BMJ Open Ophthalmology

AI-based assessment of Clinical Activity Score and detection of active thyroid eye disease using facial images: validation of Glandy CAS

Kyubo Shin,[1] Jin-Sook Yoon ,[2] Jongchan Kim,[1] Jaemin Park,[1] Hyun Young Park,[2] Namju Kim ,[3] Min Joung Lee ,[4] Ho-Kyung Choung,[5] JaeSang Ko ,[2] Jae Hoon Moon [1,6,7]

To cite: Shin K, Yoon J-S, Kim J, *et al*. AI-based assessment of Clinical Activity Score and detection of active thyroid eye disease using facial images: validation of Glandy CAS. *BMJ Open Ophthalmology* 2025;**10**:e002264. doi:10.1136/bmjophth-2025-002264

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjophth-2025-002264).

KS and J-SY are joint first authors.

Received 10 April 2025
Accepted 23 August 2025

For numbered affiliations see end of article.

Correspondence to
Dr Jae Hoon Moon; jaehoon.moon@snu.ac.kr and

Dr JaeSang Ko; jaesangko@yonsei.ac.kr

ABSTRACT

Purpose The Clinical Activity Score (CAS) is widely used to assess thyroid eye disease (TED) activity but can vary based on the evaluator's expertise. We developed and externally validated Glandy CAS, a machine learning (ML)-assisted system for detecting active TED (CAS ≥3) using digital facial images. This clinical trial aimed to gain approval from the Korea Ministry of Food and Drug Safety (KMFDS) for this Software as a Medical Device (SaMD).

Methods This is a clinical trial based on the retrospective cohort. Glandy CAS analysed 756 photos of patients with TED, classifying them as having active or inactive TED. Its diagnostic performance was compared with that of three general ophthalmologists (less than 5 years of experience), using the F1 score. The reference CAS was determined by an oculoplastic specialist.

Results Active TED was detected in 207 of 756 patients. Glandy CAS achieved a sensitivity of 87.9%, specificity of 95.8% and an F1 score of 0.88. In comparison, general ophthalmologists had a sensitivity of 60.4%, specificity of 83.0% and an F1 score of 0.57. Glandy CAS predicted CAS within 1 point of the reference score in 82.3% of cases, with a mean absolute error of 0.83.

Conclusions Glandy CAS, an ML-assisted system for detecting active TED using facial images, showed high accuracy and outperformed general ophthalmologists. This system can consistently and accurately assess disease activity, facilitating early detection and timely treatment of active TED. Based on this clinical trial, the SaMD received KMFDS approval (Product Licence No., 24–93).

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Thyroid eye disease (TED) activity is traditionally assessed using the Clinical Activity Score (CAS), which can vary due to evaluator subjectivity. Accurate assessment requires experienced ophthalmologists, which limits accessibility and consistency. Machine learning (ML) systems have shown potential in medical diagnostics but lack external validation for TED activity assessment using digital facial images.

WHAT THIS STUDY ADDS

⇒ This study validates Glandy CAS, a ML-assisted system, demonstrating superior diagnostic performance compared with general ophthalmologists for detecting active TED. Glandy CAS achieved high sensitivity (87.9%) and specificity (95.8%) while closely aligning with the evaluations of experienced oculoplastic specialists, offering reliable assessments of TED activity from digital facial images.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Glandy CAS provides a standardised and accessible tool for TED assessment, reducing interobserver variability and enabling remote evaluations through telemedicine. Its approval as a Software as a Medical Device marks a significant step towards integrating artificial intelligence into routine clinical practice, potentially improving early detection, treatment initiation and patient outcomes in TED management.

INTRODUCTION

Thyroid eye disease (TED) is the most prevalent extrathyroidal complication of autoimmune thyroid disorders, such as Graves' disease and autoimmune thyroiditis. The clinical signs of TED are apparent in around 30–50% of individuals with Graves' disease. The management of TED relies on evaluating its clinical activity and severity using standardised criteria, and immunosuppressive therapy, such as intravenous glucocorticoids, is primarily considered for patients with inflammatory active TED of moderate-to-severe severity.[1] However, if timely intervention is not administered during the active phase of TED, the effectiveness of anti-inflammatory treatment is limited, potentially leading to more severe ocular deformities and functional impairments.[2 3] Thus, the early identification of the patients with active TED and the prompt initiation of appropriate treatment are crucial for enhancing TED prognosis.[1–3] American Thyroid Association and European Thyroid

