Thyroid dysfunction diagnosis from routine laboratory tests based on machine learning

Min Hu¹, Ph.D., Chikashi Asami¹#, BS., Hiroshi Iwakura², M.D., Ph.D., Yasuyo Nakajima³,
M.D., Ph.D., Ryousuke Sema¹, MBA., Tsuyoshi Kikuchi¹, MBA., Koji Sakamaki⁵, M.S.,
Takumi Kudo⁴, M.D., Ph.D., Masanobu Yamada³, M.D., Ph.D., Takashi Akamizu⁴, M.D., Ph.D.,
Yasubumi Sakakibara¹, Ph.D.

¹Cosmic Corporation Co., Ltd., Tokyo, Japan
²Wakayama Medical University Hospital, Wakayama, Japan
³Gunma University Hospital, Gunma, Japan
⁴Kuma Hospital, Hyogo, Japan
⁵Hidaka Hospital, Gunma, Japan
⁶Keio University, Tokyo, Japan

#These authors contributed equally to this work.

Running title

AI-based system for thyroid dysfunction diagnosis

Keywords: misdiagnosed thyroid dysfunction; hyperthyroidism; hypothyroidism; machine
learning
Abstract

Approximately 2.4 million patients need treatment for thyroid disease, including Graves’ disease and Hashimoto’s disease, in Japan. However, only 450,000 of them are receiving treatment, and many patients with thyroid dysfunction remain largely overlooked. In this retrospective study, we aimed to screen patients with hyperthyroidism and hypothyroidism who would greatly benefit from prompt medical treatment, and examined routine laboratory finding data and machine learning algorithms to investigate whether such accurate and robust screening is possible to prevent overlooking and misdiagnosing thyroid dysfunction. We succeeded in developing a machine learning method to construct the classification model for detecting hyperthyroidism and hypothyroidism in patients using 11 routine laboratory tests. We collected electronic health record and medical checkup data from four hospitals in Japan. As a result of cross-validation and external evaluation, we achieved a high classification accuracy for the hyperthyroidism and hypothyroidism models.
Introduction

Thyroid dysfunction is a leading endocrine disorder with major health implications, including an increased risk of heart disease and hypercholesterolemia. One of the greatest challenges in thyroid dysfunction treatment is to prevent overlooking and misdiagnosing these diseases. Thyroid hormone excess and deficiency are frequently misunderstood and are too often overlooked and misdiagnosed (1). For hyperthyroidism, the diagnosis may be delayed or missed because some symptoms can be easily attributed to other conditions such as stress (2), and often mistaken for cardiac disease or gastrointestinal malignancies. Hypothyroidism can present with nonspecific constitutional and neuropsychiatric complaints (3), and patients with hypothyroidism are often misdiagnosed as dementia, cardiac disease, liver disease, or hyperlipidemia, and hence not given the proper treatment (4). The American Association of Clinical Endocrinologists has estimated that in the United States, approximately 4.78% of the population has misdiagnosed thyroid dysfunction (5). Another study argues that it can be calculated that approximately 15 million adults have unrecognized thyroid disease (6). In Japan, it is estimated that approximately 2.4 million patients need treatment for thyroid disease (7). However, only approximately 450,000 of them are receiving treatment. Thus, patients with thyroid dysfunction are frequently overlooked and misdiagnosed (6,7).

Hyperthyroidism is the condition that occurs due to excessive production of thyroid hormones. The first step to diagnose hyperthyroidism is to measure free thyroxine (FT4) and free triiodothyronine (FT3) thyroid hormones and thyroid-stimulating hormone (TSH) (6). In contrast, hypothyroidism is a condition in which serum thyroid hormones decrease. Typical diseases of hypothyroidism include Hashimoto's disease and are diagnosed by anti-thyroid antibody tests such as anti-thyroid peroxidase antibody (TPO) and anti-thyroglobulin antibody.
Despite their clinical significance, thyroid function tests and anti-thyroid antibody tests were not included in the Japanese national health checkups.

As popular and effective approaches to predictive analytics, machine learning is highly regarded due to their success in diagnosis, prediction, and choice of treatment. Recently, an emerging technique in the field of medical informatics has employed machine learning to accurately derive insights from medical records to support clinical screening and predict misdiagnosed disease (8). For instance, there is a study that emphasized the superiority of machine learning technology for predicting cardiovascular risk from routine clinical data (9). In another study, the incidence of myocardial infarction or cerebral infarction was predicted using the results of health checkup (10). Numerous studies have also attempted to assess the efficacy of detecting misdiagnosed diseases, including thyroid dysfunction (11-17). Aoki et al. (16,17) found that there were strong, multiple correlations between the set of routine clinical parameters and FT4 in patients with both overt hyperthyroidism and overt hypothyroidism. These studies used pattern recognition methods such as neural networks and predicted the likelihood of thyroid dysfunction from a set of routine clinical tests.

