European Society of Cardiology

European Heart Journal - Digital Health (2022) **3**, 255–264
https://doi.org/10.1093/ehjdh/ztac013

ORIGINAL ARTICLE

Electrocardiographic biomarker based on machine learning for detecting overt hyperthyroidism

Byungjin Choi[1]†, Jong-Hwan Jang[2,3]† Minkook Son[4], Min Sung Lee[2], Yong-Yeon Jo[2], Ja Young Jeon[5], Uram Jin[6], Moonseung Soh[6], Rae Woong Park[1,7]* and Joon-myoung Kwon [2,8,9]*

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea; [2]Department of Medical Research, Medical AI Co., Seoul, Republic of Korea; [3]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Yongin, Republic of Korea; [4]Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea; [5]Department of Endocrinology and Metabolism, Ajou University School of Medicine, Suwon, Republic of Korea; [6]Department of Cardiology, Ajou University School of Medicine, Suwon, Republic of Korea; [7]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea; [8]Department of Emergency Medicine, Mediplex Sejong Hospital, Incheon, Republic of Korea; and [9]Artificial Intelligence and Big Data Research Center, Sejong Medical Research Institute, Bucheon, Republic of Korea

Received 6 December 2021; revised 16 March 2022; accepted 28 March 2022; online publish-ahead-of-print 20 April 2022

Aims Although overt hyperthyroidism adversely affects a patient's prognosis, thyroid function tests (TFTs) are not routinely conducted. Furthermore, vague symptoms of hyperthyroidism often lead to hyperthyroidism being overlooked. An electrocardiogram (ECG) is a commonly used screening test, and the association between thyroid function and ECG is well known. However, it is difficult for clinicians to detect hyperthyroidism through subtle ECG changes. For early detection of hyperthyroidism, we aimed to develop and validate an electrocardiographic biomarker based on a deep learning model (DLM) for detecting hyperthyroidism.

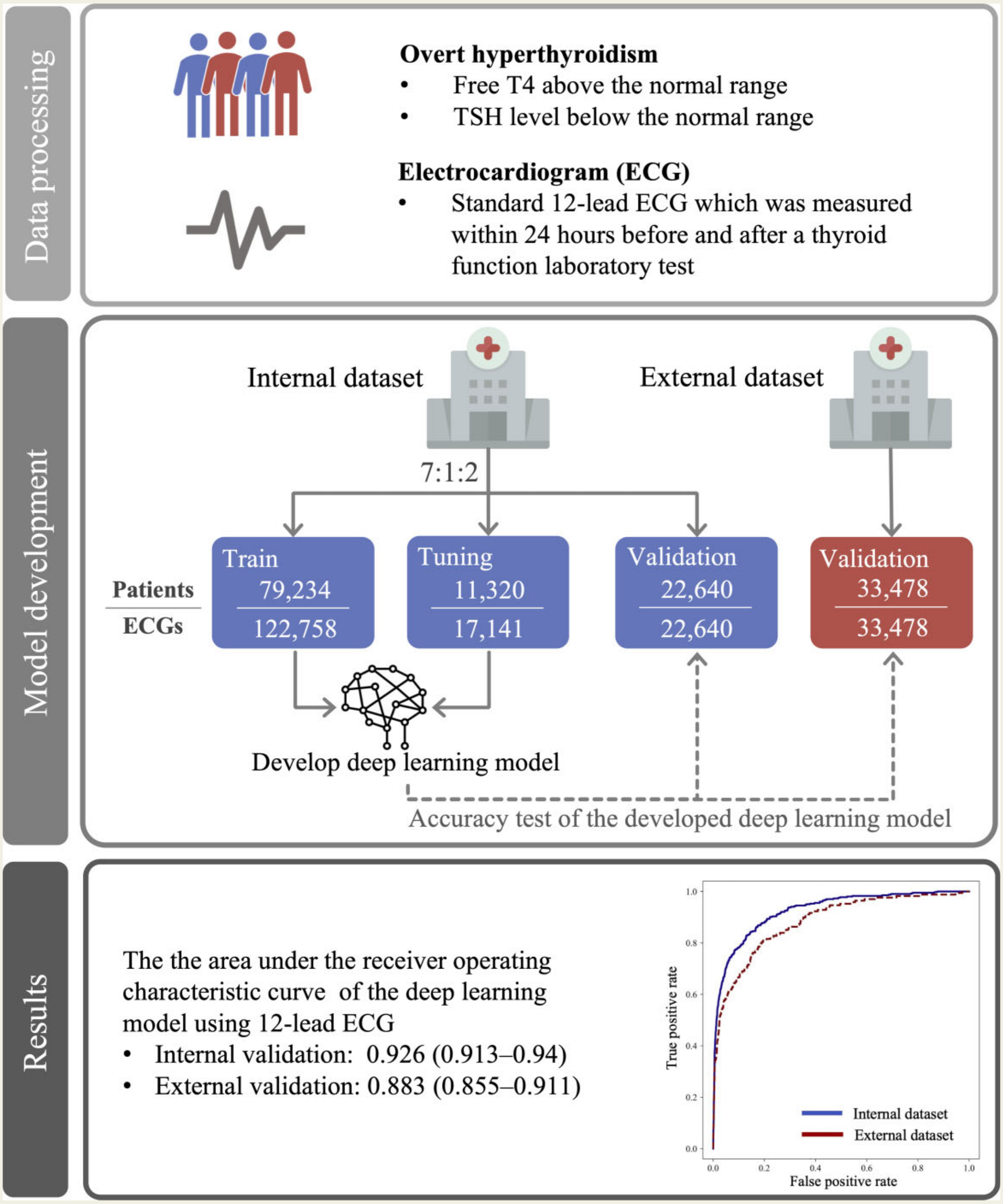
Methods and results This multicentre retrospective cohort study included patients who underwent ECG and TFTs within 24 h. For model development and internal validation, we obtained 174 331 ECGs from 113 194 patients. We extracted 48 648 ECGs from 33 478 patients from another hospital for external validation. Using 500 Hz raw ECG, we developed a DLM with 12-lead, 6-lead (limb leads, precordial leads), and single-lead (lead I) ECGs to detect overt hyperthyroidism. We calculated the model's performance on the internal and external validation sets using the area under the receiver operating characteristic curve (AUC). The AUC of the DLM using a 12-lead ECG was 0.926 (0.913–0.94) for internal validation and 0.883(0.855–0.911) for external validation. The AUC of DLMs using six and a single-lead were in the range of 0.889–0.906 for internal validation and 0.847–0.882 for external validation.

Conclusion We developed a DLM using ECG for non-invasive screening of overt hyperthyroidism. We expect this model to contribute to the early diagnosis of diseases and improve patient prognosis.