Association published a consensus statement recommending that endocrinologists should be familiar with the evaluation of the activity and severity of TED because most patients with TED develop eye manifestations while being treated for hyperthyroidism and under the care of endocrinologists.[4]

Clinical Activity Score (CAS)[1] is the most widely used scoring system for assessing TED activity. It consists of seven binary (yes/no) items, with a score of 3 or higher indicating active TED.[5] These items focus on inflammatory symptoms and signs, such as orbital redness, swelling and pain.[5] A 10-item CAS is also sometimes used for follow-up patients. It includes additional points for increase of at least 2 mm in proptosis, decrease of at least 8° in any duction and decrease of visual acuity by two lines.[6] While CAS is a widely used tool for assessing disease activity in TED, its evaluation remains inherently subjective, often leading to interobserver variability despite standardisation efforts such as the Atlas proposed by the European Group on Graves' Orbitopathy (EUGOGO).[7 8] To address this challenge, our previous studies introduced a machine learning (ML)-assisted system designed to assess CAS from digital facial images, replicating the clinical evaluation process typically conducted by ophthalmologists.[9 10] Our system demonstrated clinically reasonable performance in evaluating TED inflammatory activity.

In this study, we validated the performance of the system using an external database and compared it with the accuracy of assessments made by general ophthalmologists with less than 5 years of clinical experience. This study was designed as a confirmatory clinical trial to obtain approval from the Korea Ministry of Food and Drug Safety (KMFDS) for a Software as Medical Device (SaMD).

METHODS

Study participants and facial images

This single-centre, retrospective study used digital facial image data sets and medical records from 756 patients with TED who visited the outpatient clinic of the ophthalmology department at Severance Hospital from August 2013 to July 2023. Patients aged 18 years or older with TED whose CAS was evaluated by oculoplastic specialists were eligible for inclusion. Their digital facial images were obtained at 1 m using a digital single-lens reflex camera (Canon EOS 650D or 800D, 2976×1984 resolution, Tokyo, Japan). The participant was instructed to immobilise his or her head with the eyes in the primary position, and the head position was examined to confirm the absence of obvious tilt or chin-up or chin-down position. All photographs were taken in the same clinical examination room under standardised conditions. A built-in flash was used to ensure consistent lighting, with ISO set to automatic and white balance set to flash mode. No postprocessing, including brightness adjustment or colour correction, was applied to the images.

All experiments were performed in accordance with relevant guidelines and regulations. All data were analysed exclusively within Severance Hospital, and neither the medical record information nor the subjects' facial images were taken outside the institution. The design and outcome measures of this study were approved by the South Korean Ministry of Food and Drug Safety for clinical trials as part of medical device approval process (Approval Number: 1529).

CAS evaluation

In this study, the CAS evaluation was based on the 7-point criteria recommended by the EUGOGO.[5 11] The CAS consists of seven items, each answered with a 'yes' or 'no'. These items assess spontaneous retrobulbar pain, pain on attempted up and down gaze, redness of the eyelids, redness of the conjunctiva, swelling of the eyelids, inflammation of the caruncle and/or plica and conjunctival oedema. The total score is the sum of all items, with a maximum score of 7 and active TED was defined as a CAS of 3 or higher.

In this study, CAS was evaluated by three different entities. First, oculoplastic specialists (J-SY and JK), who directly examined the patients during the clinical visits, assessed the CAS. Second, three general ophthalmologists with less than 5 years of clinical experience evaluated the CAS based on facial images taken at the time of the clinical visit. These general ophthalmologists had completed a 4-year ophthalmology residency programme and obtained board certification from the Korean Ophthalmological Society. At the time of the study, each had fewer than 5 years of postcertification clinical experience as a practising specialist. Lastly, the CAS was also assessed using the Glandy CAS, an ML-assisted system developed in our previous study,[9] which evaluated the CAS based on facial images taken at the time of the clinical visit. Each general ophthalmologist independently interpreted the patients' facial images and rated the five items relating to orbital swelling and redness by referring to a standard photographic colour atlas.[8] For the two items relating to orbital pain, information in the medical records was used in CAS evaluation by general ophthalmologists and Glandy CAS. The reference values for evaluating the performance of general ophthalmologists and Glandy CAS were in-person CAS evaluation by oculoplastic specialists.