Despite such great efforts, there are still several concerns on the machine learning application to diagnosis of disease. Those includes the issues of data cleansing, missing value completion, dysfunction labeling criterion, integration of multiple hospital datasets, validation and interpretation of machine learning model. In this study, we developed an explainable artificial intelligence diagnosis support system using machine learning algorithms to identify thyroid dysfunction with routine clinical data to improve medical screening and prevent overlooking and misdiagnosing thyroid dysfunction. Our study addresses those concerns on the machine learning application and provides some possible solutions.
We devised two criteria for dysfunction labeling of data: thyroid test criterion and prescription criterion. Thyroid test criterion, which includes the thyroid function tests TSH and FT4, can be used to clearly model the overt and subclinical thyroid dysfunction. However, both TSH and FT4 tests are required so that the number of available data tends to be smaller. More data are available through prescription criterion based on the presence or absence of doctor’s prescriptions, though it can lead to a problem of confounding overt and subclinical thyroid dysfunction with euthyroidism. Second, we integrated data from four hospitals including electronic medical record of Wakayama Medical University hospital, Gunma University hospital, and Kuma hospital, and annual medical checkup data of Hidaka hospital. Among the four hospitals, a machine learning model was trained and evaluated via cross-validation by combining patient data of Wakayama Medical University hospital and Gunma University hospital with medical checkup data of healthy individuals in Hidaka hospital. Furthermore, electronic medical record data of Kuma hospital was used as the external evaluation for the trained models. Third, we examined four typical machine learning algorithms for the structured data: gradient boosting decision tree, support vector machines and neural networks used in related studies, as well as logistic regression, which is a common tool in medical studies. Fourth, in terms of the input feature used in machine learning models, features including AST (aspartate aminotransferase), ALT (alanine aminotransferase), γ-GTP, total cholesterol, hemoglobin, red blood cell count (RBC), creatinine, and sex were selected from the health checkup test list specific in Japan. alkaline phosphatase (ALP), uric acid (UA), and UA to serum creatinine (S-Cr) ratio were further added, and hence totally 11 features were used. To further verify the performance of models depending on the set of input features, we trained and evaluated models in the case
limited to five routine tests including AST, ALT, γ-GTP, total cholesterol, and sex. Finally, all 24 laboratory findings available in this study were also applied and validated.

**Methods**

**Data source**

In the present study, we acquired laboratory finding datasets from different clinical university medical institutions in Japan, including Wakayama Medical University Hospital, Gunma University Hospital, Hidaka Hospital, and Kuma Hospital. The anonymized electronic medical records include age, sex, diagnosis codes for insurance billing, prescribed drugs, and biochemical test results. The institutional ethical review boards of the three institutions at which the study was conducted gave their approval.

A sample of 176,727 subjects in total were included in our study, aged between 13 and 88 and from different regions in Japan between 2004 and 2019, as illustrated in Table 1. Among the four institutions, Wakayama Medical University hospital and Gunma University hospital are hospitals affiliated with a medical college, Hidaka hospital is a regional medical care support hospital, and Kuma hospital is a hospital specialized on thyroid diseases. The data of the 176,727 subjects consisted of doctor evaluations, prescriptions, clinical examinations, and laboratory findings. The doctor evaluations addressed medical history, medication use, and differential diagnosis, among other topics. If a subject was prescribed medication, the name and dose of the prescription were recorded. The examinations involved anthropometric measurements and laboratory tests, among others. The institutional ethical review boards of the three institutions at which the study was conducted gave their approval (Approval Number of Wakayama Medical University Hospital: 2301, Hidaka Hospital: 257, Gunma University Hospital: HS2018-245)
Table 1. Summary of the data from each institution

<table>
<thead>
<tr>
<th>Institution</th>
<th>Wakayama Medical University</th>
<th>Gunma University</th>
<th>Hidaka Hospital</th>
<th>Kuma Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of prescriptions</td>
<td>8,249,286</td>
<td>34,561,268</td>
<td>23,450</td>
<td>61,590</td>
</tr>
<tr>
<td>Number of patients</td>
<td>14,249</td>
<td>27,133</td>
<td>10,482</td>
<td>124,863</td>
</tr>
<tr>
<td>Average age</td>
<td>60.9</td>
<td>51.7</td>
<td>47.7</td>
<td>50.3</td>
</tr>
<tr>
<td>Male/female ratio</td>
<td>1.03 (5,888/5,723)</td>
<td>0.53 (8,143/15,296)</td>
<td>1.82 (15,125/8,325)</td>
<td>0.21</td>
</tr>
</tbody>
</table>