* Corresponding authors. Email: veritas@ajou.ac.kr (R.W.P.); Email: kwonjm@sejongh.co.kr (J.-m.K.)

†The first two authors contributed equally to the study.

Graphical Abstract



Overt hyperthyroidism

- Free T4 above the normal range
- TSH level below the normal range

Electrocardiogram (ECG)

- Standard 12-lead ECG which was measured within 24 hours before and after a thyroid function laboratory test

The the area under the receiver operating characteristic curve of the deep learning model using 12-lead ECG

- Internal validation: 0.926 (0.913–0.94)
- External validation: 0.883 (0.855–0.911)

Keywords Hyperthyroidism • Deep learning • Electrocardiography • Artificial intelligence

Introduction

Hyperthyroidism is a global healthcare issue, and the worldwide prevalence of overt hyperthyroidism ranges from 0.2% to 1.3%.[1,2] Untreated hyperthyroidism increases the risk of morbidity, including cardiac arrhythmia, stroke, and heart failure, and causes emergent life-threatening complications such as a thyroid storm.[3,4] Early and effective treatment of hyperthyroidism can prevent irreversible complications and mortality.[3,4]

Because, sometimes, the symptoms of hyperthyroidism are vague, it is challenging to detect hyperthyroidism based only on the patient's medical history and a physical examination.[1] The first-line test for diagnosing hyperthyroidism is a thyroid function test (TFT), including thyroid-stimulating hormone (TSH), free thyroxine(fT4), and triiodothyronine (T3).[5] However, TFT is an invasive test requiring blood sampling. In addition, TFT requires expensive infrastructure, including analysis devices and biochemical reagents, and is difficult to use in low-income countries that have a high prevalence of undiagnosed thyroid disease. An electrocardiogram (ECG), in contrast, is an inexpensive, non-invasive test that is one of the most used diagnostic tests and can be conducted using various wearable and lifestyle devices.

Several studies have focused on thyroid function and ECG changes such as sinus tachycardia, increased atrial arrhythmia, and changes in the QT interval.[6–8] However, no studies have attempted to detect hyperthyroidism using ECG. Recently, several deep learning models (DLMs) that predict the status of patients using only ECG have been developed.[9,10] If hyperthyroidism screening is possible only with non-invasive ECG without a separate blood test, early diagnosis and intervention can be facilitated. Moreover, if this model can detect even minute changes in the ECG of pre-hyperthyroidism, this model can be used not only to detect but also to predict future hyperthyroidism, which will improve the patient's prognosis.

For this purpose, we aimed to develop a DLM for detecting hyperthyroidism using an ECG and validated its performance, robustness, and value as a biomarker for future hyperthyroidism.

Methods

Data preparation

We conducted a retrospective multicentre cohort study to develop and validate a DLM that detects hyperthyroidism using ECG. As shown in *Figure 1*, the study population included patients who visited the hospital and underwent at least one standard 12-lead ECG and at least one TFT within 24 h before and after the index ECG.

To develop the DLM and conduct internal performance tests, we obtained eligible patient data from a tertiary hospital in Republic of Korea (Ajou University Medical Center, AUMC) from 1 January 1994 to 31 December 2020. To validate the robustness of the developed DLM, we conducted an external performance test using a dataset from a community-based secondary hospital (Incheon Sejong Hospital, ISH) from 1 March 2017 to 31 May 2021. The two hospitals exist in geographically, administratively separate areas, and belong to different foundations.

In both hospitals, we obtained the TFT values, age, and sex from electronic medical records and extracted ECG data with a sampling rate of 500 Hz that was stored in the MUSE Cardiology Information System (GE Healthcare, Wisconsin, USA). All patients whose sex or age data were missing were excluded. Less than 0.1% of the patients were excluded under the above conditions.

The study was approved by the institutional review boards (IRBs) of AUMC (AJIRB-MED-MDB-21-362) and ISH (ISH-2021-0282). The IRBs waived the need for informed consent because of the retrospective nature of the study, the fully anonymized dataset applied, and minimal risk to the patient.

Endpoint

The study endpoint was overt hyperthyroidism. Overt hyperthyroidism was defined as free T4 above the normal range and TSH levels below the normal range in radioimmunoassay.[5] However, TFT measurement values are sensitive to equipment errors, and thus, the existing hyperthyroidism study applied a different normal range for each hospital and equipment.[8,11] According to references, we defined the normal range based on the recommended range from the Department of Diagnostic Laboratory Medicine of each hospital. The detailed normal ranges for each hospital and equipment are described in Supplementary material online, *S1*, and the distribution of TSH and T4 by equipment is illustrated in Supplementary material online, *S2*.

Model development

We divided the AUMC dataset into development, model tuning, and internal validation datasets in a 7:1:2 ratio based on the patient. In the model development and tuning dataset, all multiple ECG–TFT pairs for each patient are included in the study. However, in the internal and external validation sets, we randomly selected only one ECG–TFT pair per patient. We developed a model on the development dataset and tested it on an internal validation set. To prove the robustness of the DLM, we performed an external validation on the ISH dataset. We illustrated the study flow chart as *Figure 1*.

To maximize the utility of the DLM in various settings, we used only 12-lead ECG waveform data as a predictor variable in the DLM. Furthermore, for applicability in the wearable environment, we also developed a DLM that uses six partial limb leads (lead I, II, III, aVL, aVR, and aVF), six precordial leads (V1, V2, V3, V4, V5, and V6), or a single-lead (lead I) ECG.

We removed baseline wander noise <0.5 Hz using a Butterworth pass filter and normalized the data using standard normalization.[12] Because of the frequent noise at the start and end, we removed the data from both sides of the ECG by 0.9 s to create 4096-size signal data and used it as an input for the DLM.

Figure 2 shows the architecture of the DLM. We utilized six residual blocks of ResNet.[13] Each residual block consists of a convolutional neural network, batch normalization, ReLU activation function, and a dropout layer. Detailed information about the model structure, such as hyperparameters, is provided in Supplementary material online, *S3*.

After model development, we applied gradient-weighted class activation mapping (gradCAM).[14] Using the gradient of weights from the final convolutional layer of the model, gradCAM explains the model by notifying which part of the ECG contributes to predictions and making the importance of each area into a visual representation.

Performance evaluation

To prove the performance of the DLM, we compared the probability of the model with the presence of hyperthyroidism in the internal and external validation datasets. Therefore, we used the area under the receiver operating characteristic curve (AUC). In addition, we calculated the sensitivity, specificity, positive predictive value, and negative predictive value with a cut-off point from Youden's J statistics in the development dataset.[15] The 95% confidence intervals of the AUC were determined using the Sun-Su optimization of the De-long method.[16]

Sensitivity analysis

To prove the robustness of the DLM, we conducted several sensitivity analyses.