Glandy CAS, a ML-assisted system evaluating CAS with digital facial images

In our previous study, we developed a ML-assisted system that mimics an ophthalmologist's decision-making process for evaluating the CAS in TED.[9] To construct the system, we trained artificial intelligence (AI) models using digital facial photographs from 1020 patients with varying degrees of inflammatory activity. For each image, the reference CAS was determined by three board-certified oculoplastic specialists, each with over 15 years of clinical experience at tertiary referral hospitals. The specialists independently assessed the five signs of orbital swelling and redness using standardised criteria recommended

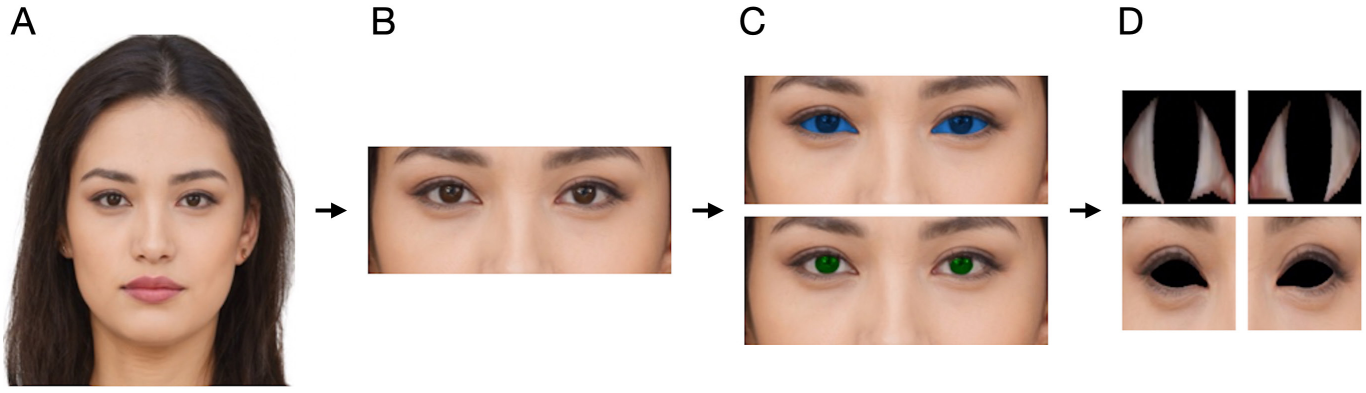

Figure 1 Image preprocessing module. The preprocessing module extracts diagnostic features from input images (A) in three steps: (B) face detection and cropping to isolate the ocular region, (C) segmentation of the eye and corneal regions, (D) generation of region-specific model inputs and preparation of image patches for sign-specific classifiers. This process enables the system to focus on relevant anatomical areas associated with each CAS-related sign. All facial images used in this figure are synthetically generated and do not depict real individuals. CAS, Clinical Activity Score.

by the EUGOGO and resolved any discrepancies through discussion to reach a consensus. Their assessments were then used as labels to train the AI models.[5 8 9 11]

The ML-assisted system was structured to reflect the three-step diagnostic logic employed by ophthalmologists: (1) evaluate the presence or absence of each inflammatory sign in both eyes, (2) determine whether the patient exhibits each sign if at least one eye is positive and (3) compute the CAS and determine TED activity status. The system architecture includes (1) a preprocessing module that detects facial landmarks and crops regions of interest (figure 1), (2) five AI models—each dedicated to detecting one of the CAS-related signs. Each of these sign-specific classifiers was based on the ResNet-18 architecture.[12] Finally, (3) a CAS prediction module that integrates these outputs with two symptom-based inputs (retrobulbar pain and gaze-evoked pain). A detailed structural diagram is shown in figure 2.

In a subsequent study,[10] we enhanced the robustness of the system across different image acquisition devices by incorporating a synthetic image augmentation strategy. This enabled the system to maintain high diagnostic accuracy even when applied to facial photographs taken with lower-end smartphones, outperforming junior ophthalmologists in comparative settings. In this current study, we used the previously established ML-assisted system[9 10] with two modifications: (1) the model was retrained with an updated output structure to predict both individual and consensus decisions of three expert graders and (2) we implemented a server-side image quality assurance step to ensure the suitability of input images. Full technical details, including model architecture, training

