2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Identifying Important Features for Clinical Diagnosis of Thyroid Disorder

Sowmya Balasubramanian
Department of Computer Science
University of Victoria
Victoria, Canada
sowmyab@uvic.ca

Venkatesh Srinivasan
Department of Computer Science
University of Victoria
Victoria, Canada
srinivas@uvic.ca

Alex Thomo
Department of Computer Science
University of Victoria
Victoria, Canada
thomo@uvic.ca

***Abstract*—Abnormal production of thyroid hormones in our body causes thyroid disorders such as hypothyroidism, hyperthyroidism, Hashimoto's disease, Graves' disease, and thyroid nodules. Undiagnosed thyroid disorders can affect the quality of life of an individual both physically and mentally. Thyroid disorders are common but sometimes become difficult to diagnose since the symptoms can be easily associated with other health conditions. Clinicians identify thyroid disorders by measuring the levels of thyroid hormones in our blood stream. This work aims to help clinicians by carefully investigating if thyroid diagnosis improves when all important features (a complete thyroid panel) is measured as opposed to a select few.**

Much of previous work has focused on the performance of classifiers, supervised and unsupervised, for the prediction of this disorder. Departing from this tradition, we focus on the concept of feature importance and its clinical implications. We identify the top-4 important features that predict the presence of thyroid disorder and show that these can be measured by clinicians cost-effectively. We also identify the pitfalls of current clinical practice of not checking the entire thyroid panel, prevalent in many countries with universal health care. Finally, we show that our results are quite robust and are unlikely to change with the choice of classifier or due to the inherent nature of a dataset in hand like imbalance.

***Index Terms*—thyroid disease, data mining, machine learning, feature importance**

I. Introduction

The endocrine system is a collection of glands that produce the hormones that control nearly all the important biological processes of the human body. These glands, situated in different parts of our body, include the hypothalamus, pituitary gland, and pineal gland in the brain; the thyroid and parathyroid glands in the neck; the thymus gland between the lungs; the adrenal gland on top of the kidneys; the pancreas behind the stomach, and the ovaries or testes in the pelvic region [1]. All these glands work together to help us lead a normal and healthy life.

The thyroid gland is a butterfly shaped gland that releases three thyroid hormones into the blood stream, levothyroxine (T4), triiodothyronine (T3) and calcitonin that regulate the body's metabolic rate, controlling heart, muscle and digestive function, brain development and bone maintenance [2]. Optimal supply of iodine from our diet aids in the correct functioning of the thyroid gland [3]. Malfunction of this gland results in thyroid disorders causing extreme fatigue, trouble sleeping, enlarged thyroid gland (goiter), vision problems, muscle weakness, weight gain or weight loss, intolerance to heat or cold and many more symptoms [4].

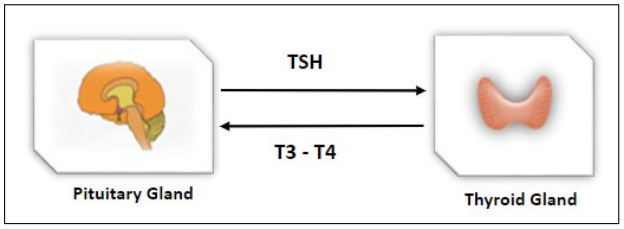


Figure 1. Pituitary Gland and Thyroid Gland: Relationship

Pituitary gland is a pea-sized gland located at the base of the brain below the hypothalamus that regulates other endocrine glands to release hormones [5]. Pituitary gland signals the amount of T3 and T4 the thyroid gland needs to release into the bloodstream by producing a regulating hormone called thyroid-stimulating hormone (TSH) [5] as shown symbolically in Fig. 1. A normal healthy thyroid shows an optimal balance of TSH, T3, and T4 hormones. However, when thyroid does not produce the right amount of hormones it causes a thyroid disorder or disease. Hypothyroidism is a condition of having too much TSH in the bloodstream indicating that the thyroid gland is not making enough T3 or T4 [4]. Hyperthyroidism is a condition of having low TSH levels showing that the thyroid gland is producing too much T3 and T4 [4].

Thyroid disease is one of the most common health conditions in adults that often remains undiagnosed. Thyroid Foundation of Canada website indicates that 1 in 10 Canadians suffer from a thyroid condition and 50% of them are undiagnosed [6]. American Thyroid Association estimates that around 20 million Americans have some form of thyroid disease and around 60 percent have not been diagnosed [7]. An undiagnosed variation in the level of thyroid hormones can have phenomenal effect on one's physical and mental health.