The K-nearest neighbor (KNN) algorithm was used to predict and complement the missing values, with k set to 3 in the data filling process. A previous study (11) has reported KNN to substantially increase the number of applicable subjects. Compared with missing value deletion, it is easily applied, performs well for nonparametric datasets and provides a larger sample size. Furthermore, since the age and sex distributions were different among the institutions, as shown in Table 1, we also conducted random under sampling to fix the gaps in these differences. From this dataset, the model was constructed using the thyroid patient data from Wakayama Medical University and Gunma University and the data of control groups from Hidaka hospital, and was evaluated using cross-validation. To validate on external data, the model was also evaluated on the dataset of Kuma hospital.

Construction of machine learning model
As shown in Table 2, four verification items were devised in this study to improve the performance of our machine learning model. The criteria of data labeling and the combination of multiple institutions were evaluated at first. Then four different machine learning algorithms and three sets of input features were evaluated to achieve the best performance of our thyroid dysfunction classification models.

<table>
<thead>
<tr>
<th>No.</th>
<th>Verification item</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Training data labeling</td>
<td>Thyroid function test criterion</td>
</tr>
<tr>
<td>2</td>
<td>Institution combination (for patient data and control group data)</td>
<td>Institution combination 1 (Inst. comb. 1)</td>
</tr>
<tr>
<td>3</td>
<td>Machine learning algorithm</td>
<td>GBDT</td>
</tr>
<tr>
<td>4</td>
<td>Input features</td>
<td>Feature set 1</td>
</tr>
</tbody>
</table>

**Data labeling criterion**

According to the guidelines of Japanese Society of Laboratory Medicine for the diagnosis of hyperthyroidism and hypothyroidism, if the disorder is suspected from the clinical findings, first the thyroid function test (TSH and FT4 measurement) is conducted, from which the disorder is classified into three categories, hyperthyroidism, hypothyroidism, and euthyroidism (5).

Therefore, we devised and compared the performance of two data labeling criteria.

We firstly devised the labeling criterion by using the result of the thyroid function test as a reference (hereinafter referred to as the “thyroid function test criterion”). Specifically, in the dataset of Wakayama Medical University, FT4 and TSH were measured with the ECLusys kits.
TSH < 0.5 and FT4 > 1.7 was defined as overt hyperthyroidism, TSH < 0.5 and 0.9 ≤ FT4 ≤ 1.7 as subclinical hyperthyroidism, TSH > 5.0 and FT4 < 0.9 as overt hypothyroidism, and TSH > 5.0 and 0.9 ≤ FT4 ≤ 1.7 as subclinical hypothyroidism (TSH unit: μIU/mL; FT4 unit: ng/dL). In the dataset of Gunma University, in which FT4 and TSH were measured with the Architect kit, TSH < 0.35 and FT4 > 1.48 was defined as overt hyperthyroidism, TSH < 0.35 and 0.7 ≤ FT4 ≤ 1.48 as subclinical hyperthyroidism, TSH > 4.94 and FT4 < 0.7 as overt hypothyroidism, and TSH > 4.94 and 0.7 ≤ FT4 ≤ 1.48 as subclinical hypothyroidism. In this study, overt and subclinical hyperthyroid patients are collectively referred to as the hyperthyroidism group, and overt and subclinical hypothyroid patients are collectively referred to as hypothyroidism group.

Data for the control group were extracted from the third institution, Hidaka hospital, which consisted of the test results from regular medical examinations. We extracted comprehensive medical examination data for subjects who did not have any symptoms suggesting thyroid dysfunction or abnormal values in the laboratory tests of thyroid-stimulating hormone (TSH) and serum free T4 (fT4). The normal ranges were set to 0.34-3.88 μIU/mL for TSH and 0.95-1.74 ng/dL for fT4. Random under sampling was conducted for the control group in such a way that the sample size of the control group was equivalent to the size of the hyperthyroidism and hypothyroidism groups. The thyroid function test criterion required both TSH and FT4 test results, but a smaller number of patient records tended to have both of these levels. Therefore, as an alternative solution, we devised another criterion of labeling the training data according to the presence of prescription (hereinafter referred to as the “prescription criterion”) for thyroid disorder. Specifically, the procedure of prescription criterion satisfies the following conditions:

(a) it includes patient records with standard prescribed medications for thyroid dysfunction (including thiamazole, propylthiouracil, and potassium iodide for the hyperthyroidism group, and...
levothyroxin and thyronamine for the hypothyroidism group) obtained on the patient’s first visits, (b) the patient is not diagnosed with thyroid nodules, (c) patient records contain laboratory findings obtained within four weeks after the patient's first prescription, and (d) exclude records with missing values of more than half of our selected features. Since the age distributions were different among the institutions, as shown in Table 1, we also conducted data under sampling to fix the gaps in these differences.

