopathy

Supplementary Figure 1(A) is a 2D plot of Age vs. Count that shows the population is dense around 20-35 years, which is a relatively young population who suffer from DTC. In Supplementary Figure 1(B), the population shows predominance of non-smokers who suffer from DTC. Supplementary Figure 1(C) shows that most of the population does not have any kind of adenopathy but suffers from DTC.



Supplementary Figure 2. Outlier plots of variables

In this plot, Unnamed is the non-variable first column showing 383 cases and can be neglected while “A-M” are thirteen variables not showing any outlier. That is why ACA is the correct choice for clustering the DTC dataset.