Clinicians use a simple blood test to diagnose thyroid disorders. It measures the amount of thyroid hormones in the blood stream and reports if the person has a thyroid disorder or not. Such a test can either measure the complete thyroid panel (TSH, FT4 (Free T4) and FT3 (Free T3)) or only the TSH as

IEEE/ACM ASONAM 2022, November 10-13, 2022
http://dx.doi.org/10.1145/XXXXXXX.XXXXXXX
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

a cost saving way. We observed that the clinical practice of measuring TSH only is common in countries with universal health care [8], [9]. In order to understand the efficacy of these tests and provide clinicians with information they can trust, we initiate a systematic study of feature importance in the diagnosis of thyroid disorder.

This work develops a predictive model to facilitate early diagnosis of this disease. Using data mining techniques, we identify important predictors of thyroid disease that enable clinicians to diagnose this thyroid disease effectively. Our main contributions are as follows:

1) Using an explainable classifier (CART), we show that the top-4 features for thyroid diagnosis are FTI, TSH, TT4 and T3 (See Table III) which can be measured cost effectively using a simple blood test.
2) We also show that a clinical test in which TSH is measured in isolation is not sufficient leading to possible misdiagnosis of the disorder.
3) We perform experiments to demonstrate the robustness of our conclusions using different approaches, including the use of a new classifier (Random Forest), principal component analysis, and methods to handle the imbalance in our dataset.
4) Finally, we also shed light on an ambiguity in the existing literature when defining class labels for the thyroid dataset we use and provide explanations to resolve it.

II. Related Work

There is extensive literature analyzing the performance of various classification methods, supervised and unsupervised, for the diagnosis of thyroid dysfunction.

Early research focused on the use of neural networks for this task. Sharpe et al. [10] initiated the use of ANN for the diagnosis of thyroid disorder. Zhang et al. [11] showed that neural networks are more robust to sampling variation compared to traditional Bayesian classifiers. Hoshi et al. [12] analyzed the performance of two neural networks: self organizing maps and Bayesian regularized neural networks in the study of thyroid function and showed both to be useful. Temurtas [13] compared the performance of three different types of neural networks, multilayer (MLNN), probabilistic (PNN), and learning vector quantization (LVQ-NN) for thyroid disease diagnosis with PNN giving the best result. Saiti et al. [14] investigated the performance of genetic algorithms, based on combining SVM with PNN, for this task.

Liu et al. [15] built a classifier based on fuzzy K-nearest neighbour (kNN) with a strong performance. Chen et al. [16] explore the the performance of a hybrid system based on support vector machines for thyroid disease diagnosis. Li et al. [17] designed a computer aided diagnosis system based on principle component analysis (PCA) and extreme learning machine (ELM) to assist the task of thyroid disease diagnosis. In the recent years, [18]–[21] has focused on analyzing the effectiveness of simple decision tree algorithms for this prediction task. Please see [22] for a detailed survey of all the research on prediction of thyroid disease.

III. Dataset and Methods

A. Thyroid Dataset Description

This work uses the thyroid disease dataset available in KEEL repository [23]. KEEL cites as its source one of the databases available in the UCI repository [24]. The dataset in the UCI repository was contributed by J. R. Quinlan. The dataset, obtained from Daimler-Benz, contains 7200 instances and 21 features, of which 6 are continuous and 15 are binary datatype, and has no missing data. We note that the class label definitions are different in KEEL [23] and UCI [24] thyroid dataset. Table I summarizes class label definitions as reported in KEEL identifying **hypothyroid (93%) as majority class**. Table II summarizes the UCI class label definitions that reports **normal (93%) as the majority class.**

Table I
Class Label Definitions - KEEL Repository

Class Label	Description	No. of Records	%
1	Normal	166	2%
2	Hyperthyroid	368	5%
3	Hypothyroid	6666	93%

Among prior research works, KEEL [23] is used by [25], [26] and UCI [24] is used by [27]–[30]. The KEEL [23] and UCI [24] thyroid datasets having different class variable definitions pose a potential challenge regarding which one to use. We overcome this challenge by preprocessing and analyzing both the KEEL [23] and UCI [24] datasets carefully to help obtain more insights about the thyroid data in hand to aid with our decision making. See Section III-C below.

Table II
Class Label Definitions - UCI Repository

Class Label	Description	No. of Records	%
1	Hyperthyroid	166	2%
2	Hypothyroid	368	5%
3	Normal	6666	93%