We divided subgroups according to gender and age. Then, we observed the model performance for each group. The criteria for the subgroups were male and female (sex) and –40, 40–50, 50–60, 50–60, 60–70, and 70– (age).

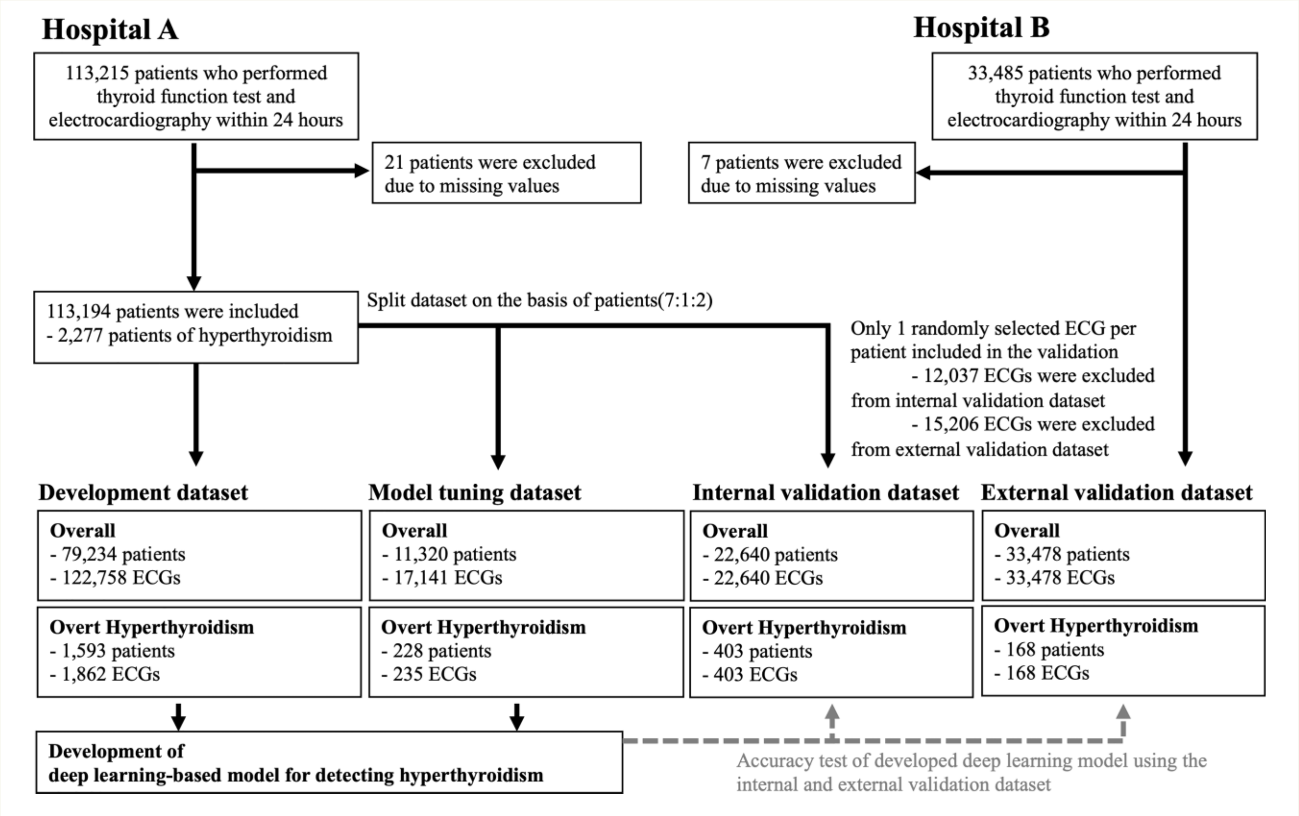

Hospital A
113,215 patients who performed thyroid function test and electrocardiography within 24 hours
21 patients were excluded due to missing values
113,194 patients were included
- 2,277 patients of hyperthyroidism
Split dataset on the basis of patients(7:1:2)
Hospital B
33,485 patients who performed thyroid function test and electrocardiography within 24 hours
7 patients were excluded due to missing values
Only 1 randomly selected ECG per patient included in the validation
- 12,037 ECGs were excluded from internal validation dataset
- 15,206 ECGs were excluded from external validation dataset
Development dataset
Overall
- 79,234 patients
- 122,758 ECGs
Overt Hyperthyroidism
- 1,593 patients
- 1,862 ECGs
Model tuning dataset
Overall
- 11,320 patients
- 17,141 ECGs
Overt Hyperthyroidism
- 228 patients
- 235 ECGs
Internal validation dataset
Overall
- 22,640 patients
- 22,640 ECGs
Overt Hyperthyroidism
- 403 patients
- 403 ECGs
External validation dataset
Overall
- 33,478 patients
- 33,478 ECGs
Overt Hyperthyroidism
- 168 patients
- 168 ECGs
Development of deep learning-based model for detecting hyperthyroidism
Accuracy test of developed deep learning model using the internal and external validation dataset

Figure 1 Study flow chart. To develop the deep learning model and conduct internal performance tests, we obtained patient data from a tertiary teaching hospital (Ajou University Medical Center, Hospital A) in Republic of Korea, from 1 January 1994 to 31 December 2020. We also conducted an external performance test using a dataset from a community-based secondary hospital (Incheon Sejong Hospital, Hospital B) from 1 March 2017 to 31 May 2021. The study population included patients who visited the hospital and underwent at least one standard 12-lead electrocardiogram and at least one thyroid function test within 24 h before and after the index electrocardiogram. In both hospitals, we obtained thyroid function test values, age, sex, and digitally stored electrocardiogram data. We excluded 21 and 7 patients, respectively, whose sex or age information was missing. As a result, 113 194 and 33 478 patients in Hospitals A and B, respectively, were included. For development, tuning, and internal validation, we divided the data from Hospital A by the patient (7:1:2).

Since our dataset consists of TFTs obtained from different equipment, we validated that the model is robust regardless of the radioimmunoassay equipment. So, we divided the internal and external validation dataset by equipment (Roche Elecsys, Siemens ADVIA, Architect i2000SR), and calculated model performance for each group.

In addition, the performance of a DLM is sensitive to the distribution of the dataset. To validate the robustness of our model, we sampled the data to obtain a uniform distribution for TSH and T4 levels. We divided the entire data set into 10, 25, 50, and 100 groups based on TSH and T4 levels. Finally, the AUCs of the DLM were recalculated. We bootstrapped 1000 times to calculate the average AUC and the confidence interval of the AUC.

Also, for validating the robustness of the model in patients without symptoms, we extracted ECG–TFT pairs from the internal validation dataset which were measured in regular national health examinations without specific medical complaints. Then, we observed the model performance in the health examination subgroup.

