

1 **Thyroid dysfunction diagnosis from routine laboratory tests based on machine learning**

2 Min Hu<sup>1#</sup>, Ph.D., Chikashi Asami<sup>1#</sup>, BS., Hiroshi Iwakura<sup>2</sup>, M.D., Ph.D., Yasuyo Nakajima<sup>3</sup>,  
3 M.D., Ph.D., Ryouyuke Sema<sup>1</sup>, MBA., Tsuyoshi Kikuchi<sup>1</sup>, MBA., Koji Sakamaki<sup>5</sup>, M.S.,  
4 Takumi Kudo<sup>4</sup>, M.D., Ph.D., Masanobu Yamada<sup>3</sup>, M.D., Ph.D., Takashi Akamizu<sup>4</sup>, M.D., Ph.D.,  
5 Yasubumi Sakakibara<sup>1,6</sup>, Ph.D.

6

7 <sup>1</sup>Cosmic Corporation Co., Ltd., Tokyo, Japan

8 <sup>2</sup>Wakayama Medical University Hospital, Wakayama, Japan

9 <sup>3</sup>Gunma University Hospital, Gunma, Japan

10 <sup>4</sup>Kuma Hospital, Hyogo, Japan

11 <sup>5</sup>Hidaka Hospital, Gunma, Japan

12 <sup>6</sup>Keio University, Tokyo, Japan

13 <sup>#</sup>These authors contributed equally to this work.

14

15 **Running title**

16 AI-based system for thyroid dysfunction diagnosis

17 **Keywords:** misdiagnosed thyroid dysfunction; hyperthyroidism; hypothyroidism; machine  
18 learning

19

20 **Abstract**

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

33

## 34 **Introduction**

35 Thyroid dysfunction is a leading endocrine disorder with major health implications, including an  
36 increased risk of heart disease and hypercholesterolemia. One of the greatest challenges in  
37 thyroid dysfunction treatment is to prevent overlooking and misdiagnosing these diseases.  
38 Thyroid hormone excess and deficiency are frequently misunderstood and are too often  
39 overlooked and misdiagnosed (1). For hyperthyroidism, the diagnosis may be delayed or missed  
40 because some symptoms can be easily attributed to other conditions such as stress (2), and often  
41 mistaken for cardiac disease or gastrointestinal malignancies. Hypothyroidism can present with  
42 nonspecific constitutional and neuropsychiatric complaints (3), and patients with hypothyroidism  
43 are often misdiagnosed as dementia, cardiac disease, liver disease, or hyperlipidemia, and hence  
44 not given the proper treatment (4). The American Association of Clinical Endocrinologists has  
45 estimated that in the United States, approximately 4.78% of the population has misdiagnosed  
46 thyroid dysfunction (5). Another study argues that it can be calculated that approximately 15  
47 million adults have unrecognized thyroid disease (6). In Japan, it is estimated that approximately  
48 2.4 million patients need treatment for thyroid disease (7). However, only approximately 450,000  
49 of them are receiving treatment. Thus, patients with thyroid dysfunction are frequently  
50 overlooked and misdiagnosed (6,7).

51       Hyperthyroidism is the condition that occurs due to excessive production of thyroid  
52 hormones. The first step to diagnose hyperthyroidism is to measure free thyroxine (FT4) and free  
53 triiodothyronine (FT3) thyroid hormones and thyroid-stimulating hormone (TSH) (6). In  
54 contrast, hypothyroidism is a condition in which serum thyroid hormones decrease. Typical  
55 diseases of hypothyroidism include Hashimoto's disease and are diagnosed by anti-thyroid  
56 antibody tests such as anti-thyroid peroxidase antibody (TPO) and anti-thyroglobulin antibody

57 (TgAb) (5). Despite their clinical significance, thyroid function tests and anti-thyroid antibody  
58 tests were not included in the Japanese national health checkups.

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

72 Despite such great efforts, there are still several concerns on the machine learning  
73 application to diagnosis of disease. Those includes the issues of data cleansing, missing value  
74 completion, dysfunction labeling criterion , integration of multiple hospital datasets, validation  
75 and interpretation of machine learning model. In this study, we developed an explainable  
76 artificial intelligence diagnosis support system using machine learning algorithms to identify  
77 thyroid dysfunction with routine clinical data to improve medical screening and prevent  
78 overlooking and misdiagnosing thyroid dysfunction. Our study addresses those concerns on the  
79 machine learning application and provides some possible solutions.

80 We devised two criteria for dysfunction labeling of data: thyroid test criterion and  
81 prescription criterion. Thyroid test criterion, which includes the thyroid function tests TSH and  
82 FT4, can be used to clearly model the overt and subclinical thyroid dysfunction. However, both  
83 TSH and FT4 tests are required so that the number of available data tends to be smaller. More  
84 data are available through prescription criterion based on the presence or absence of doctor's  
85 prescriptions, though it can lead to a problem of confounding overt and subclinical thyroid  
86 dysfunction with euthyroidism. Second, we integrated data from four hospitals including  
87 electronic medical record of Wakayama Medical University hospital, Gunma University hospital,  
88 and Kuma hospital, and annual medical checkup data of Hidaka hospital. Among the four  
89 hospitals, a machine learning model was trained and evaluated via cross-validation by combining  
90 patient data of Wakayama Medical University hospital and Gunma University hospital with  
91 medical checkup data of healthy individuals in Hidaka hospital. Furthermore, electronic medical  
92 record data of Kuma hospital was used as the external evaluation for the trained models. Third,  
93 we examined four typical machine learning algorithms for the structured data: gradient boosting  
94 decision tree, support vector machines and neural networks used in related studies, as well as  
95 logistic regression, which is a common tool in medical studies. Fourth, in terms of the input  
96 feature used in machine learning models, features including AST (aspartate aminotransferase),  
97 ALT (alanine aminotransferase),  $\gamma$ -GTP, total cholesterol, hemoglobin, red blood cell count  
98 (RBC), creatinine, and sex were selected from the health checkup test list specific in Japan.  
99 alkaline phosphatase (ALP), uric acid (UA), and UA to serum creatinine (S-Cr) ratio were  
100 further added, and hence totally 11 features were used. To further verify the performance of  
101 models depending on the set of input features, we trained and evaluated models in the case

102 limited to five routine tests including AST, ALT,  $\gamma$ -GTP, total cholesterol, and sex. Finally, all  
103 24 laboratory findings available in this study were also applied and validated.