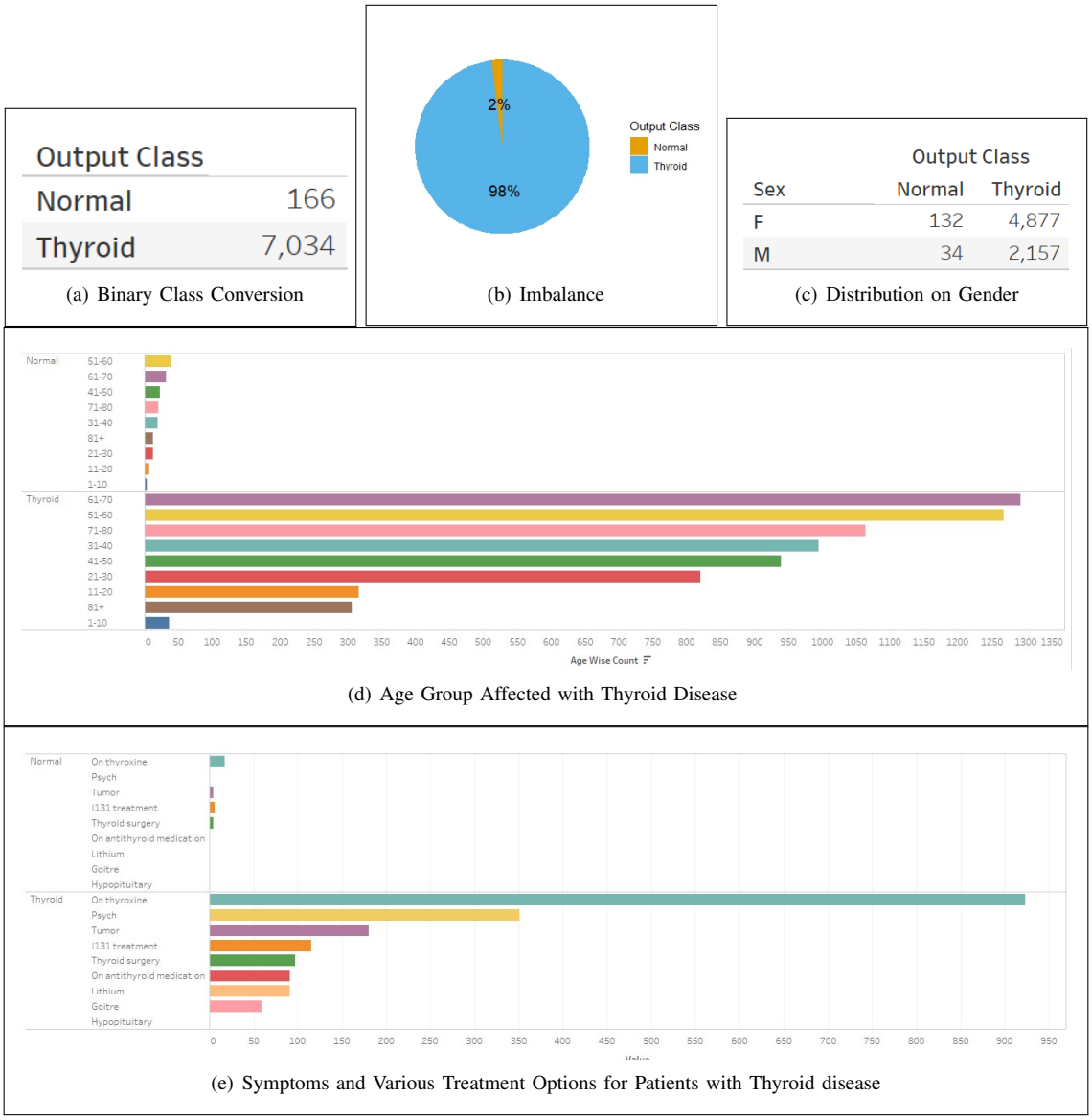
Table III elaborates on the acronyms for the most relevant variables used in KEEL [23] and UCI [24] thyroid dataset.

Table III
Thyroid Dataset Abbreviations

S.No	Acronyms	Description
1	TSH	Thyroid Stimulating Hormone
2	FTI	Free T4 Index (relates to FT4)
3	TT4	Total Thyroxine (relates to FT4)
4	T3	Triiodothyronine (relates to FT3)

B. Data Preprocessing

The primary objective of our work is to establish the presence or absence of a thyroid disorder while identifying important features that support clinical diagnosis. Keeping our objective in mind and the ambiguity in class label description found in previous literature as described earlier in Sec. III-A,



(a) Binary Class Conversion

(b) Imbalance

(c) Distribution on Gender

(d) Age Group Affected with Thyroid Disease

(e) Symptoms and Various Treatment Options for Patients with Thyroid disease

Figure 2. Descriptive Statistics: KEEL Thyroid Dataset

we decided to transform KEEL [23] and UCI [24] datasets to a binary classification problem. This is done by combining class labels hyperthyroid and hypothyroid into a new class label "1-Thyroid" indicating presence of "thyroid disease" and Class label "0-Normal" identifies normal samples without "thyroid disease". Dataset descriptions with the new binary class labels are outlined in Tables IV and V.

Table IV
KEEL Repository - Binary Classifications

Class Label	Description	No. of Records	%
0	Normal	166	2%
1	Thyroid	7034	98%

This decision not only left KEEL [23] and UCI [24] datasets heavily imbalanced as noticed in Table IV and Table V but also raised another difficult question as to which choice of class label descriptions is convincing for us to utilize. This challenge is overcome by analysing KEEL [23] and UCI [24] thyroid dataset. Among many observations, insights provided by Fig. 2(e) in Sec. III-C helped us choose the dataset to use for this work. Fig. 2(e) reports a very high number of patients getting treated with various thyroid treatments that are correctly classified as "1-Thyroid" by KEEL [23] making it the preferred choice for this work.