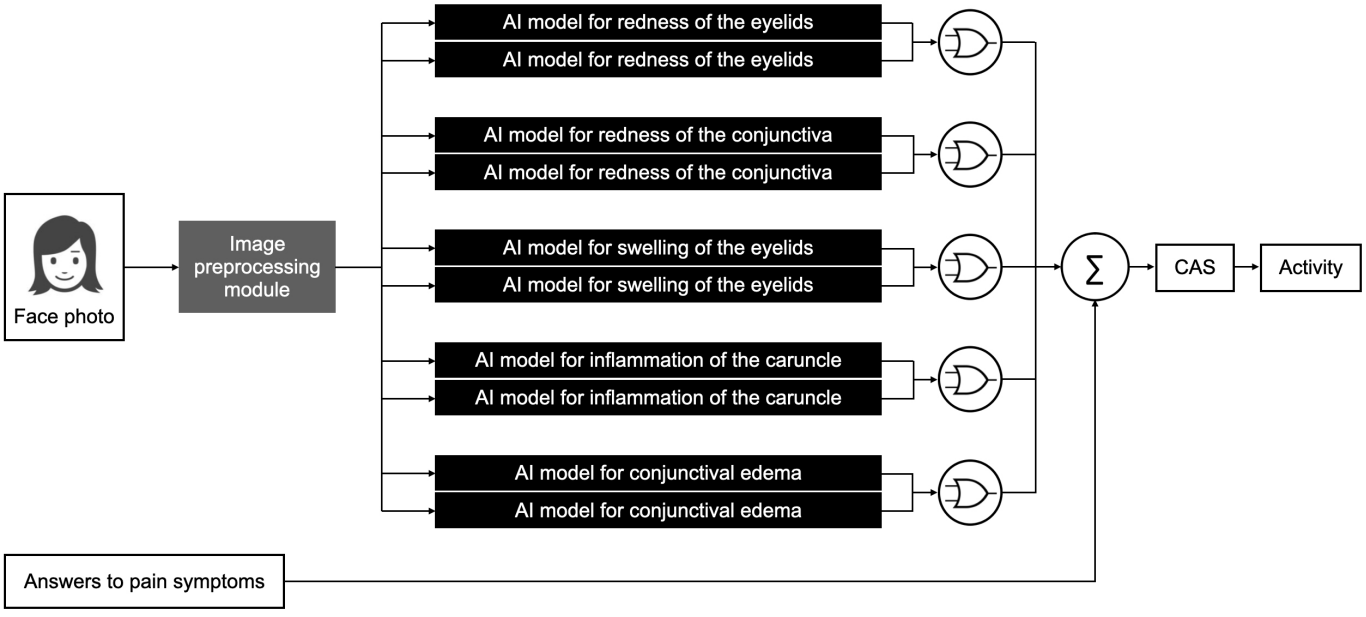

Figure 2 Structural diagram of ML-assisted system to predict active TED. AI, artificial intelligence; CAS, Clinical Activity Score; ML, machine learning; TED, thyroid eye disease.

procedures and quality metrics, are provided in the online supplemental materials.

To ensure data reliability, we incorporated an image quality assurance process prior to inference. The original system includes a two-step quality evaluation process. The first step, conducted on the client side (eg, smartphone or tablet), performs real-time assessment of photographing conditions during image capture. It checks whether the subject is smiling, whether facial rotation angles (pitch, yaw, roll) are within acceptable ranges, and whether the face's size and position meet predefined criteria. If any condition is not satisfied, image acquisition is halted and feedback is provided to the user. The second step, executed on the server side, evaluates whether the captured image meets standardised quality metrics (eg, ISO/IEC TR 29794-5[13]), verifies consistency in facial region and eye positioning, ensures corneal boundaries are clear and confirms that both eyes exhibit a sufficient margin reflex distance (MRD1). In this study, since the dataset consisted of precaptured images, the real-time client-side assessment was not applicable. However, the second, server-side quality evaluation was still implemented to ensure the appropriateness of the images for reliable CAS prediction.

Biochemical measurements

The serum levels of free thyroxine (T4) and thyroid stimulating hormone (TSH) were assessed using chemiluminescence immunoassays from Abbott. The normal reference ranges for free T4 and TSH were 0.80–1.23 ng/dL and 0.41–4.30 mIU/L, respectively. The presence of anti-TSH receptor antibodies was tested through an immunoassay provided by Roche, with the cut-off for considering it positive being >1.75 IU/L. Lab results obtained within 1 month of the facial photograph were included in the analysis. In this study, out of 756 participants, data on serum free T4 were available for 434 individuals, data on TSH were available for 435 individuals and data on anti-TSH receptor antibodies were available for 406 individuals.