We also tested the robustness of DLM from the effects of the anti-thyroid agents (ATAs) such as methimazole, carbimazole, and propylthiouracil. We calculated DLM performances for each subgroup with a different drug history. We confirmed the change in model performance by obtaining a subgroup from the internal validation dataset excluding patients treated with an ATA within 3 months before TFT measurement and a subgroup excluding patients with any ATA prescription record before TFT measurement

Also, we checked whether the ATA itself affects the predicted probability of the DLM. We selected patients with overt hyperthyroidism (initial TFT) in the internal validation set. And we extracted patients who did not have a drug record before initial TFT and had a drug prescription, follow-up ECG–TFT pair within one to 12 months after initial TFT. If there are multiple follow-up pairs, the pair closest to the initial TFT was selected. Then, we illustrated the change in the predicted probability of the DLM at initial and follow-up.

Sub-analysis of model

We hypothesized that the ECGs have subtle changes in the pre-hyperthyroidism period and that the DLM can predict the development of overt hyperthyroidism by detecting such vague changes. To prove this hypothesis, we conducted a sub-analysis of patients with initial normal TFT and follow-up TFT, at least 4 weeks apart from the initial TFT, from among the internal validation datasets.

We used Youden's J statistics with the development dataset to determine cut-off points,[15] defined a group with a probability of the DLM greater than or equal to the cut-off as a high-risk group and defined

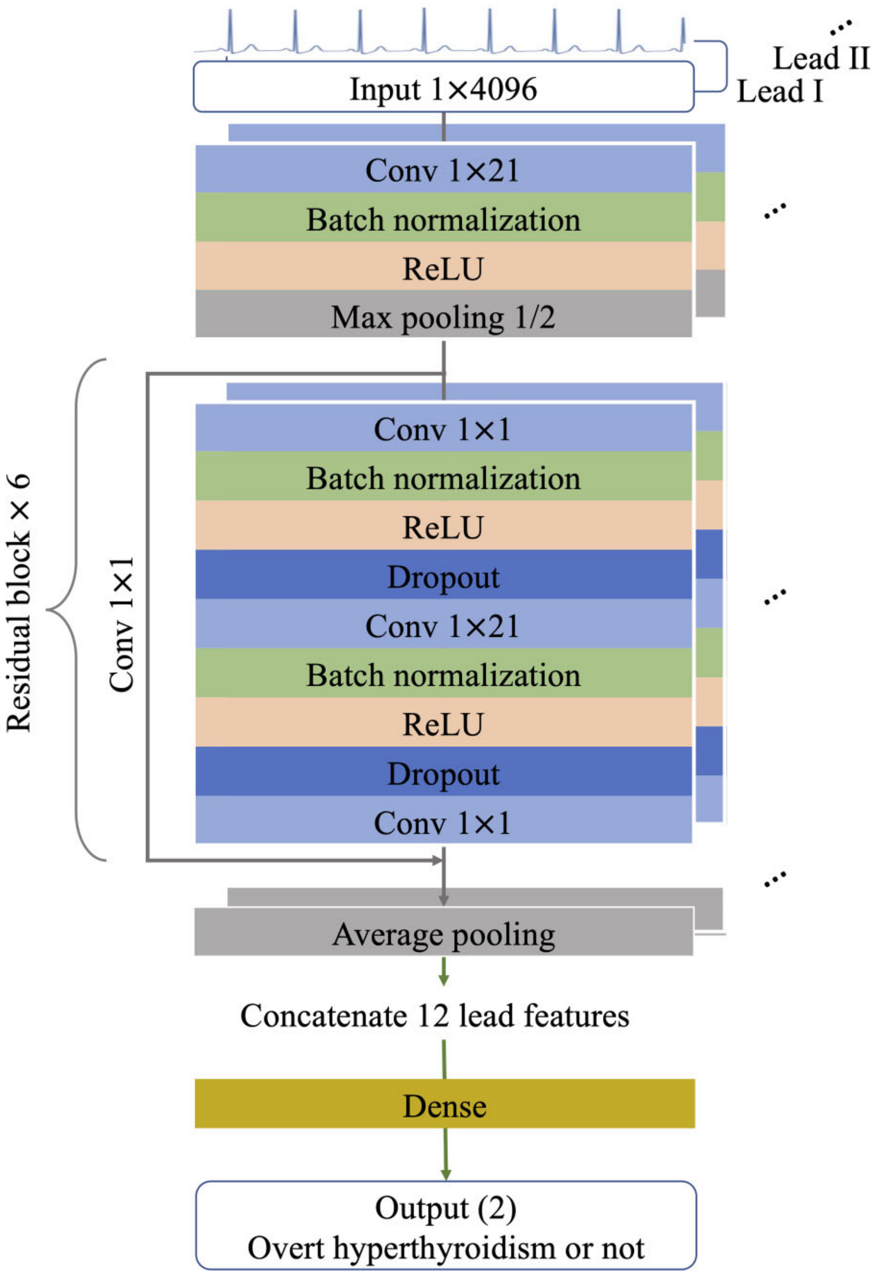


Figure 2 Architecture of deep learning model for detecting hyperthyroidism. We used only 12-lead electrocardiogram waveform data as a predictor variable in the deep learning model. We removed baseline wander noise <0.5 Hz for preprocessing using a Butterworth pass filter and normalized the data by standard normalization.[12] We trimmed both sides by 0.9 s to remove noise at the start and end of the electrocardiogram. Finally, we used a 4096-size vector as the input for the deep learning model. We utilized six residual blocks of the ResNet.[13] Each residual block consists of a convolutional neural network, batch normalization, ReLU activation function, and a dropout layer. The filter size was set to 21. Six residual blocks were used, and the length of the input was reduced by half every time the three residual blocks passed. Each electrocardiogram lead had a different ResNet model. At the end of the residual block, the outputs were channel-wise average pooled and concatenated with all outputs from each lead. Using the concatenated output, dense layers finally predict the probability (0–1) of overt hyperthyroidism. The above hyperparameters were selected through GridSearch.

others as a low-risk group. According to our theory, there were more ECG changes in the high-risk group, even in patients with normal thyroid function. Accordingly, the future overt hypertension incidence rate should be higher than that in the low-risk group.

Using Kaplan–Meier estimation, we analysed whether patients in the high-risk and low-risk groups developed new overt hyperthyroidism within 36 months and compared the statistical differences between the two survival curves using a log-rank test.

Statistical analysis

For baseline characteristics, continuous variables are presented as mean values and standard deviations and compared using the unpaired Student's *t*-test or Mann–Whitney *U* test. Categorical variables were described as percentages and were compared using the χ^2 test.

Neurokit2 0.1.4, Pytorch 1.8, and Python 3.6 were used for signal preprocessing and DLM development. ROCR and R 4.0.3 were used to visualize the results.