104

## 105 **Methods**

### 106 *Data source*

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

113 A sample of 176,727 subjects in total were included in our study, aged between 13 and 88  
114 and from different regions in Japan between 2004 and 2019, as illustrated in Table 1. Among the  
115 four institutions, Wakayama Medical University hospital and Gunma University hospital are  
116 hospitals affiliated with a medical college, Hidaka hospital is a regional medical care support  
117 hospital, and Kuma hospital is a hospital specialized on thyroid diseases. The data of the  
118 176,727 subjects consisted of doctor evaluations, prescriptions, clinical examinations, and  
119 laboratory findings. The doctor evaluations addressed medical history, medication use, and  
120 differential diagnosis, among other topics. If a subject was prescribed medication, the name and  
121 dose of the prescription were recorded. The examinations involved anthropometric  
122 measurements and laboratory tests, among others. The institutional ethical review boards of the  
123 three institutions at which the study was conducted gave their approval (Approval Number of  
124 Wakayama Medical University Hospital: 2301, Hidaka Hospital: 257, Gunma University  
125 Hospital : HS2018-245)

126

127

128

129

Table 1. Summary of the data from each institution

Institution	Wakayama Medical University	Gunma University	Hidaka Hospital	Kuma Hospital
Number of prescriptions	8,249,286	34,561,268	23,450	61,590
Number of patients	14,249	27,133	10,482	124,863
Average age	60.9	51.7	47.7	50.3
Male/female ratio	1.03 (5,888/5,723)	0.53 (8,143/15,296)	1.82 (15,125/8,325)	0.21
Data period	2010-2018	2004-2019	2004-2007	2007-2020

130

131

132

133

134

135

136

137

138

139

140

141

142

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***

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

148 Table 2. List of verification items

No.	Verification item	Option			
1	Training data labeling	Thyroid function test criterion	Prescription criterion		
2	Institution combination (for patient data and control group data)	Institution combination 1 (Inst. comb. 1)	Institution combination 2 (Inst. comb. 2)	Institution combination 3 (Inst. comb. 3)	External
3	Machine learning algorithm	GBDT	SVM	Logistic regression	ANN
4	Input features	Feature set 1	Feature set 2	Feature set 3	

149  
150

### 151 ***Data labeling criterion***

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

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

157 We firstly devised the labeling criterion by using the result of the thyroid function test as  
158 a reference (hereinafter referred to as the “thyroid function test criterion”). Specifically, in the  
159 dataset of Wakayama Medical University, FT4 and TSH were measured with the ECLusys kits.

160 TSH  $< 0.5$  and FT4  $> 1.7$  was defined as overt hyperthyroidism, TSH  $< 0.5$  and  $0.9 \leq \text{FT4} \leq 1.7$   
161 as subclinical hyperthyroidism, TSH  $> 5.0$  and FT4  $< 0.9$  as overt hypothyroidism, and TSH  $> 5.0$   
162 and  $0.9 \leq \text{FT4} \leq 1.7$  as subclinical hypothyroidism (TSH unit:  $\mu\text{IU/mL}$ ; FT4 unit:  $\text{ng/dL}$ ). In the  
163 dataset of Gunma University, in which FT4 and TSH were measured with the Architect kit, TSH  
164  $< 0.35$  and FT4  $> 1.48$  was defined as overt hyperthyroidism, TSH  $< 0.35$  and  $0.7 \leq \text{FT4} \leq 1.48$   
165 as subclinical hyperthyroidism, TSH  $> 4.94$  and FT4  $< 0.7$  as overt hypothyroidism, and TSH  $>$   
166  $4.94$  and  $0.7 \leq \text{FT4} \leq 1.48$  as subclinical hypothyroidism. In this study, overt and subclinical  
167 hyperthyroid patients are collectively referred to as the hyperthyroidism group, and overt and  
168 subclinical hypothyroid patients are collectively referred to as hypothyroidism group.

169 Data for the control group were extracted from the third institution, Hidaka hospital,  
170 which consisted of the test results from regular medical examinations. We extracted  
171 comprehensive medical examination data for subject who did not have any symptoms suggesting  
172 thyroid dysfunction or abnormal values in the laboratory tests of thyroid-stimulating hormone  
173 (TSH) and serum free T4 (fT4) The normal ranges were set to  $0.34\text{-}3.88 \mu\text{IU/mL}$  for TSH and  
174  $0.95\text{-}1.74 \text{ng/dL}$  for fT4. Random under sampling was conducted for the control group in such a  
175 way that the sample size of the control group was equivalent to the size of the hyperthyroidism  
176 and hypothyroidism groups. The thyroid function test criterion required both TSH and FT4 test  
177 results, but a smaller number of patient records tended to have both of these levels. Therefore, as  
178 an alternative solution, we devised another criterion of labeling the training data according to the  
179 presence of prescription (hereinafter referred to as the “prescription criterion”) for thyroid  
180 disorder. Specifically, the procedure of prescription criterion satisfies the following conditions:  
181 (a) it includes patient records with standard prescribed medications for thyroid dysfunction  
182 (including thiamazole, propylthiouracil, and potassium iodide for the hyperthyroidism group, and

183 levothyroxin and thyronamine for the hypothyroidism group) obtained on the patient's first  
184 visits, (b) the patient is not diagnosed with thyroid nodules, (c) patient records contain laboratory  
185 findings obtained within four weeks after the patient's first prescription, and (d) exclude records  
186 with missing values of more than half of our selected features. Since the age distributions were  
187 different among the institutions, as shown in Table 1, we also conducted data under sampling to  
188 fix the gaps in these differences.