Statistical analysis

Data with a normal distribution are presented as mean±SD, while those with a non-normal distribution are presented as median (IQR). The sensitivity and specificity of the ML-assisted system in identifying active TED were calculated using reference diagnoses provided by oculoplastic specialists who assessed the patients during their clinical visits. The F1 score, which is the harmonic mean of sensitivity and positive predictive value, was used to evaluate the performance of both the ML-assisted system and the evaluators.

In this study, the criteria for obtaining approval from the KMFDS for Glandy CAS were defined as demonstrating superior diagnostic performance in classifying active and inactive TED using facial images compared with the facial image interpretation performance of general ophthalmologists with less than 5 years of clinical experience, as measured by the F1 score. Superiority was defined as having the F1 score of Glandy CAS exceed that of general ophthalmologists by at least 0.0555, with the lower bound of the 95% CI for the F1 score of Glandy CAS being higher than the F1 score of general ophthalmologists. Based on the analysis of data from previous studies, the F1 score of Glandy CAS was 0.70, while the median F1 score of general ophthalmologists was 0.6445. Therefore, in this study, if the difference in F1 scores is greater than in the previous studies and the lower bound of the CI for the F1 score of Glandy CAS is higher than 0.6445, the diagnostic performance of Glandy CAS in identifying active TED was considered superior to that of general ophthalmologists.

Table 1 Baseline characteristics

Age (year)	**39.6±11.5**
Sex (male/female)	163/593
Thyroid function test*	
Free T4 (ng/dL)	1.32±1.01
TSH (mIU/L)	0.19 (0.01–1.925)
Anti-TSH receptor antibody*	5.08 (1.74–14.68)
CAS (total score)	1.4±1.7
Spontaneous retrobulbar pain (yes/no)	108/648
Pain on attempt up and down gaze (yes/no)	62/694
Redness of eyelids (yes/no)	139/617
Redness of conjunctiva (yes/no)	177/579
Swelling of eyelids (yes/no)	267/489
Inflammation of the caruncle and/or plica (yes/no)	67/689
Conjunctival oedema (yes/no)	53/703

Data are expressed as mean±SD or median (IQR).
*Data on serum free T4, TSH and anti-TSH receptor antibodies were available for 434, 435 and 406 individuals, respectively.
CAS, Clinical Activity Score; T4, thyroxine; TSH, thyroid stimulating hormone.

Patient and public involvement

There was no scope for involving patients or the public in the design, conduct, reporting or dissemination plans of this study.

RESULTS

Participant characteristics

Table 1 outlines the baseline characteristics of the 756 participants in this study. The average age was 39.6±11.5 years, and 78.4% of the participants were female. Anti-TSH receptor antibodies were detected (>1.75 U/mL) in 75.4% (306 of 406 participants whose anti-TSH receptor antibody data were available) of the patients. The mean CAS was 1.3±1.7, with eyelid swelling being the most common inflammatory symptom or sign. Active thyroid-associated orbitopathy (TAO) (CAS ≥3) was

Table 2 Diagnostic performance of general ophthalmologists and Glandy CAS for detecting active TED

	Sensitivity (95% CI), %	Specificity (95% CI), %	F1 score (95% CI)	AUC (95% CI)
General ophthalmologist 1	94.2 (91.1 to 97.2)	55.6 (51.4 to 59.9)	0.60 (0.56 to 0.65)	0.749 (0.721 to 0.775)
General ophthalmologist 2	58.9 (52.4 to 65.9)	94.7 (92.9 to 96.5)	0.68 (0.63 to 0.74)	0.768 (0.743 to 0.803)
General ophthalmologist 3	28.0 (22.3 to 34.1)	98.7 (97.8 to 99.6)	0.43 (0.36 to 0.50)	0.634 (0.605 to 0.665)
Average of general ophthalmologists	60.4 (56.5 to 64.3)	83.0 (81.3 to 84.8)	0.57 (0.53 to 0.61)	0.717 (0.696 to 0.739)
Glandy CAS	87.9 (82.7 to 92.0)	95.8 (92.9 to 96.7)	0.88 (0.84 to 0.90)	0.915 (0.891 to 0.939)

AUC, area under the curve; CAS, Clinical Activity Score; TED, thyroid eye disease.

observed in 207 patients (27.4%), and highly active TAO (CAS ≥5) in 41 patients (5.4%).