In machine learning, a control group is generally used as negative label. Since hyperthyroidism and hypothyroidism are thyroid dysfunction, both often express similar symptoms and effects on some routine laboratory findings (e.g. Hb is decreased in both hyperthyroidism and hypothyroidism patients). Therefore, we consider the confounding of hyperthyroidism and hyperthyroidism as “crosstalk” and refined the labeling criteria in such a way that the negative label is set as both the healthy subjects of the control group and the patients of the opposite type of thyroid dysfunction. For instance, in the data labeling process of the hyperthyroidism classification model, hyperthyroidism group was set as positive label whereas both healthy subjects of the control group and hypothyroidism patients were set as negative label.

**Integrating multiple hospital datasets**

The demographics were different among the three institutions from different districts. To investigate the effect of integrating three hospital datasets, we explored three combinations of the datasets to increase the generalization ability of our models. Specifically, three options on datasets, namely, thyroid dysfunction group data from both Wakayama Medical University and Gunma University and control group data from Hidaka hospital (referred to as Inst. comb. 1),
thyroid dysfunction group data from Wakayama Medical University and control group data from Hidaka hospital (referred to as Inst. comb. 2), and thyroid dysfunction group data from Gunma University and control group data from Hidaka hospital (referred to as Inst. comb. 3), were set to train and evaluate the models.

**Machine learning algorithms**

Four representative machine learning algorithms were applied and evaluated of the performance on thyroid dysfunction classification:

- Gradient boosting decision tree (GBDT), as proposed by Friedman (18), produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It is based on a machine learning technique that consists of an “ensemble” family of algorithms, creates multiple models (called weak learners), and combines them to increase the prediction accuracy. The main idea of this technique is to build a set of decision trees and use them to classify a new case. Each decision tree is generated using randomly selected variable subsets from all feature variables and a randomly selected subset of data combined by bootstrapping (19). In this study, we employed the most accurate algorithm, called CATBoost (20), in the GBDT family.

- The artificial neural network (ANN) is a well-established classification technique that is widely used in pattern recognition studies. In general, an ANN consists of 3 layers: an input layer that receives information, a hidden layer that processes information, and an output layer that calculates the results (21). In the present study, a standard feed-forward ANN was applied due to its relative simplicity and stability.
Support vector machine (SVM) is a supervised machine learning technique that is widely used in pattern recognition and classification problems (22). In the approach of this method, each data sample is a vector whose dimensions are equal to the number of features to be considered, and the SVM creates a hyperplane that separates samples into two categories. The induced hyperplane is constructed to maximize its distance from the samples of both classes. This algorithm achieves high classification performance by using special nonlinear functions called kernels to transform the input space into a multidimensional space (22). In this study, the radial basis function kernel is used.

Logistic regression is a statistical classifier that provides the probability for predicting the labeled class of categorical type by using a number of attributes. Logistic regression is frequently used to examine the risk relationship between disease and exposure, with the ability to test for statistical interaction and control for multi-variable confiding (23). It is a linear model and used as the baseline model for the performance comparison.

**Explanatory features (variables) for machine learning**

Features from a subject’s record were designed to sufficiently explain factors that were related to thyroid dysfunction. We used 11 variables as explanatory variables in this study as the first experimented set of features (referred to as Feature set 1) in this study, of which eight tests are tests measured in routine health checkup: sex, AST, ALT, γ-GTP, total cholesterol, Hemoglobin (Hb), RBC, and creatinine (S-Cr). In addition, since ALP, UA, and S-Cr ratio are reported to be highly relevant to thyroid dysfunction (24, 25), these were added to the above items. We also included UA/S-Cr ratio in this study considering that the reduction of S-Cr has been reported in hyperthyroidism, while UA has not been confirmed to fluctuate with thyroid dysfunction. To discriminate hyperthyroidism with renal dysfunction, which usually leads to the rise of both S-Cr
and UA, we introduced UA/S-Cr ratio as one of the features to improve the classification performance. 11 tests (Feature set 1) in total were used as features to train machine learning models in this study.

With an aim to quantify the necessity of each of the 11 tests mentioned above, the performance of five items (referred to as Feature set 2) out of the 11 tests was checked. Feature set 2 excluded three items, Hb, S-Cr, and RBC, which are the tests measured only at the doctor’s discretion.