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

199

### 200 ***Integrating multiple hospital datasets***

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

206 thyroid dysfunction group data from Wakayama Medical University and control group data from  
207 Hidaka hospital (referred to as Inst. comb. 2), and thyroid dysfunction group data from Gunma  
208 University and control group data from Hidaka hospital (referred to as Inst. comb. 3), were set to  
209 train and evaluate the models.

210

### 211 ***Machine learning algorithms***

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

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

223       The artificial neural network (ANN) is a well-established classification technique that is  
224 widely used in pattern recognition studies. In general, an ANN consists of 3 layers: an input layer  
225 that receives information, a hidden layer that processes information, and an output layer that  
226 calculates the results (21). In the present study, a standard feed-forward ANN was applied due to  
227 its relative simplicity and stability.

228 Support vector machine (SVM) is a supervised machine learning technique that is widely  
229 used in pattern recognition and classification problems (22). In the approach of this method, each  
230 data sample is a vector whose dimensions are equal to the number of features to be considered,  
231 and the SVM creates a hyperplane that separates samples into two categories. The induced  
232 hyperplane is constructed to maximize its distance from the samples of both classes. This  
233 algorithm achieves high classification performance by using special nonlinear functions called  
234 kernels to transform the input space into a multidimensional space (22). In this study, the radial  
235 basis function kernel is used.

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

241  
242 ***Explanatory features (variables) for machine learning***

243 Features from a subject's record were designed to sufficiently explain factors that were related to  
244 thyroid dysfunction. We used 11 variables as explanatory variables in this study as the first  
245 experimented set of features (referred to as Feature set 1) in this study, of which eight tests are  
246 tests measured in routine health checkup: sex, AST, ALT,  $\gamma$ -GTP, total cholesterol, Hemoglobin  
247 (Hb), RBC, and creatinine (S-Cr). In addition, since ALP, UA, and S-Cr ratio are reported to be  
248 highly relevant to thyroid dysfunction (24, 25), these were added to the above items. We also  
249 included UA/S-Cr ratio in this study considering that the reduction of S-Cr has been reported in  
250 hyperthyroidism, while UA has not been confirmed to fluctuate with thyroid dysfunction. To  
251 discriminate hyperthyroidism with renal dysfunction, which usually leads to the rise of both S-Cr

252 and UA, we introduced UA/S-Cr ratio as one of the features to improve the classification  
253 performance. 11 tests (Feature set 1) in total were used as features to train machine learning  
254 models in this study.

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

259

### 260 ***Model validation***

261 Cross-validation was applied to evaluate the performance of our machine learning method in  
262 classifying patients. The evaluation was conducted by extracting 9/10 training data and 1/10 test  
263 data by conducting 10-fold cross-validation. This was repeated 10 times to extract the training  
264 and test data uniformly, and the average and standard deviation of each evaluation score of each  
265 time were calculated. During the model training and test process, we avoided including the same  
266 subject to both training dataset and test dataset. The following measures were used for the  
267 performance evaluation criteria: area under the receiver operating characteristic curve (AUROC),  
268 area under the precision-recall curve (AUPRC), sensitivity is defined by  $TP/(TP+FN)$ , and  
269 specificity is defined by  $TN/(TN+FP)$ , where TP is the number of true positives, TN is the  
270 number of true negatives, FP is the number of false positives, FN is the number of false  
271 negatives. Note that the cutoff value for classifying as positive or negative is determined by  
272 Youden index (26). Finally, the AUROC performance difference between models was verified as  
273 statistically significant by the Wilcoxon signed-rank test.

274 In addition, the data of Kuma hospital were employed as an external validation. The  
275 model was constructed using the hyperthyroidism group and the hypothyroidism group of  
276 Wakayama Medical University and Gunma University and the control group of Hidaka hospital  
277 as the training data. The model was evaluated using the hyperthyroidism group and  
278 hypothyroidism group of Kuma hospital and the control group of Hidaka hospital (referred to as  
279 External).

280

### 281 ***Classification of subclinical thyroid dysfunction***

282 In the guideline of Japan Thyroid Association (27), subclinical hypothyroidism is defined as  
283 when FT4 is within the normal limit but the TSH measured is higher than normal .On the other  
284 hand subclinical hyperthyroidism is defined as when FT4 is within normal limit and TSH is  
285 lower than normal. Compared to the overt thyroid dysfunction where both TSH and FT4 are out  
286 of the standard ranges, it is difficult to classify subclinical thyroid dysfunction. This study  
287 evaluated the classification performance of the machine learning model by using subclinical  
288 standards in the thyroid function test criterion labeling method. We further extended the feature  
289 set in the attempt of improving model performance and selected 24 tests (referred to as Feature  
290 set 3), which was the all the laboratory tests available in this study<sup>1</sup>.

291

### 292 ***Feature importance***

---

<sup>1</sup> Feature set 3 includes sex, AST, ALT,  $\gamma$ -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.

293 To further understand how each feature contributes to the classification of patients in our model,  
294 we introduced feature importance. Feature importance represents the factor by which the model  
295 error is increased compared to the original model error. In the decision tree-based machine  
296 learning algorithms, including GBDT, impurities and the features at which the node is split are  
297 recorded for all the nodes when the decision tree learning is finished, and the decision tree  
298 calculates the features importance using this information (19).