Table V
UCI Repository - Binary Classifications

Class Label	Description	No. of Records	%
0	Normal	6666	98%
1	Thyroid	534	2%

On the other hand, UCI [24] classifies these patients as

Table VI
TREATMENT OPTIONS - UCI REPOSITORY

Class	Thyroxine	I131	Surgery	Antithyroid	Lithium
0 - Normal	923	109	97	99	85
1- Thyroid	17	12	4	4	6

"0-Normal" despite getting treated for thyroid disease. This causes confusion and raises the following question for which we could *not* find a convincing answer - why are numerous patients classified as "0-Normal" by the UCI [24] dataset receiving thyroid treatment as reported in Table VI?

C. Descriptive Statistics

Data analysis convinced us to use the binary version of KEEL [23] as defined in Table IV for this work. This dataset is referred as KEEL dataset in the rest of this paper.

Fig 2 presents various observations of KEEL dataset. Fig. 2(a) and 2(b) show the binary class distribution and severe imbalance with thyroid at 98% and normal at 2%. Fig. 2(c) clarifies that women are more susceptible to this disease than men. Fig. 2(d) exhibits that "thyroid disease" is a progressive disease that predominantly occurs between the age of 30 and 70 and can even affect children less than 10 years.

Fig. 2(e) highlights the small number of outliers, those who are among the "0-Normal" patients but are treated using surgery or medications for the presence of thyroid disease. This project includes those outliers (a select few) without eliminating them since the samples identified as "0-Normal" are already low in KEEL dataset. Fig 2(e) also highlights the various symptoms and treatments among patients with thyroid disease conveying thyroxine as the most common treatment and not all thyroid patients end up having a goitre making this symptom less common.

D. Feature importance

This work focuses on understanding feature importance in supervised and unsupervised setting and its clinical relevance.

1) Supervised Learning using CART & Random Forest: Supervised learning is a machine learning (ML) approach where a model gets trained to classify labeled datasets and later apply them to predict outcome accurately [31]. Prior works [28]–[30] using [24] have established that decision tree classifiers achieve strong performance metrics. Furthermore, an added advantage of these classifiers is that they are highly interpretable. Therefore, Classification and Regression Trees (CART) is an excellent choice for the KEEL dataset and is our algorithm of choice for supervised learning. The CART model built using the training data is represented as a binary tree in which the internal nodes are labeled by features and the leaves by the class variable. Given the tree representation of the CART model, the important features are easily identified as they label the internal nodes at top levels of the tree.

Random Forest is another ML approach that is used to solve classification and regression problems. As opposed to CART that builds one decision tree, Random forest algorithm is a collection of decision trees. The algorithm bases its prediction on the resulting "forest" of decision trees by taking the majority (or mean) of the output from various trees. Intuitively, a decision based on collection of trees increases the accuracy of the outcome. CART and Random Forest have a built-in feature importance algorithm that uses Gini importance or mean decrease impurity.

2) Unsupervised Learning using Principle Component Analysis (PCA): Unsupervised learning is a ML approach used to analyze unlabeled datasets and find patterns [31]. Analyzing KEEL dataset by eliminating class labels not only help with identifying the top-4 important features but also validate the results provided by supervised approach. We choose this approach having found ambiguity in literature about the class label definition as described earlier in Sec. III-B.

PCA is an extensively used method for reducing the dimensionality of the feature set and is suited to work well on the continuous variables in a dataset. KEEL dataset has six continuous variables: Age, TSH, FTI, T3, TT4, T4U. PCA internally reduces the dimensions of a multivariate data to six principal components (PC), that can be visualized graphically, with minimal loss of information. PCA also reveals the data attributes contributing to each of these dimensions. The features contributing to principal dimension 1 (PC1) are considered important as it is the dimension with largest variance [32].

E. Metrics

In this work, we use several well known metrics to evaluate the performance of a classifier. Let TP, TN, FP and FN represent the True Positive, True Negative, False Positive and False Negative predictions of a chosen classifier. For example, false positives are cases that are actually negative but the model incorrectly labels as positive, or in our example, the model classifies a person as having a thyroid condition when they are actually normal.