Model validation

Cross-validation was applied to evaluate the performance of our machine learning method in classifying patients. The evaluation was conducted by extracting 9/10 training data and 1/10 test data by conducting 10-fold cross-validation. This was repeated 10 times to extract the training and test data uniformly, and the average and standard deviation of each evaluation score of each time were calculated. During the model training and test process, we avoided including the same subject to both training dataset and test dataset. The following measures were used for the performance evaluation criteria: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), sensitivity is defined by TP/(TP+FN), and specificity is defined by TN/(TN+FP), where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives. Note that the cutoff value for classifying as positive or negative is determined by Youden index (26). Finally, the AUROC performance difference between models was verified as statistically significant by the Wilcoxon signed-rank test.
In addition, the data of Kuma hospital were employed as an external validation. The model was constructed using the hyperthyroidism group and the hypothyroidism group of Wakayama Medical University and Gunma University and the control group of Hidaka hospital as the training data. The model was evaluated using the hyperthyroidism group and hypothyroidism group of Kuma hospital and the control group of Hidaka hospital (referred to as External).

Classification of subclinical thyroid dysfunction

In the guideline of Japan Thyroid Association (27), subclinical hypothyroidism is defined as when FT4 is within the normal limit but the TSH measured is higher than normal. On the other hand subclinical hyperthyroidism is defined as when FT4 is within normal limit and TSH is lower than normal. Compared to the overt thyroid dysfunction where both TSH and FT4 are out of the standard ranges, it is difficult to classify subclinical thyroid dysfunction. This study evaluated the classification performance of the machine learning model by using subclinical standards in the thyroid function test criterion labeling method. We further extended the feature set in the attempt of improving model performance and selected 24 tests (referred to as Feature set 3), which was the all the laboratory tests available in this study.

Feature importance

Feature set 3 includes sex, AST, ALT, γ-GTP, Total cholesterol, RBC, hemoglobin, uric acid, S-Cr, uric acid/S-Cr ratio, ALP, albumin-globulin ratio, albumin, blood urea nitrogen, C-reactive protein, hematocrit, lactate dehydrogenase, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, platelet count, total bilirubin, total protein, white blood count.
To further understand how each feature contributes to the classification of patients in our model, we introduced feature importance. Feature importance represents the factor by which the model error is increased compared to the original model error. In the decision tree–based machine learning algorithms, including GBDT, impurities and the features at which the node is split are recorded for all the nodes when the decision tree learning is finished, and the decision tree calculates the features importance using this information (19).

Results

Model validation

Table 3 is a summary of the performance results of the machine learning model constructed in this study. As the result of 10-fold cross-validation, as shown in No. I of Table 3, the best classification model for overt hyperthyroidism achieved an accuracy of AUROC = 92.4%, sensitivity = 83.3%, and specificity = 90.9%. The best classification model for overt hypothyroidism achieved an accuracy of AUROC = 90.5%, sensitivity = 84.4%, and specificity = 86.4%. In the external evaluation, as shown in No. IX of Table 3, the classification model for overt hyperthyroidism achieved an accuracy of AUROC = 96.3%, and the classification model for overt hypothyroidism achieved an accuracy of AUROC = 92.9%. As shown in No. XI of Table 3, the classification model for subclinical hyperthyroidism achieved an accuracy of AUROC = 73.8%, and the classification model for subclinical hypothyroidism achieved an accuracy of AUROC = 75.2%.

The result of comparing different labeling criteria is shown in No. I and II of Table 3. When the prescription criterion was applied as the labeling criterion, the accuracy of the hyperthyroidism classification model achieved AUROC = 88.2%, and that of the hypothyroidism classification model achieved AUROC = 82.4%. On the other hand, as shown in No. I, when the
thyroid function test criterion was used, the accuracy of the hyperthyroidism classification model achieved AUROC = 92.4%, and that of the hypothyroidism classification model achieved AUROC = 90.5%. The model trained on the thyroid function test criterion data achieved a superior performance, which was statistically significant by the Wilcoxon test at p-value 0.05.

The result of comparing models built on different institution combinations is shown in No. I, III, and IV of Table 3, as the highest performance was obtained when institution combination 1 was used as training set, and the accuracy of the hyperthyroidism classification model achieved AUROC = 92.4%, while that of the hypothyroidism classification model achieved AUROC = 90.5%.