Results

Baseline characteristics

The eligible populations were 113 215 and 33 485 patients with AUMC and ISH, respectively. We excluded patients 21 and 7 from AUMC and ISH, respectively, because of missing clinical information. As a result, 113 194 patients with AUMC and 33 478 patients with ISH were included in the study. Among them, 2277 and 168 patients had hyperthyroidism, respectively. After the AUMC dataset was randomly divided in a ratio of 7:1:2 by the patient, the DLM was developed using the development and tuning dataset, which included 139 899 ECGs of 90 554 patients from AUMC. The internal performance test was conducted using the internal validation dataset of 22 640 ECGs from 22 640 patients from the AUMC. The external performance test was conducted using the external validation dataset containing 33 478 ECGs from 33 478 patients with ISH. A detailed study flow chart is shown in *Figure 1*.

The baseline characteristics of the development cohort (AUMC, $n = 113\,194$) and external validation cohort (ISH, $n = 33\,478$) used in this study are shown in *Table 1*. Sex, age, and prevalence of hyperthyroidism showed statistically significant differences between hospitals ($P < 0.001$). Patients with hyperthyroidism had more tachycardia and prolonged QT intervals ($P < 0.001$). Patients with hyperthyroidism had more rightward axis deviation of the P, R, and T wave axis and shorter QRS duration, from the beginning of the Q wave to the end of the S wave ($P < 0.001$).

Performance of deep learning model

During the internal and external validation tests, the AUCs of the DLM for the detection of hyperthyroidism using 12-lead ECG were 0.926 (0.913–0.94) and 0.883 (0.855–0.911), AUCs of the DLM using six limb leads were 0.906 (0.89–0.923) and 0.867 (0.835–0.899), and AUCs of the DLM using six precordial leads were 0.899 (0.881–0.918) and 0.882 (0.854–0.909), respectively. The lowest-performing DLM was a single-lead model with an AUC of 0.889 (0.873–0.906) for internal validation and 0.847 (0.821–0.874) for external validation. The detailed performance of the DLM for detecting hyperthyroidism using the 12-, 6-, and single-lead ECGs is shown in *Figure 3*.

After model development, we applied gradCAM to identify the ECG regions that were critical for model prediction. In the activation map, the DLM focused on the area between the T and R peaks to determine the presence of overt hyperthyroidism. A detailed figure of the gradCAM is depicted in Supplementary material online, *S4*.

Sensitivity analysis of deep learning model

In the sensitivity analysis of demographics, under 60 years of age, the model showed an AUC of 0.873 (0.761–0.985) or higher in all age groups and genders. The lowest model performance for all genders and age groups was for males aged 60–69 years. The model performance for all subgroups is described in Supplementary material online, *S5*.

In the sensitivity analysis of TFT equipment, the DLM shows AUC 0.926 (0.908–0.945) in the dataset from Roche Elecsys equipment, AUC = 0.895 (0.868–0.922) in Siemens ADVIA dataset, and

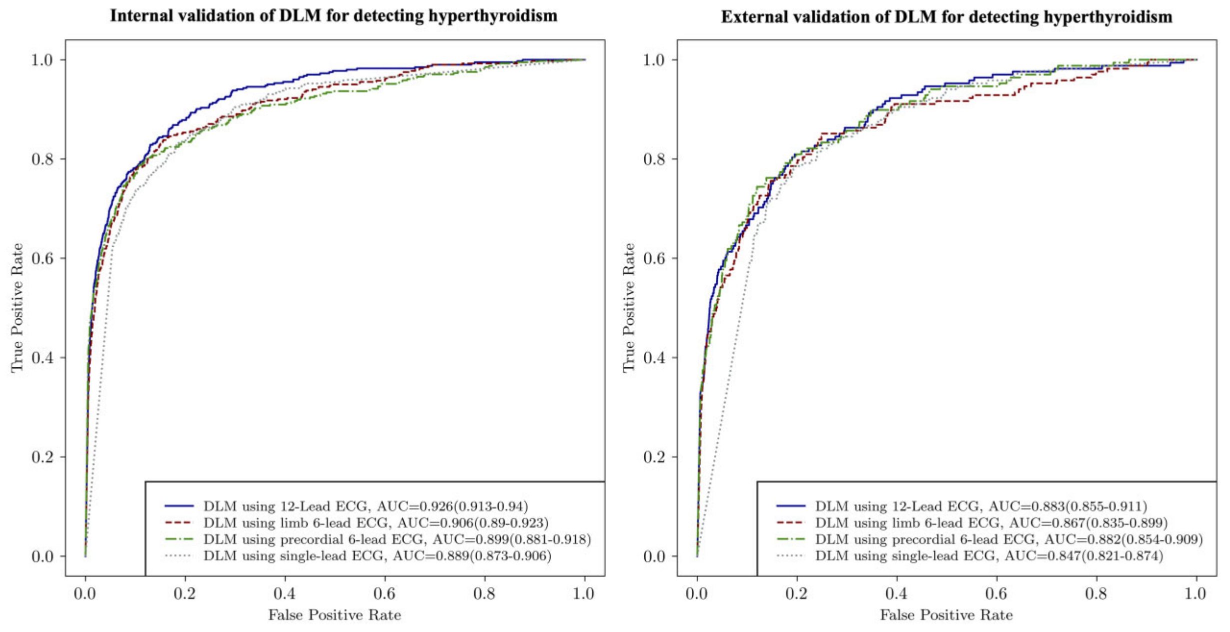
Table 1 **Baseline characteristics of cohorts**

Characteristic	Development and internal validation dataset $n = 113\,194$			External validation dataset $n = 33\,478$			*P*
	Non-hyperthyroidism	**Hyperthyroidism**	P^a	**Non-hyperthyroidism**	**Hyperthyroidism**	P^b	
Study population, *n* (%)	111 195 (98.2)	1999 (1.8)		33 313 (99.5)	165 (0.5)		<0.001
Age, year, mean (SD)	43.95 (13.41)	40.64 (14.38)	<0.001	55.37 (15.48)	51.51 (14.93)	0.001	<0.001
Male, *n*, (%)	56 808 (51.1)	638 (31.9)	<0.001	15 960 (47.9)	56 (33.9)	<0.001	<0.001
Heart rate, b.p.m. (%)	66.47 (12.75)	89.37 (19.61)	<0.001	70.89 (16.20)	97.59 (24.19)	<0.001	<0.001
PR interval, ms, mean (SD)	158.46 (23.63)	145.83 (25.51)	<0.001	166.97 (26.47)	151.08 (27.45)	<0.001	<0.001
QRS duration, ms, mean (SD)	93.07 (12.00)	86.24 (11.37)	<0.001	94.36 (15.02)	89.67 (13.46)	<0.001	<0.001
QTc interval, ms, mean (SD)	418.46 (23.42)	427.41 (32.29)	<0.001	433.44 (31.63)	442.88 (31.23)	<0.001	<0.001
P axis, mean (SD)	47.38 (23.68)	50.37 (24.46)	<0.001	44.34 (27.65)	48.93 (31.43)	0.045	<0.001
R axis, mean (SD)	49.42 (32.13)	54.86 (28.28)	<0.001	40.51 (39.55)	46.27 (33.08)	0.062	0.002
T axis, mean (SD)	39.83 (23.53)	46.32 (25.33)	<0.001	39.38 (38.21)	47.95 (47.98)	0.004	<0.001

Hospital A denotes Ajou University Medical Center (AUMC); Hospital B denotes Incheon Sejong Hospital (ISH).