299

## 300 **Results**

### 301 *Model validation*

302 Table 3 is a summary of the performance results of the machine learning model constructed in  
303 this study. As the result of 10-fold cross-validation, as shown in No. I of Table 3, the best  
304 classification model for overt hyperthyroidism achieved an accuracy of AUROC = 92.4%,  
305 sensitivity = 83.3%, and specificity = 90.9%. The best classification model for overt  
306 hypothyroidism achieved an accuracy of AUROC = 90.5%, sensitivity = 84.4%, and specificity  
307 = 86.4%. In the external evaluation, as shown in No. IX of Table 3, the classification model for  
308 overt hyperthyroidism achieved an accuracy of AUROC = 96.3%, and the classification model  
309 for overt hypothyroidism achieved an accuracy of AUROC = 92.9%. As shown in No. XI of  
310 Table 3, the classification model for subclinical hyperthyroidism achieved an accuracy of  
311 AUROC = 73.8%, and the classification model for subclinical hypothyroidism achieved an  
312 accuracy of AUROC = 75.2%.

313 The result of comparing different labeling criteria is shown in No. I and II of Table 3.  
314 When the prescription criterion was applied as the labeling criterion, the accuracy of the  
315 hyperthyroidism classification model achieved AUROC = 88.2%, and that of the hypothyroidism  
316 classification model achieved AUROC = 82.4%. On the other hand, as shown in No. I, when the

317 thyroid function test criterion was used, the accuracy of the hyperthyroidism classification model  
318 achieved AUROC = 92.4%, and that of the hypothyroidism classification model achieved  
319 AUROC = 90.5%. The model trained on the thyroid function test criterion data achieved a  
320 superior performance, which was statistically significant by the Wilcoxon test at p-value 0.05.  
321 The result of comparing models built on different institution combinations is shown in No. I, III,  
322 and IV of Table 3, as the highest performance was obtained when institution combination 1 was  
323 used as training set, and the accuracy of the hyperthyroidism classification model achieved  
324 AUROC = 92.4%, while and that of the hypothyroidism classification model achieved AUROC  
325 = 90.5%.

326         Among the four machine learning algorithms used in this study, including GBDT, SVM,  
327 logistic regression, and ANN, the highest performance was obtained when the GBDT method  
328 was applied as shown in No. I, V, VI, and VII of Table 3. The accuracy of the hyperthyroidism  
329 classification model achieved AUROC = 92.4%, while that of the hypothyroidism classification  
330 model achieved AUROC = 90.5%, which were statistically significant at p-value 0.05 by the  
331 Wilcoxon test. After comparing the performance of different feature sets, as shown in I and VIII  
332 of Table 3, when the feature set 3 was applied, the accuracy of the hyperthyroidism classification  
333 model was reduced to AUROC = 87.4%, and the performance of the hypothyroidism  
334 classification model was reduced to AUROC = 85.5%, which shows significant differences by  
335 the Wilcoxon test at p-value 0.05.

Table 3. Results of validation on different models

No.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	
Training	Data labeling	Thyroid function test criterion	Prescription criterion	Thyroid function test criterion								
	Institution combination	Inst. comb. 1	Inst. comb. 1	Inst. comb. 2	Inst. comb. 3	Inst. comb. 1						
Validation	Machine learning algorithm	GBDT		GBDT	SVM	Logistics regression	ANN	GBDT				
	Input features	Feature set 1	Feature set 1								Feature set 2	Feature set 1
Labeling criteria												
Thyroid function test criterion												
Overt/Subclinical												
Overt												
Subclinical												
Institution combination												
	Inst. comb. 1	Inst. comb. 1										
	AUROC	92.4±3.2%	88.2±3.2%	91.4±3.0%	90.7±4.1%	85.7±5.2%	86.1±4.6%	86.7±4.0%	87.4±4.2%	96.3±0.3%	73.0±5.7%	73.8±6.3%
Hyperthyroidism	AUPRC	88.5±5.2%	82.8±4.9%	87.4±3.3%	85.7±6.9%	80.7±5.5%	79.8±4.5%	81.5±4.5%	80.0±7.3%	99.1±0.1%	56.6±8.3%	57.6±7.9%
	Sensitivity	83.3±9.0%	77.5±8.5%	81.9±9.2%	88.4±7.8%	70.8±11.9%	76.4±11.4%	73.5±9.3%	82.3±6.0%	87.7±1.8%	71.7±12.4%	78.7±11.2%
	Specificity	90.9±6.8%	86.4±6.2%	89.8±5.0%	80.5±12.5%	91.4±4.6%	86.0±5.8%	87.8±5.6%	83.9±7.7%	93.5±1.1%	66.6±14.0%	61.6±16.3%
Hypothyroidism	AUROC	90.5±3.6%	82.4±5.3%	85.9±4.9%	86.5±3.5%	77.9±4.4%	81.4±5.7%	79.8±5.3%	85.5±3.1%	92.9±0.7%	69.1±4.1%	75.2±3.3%
	AUPRC	84.1±5.7%	71.3±9.9%	77.1±7.5%	79.0±5.8%	69.5±6.8%	71.3±6.8%	70.3±7.3%	78.2±4.9%	75.7±2.7%	51.3±4.2%	59.9±7.1%
	Sensitivity	84.4±9.6%	77.4±8.9%	83.6±4.6%	80.1±10.1%	66.9±14.2%	79.5±8.8%	73.5±8.8%	76.0±7.2%	87.1±3.3%	68.2±5.8%	77.7±10.4%
	Specificity	86.4±5.2%	77.0±11.7%	78.3±8.8%	81.0±9.2%	78.4±14.9%	73.1±10.7%	77.1±8.4%	82.9±9.0%	84.2±3.7%	64.3±16.4%	64.3±13.6%

The mean and standard deviation for the 10 folds are shown in each score.