Diagnostic performance of the ML-assisted system for detecting active TAO

Table 2 presents the performance of three general ophthalmologists and the Glandy CAS in detecting active TED through the analysis of digital facial images. The general ophthalmologists showed varying performance, with one exhibiting high sensitivity but low specificity, and others showing the opposite trend—indicating a tendency to either overdiagnose or underdiagnose active TED. In contrast, Glandy CAS demonstrated both high sensitivity (87.9%; 95% CI, 82.7 to 92.0) and specificity (95.8%; 95% CI, 92.9 to 96.7), achieving a higher F1 score (0.88; 95% CI, 0.84 to 0.90) than the general ophthalmologists.

Estimation of CAS

Table 3 shows the CAS prediction performance of Glandy CAS. Our ML-assisted system accurately predicted the CAS within 1 point of the in-person CAS evaluation by oculoplastic specialists in 82.3% of cases. Glandy CAS classified 95.8% of cases with a CAS evaluation of 5 or higher by oculoplastic specialists (indicating highly active TED) as active TED. Additionally, it correctly identified 96.2% of cases with a CAS evaluation of 2 or lower (clearly inactive TED) as inactive TED. The mean absolute error between Glandy CAS and the reference CAS was 0.83, indicating that on average, the system's predictions deviated from expert evaluations by less than one point on the 7-point CAS scale.

Table 3 Estimation of the CAS by Glandy CAS using the facial images

		Estimated CAS							
		0	1	2	3	4	5	6	7
Reference CAS	0	124	167	56	2	2	0	0	0
	1	22	64	25	11	3	0	0	0
	2	10	32	22	8	1	0	0	0
	3	0	2	7	34	15	9	0	0
	4	0	1	13	37	20	15	6	0
	5	0	0	2	11	9	6	3	2
	6	0	0	0	2	1	4	5	0
	7	0	0	0	0	0	0	1	2

The reference CAS was in-person CAS evaluation of oculoplastic specialists during the clinical visit. Cells with bold borders indicate results where the difference between the estimated CAS and reference CAS was ≤1. Cells with a grey fill indicate exact agreement between the estimated CAS and reference CAS.
CAS, Clinical Activity Score.

DISCUSSION

This study evaluated the diagnostic performance of the Glandy CAS, a ML-assisted system, in identifying active TED using digital facial images. Our results indicate that Glandy CAS exhibits superior accuracy in diagnosing active TED compared with general ophthalmologists with less than 5 years of clinical experience. Accordingly, Glandy CAS has been approved by the KMFDS as a SaMD (Product Licence No., 24–93). Furthermore, the high concordance between the active TED detection by Glandy CAS and the in-person diagnoses made by specialists who examined the patients is highly encouraging. These findings have significant implications for the clinical management of TED, particularly in settings where specialist expertise is limited. Considering that the reference CAS data used for both training and performance validation of the Glandy CAS AI model were not derived from in-person clinical examinations but rather from the interpretation of facial images by experienced oculoplastic specialists,[9 10] these results are particularly intriguing. This is likely because all three oculoplastic specialists who interpreted the facial images to build the reference CAS data have over 15 years of clinical experience, and their consensus was used as the reference CAS. Moreover, one of these specialists had directly examined the patients whose facial images were used, which likely increased the reliability of the reference CAS, contributing to these favourable outcomes. The ability of Glandy CAS to predict CAS within one point of the reference score in 82.3% of cases suggests that this system can provide reliable assessments comparable to those of experienced specialists. Additionally, the classification of 95.8% of highly active TED cases with a CAS of 5 or higher as active TED, and the classification of clearly

inactive TED cases with a CAS of 2 or lower as inactive TED, support the reliability and safety of this ML-assisted system when deployed in actual clinical settings.