Sensitivity is the fraction of actual positives which got predicted as a (true) positive. Sensitivity is also referred to as *Recall* or true positive rate. We have the following definition:

$$Sensitivity(or)Recall = \frac{TP}{TP + FN}$$

Specificity or true negative rate is defined as the fraction of actual negatives, which got predicted as a negative. That is,

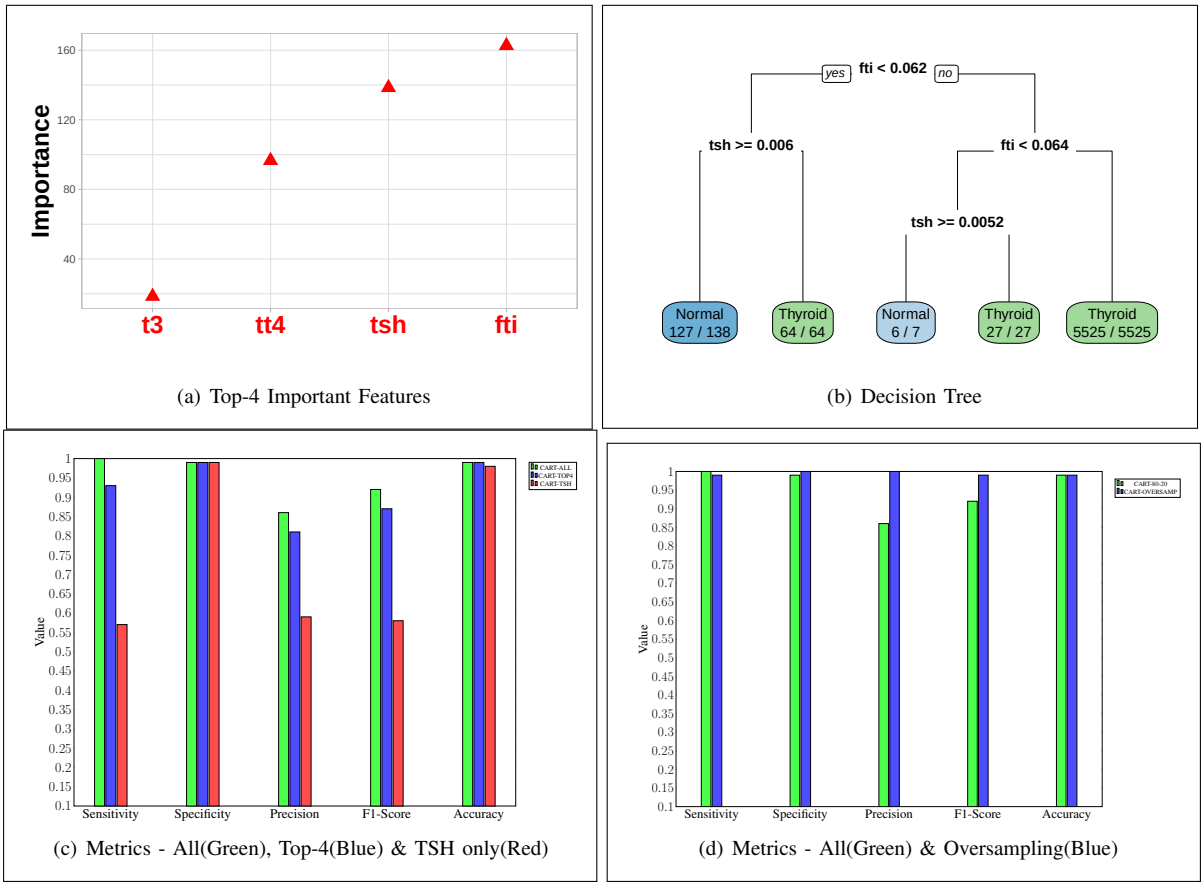
$$Specificity = \frac{TN}{TN + FP}$$

Precision measures the quality of the positive predictions of the model.

$$Precision = \frac{TP}{TP + FP}$$

Accuracy measures the fraction of correct predictions by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$



(a) Top-4 Important Features

(b) Decision Tree

(c) Metrics - All(Green), Top-4(Blue) & TSH only(Red)

(d) Metrics - All(Green) & Oversampling(Blue)

Figure 3. Supervised Learning: CART Results

When we want to find an optimal blend of precision and recall we can combine the two metrics using the *F1 score*. The *F1 score* is the harmonic mean of precision and recall taking both metrics into account as follows:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

IV. Results and Discussion

This section summarizes our results and observations on KEEL dataset obtained using supervised learning (CART and Random Forest), unsupervised learning (PCA) as outlined in Sec. III-D. We refer the reader to Table III for the description of terminology used in this section.

RQ1 : Among the 21 different features, what are the top-4 predictors for clinical thyroid disease diagnosis? Can these be measured cost effectively?