Among the four machine learning algorithms used in this study, including GBDT, SVM, logistic regression, and ANN, the highest performance was obtained when the GBDT method was applied as shown in No. I, V, VI, and VII of Table 3. The accuracy of the hyperthyroidism classification model achieved AUROC = 92.4%, while that of the hypothyroidism classification model achieved AUROC = 90.5%, which were statistically significant at p-value 0.05 by the Wilcoxon test. After comparing the performance of different feature sets, as shown in I and VIII of Table 3, when the feature set 3 was applied, the accuracy of the hyperthyroidism classification model was reduced to AUROC = 87.4%, and the performance of the hypothyroidism classification model was reduced to AUROC = 85.5%, which shows significant differences by the Wilcoxon test at p-value 0.05.
The mean and standard deviation for the 10 folds are shown in each score:

<table>
<thead>
<tr>
<th>Institution combination</th>
<th>Inst. comb. 1</th>
<th>Inst. comb. 2</th>
<th>Inst. comb. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overt/Subclinical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUROC</td>
<td>88.5±7.4%</td>
<td>84.5±6.9%</td>
<td>83.2±6.4%</td>
</tr>
<tr>
<td>AUPRC</td>
<td>77.4±5.2%</td>
<td>73.4±4.9%</td>
<td>76.8±4.9%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>83.3±9.0%</td>
<td>77.5±8.5%</td>
<td>81.9±9.2%</td>
</tr>
<tr>
<td>Specificity</td>
<td>90.9±6.8%</td>
<td>86.4±6.2%</td>
<td>89.8±5.0%</td>
</tr>
<tr>
<td>AUROC</td>
<td>79.0±3.5%</td>
<td>85.0±4.6%</td>
<td>94.9±5.7%</td>
</tr>
<tr>
<td>AUPRC</td>
<td>77.4±4.9%</td>
<td>84.5±5.2%</td>
<td>83.2±6.4%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>83.3±9.0%</td>
<td>80.8±10.1%</td>
<td>70.8±11.9%</td>
</tr>
<tr>
<td>Specificity</td>
<td>90.9±6.8%</td>
<td>86.4±6.2%</td>
<td>89.8±5.0%</td>
</tr>
<tr>
<td>AUROC</td>
<td>79.0±3.5%</td>
<td>85.0±4.6%</td>
<td>94.9±5.7%</td>
</tr>
<tr>
<td>AUPRC</td>
<td>77.4±4.9%</td>
<td>84.5±5.2%</td>
<td>83.2±6.4%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>83.3±9.0%</td>
<td>80.8±10.1%</td>
<td>70.8±11.9%</td>
</tr>
<tr>
<td>Specificity</td>
<td>90.9±6.8%</td>
<td>86.4±6.2%</td>
<td>89.8±5.0%</td>
</tr>
<tr>
<td>AUROC</td>
<td>79.0±3.5%</td>
<td>85.0±4.6%</td>
<td>94.9±5.7%</td>
</tr>
</tbody>
</table>

Table 3: Results of validation on different models.
The model with the best performance was evaluated using the external dataset for Kuma Hospital, as shown in No. IX of Table 3. High classification performance was achieved using the external data: AUROC = 96.3%, sensitivity = 87.7%, and specificity = 93.5% for the hyperthyroidism classification model and AUROC = 92.9%, sensitivity = 75.7%, and specificity = 87.1% for the hypothyroidism classification model. No. X and XI of Table 3 show that using feature set 3 improved the classification performance: for subclinical thyroid dysfunction, AUROC = 73.8%, sensitivity = 78.7%, and specificity = 61.6%; for hypothyroidism, AUROC = 75.2%, sensitivity = 59.9%, and specificity = 77.7%. In particular, the significance of the hypothyroidism classification models was statistically confirmed by the Wilcoxon test at p-value 0.05.

**Feature Importance**

The features importance of each model was examined using the feature set 1. The left picture of Figure 1 shows the features importance of the overt hyperthyroidism classification model and the overt hypothyroidism classification model. The three most important features in the overt hyperthyroidism model were ALP, UA/S-Cr ratio, and total cholesterol. The three most important features in the overt hypothyroidism model were AST, total cholesterol, and RBC. On the other hand, the right picture of Figure 1 shows the features importance of the subclinical hyperthyroidism classification model and the subclinical hypothyroidism classification model. The three most important features in the subclinical hyperthyroidism model were ALP, UA/S-Cr ratio, and S-Cr, and the three most important features in the subclinical hypothyroidism model were total cholesterol, AST, and RBC. For both overt and subclinical disease, ALP and S-Cr
were the top related features in the hyperthyroidism classification model, and total cholesterol, AST, and RBC were the top features in the hypothyroidism classification model.