Traditional evaluation of TED activity heavily relies on the expertise of experienced ophthalmologists, which can introduce variability and subjectivity into the assessment process.[7] Moreover, most patients with TED manifest eye symptoms during treatment for Graves' disease or hypothyroidism under endocrinological care, subsequently necessitating referral to ophthalmology for TED treatment. In this process, if an endocrinologist is not familiar with CAS evaluation, timely referral for ophthalmological assessment becomes challenging. Even among ophthalmologists, as observed in the results of this study, the accuracy of CAS evaluation varies depending on clinical experience. The general ophthalmologists set as comparators for the ML-assisted system were board-certified ophthalmologists who had completed their residency training at a tertiary hospital but were not oculoplastic specialists, with less than 5 years of clinical experience. To reflect real-world clinical practice, although no additional training on image interpretation was provided specifically for this study, the participants were instructed to refer to the atlas for CAS evaluation published by EUGOGO.[8] On the contrary, Glandy CAS offers a standardised and objective approach, potentially reducing interobserver variability and improving the consistency of TED assessments across different clinical settings. Additionally, the use of digital facial images for assessment aligns with telemedicine practices, enabling remote patient evaluation and serving as a valuable tool in regions with limited access to specialised care. In our recent study, we reported achieving clinically reliable accuracy in active TED detection by training Glandy CAS to analyse not only studio-acquired photographs but also selfie images captured with smartphone cameras.[10] This suggests the possibility of extending usability to continuously monitor TED activity not only during telemedicine sessions using specific devices but also through patients' smartphones. Consistent and accurate CAS evaluation and continuous monitoring using such ML-assisted systems will contribute to the early diagnosis and timely treatment of active TED requiring intervention, thereby improving patient outcomes.

In the present study, we focused on evaluating CAS using the 7-item criteria. Although not all included patients were new, the three additional items of the 10-item CAS are designed specifically for follow-up assessments that require comparison with previous examinations. Since our data set consisted of single-time point facial images, it was not feasible to evaluate these follow-up items. Furthermore, two of the three additional signs—extraocular muscle restriction and visual acuity decrease—cannot be reliably assessed from facial photographs. Therefore, the 7-item CAS was used as the primary measure of disease activity, which is applicable to both new and follow-up patients. In future work, our previously developed AI tools for assessing lid retraction and proptosis from facial images[14] may support longitudinal monitoring of TED progression using digital images.

In recent years, AI has been increasingly used to improve the management of TED, covering areas such as diagnosis, staging, grading and treatment planning.[15] However, most current studies have focused on using orbital CT or MRI scans to diagnose TED,[16] measure exophthalmos,[17 18] evaluate inflammatory activity[19–21] or predict treatment responses.[22] As a result, patients still need to undergo expensive imaging tests, making it difficult to move away from conventional TED management. Some studies have explored using facial images for TED diagnosis. Karlin *et al*[23] introduced a deep learning-based classifier designed to detect the presence of TAO using facial images. However, their approach did not specifically target the active phase of TAO, which is typically assessed through CAS. Their classifier, an ensemble of ResNet-18 networks, leveraged facial disfigurement as a key feature of TAO, achieving an accuracy of 89.2% (sensitivity: 93.4%; specificity: 86.9%). Similarly, Huang *et al*[24] developed a diagnostic system employing a deep learning-based semantic segmentation network to identify ocular dyskinesia and a classification network to recognise other TAO-related signs. Their model's performance, measured by area under the curve, ranged from 0.60 to 0.94 depending on the specific clinical sign. Glandy CAS estimates the CAS, the best-validated scoring system for assessing TED activity, and detects inflammatory active TED. Moreover, the current study demonstrated that the CAS evaluation performance of Glandy CAS closely matched the evaluations performed by oculoplastic specialists who examined the patients in person. Therefore, Glandy CAS has the significant advantage of performing clinical tasks consistently at an expert level, making it highly useful in clinical practice.

While the results of this study are promising, there are some limitations to consider. This study is an external validation aimed at examining the active TED detection performance of Glandy CAS. However, the study's retrospective design and the use of a single-centre data set may limit the generalisability of the findings. Future research should focus on prospective, multicentre studies to validate the performance of Glandy CAS and its clinical utility across diverse populations and clinical settings. Specifically, as Glandy CAS has been trained solely on facial images of the Korean population thus far, its performance across other ethnicities remains uncertain. Therefore, it is necessary to enhance the system by training it with facial images and clinical data from various ethnic groups.

In conclusion, Glandy CAS represents a significant advancement in the assessment of TED activity using digital facial images. Its superior diagnostic performance compared with general ophthalmologists highlights its potential to improve the early detection and timely management of active TED, ultimately enhancing patient outcomes. The approval of Glandy CAS by the KMFDS underscores its clinical value and paves the way for its

incorporation into standard clinical practice. Future research should continue to explore its applications and refine its algorithms to further optimise its diagnostic capabilities.

Code availability

The underlying algorithm is copyrighted by THYROSCOPE INC. and will not be available to the public. The authors agree to apply the algorithm to data provided by other academic researchers on their behalf for research purposes only.