[a] The alternative hypothesis for this *P*-value was that there was a difference between hyperthyroidism and overt hyperthyroidism.

[b] The alternative hypothesis for this *P*-value was that there was a difference between Hospital A (the development and internal validation data group) and Hospital B (external validation group) for each variable.



	Internal validation					External validation				
	AUC(95%CI)	SEN(95%CI)	SPE(95%CI)	PPV(95%CI)	NPV(95%CI)	AUC(95%CI)	SEN(95%CI)	SPE(95%CI)	PPV(95%CI)	NPV(95%CI)
DLM using 12-lead ECG	0.926 (0.913-0.94)	0.826 (0.816-0.838)	0.870 (0.869-0.872)	0.103 (0.100-0.107)	0.996 (0.996-0.997)	0.883 (0.855-0.911)	0.804 (0.766-0.820)	0.805 (0.804-0.846)	0.020 (0.019-0.024)	0.999 (0.998-0.999)
DLM using limb 6-lead ECG	0.906 (0.89-0.923)	0.830 (0.826-0.851)	0.838 (0.837-0.850)	0.086 (0.083-0.092)	0.996 (0.996-0.997)	0.867 (0.835-0.899)	0.750 (0.726-0.855)	0.852 (0.749-0.875)	0.025 (0.016-0.028)	0.998 (0.998-0.999)
DLM using chest 6-lead ECG	0.899 (0.881-0.918)	0.791 (0.763-0.832)	0.879 (0.878-0.899)	0.109 (0.105-0.132)	0.996 (0.995-0.996)	0.882 (0.854-0.909)	0.738 (0.722-0.780)	0.880 (0.860-0.887)	0.030 (0.026-0.032)	0.999 (0.998-0.999)
DLM using single-lead ECG	0.889 (0.873-0.906)	0.828 (0.741-0.839)	0.817 (0.815-0.898)	0.077 (0.075-0.123)	0.996 (0.995-0.996)	0.847 (0.821-0.874)	0.779 (0.751-0.801)	0.805 (0.803-0.832)	0.019 (0.019-0.022)	0.999 (0.998-0.999)

Figure 3 Performance of deep learning-based model for detecting hyperthyroidism. To maximize the utility of the deep learning model in various settings, we modified the 12-lead deep learning model to partially lead deep learning model, using only six limb leads (lead I, II, III, aVL, aVR, and aVF), six precordial leads (V1, V2, V3, V4, V5, and V6), or a single-lead (lead I) electrocardiogram. We validated the developed models using internal validation and external validation sets. The performance metrics were the area under the receiver operating characteristic curve, sensitivity, specificity, negative predictive value, and positive predictive value. To calculate the sensitivity, specificity, positive predictive value, and negative predictive value, we calculated the cut-off point using Youden's J statistics in the development dataset and applied the cut-off point to internal and external datasets.[15] The 95% confidence intervals were determined using the Sun-Su optimization of the De-long method.[16]

AUC = 0.883 (0.855–0.911) in the dataset from the Architect i2000SR, external validation dataset. We illustrated performances divided by equipment as Supplementary material online, *S6*.

Also, when a uniform distribution was created based on TSH, the mean AUC ranged from 0.913 to 0.925 in the internal validation dataset and from 0.874 to 0.882 in the external validation dataset. When a uniform distribution was achieved based on T4, the mean AUC ranged from 0.869 to 0.925 in the internal validation dataset and from 0.867 to 0.883 in the external validation dataset. We illustrated performances of DLM in uniform distribution as Supplementary material online, *S7*.

When we tested our model in the dataset from the regular health examination (*n* = 155), the DLM shows AUC = 0.869 (0.636–0.999). We illustrated performances of DLM as below and added figure as Supplementary material online, *S8*.

In the analysis of subgroups according to drug use history, the subgroup with 22 396 patients excluding patients who were using ATA showed AUC 0.918 (0.901–0.934). The model maintained a high performance of 0.919 (0.902–0.935) in the dataset except for 244 patients who had any ATA history. We illustrated performances of DLM in patients with a different drug history as Supplementary material online, *S9*.

When we inspect the change of predicted probability after administration of ATA, the overall predicted probability of the model was decreased after using ATA (0.890–0.485). However, even with the prescription of ATA, if overt hyperthyroidism remains, the model probability is not reduced (0.875–0.875). Anti-thyroid agent administration itself without TFT change does not affect predicted model probability. We illustrated the change of predicted probability after ATA as Supplementary material online, *Figure S10*.