Furthermore, the features importance in the subclinical hyperthyroidism and subclinical hypothyroidism classification models using the feature set 3 was conducted. As shown in Figure 2, ALP and the UA/S-Cr ratio were among the three most important features in the subclinical hyperthyroidism classification model when the feature set 1 was used, as well as when the feature set 3 was used. If the five most important features were considered, MCV and MCH, two features added to the feature set 3, were included. These findings suggest that these two features are also likely to be effective in hyperthyroidism classification. On the other hand, as shown on right side of Figure 2, a difference was seen in the subclinical hypothyroidism classification model when the feature set 1 was used vs. when the feature set 3 was used. The three most important features in the model that used the feature set 1 were total cholesterol, AST, and RBC, whereas the three most important features in the model that used the feature set 3 were total protein, total cholesterol, AST, and the UA/S-Cr ratio. These findings suggest that total protein is likely to be effective in classifying subclinical hypothyroidism.
Figure 1. Comparison of feature importance between overt and subclinical thyroid dysfunction classification models

Figure 2. Comparison of feature importance between models built on the Feature set 1 and Feature set 3
Discussion

Feature importance

The correlation of routine laboratory tests such as ALP, S-Cr, UA, and RBC, etc. with thyroid dysfunction has been pointed out in many previous studies. According to studies on the relationship between thyroid dysfunction and liver function (28, 29), a correlation was confirmed between the increase in ALP and hyperthyroidism, as the ALP value was significantly higher when bone metabolism increases in Graves’s disease, which is a typical disorder of hyperthyroidism. Sönmez (30) examined data from 433 patients and reported that S-Cr in the hyperthyroidism group was significantly lower than in the euthyroid group. TSH and S-Cr were also reported to have a significantly negative correlation with overt hypothyroidism (31).

Dorgalaleh (32) suggested that thyroid dysfunction directly affects most of the blood values, including RBC, and health professionals must pay attention to such effects. The correlation between hypothyroidism and hyperuricemia has also been confirmed by in multiple studies (33, 34).

Comparison with related studies

Several previous studies revealed promising results from the use of machine learning approaches for predicting thyroid dysfunction (16, 17).

Similar to the present study, Aoki’s (17) study used pattern recognition methods such as neural networks to predict the likelihood of thyroid dysfunction from a set of routine test parameters such as ALP, S-Cr, and TC. Their results suggested that most patients with overt thyroid dysfunction could be screened by using a set of routine clinical data without measuring thyroid hormone levels. The correct rate of 91.3% was reported in the hyperthyroidism classification model, and the correct rate of 90.0% was reported in the hypothyroidism
classification model. Their results suggested that there is a high correlation between a set of routine laboratory tests and thyroid dysfunction. However, the model verification of these studies used the leave-one-out method instead of cross-validation and used the correct rate as the indicator instead of AUROC. Thus, the model evaluation was considered insufficient.

Unlike the present study, one drawback of these previous studies is that those have not considered crosstalk in the data labeling process. For hyperthyroidism classification in this study, the hyperthyroidism group was used as a positive label, and both the control and hypothyroidism groups were negatively labeled. For the hypothyroidism classification in this study, the hypothyroidism group was used as a positive label, whereas both the control and hyperthyroidism groups were negatively labeled (referred to as “crosstalk on”). On the other hand, related studies (16, 17) performed classification by setting thyroid dysfunction patients (with hyperthyroidism or hypothyroidism) as positive label and only control group as negative label (referred to as “crosstalk off”). Therefore, we evaluated the performance of the models with similar settings as these studies. As shown in A-1 column of Table 4, when only control group was labeled negative in both the training data and validation data, a high classification performance of AUROC = 94.9% and AUROC = 91.3% was achieved in the classification of overt hyperthyroidism and overt hypothyroidism, respectively. However, as shown in A-2 column of Table 4, when both control group and hypothyroidism group were labeled negative in the validation data of overt hyperthyroidism and when both control group and hyperthyroidism group were labeled negative in the validation data of overt hypothyroidism, the classification performance was reduced to AUROC = 78.5% and AUROC = 68.1%, respectively. The classification performance dropped significantly in the models in which crosstalk was not considered during the negative labeling process.
Table 4. Evaluation result obtained without considering crosstalk

<table>
<thead>
<tr>
<th>No.</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thyroid function test criterion</td>
<td>Overt + subclinical</td>
</tr>
<tr>
<td></td>
<td>Negative label setting</td>
<td>Crosstalk off</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hyperthyroidism</th>
<th>Hypothyroidism</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUROC</td>
<td>94.9±2.4%</td>
<td>91.5±4.0%</td>
</tr>
<tr>
<td></td>
<td>78.5±3.1%</td>
<td>68.1±3.1%</td>
</tr>
</tbody>
</table>

The mean and standard deviation for the 10 folds are shown in AUROC scores.

Limitations

In the current study, subjects under medication may be included in the data extraction process of this study. Though we extracted only the laboratory tests at each subject’s first visit to avoid including the influence of thyroid dysfunction treatment, some subjects might be already on medication before being referred to the hospitals in our study. These subjects on medications may have an unexpected impact on the models we built in this study.