337 The model with the best performance was evaluated using the external dataset for Kuma  
338 Hospital, as shown in No. IX of Table 3. High classification performance was achieved using the  
339 external data: AUROC = 96.3%, sensitivity = 87.7%, and specificity = 93.5% for the  
340 hyperthyroidism classification model and AUROC = 92.9%, sensitivity = 75.7%, and specificity  
341 = 87.1% for the hypothyroidism classification model. No. X and XI of Table 3 show that using  
342 feature set 3 improved the classification performance: for subclinical thyroid dysfunction,  
343 AUROC = 73.8%, sensitivity = 78.7%, and specificity = 61.6%; for hypothyroidism, AUROC =  
344 75.2%, sensitivity = 59.9%, and specificity = 77.7%. In particular, the significance of the  
345 hypothyroidism classification models was statistically confirmed by the Wilcoxon test at p-value  
346 0.05.

347

#### 348 ***Feature Importance***

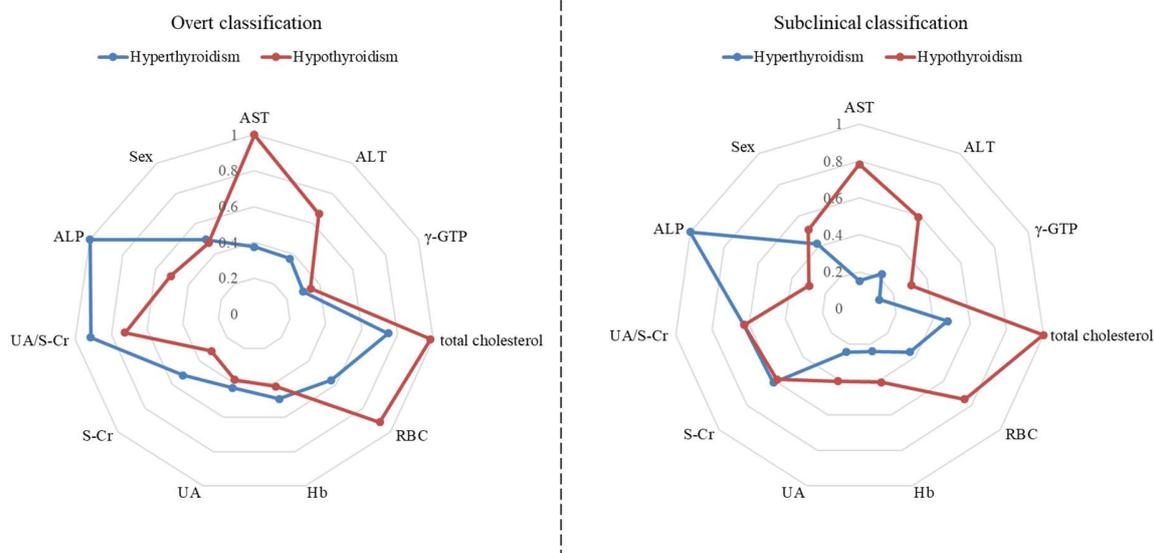
349 The features importance of each model was examined using the feature set 1. The left picture of  
350 Figure 1 shows the features importance of the overt hyperthyroidism classification model and the  
351 overt hypothyroidism classification model. The three most important features in the overt  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

359 were the top related features in the hyperthyroidism classification model, and total cholesterol,  
360 AST, and RBC were the top features in the hypothyroidism classification model.

361 Furthermore, the features importance in the subclinical hyperthyroidism and subclinical  
362 hypothyroidism classification models using the feature set 3 was conducted. As shown in Figure  
363 2, ALP and the UA/S-Cr ratio were among the three most important features in the subclinical  
364 hyperthyroidism classification model when the feature set 1 was used, as well as when the  
365 feature set 3 was used. If the five most important features were considered, MCV and MCH, two  
366 features added to the feature set 3, were included. These findings suggest that these two features  
367 are also likely to be effective in hyperthyroidism classification. On the other hand, as shown on  
368 right side of Figure 2, a difference was seen in the subclinical hypothyroidism classification  
369 model when the feature set 1 was used vs. when the feature set 3 was used. The three most  
370 important features in the model that used the feature set 1 were total cholesterol, AST, and RBC,  
371 whereas the three most important features in the model that used the feature set 3 were total  
372 protein, total cholesterol, AST, and the UA/S-Cr ratio. These findings suggest that total protein is  
373 likely to be effective in classifying subclinical hypothyroidism.