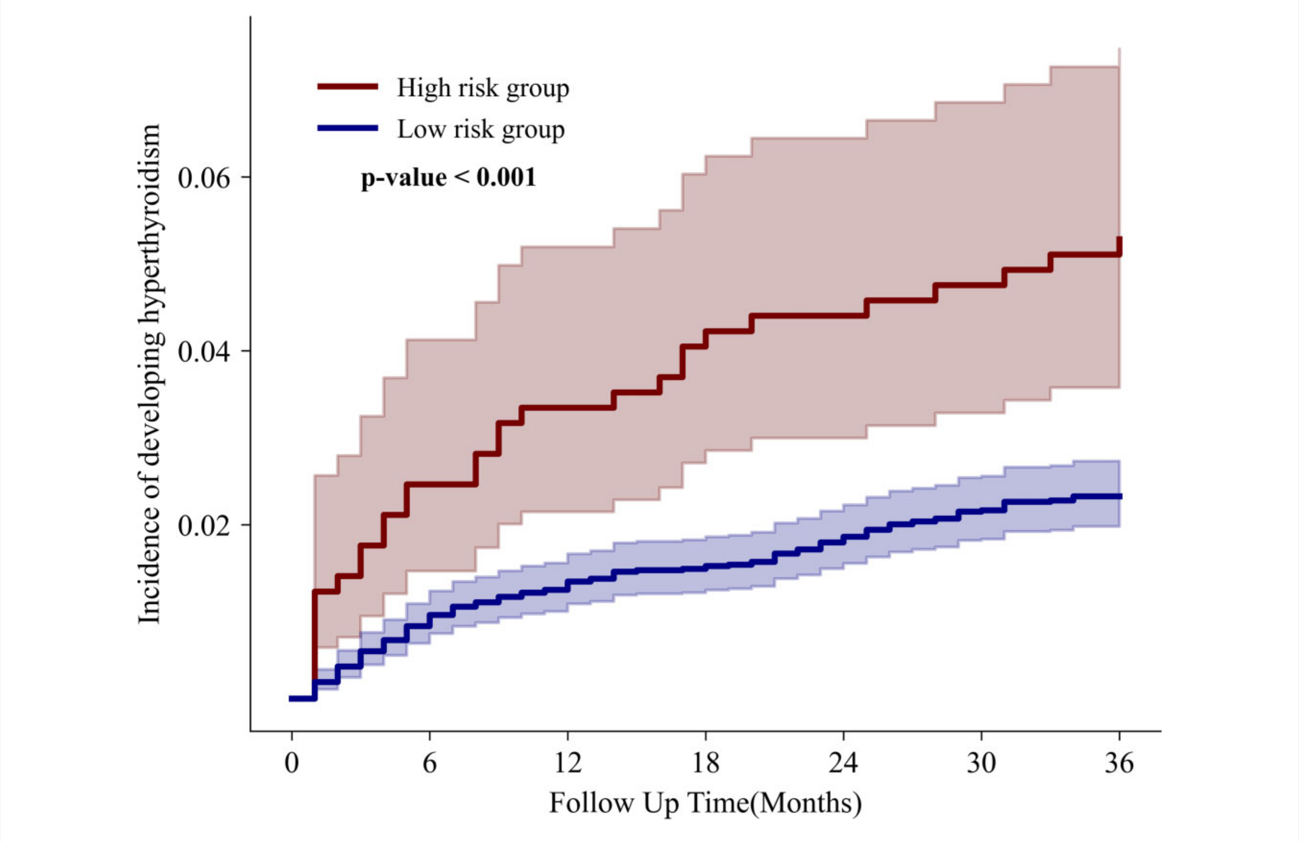


Figure 4 Cumulative hazard of developing hyperthyroidism in patients with an initially normal. We hypothesized that, even in normal patients, a high probability of the model could be a biomarker for predicting overt hyperthyroidism. To prove this hypothesis, we conducted a sub-analysis of 6800 patients with initial normal thyroid function test and follow-up thyroid function test, at least 4 weeks apart from the initial thyroid function test, from among the internal validation datasets. We used Youden's J statistics with the development dataset to determine cut-off points,[15] defined a group with a probability of deep learning model greater than or equal to the cut-off as a high-risk group and defined others as a low-risk group. Using Kaplan–Meier estimation, we analysed whether patients in the high-risk ($n = 6323$) and low-risk groups ($n = 568$) developed new overt hyperthyroidism within 36 months and compared the statistical differences between the two survival curves using a log-rank test. Thirty people in the high-risk group show overt hyperthyroidism, and 145 people in the low-risk group show overt hyperthyroidism. The incidence rate of overt hyperthyroidism is significantly higher in the high-risk group (5.28% vs. 2.33%, $P < 0.001$).

Prediction of hyperthyroidism in normal thyroid function patients

A sub-analysis was conducted using data from 6800 patients with normal TFT and follow-up TFT results. Among the sub-analysis patients, patients were divided into 6323 high-risk and 568 low-risk patient groups according to the DLM probability. As shown in *Figure 4*, 30 people in the high-risk group show overt hyperthyroidism, and 145 people in the low-risk group show overt hyperthyroidism. The incidence rate of overt hyperthyroidism is significantly higher in the high-risk group (5.28% vs. 2.33%, $P < 0.001$).

Discussion

Using 113 194 ECG–TFT pairs in a tertiary teaching hospital, we developed a DLM to predict overt hyperthyroidism and externally validated the model using 33 478 ECG–TFT pairs in a community-based secondary hospital. The performance of the DLM model using 12-lead ECG showed excellent performance in both internal (AUC 0.926) and external validations (AUC 0.883). When we limited the input ECG to the partial lead and performed external validation, consistently high performance was observed in limb six leads (AUC 0.867), precordial six leads (AUC 0.882), and even in the single-lead model (AUC 0.847). In addition, several sensitivity analyses about demographics, equipment, distribution, and drug history prove the robustness of the model. To our knowledge, this study is the first to propose using a DLM to evaluate overt hyperthyroidism using ECG.

Although the treatment for hyperthyroidism is well established, if not detected early, it can cause irreversible damage.[3,4] Therefore, early diagnosis through screening tests is necessary. In addition to screening purposes, repeated monitoring of thyroid function is essential for patients treated with anti-thyroid medicine, radioactive iodine ablation, or surgical thyroidectomy.[17–19] The most typical procedure for evaluating thyroid function is TFT.[5] However, TFT is invasive because it requires a blood sample. In our study, we developed a model that predicts hyperthyroidism using only a non-invasive standard 12-lead ECG. The proposed model demonstrated excellent performance in multicentre data. In addition, the model required only

ECG, which is one of the most commonly performed diagnostic tests in hospitals. Because of this, the developed model can be applied without any additional burden on the patient or practitioner.

The model developed in this study maintains high performance even in a situation with a partial lead. Recently, several wearable devices supported single-lead ECG measurements, and several models that detect various health statuses through wearable ECG have been developed.[20–22] As our model also shows good performance using only a single-lead, it is necessary to validate whether it can be applied to wearable ECG to enable continuous and non-invasive thyroid function monitoring.

Sub-analysis shows that the model's probability can be used as a biomarker for predicting future hyperthyroidism in patients with normal thyroid function. Among normal thyroid patients, those classified as a high-risk group by the model probability had a significantly higher incidence rate of future overt hyperthyroidism than those in the low-risk group (5.28% vs. 2.33%, $P < 0.001$). In the gradCAM analysis for DLM interpretation, we observed that DLM focused on the ECG changes known in hyperthyroidism. Accordingly, we infer that the high predicted probability from the model suggests that the high-risk group has a higher probability of minor thyroid dysfunction which can be turned into overt hyperthyroidism. However, DLM's black box characteristics make that our hypothesis cannot be convinced, additional research on DLM interpretability in the ECG part will be required further.