Another limitation of this study is that the hypothyroidism classification models exhibited lower performance than the hyperthyroidism classification models. This result is attributed to differences in the respective serum hormones and underlying molecular mechanisms (35). The various nonspecific symptoms of hypothyroidism may not manifest simultaneously, resulting its subclinical rate larger than that of hyperthyroidism. In addition, patients with hypothyroidism such as Hashimoto’s thyroiditis are dependent upon long-term levothyroxine treatment, which may affect the manifestation of routine laboratory findings.
Furthermore, in the external evaluation of this study, the subclinical classification model showed lower overall results than the overt classification models. Among subclinical thyroid dysfunctions, the cause of subclinical hypothyroidism is associated with chronic thyroiditis (Hashimoto's disease), of which approximately 60-80% of cases are related thyroid autoantibodies (36). On the other hand, the causes of subclinical thyrotoxicosis are classified into extrinsic overdose of thyroid hormone drugs, and endogenous hyperthyroidism such as Graves' disease (37). Most of the subclinical thyroid dysfunctions such as subclinical thyrotoxicosis and subclinical hypothyroidism have no subjective symptoms and are usually considered to be transient (38, 39). Performance may have been limited due to the fact that symptoms of subclinical thyroid dysfunction are usually minor compared to overt thyroid dysfunction, and the phenotype of subclinical thyroid dysfunction may not be reflected in the results of routine laboratory examination.

**Conclusion**

This study evaluated the screening method to discriminate hyperthyroidism and hypothyroidism from the electronic medical records or routine laboratory finding data from health checkups using a machine learning method with an aim to prevent missed diagnosis of thyroid dysfunction. This is a versatile new screening method that was successfully developed from a machine learning model construction method to discriminate patients with hyperthyroidism and hypothyroidism using 11 features. High accuracy was achieved in the discrimination of evident hyperthyroidism or hypothyroidism, although the discrimination accuracy of subclinical hyperthyroidism or hypothyroidism was not satisfactory, these alerts can be useful for non-specialists for thyroid diseases.
It is expected that the quality of life of patients will improve by applying the model developed in this study. If thyroid dysfunction is screened using our method in healthcare facilities, including hospitals and health checkup facilities, prompt and accurate diagnostic support can be provided from only routine laboratory tests.

References


(36) Cooper DS, Biondi B : Subclinical thyroid disease. Lancet 2012 ; 379 : 1142-1154

(37) Biondi B, Cooper DS : The clinical significance of subclinical thyroid dysfunction. Endocr Rev 2008 ; 29 :76-131
(38) Fatourechi V. Subclinical hypothyroidism: an update for primary care physicians. Mayo

(39) Biondi B, et al. Endogenous subclinical hyperthyroidism affects quality of life and
cardiac morphology and function in young and middle-aged patients. The Journal of
Clinical Endocrinology & Metabolism, 2000, 85.12: 4701-4705.

List of Abbreviations

TSH: thyroid-stimulating hormone
FT3: free triiodothyronine
FT4: free triiodothyronine
GBDT: gradient boosting decision tree
ANN: artificial neural network
SVM: support vector machine
AUROC: area under the receiver operating characteristic curve
AUPRC: area under the precision-recall curve
ALP: alkaline phosphatase
UA: uric acid
S-Cr: serum creatinine
AST: glutamic aspartate transaminase
ALT: alanine aminotransferase
RBC: red blood cell count

Declarations
Data Availability

The data that support the findings of this study are available from Japan Thyroid Association but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Japan Thyroid Association.

Correspondence address

Min Hu
Koishikawa 2-7-3 Tomisaka Building, Bunkyo-ku, Tokyo, Japan
E-mail: m.hu@cosmic-jpn.co.jp

Yasubumi Sakakibara
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan
E-mail: yasu@bio.keio.ac.jp

Competing Interests

YS is a paid scientific advisory board of Cosmic corporation co., ltd. The other authors declare no competing financial interests.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions
MH, CA: implemented the software, analyzed the data, and co-wrote the paper. MY, TA: supervised the research, and collected the medical data. HI, YN, KS, TK: aided in the feature sets designing and interpreting the results of the models, as well as collected the medical data. RS, TM: contributed to the design of the research, and to the writing of the manuscript. YS: designed and supervised the research, analyzed the data, and co-wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This study would not have been possible without the exceptional support of Dr. Akira Miyauchi, who shared insightful comments on this project and provided the opportunity to validate the models on external dataset and improved this study in innumerable ways. Dr. Masako Akuzawa and Dr. Yoshitaka Ando facilitated this project to accessing the dataset of control group in Hidaka Hospital, which significantly improved the generalization performance of the thyroid dysfunction classification models built in this project.