374

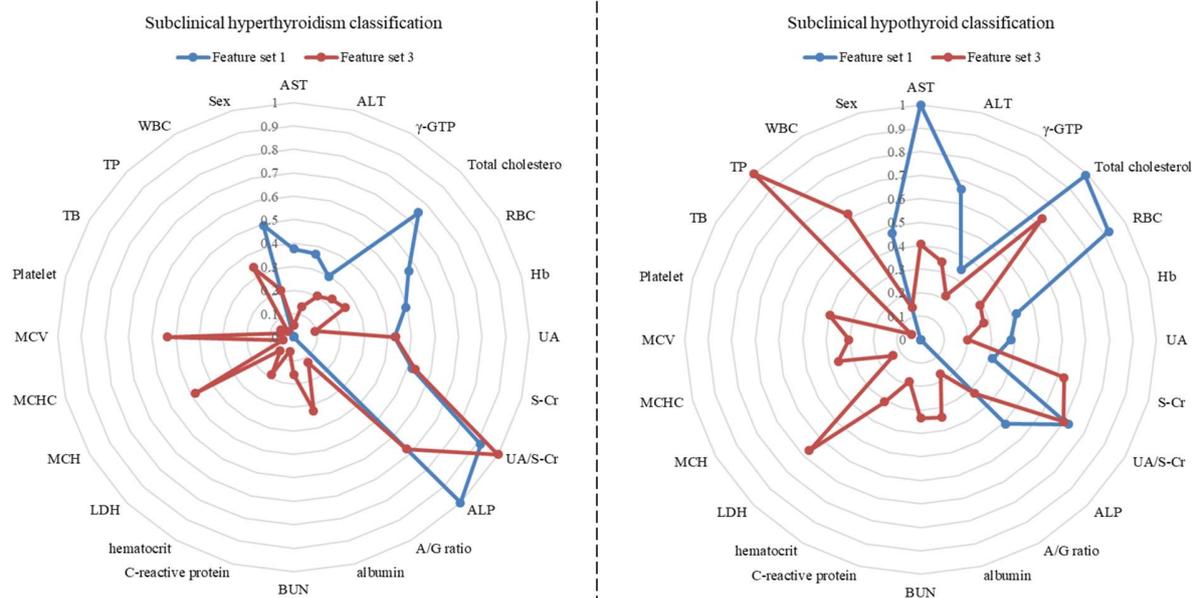
375



376

377  
378  
379

Figure 1. Comparison of feature importance between overt and subclinical thyroid dysfunction classification models



380

381

382

383

384

385

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

386 **Discussion**

387 ***Feature importance***

388 The correlation of routine laboratory tests such as ALP, S-Cr, UA, and RBC, etc. with thyroid  
389 dysfunction has been pointed out in many previous studies. According to studies on the  
390 relationship between thyroid dysfunction and liver function (28, 29), a correlation was confirmed  
391 between the increase in ALP and hyperthyroidism, as the ALP value was significantly higher  
392 when bone metabolism increases in Graves's disease, which is a typical disorder of  
393 hyperthyroidism. Sönmez (30) examined data from 433 patients and reported that S-Cr in the  
394 hyperthyroidism group was significantly lower than in the euthyroid group. TSH and S-Cr were  
395 also reported to have a significantly negative correlation with overt hypothyroidism (31).  
396 Dorgalaleh (32) suggested that thyroid dysfunction directly affects most of the blood values,  
397 including RBC, and health professionals must pay attention to such effects. The correlation  
398 between hypothyroidism and hyperuricemia has also been confirmed by in multiple studies (33,  
399 34).

400

401 ***Comparison with related studies***

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

404         Similar to the present study, Aoki's (17) study used pattern recognition methods such as  
405 neural networks to predict the likelihood of thyroid dysfunction from a set of routine test  
406 parameters such as ALP, S-Cr, and TC. Their results suggested that most patients with overt  
407 thyroid dysfunction could be screened by using a set of routine clinical data without measuring  
408 thyroid hormone levels. The correct rate of 91.3% was reported in the hyperthyroidism  
409 classification model, and the correct rate of 90.0% was reported in the hypothyroidism

410 classification model. Their results suggested that there is a high correlation between a set of  
411 routine laboratory tests and thyroid dysfunction. However, the model verification of these studies  
412 used the leave-one-out method instead of cross-validation and used the correct rate as the  
413 indicator instead of AUROC. Thus, the model evaluation was considered insufficient.

414 Unlike the present study, one drawback of these previous studies is that those have not  
415 considered crosstalk in the data labeling process. For hyperthyroidism classification in this study,  
416 the hyperthyroidism group was used as a positive label, and both the control and hypothyroidism  
417 groups were negatively labeled. For the hypothyroidism classification in this study, the  
418 hypothyroidism group was used as a positive label, whereas both the control and  
419 hyperthyroidism groups were negatively labeled (referred to as “crosstalk on”). On the other  
420 hand, related studies (16, 17) performed classification by setting thyroid dysfunction patients  
421 (with hyperthyroidism or hypothyroidism) as positive label and only control group as negative  
422 label (referred to as “crosstalk off”). Therefore, we evaluated the performance of the models with  
423 similar settings as these studies. As shown in A-1 column of Table 4, when only control group  
424 was labeled negative in both the training data and validation data, a high classification  
425 performance of AUROC = 94.9% and AUROC = 91.3% was achieved in the classification of  
426 overt hyperthyroidism and overt hypothyroidism, respectively. However, as shown in A-2  
427 column of Table 4, when both control group and hypothyroidism group were labeled negative in  
428 the validation data of overt hyperthyroidism and when both control group and hyperthyroidism  
429 group were labeled negative in the validation data of overt hypothyroidism, the classification  
430 performance was reduced to AUROC = 78.5% and AUROC = 68.1%, respectively. The  
431 classification performance dropped significantly in the models in which crosstalk was not  
432 considered during the negative labeling process.

433

434

Table 4. Evaluation result obtained without considering crosstalk

No.		A-1	A-2
Training	Thyroid function test criterion	Overt + subclinical	
	Negative label setting	Crosstalk off	
Validation	Thyroid function test criterion	Overt	
	Negative label setting	Crosstalk off	Crosstalk on
Hyperthyroidism	AUROC	94.9±2.4%	78.5±3.1%
Hypothyroidism	AUROC	91.3±4.0%	68.1±3.1%

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

436

### 437 **Limitations**

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

443

444 Another limitation of this study is that the hypothyroidism classification models exhibited  
445 lower performance than the hyperthyroidism classification models. This result is attributed to  
446 differences in the respective serum hormones and underlying molecular mechanisms (35). The  
447 various nonspecific symptoms of hypothyroidism may not manifest simultaneously, resulting its  
448 subclinical rate larger than that of hyperthyroidism. In addition, patients with hypothyroidism  
449 such as Hashimoto's thyroiditis are dependent upon long-term levothyroxine treatment, which  
450 may affect the manifestation of routine laboratory findings.