Nevertheless, this study has some limitations. First, the model trained only in tertiary teaching hospitals may not be robust to diverse population groups. To prove the robustness of the model, we performed an external validation in a community-based secondary hospital. Even though the baseline characteristics of the external validation cohort were significantly different from those of the development cohort, model performance remained with minimal degradation, from AUC 0.926 to 0.883. Second, because the DLM is a black box, the detailed prediction process of the model is unknown. To mitigate this, we adopted gradCAM to identify which part of the ECG the model considered necessary. GradCAM shows that the model mainly investigates the area between the T and R peaks to predict hyperthyroidism, which seems to be consistent with ECG changes known in hyperthyroidism, such as elevated heart rate and shortening of the PR and QRS interval.[6–8] However, gradCAM only informs about the position of ECG but does not tell why the position is essential. Therefore, further studies on explainable ECG DLMs should be conducted. Also, although the higher predicted probability is significantly associated with higher overt hyperthyroidism, its role as a biomarker in ECG in randomized patients is somewhat limited. Considering our dataset only includes patients who underwent ECG and TFT, symptoms of patients might have intervened in physicians' decision on whether to perform the measurement. Although we conducted subgroup analysis in the health examination subgroup to compensate for the above bias and showed that our model maintains high performance, prospective validation or clinical trial will be required before the clinical application of the model.

Conclusion

We developed a DLM to detect overt hyperthyroidism using ECG and a validated model in multiple centres. To our knowledge, this study is the first to propose a DLM to detect overt hyperthyroidism using ECG. The model showed excellent performance in both internal and external validations and showed excellent performance even when only six or one lead out of 12 electrodes were used. We anticipate that non-invasive hyperthyroidism screening can be performed using our model. We expect that this model will contribute to the early diagnosis of diseases and improve patient prognosis.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Authors' contributions

B.C. and J.-H.J. performed the data analysis and verified the clinical coding. B.C., J.-H.J., M.S. (the Gwangju Institute of Science and Technology), Y.-Y.J., and M.S.L. contributed to the study idea, design, and data collection performed data analysis and contributed to subsequent drafts. J.Y.J., U.J., and M.S. (Ajou University Medical Center) contributed to data collection and revised the manuscript. J.-m.K. and R.W.P. are the principal investigators and contributed to the study idea and design, data analysis, verified the clinical coding, and contributed to subsequent drafts.

Funding

This research was funded by the Bio Industrial Strategic Technology Development Program (20003883, 20005021) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health &Welfare, Republic of Korea (grant number: HR16C0001).

Conflict of interest: B.C., J.-Y.Y., U.J., M.S. (Ajou University Medical Center), R.W.P., M.S. (Gwangju Institute of Science and Technology) declare no competing interests. J.-m.K. is a co-founder, and J.-H.J., Y.-Y.J., and M.S.L. are researchers of Medical AI Co., a medical artificial intelligence company. There are no products in development or marketed products to declare. This does not alter our adherence to the journal.

Data availability

The data underlying this article will be shared at a reasonable request with the corresponding author.

References

1. Taylor PN, Albrecht D, Scholz A, Gutierrez-Buey G, Lazarus JH, Dayan CM, Okosieme OE. Global epidemiology of hyperthyroidism and hypothyroidism. *Nat Rev Endocrinol* 2018;**14**:301–316.
2. Toft AD. General cardiology: thyroid disease and the heart. *Heart* 2000;**84**:455–460.
3. Okosieme OE, Taylor PN, Evans C, Thayer D, Chai A, Khan I, Draman MS, Tennant B, Geen J, Sayers A, French R, Lazarus JH, Premawardhana LD, Dayan CM. Primary therapy of Graves' disease and cardiovascular morbidity and mortality: a linked-record cohort study. *Lancet Diabetes Endocrinol* 2019;**7**:278–287.
4. Danzi S, Klein I. Thyroid hormone and the cardiovascular system. *Med Clin N Am* 2012;**96**:257–268.
5. Dayan CM. Interpretation of thyroid function tests. *Lancet* 2001;**357**:619–624.
6. Tribulova N, Knezl V, Shainberg A, Seki S, Soukup T. Thyroid hormones and cardiac arrhythmias. *Vasc Pharmacol* 2010;**52**:102–112.

7. Colzani RM, Emdin M, Conforti F, Passino C, Scarlattini M, Iervasi G. Hyperthyroidism is associated with lengthening of ventricular repolarization. *Clin Endocrinol* 2001;**55**:27–32.
8. Tayal B, Graff C, Selmer C, Kragholm KH, Kihlstrom M, Nielsen JB, Olsen A-MS, Pietersen AH, Holst AG, Søgaard P, Christiansen CB, Faber J, Gislason GH, Torp-Pedersen C, Hansen SM. Thyroid dysfunction and electrocardiographic changes in subjects without arrhythmias: a cross-sectional study of primary healthcare subjects from Copenhagen. *BMJ Open* 2019;**9**:e023854.
9. Galloway CD, Valys AV, Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, Albert DE, Attia ZI, Carter RE, Asirvatham SJ, Ackerman MJ, Noseworthy PA, Dillon JJ, Friedman PA. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiol* 2019;**4**:428–436.
10. Kwon J-M, Cho Y, Jeon K-H, Cho S, Kim K-H, Baek SD, Jeung S, Park J, Oh B-H. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *Lancet Digit Health* 2020;**2**:e358–e367.
11. Andersen SL, Christensen PA, Knøsgaard L, Andersen S, Handberg A, Hansen AB, Vestergaard P. Classification of thyroid dysfunction in pregnant women differs by analytical method and type of thyroid function test. *J Clin Endocrinol Metab* 2020; **105**:e4012–e4022.
12. Selesnick IW, Burrus CS. Generalized digital Butterworth filter design. *IEEE Trans Signal Process* 1998;**46**:1688–1694.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition in Las Vegas, USA. IEEE; 2016.
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy. IEEE; 2017, p.618–626.
15. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* 2005;**16**:73–81.
16. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;**44**:837.
17. Bhatt N, Taufique Z, Kamen E, Wang B, Concert C, Li Z, Hu K, Givi B. Improving thyroid function monitoring in head and neck cancer patients: a quality improvement study. *Laryngoscope* 2020;**130**:E573–E579.
18. Iglesias P, Acosta M, Sánchez R, Fernández-Reyes MJ, Mon C, Diez JJ. Ambulatory blood pressure monitoring in patients with hyperthyroidism before and after control of thyroid function. *Clin Endocrinol* 2005;**63**:66–72.
19. Fraser WD, Biggart EM, O'Reilly DS, Gray HW, McKillop JH, Thomson JA. Are biochemical tests of thyroid function of any value in monitoring patients receiving thyroxine replacement? *Br Med J (Clin Res Ed)* 1986;**293**:808–810.
20. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv Prepr arXiv170701836*, 2017.
21. Galloway CD, Valys A V, Petterson FL, Gundotra VP, Treiman DL, Albert DE, Dillon JJ, Attia ZI, Friedman PA. Non-invasive detection of hyperkalemia with a smartphone electrocardiogram and artificial intelligence. *J Am Coll Cardiol* 2018;**71**:A272.
22. Perez M V, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L, Hung G, Lee J, Kowey P, Talati N, Nag D, Gummidipundi SE, Beatty A, Hills MT, Desai S, Granger CB, Desai M, Turakhia MP. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;**381**:1909–1917.