A. Supervised Learning: CART

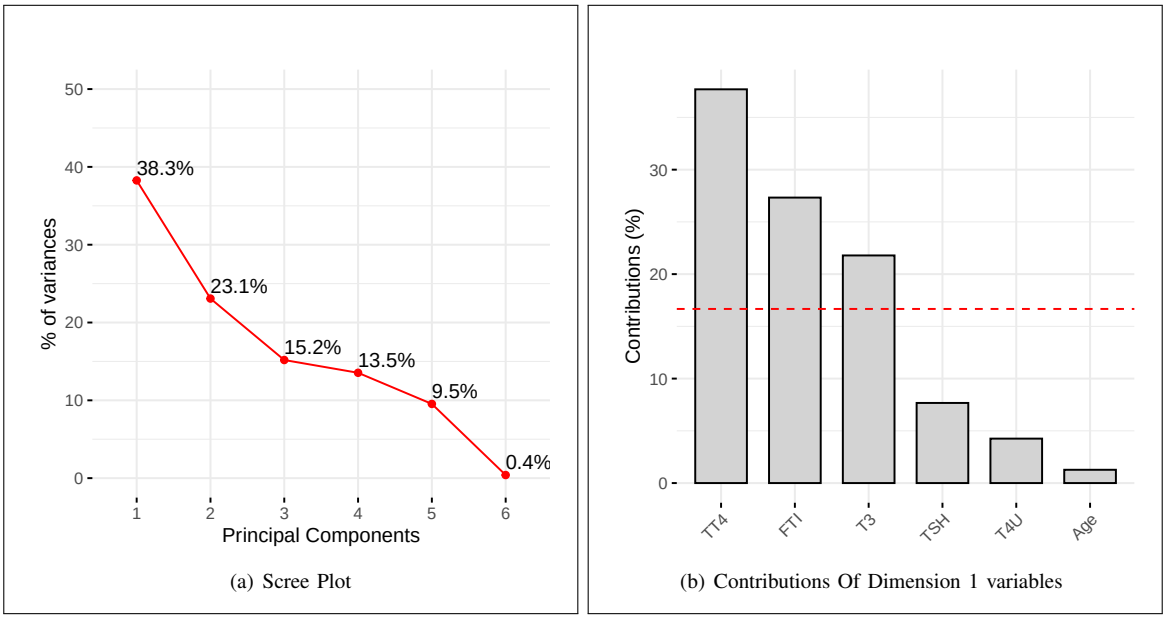
This work uses all the 21 data attributes of KEEL dataset to build a predictive CART model. Initially, our approach is to build a successful predictive model for imbalanced KEEL dataset by performing a simple 80:20 train/test split and evaluating performance metrics. We will later re-evaluate this model using oversampling and a more sophisticated model like Random Forest. Training dataset consisting of 5761 records is used to fit the chosen model (in our work, CART) enabling it learn from this data. Test dataset, with 1439 records, is used to provide an unbiased evaluation of a final model fit. The test dataset provides the gold standard used to evaluate the model and is only used once a model is completely trained.

Visualization of results of the CART generated model in Fig. 3 shows the important features used in the classification decision tree and are summarized below.

1) Fig. 3(a) identifies FTI as the most important predictor for the risk of a thyroid disease based on Gini index closely followed by TSH, TT4, and T3 in that order. These top-4 important features can be measured cost effectively by a simple blood test.
2) Resulting CART decision tree, Fig. 3(b), shows FTI as the root node of the tree followed by TSH as second level node. Note that the printed tree does not show the other two important features, TT4 and T3, due to pruning that the CART algorithm applies on the tree. Nonetheless, these two attributes are indeed deemed important by CART during the tree construction.

B. Unsupervised Learning: PCA

Results obtained using PCA are summarized in Fig. 4. PCA uses the six continuous data attributes, FTI, TT4, T3, T4U, TSH and Age from the KEEL dataset. Therefore, Scree Plot



(a) Scree Plot

(b) Contributions Of Dimension 1 variables

Figure 4. Unsupervised Learning: PCA Results

shown in Fig. 4(a) identifies six principal components (PC) and plots the eigenvalues, ordered from largest to the smallest. Fig. 4(b) displays the list of important data attributes that contribute to principal dimension 1 (PC1) (dimension with the largest eigen value and maximum variance) and their % contribution [32].

PCA results reported in Fig. 4(b) align well with RQ1 results in CART model (Fig. 3) by identifying the very same features, FTI, TSH, TT4, and T3 as the top-4 important features for predicting thyroid disease. The red dotted line in Fig. 4(b) indicates the expected average contribution. The variables above the dotted line contribute more than the average and the ones below contribute lower than the average. This also demonstrates that the class label assumptions we made in Sec. III-B does not intervene with identifying the top-4 important features.

RQ2: How effective is TSH alone in predicting the presence of thyroid disease? That is, does the prediction improve by adding other three important features to the testing palette?

To understand the significance of the top-4 important features identified by CART in Fig. 3(a), we further study the corresponding performance metrics when we consider:

- All 21 attributes
- Only the top-4 important features
- Only TSH

The metrics obtained are summarized in Fig. 3(c) using green, blue and red bars respectively. It shows that there is no noticeable loss of performance when using only top-4 attributes (the blue bars) as opposed to all 21 attributes (the green bars).