451 Furthermore, in the external evaluation of this study, the subclinical classification model  
452 showed lower overall results than the overt classification models. Among subclinical thyroid  
453 dysfunctions, the cause of subclinical hypothyroidism is associated with chronic thyroiditis  
454 (Hashimoto's disease), of which approximately 60-80% of cases are related thyroid  
455 autoantibodies (36). On the other hand, the causes of subclinical thyrotoxicosis are classified into  
456 extrinsic overdose of thyroid hormone drugs, and endogenous hyperthyroidism such as Graves'  
457 disease (37). Most of the subclinical thyroid dysfunctions such as subclinical thyrotoxicosis and  
458 subclinical hypothyroidism have no subjective symptoms and are usually considered to be  
459 transient (38, 39). Performance may have been limited due to the fact that symptoms of  
460 subclinical thyroid dysfunction are usually minor compared to overt thyroid dysfunction, and the  
461 phenotype of subclinical thyroid dysfunction may not be reflected in the results of routine  
462 laboratory examination.

463

#### 464 ***Conclusion***

465 This study evaluated the screening method to discriminate hyperthyroidism and hypothyroidism  
466 from the electronic medical records or routine laboratory finding data from health checkups  
467 using a machine learning method with an aim to prevent missed diagnosis of thyroid  
468 dysfunction. This is a versatile new screening method that was successfully developed from a  
469 machine learning model construction method to discriminate patients with hyperthyroidism and  
470 hypothyroidism using 11 features. High accuracy was achieved in the discrimination of evident  
471 hyperthyroidism or hypothyroidism, although the discrimination accuracy of subclinical  
472 hyperthyroidism or hypothyroidism was not satisfactory, these alerts can be useful for non-  
473 specialists for thyroid diseases.

474 It is expected that the quality of life of patients will improve by applying the model  
475 developed in this study. If thyroid dysfunction is screened using our method in healthcare  
476 facilities, including hospitals and health checkup facilities, prompt and accurate diagnostic  
477 support can be provided from only routine laboratory tests.

478

#### 479 **References**

- 480 (1) Garmendia Madariaga A, Santos Palacios S, Guillén-Grima F, et al. The incidence and  
481 prevalence of thyroid dysfunction in Europe: A meta-analysis. *J Clin Endocrinol Metab.*  
482 2014;99:923-931.
- 483 (2) Cooper DS. Hyperthyroidism. *Lancet.* 2003;362:459-468.
- 484 (3) Roberts CG, Ladenson PW. Hypothyroidism. *Lancet.* 2004;363:793-803.
- 485 (4) Garber JR, Cobin RH, Gharib H, et al. Clinical practice guidelines for hypothyroidism in  
486 adults: Cosponsored by the American Association of Clinical Endocrinologists and the  
487 American Thyroid Association. *Endocr Pract.* 2012;18:988-1028.
- 488 (5) Cooper DS, Ridgway EC. Thoughts on prevention of thyroid disease in the United States.  
489 *Thyroid.* 2002;12:925-929.
- 490 (6) Hamada N. The frequency of thyroid diseases that should not be overlooked in general  
491 outpatient settings. *Jap Med J.* 1995;3740:22.
- 492 (7) Japanese Ministry of Health Patient Survey Database. Available at [www.mhlw.go.jp](http://www.mhlw.go.jp).  
493 Accessed February 17, 2020.
- 494 (8) Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery.  
495 *Lancet Oncol.* 2019;20:e262-e273.
- 496 (9) Weng S.F., et al. Can machine-learning improve cardiovascular risk prediction using  
497 routine clinical data?. *PloS one*, 2017, 12.4: e0174944.

- 498 (10) Yatsuya, Hiroshi, et al. Development of a Risk Equation for the Incidence of Coronary  
499 Artery Disease and Ischemic Stroke for Middle-Aged Japanese—Japan Public Health  
500 Center-Based Prospective Study. *Circulation Journal* 80.6 (2016): 1386-1395.
- 501 (11) Chung JW, Kim WJ, Choi SB, et al. Screening for pre-diabetes using support vector  
502 machine model. Presented at the 36th Annual International Conference of the IEEE  
503 Engineering in Medicine and Biology Society. Chicago, IL, August 26-30, 2014.
- 504 (12) Soguero-Ruiz C, Mora-Jiménez I, Rojo-Alvarez JL, et al. Feature selection using Kernel  
505 component analysis for early detection of anastomosis leakage. Presented at the 2nd  
506 International Workshop on Pattern Recognition for Healthcare Analytics, Stockholm,  
507 Sweden, August 24, 2014.
- 508 (13) Kawakami J, Hoshi K, Sato W, et al. Screening of the patient with hyperthyroidism using  
509 routine test data. *J Tohoku Pharm Univ.* 2005;52:141-148.
- 510 (14) Hoshi K, Kawakami J, Sato W, et al. Assisting the diagnosis of thyroid diseases with  
511 Bayesian-type and SOM-type neural networks making use of routine test data. *Chem*  
512 *Pharm Bull (Tokyo).* 2006;54:1162-1169.
- 513 (15) Sato W, Hoshi K, Kawakami J, et al. Assisting the diagnosis of Graves' hyperthyroidism  
514 with Bayesian-type and SOM-type neural networks by making use of a set of three routine  
515 tests and their correlation with free T4. *Biomed Pharmacother.* 2010;64:7-15.
- 516 (16) Aoki S, Hoshi K, Kawakami J, et al. Assisting the diagnosis of Graves' hyperthyroidism  
517 with pattern recognition methods and a set of three routine tests parameters, and their  
518 correlations with free T4 levels: Extension to male patients. *Biomed Pharmacother.*  
519 2011;65:95-104.

- 520 (17) Aoki S, Hoshi K, Kawakami J, et al. Assisting the diagnosis of overt hypothyroidism with  
521 pattern recognition methods, making use of a set of routine tests, and their multiple  
522 correlation with total T4. *Biomed Pharmacother.* 2012;66:195-205.
- 523 (18) Liang W, Luo, S, et al. Predicting hard rock pillar stability using GBDT, XGBoost, and  
524 LightGBM algorithms. *Mathematics* 8.5 (2020): 765.
- 525 (19) Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat.*  
526 2001;29:1189-1232.
- 527 (20) Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with  
528 categorical features. In: *Advances in neural information processing systems.* 2018. p. 6638-  
529 6648.
- 530 (21) Bishop CM, Nasrabadi N. *Pattern recognition and machine learning.* *Pattern Recognit.*  
531 2006;4:738.
- 532 (22) Cortes C, Vapnik V. Support-vector networks. *Machine learning.* 1995;20(3):273–97.
- 533 (23) Hosmer DW, Lemeshow S, and Sturdivant RX. *Applied logistic regression.* John Wiley &  
534 Sons. 2013; Vol. 398.
- 535 (24) Hoshi K, Kawakami J, Sato W, Sato K, Sugawara A, Saito Y, et al. Assisting the diagnosis  
536 of thyroid diseases with Bayesian-type and SOM-type neural networks making use of  
537 routine test data. *Chem Pharm Bull* 2006;54:1162–9.
- 538 (25) Sato W, Hoshi K, Kawakami J, Sato K, Sugawara A, Saito Y, Yoshida K. Assisting the  
539 diagnosis of Graves' hyperthyroidism with Bayesian-type and SOM-type neural networks  
540 by making use of a set of three routine tests and their correlation with free T4. *Biomed*  
541 *Pharmacother.* 2010 Jan;64(1):7-15.
- 542 (26) Youden, W. J. Index for rating diagnostic tests. *Cancer.* 1950 Jan;3(1):32-5.

- 543 (27) Japan Thyroid Association Guidelines 2013:  
544 <http://www.japanthyroid.jp/doctor/guideline/japanese.html> [2021 . 1 . 19] Japanese
- 545 (28) Cooper DS, Kaplan MM, Ridgway EC, Maloof F, Daniels GH. Alkaline phosphatase  
546 isoenzyme patterns in hyperthyroidism. *Ann Intern Med.* 1979;90(2):164-168.
- 547 (29) Malik R, Hodgson H. The relationship between the thyroid gland and the liver. *QJM: An*  
548 *International Journal of Medicine.* 2002;559-569.
- 549 (30) Sönmez E, Bulur O, Ertugrul DT, Sahin K, Beyan E, Dal K. Hyperthyroidism influences  
550 renal function. *Endocrine.* 2019 Jul;65(1):144-148.
- 551 (31) Saini, Vandana, et al. Correlation of creatinine with TSH levels in overt hypothyroidism—  
552 A requirement for monitoring of renal function in hypothyroid patients?. *Clinical*  
553 *biochemistry* 45.3 (2012): 212-214.
- 554 (32) Dorgalaleh A, Mahmoodi M, Varmaghani B, et al. Effect of thyroid dysfunctions on blood  
555 cell count and red blood cell indice. *Iran J Pediatr Hematol Oncol.* 2013;3(2):73-77.
- 556 (33) Kuhlback B Creatine and creatinine metabolism in thyrotoxicosis and hypothyroidism: a  
557 clinical study. *Acta Med Scand Suppl.* 1957;331:1–70.
- 558 (34) Erickson AR, Enzenauer RJ, Nordstrom DM et al. The prevalence of hypothyroidism in  
559 gout. *Am J Med.* 1994;97:231–234.
- 560 (35) Okamura K, Nakashima T, Ueda K, et al. Thyroid disorders in the general population of  
561 Hisayama Japan, with special reference to prevalence and sex differences. *International*  
562 *Journal of Epidemiology.* 1987; 16(4), 545–549.
- 563 (36) Cooper DS, Biondi B : Subclinical thyroid disease. *Lancet* 2012 ; 379 : 1142-1154
- 564 (37) Biondi B, Cooper DS : The clinical significance of subclinical thyroid dysfunction. *Endocr*  
565 *Rev* 2008 ; 29 :76-131

566 (38) Fatourechi V. Subclinical hypothyroidism: an update for primary care physicians. Mayo  
567 Clinic Proceedings. Vol. 84. No. 1. Elsevier, 2009.

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

571

## 572 **List of Abbreviations**

573 TSH: thyroid-stimulating hormone

574 FT3: free triiodothyronine

575 FT4: free triiodothyronine

576 GBDT: gradient boosting decision tree

577 ANN: artificial neural network

578 SVM: support vector machine

579 AUROC: area under the receiver operating characteristic curve

580 AUPRC: area under the precision-recall curve

581 ALP :alkaline phosphatase

582 UA : uric acid

583 S-Cr: serum creatinine

584 AST : glutamic aspartate transaminase

585 ALT: alanine aminotransferase

586 RBC : red blood cell count

587

## 588 **Declarations**

589 ***Data Availability***

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

594

595 ***Correspondence address***

596 Min Hu

597 Koishikawa 2-7-3 Tomisaka Building, Bunkyo-ku, Tokyo, Japan

598 E-mail: [m.hu@cosmic-jpn.co.jp](mailto:m.hu@cosmic-jpn.co.jp)

599 Yasubumi Sakakibara

600 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan

601 E-mail: [yasu@bio.keio.ac.jp](mailto:yasu@bio.keio.ac.jp)

602

603 ***Competing Interests***

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

606

607 ***Funding statement***

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

610

611 ***Author Contributions***

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

618

### 619 *Acknowledgements*

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