However, there is a significant loss in performance metrics like sensitivity, precision and F1 score using TSH only (the red bars) when compared with other two options. For example, the sensitivity falls to 0.57 while it is 0.91 when top-4 important attributes are included. Similar reduction is observed for precision and F1-score as well.

This result is of clinical importance as it shows that measuring TSH alone in the blood test is not reliable and can result in misdiagnosis of thyroid disease. We note that the practice of measuring TSH alone is prevalent in many countries that have universal health care such as Canada [9].

RQ3: How much does the imbalance in the KEEL dataset influence the results for RQ1 and RQ2? More generally, how robust are our answers to RQ1 and RQ2?

We explore two different approaches to answer RQ3.

- **Over Sampling:** To address the imbalance in KEEL dataset, we use oversampling that works with minority class by replicating the observations. Fig 3(d) shows that all the performance metrics improve after use of oversampling. However, for the purpose of this work the crucial observation is that oversampling also reports the top-4 important features as FTI, TSH, TT4 and T3 which are the same as Fig. 3(a) showing that balancing the dataset does not affect our result for RQ1.
- **Random Forest:** We also test the robustness of our conclusions by revisiting RQ1 and RQ2 using a more sophisticated classifier, Random Forest. We observed that the top-4 important features are TSH, FTI, T3, TT4 in that order. That is, the top-4 important features remain the same as reported for CART in Fig. 3(a) though their relative order has changed. In addition, we summarize the performance of Random Forest on the KEEL dataset using the three scenarios described in RQ2 in Fig. 5. We again notice that using the top-4 important features gives almost the same performance as using all the variables. However using only TSH causes a significant

drop in performance for sensitivity, precision, and F1. For example, the sensitivity drops from 0.93 to 0.61 and the F1-score drops from 0.91 to 0.58.

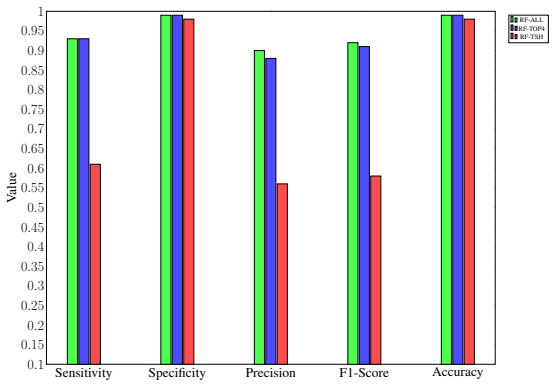


Figure 5. Random Forest Metrics: All (Green), Top-4 (Blue) & TSH only (Red)

V. Conclusions

The main objective in applying machine learning (ML) approaches to health data should be the usability of the results by the end users, the clinicians. With the help of a user study, [33] highlights the properties that clinicians look for in ML approaches such as clinical relevance, explainability and clear specification of *the important features that helped the model make a decision*. Our results in this work are geared towards meeting these requirements and gaining clinicians' trust.

We apply a rule-based ML algorithm CART (considered transparent by clinicians) to KEEL thyroid dataset. We provide clinicians with clear evidence that measuring all the important features (as opposed to a select few) is key to efficiently diagnosing thyroid disease. Specifically, through a careful data analysis in the settings of supervised and unsupervised learning, we show that testing TSH only can result in misdiagnosis but measuring the complete thyroid panel (FTI, TT4, TSH and T3) is highly effective and recommended for medical practitioners. This result is of paramount importance to diagnosis of thyroid disorder from a clinical perspective.

References

[1] B. Brody, "The endocrine system and glands of the human body," https://www.webmd.com/diabetes/endocrine-system-facts/.

[2] J. M. Hershman, "Overview of the thyroid gland," https://www.merckmanuals.com/en-ca/home/hormonal-and-metabolic-disorders/thyroid-gland-disorders/overview-of-the-thyroid-gland.

[3] M. Hye Rim Chung, "Iodine and thyroid function," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4049553/.

[4] B. T. Foundation, "What is thyroid disorder," https://www.btf-thyroid.org/what-is-thyroid-disorder.

[5] J. D. Carmichael, "Overview of the pituitary gland," https://www.merckmanuals.com/en-ca/home/hormonal-and-metabolic-disorders/pituitary-gland-disorders/overview-of-the-pituitary-gland.

[6] T. F. of Canada, "Thyroid foundation of canada," https://thyroid.ca/thyroid-disease/#:~:text=It%20is%20estimated%20that%20200,many%20as%2050%25%20are%20undiagnosed!

[7] A. T. Association, "Thyroid disease facts," https://www.thyroid.org/media-main/press-room/#:~:text=Up.

[8] B. T. Foundation, "Thyroid function tests," https://whttps://www.btf-thyroid.org/thyroid-function-tests.

[9] B. G. Canada, "Thyroid function tests," https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/bc-guidelines/thyroid-testing.

[10] P. K. Sharpe, H. E. Solberg, K. Rootwelt, and M. Yearworth, "Artificial neural networks in diagnosis of thyroid function from in vitro laboratory tests," *Clinical chemistry*, vol. 39, no. 11, pp. 2248–2253, 1993.

[11] G. P. Zhang and V. L. Berardi, "An investigation of neural networks in thyroid function diagnosis," *Health Care Management Science*, vol. 1, no. 1, pp. 29–37, 1998.

[12] K. Hoshi, J. Kawakami, M. Kumagai, S. Kasahara, N. Nishimura, H. Nakamura, and K. Sato, "An analysis of thyroid function diagnosis using bayesian-type and som-type neural networks," *Chemical and pharmaceutical bull.*, vol. 53, no. 12, pp. 1570–1574, 2005.

[13] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Systems with Applications*, vol. 36, no. 1, pp. 944–949, 2009.

[14] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid disease diagnosis based on genetic algorithms using pnn and svm," in *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*. IEEE, 2009, pp. 1–4.

[15] D.-Y. Liu, H.-L. Chen, B. Yang, X.-E. Lv, L.-N. Li, and J. Liu, "Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnostic system for thyroid disease," *J. of medical sys.*, vol. 36, no. 5, pp. 3243–3254, 2012.

[16] H.-L. Chen, B. Yang, G. Wang, J. Liu, Y.-D. Chen, and D.-Y. Liu, "A three-stage expert system based on support vector machines for thyroid disease diagnosis," *J. of medical sys.*, vol. 36 (3), pp. 1953–1963, 2012.

[17] L.-N. Li, J.-H. Ouyang, H.-L. Chen, and D.-Y. Liu, "A computer aided diagnosis system for thyroid disease using extreme learning machine," *Journal of medical systems*, vol. 36, no. 5, pp. 3327–3337, 2012.

[18] Q. Pan, Y. Zhang, M. Zuo, L. Xiang, and D. Chen, "Improved ensemble classification method of thyroid disease based on random forest," in *2016 8th International Conference on Information Technology in Medicine and Education (ITME)*, 2016, pp. 567–571.

[19] E. Turanoglu-Bekar, G. Ulutagay, and S. Kantarcı-Savas, "Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms," *Oxford Journal of Intelligent Decision and Data Sciences*, vol. 2, pp. 13–28, 2016.

[20] S. Razia, P. S. Prathyusha, N. V. Krishna, and N. S. Sumana, "A comparative study of machine learning algorithms on thyroid disease prediction," *International Journal of Engineering and Technology (UAE), ISSN No*, pp. 315–319, 2018.

[21] Y. I. Mir and S. Mittal, "Thyroid disease prediction using hybrid machine learning techniques: An effective framework," *International Journal of Scientific & Technology Research*, vol. 9, no. 02, 2020.

[22] Y. I. Mir, "Improved thyroid disease prediction model using data mining techniques with outlier detection," in *Advanced Machine Learning Approaches in Cancer Prognosis*, 2021, pp. 129–161.

[23] "Keel," https://sci2s.ugr.es/keel/dataset.php?cod=67.

[24] "UCI - ANN Dataset is the source for keel," http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease.

[25] D. Hemalatha and S. Poorani, "Supervised machine learning models for classification of thyroid data," *International Journal of Scientific & Technology research volume 9*, 2020.

[26] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, "A comparative study on thyroid disease detection using k-nearest neighbor and naive bayes classification techniques," *CSI transactions on ICT*, vol. 4, no. 2, pp. 313–319, 2016.

[27] A.H.Shahid, M.P.Singh, R.K.Raj, R.Suman, D.Jawaid, and M.Alam, "A study on label tsh, t3, t4u, tt4, fti in hyperthyroidism and hypothyroidism using machine learning techniques," *Proc. of ICCES*, 2019.

[28] W. Schiffmann, M. Joost, and R. Werner, "Synthesis and performance analysis of multilayer neural network architectures," *University of Koblenz: Institute of Physics*, 1992.

[29] ——, "Optimization of the backpropagation algorithm for training multilayer perceptrons," *Univ. of Koblenz: Institute of Physics*, 1994.

[30] S. N. Yoichi Hayashi and S. Fujisawa, "Use of the recursive-rule extraction algorithm with continuous attributes to improve diagnostic accuracy in thyroid disease," *Meiji University, Japan:Department of Computer Science*, 2016.

[31] I. C. Education, "Supervised learning," https://www.ibm.com/cloud/learn/supervised-learning.

[32] stackexchange, "Using principal component analysis (pca) for feature selection," https://stats.stackexchange.com/questions/27300/using-principal-component-analysis-pca-for-feature-selection/27310#27310.

[33] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*. PMLR, 2019, pp. 359–380.