Author affiliations
[1]THYROSCOPE INC, Ulsan, Korea (the Republic of)
[2]Department of Ophthalmology, Severance Hospital, Institute of Vision Research, Yonsei University College of Medicine, Seoul, Korea (the Republic of)
[3]Department of Ophthalmology, Seoul National University Bundang Hospital and Seoul National University College of Medicine, Seongnam, Korea (the Republic of)
[4]Department of Ophthalmology, Hallym University Sacred Heart Hospital and Hallym University College of Medicine, Chuncheon, Korea (the Republic of)
[5]Department of Ophthalmology, Seoul Metropolitan Government Seoul National University Boramae Medical Center and Seoul National University College of Medicine, Seoul, Korea (the Republic of)
[6]Department of Internal Medicine, Seoul National University Bundang Hospital and Seoul National University College of Medicine, Seongnam, Korea (the Republic of)
[7]Center for Artificial Intelligence in Healthcare, Seoul National University Bundang Hospital, Seongnam, Korea (the Republic of)

Contributors KS: writing—original draft; analysing and interpreting the data; methodology. J-SY: writing—original draft; collecting the data. JK: analysing and interpreting the data; methodology. JP: analysing and interpreting the data; methodology. HYP: analysing and interpreting the data; collecting the data. NK: review and editing; analysing and interpreting the data. MJL: review and editing; analysing and interpreting the data. H-KC: review and editing; analysing and interpreting the data. JK: conceptualisation; collecting the data; writing—review and editing. JHM: conceptualisation; collecting the data; writing—original draft; writing—review and editing. JHM and JK are the guarantor.

Funding This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number, HI22C1078), a grant of the AI-based medical system digital transformation support programme through the National IT Industry Promotion Agency (NIPA) funded by the Ministry of Science, ICT (grant number, H0906-24-1006) and a grant of the Korean Thyroid Association Clinical Research Award 2019.

Competing interests THYROSCOPE has a patent for the ML-assisted system to predict the clinical activity of TED. JHM: a stock owner of THYROSCOPE. JP: a stock owner of THYROSCOPE. J-SY: a member of the medical advisory board of THYROSCOPE. NK: a member of the medical advisory board of THYROSCOPE. JK: a member of the medical advisory board of THYROSCOPE. KS: no competing interests. JK: no competing interests. HYP: no competing interests. MJL: no competing interests. H-KC: no competing interests.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the Severance Hospital Institutional Review Board (IRB No. 1-2023-0050), and the requirement for informed consent was waived by the Severance Hospital Institutional Review Board given that the study design was based on a retrospective review of medical records and preobtained facial images.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. No data are available. The raw data including medical records and facial images are not available upon request due to ethical and legal restrictions imposed by the Severance Hospital Institutional Review Board. The original data are derived from the institutions' electronic health records and contain patients' protected health information. Deidentified data are available from Severance Hospital for researchers who meet the criteria for access to confidential data and have a data usage agreement with the hospital.

ORCID iDs
Jin-Sook Yoon http://orcid.org/0000-0002-8751-9467
Namju Kim http://orcid.org/0000-0002-4815-0804
Min Joung Lee http://orcid.org/0000-0003-4130-6092
JaeSang Ko http://orcid.org/0000-0002-3011-7213
Jae Hoon Moon http://orcid.org/0000-0001-6327-0575

REFERENCES

1 Bartalena L, Kahaly GJ, Baldeschi L, *et al*. The 2021 European Group on Graves' orbitopathy (EUGOGO) clinical practice guidelines for the medical management of Graves' orbitopathy. *Eur J Endocrinol* 2021;185:G43–67.
2 Gontarz-Nowak K, Szychlińska M, Matuszewski W, *et al*. Current Knowledge on Graves' Orbitopathy. *J Clin Med* 2020;10:16.
3 Dolman PJ. Evaluating Graves' orbitopathy. *Best Pract Res Clin Endocrinol Metab* 2012;26:229–48.
4 Burch HB, Perros P, Bednarczuk T, *et al*. Management of thyroid eye disease: a Consensus Statement by the American Thyroid Association and the European Thyroid Association. *Eur Thyroid J* 2022;11:e220189.
5 Bartalena L, Baldeschi L, Dickinson AJ, *et al*. Consensus statement of the European group on Graves' orbitopathy (EUGOGO) on management of Graves' orbitopathy. *Thyroid* 2008;18:333–46.
6 Burch HB, Perros P, Bednarczuk T, *et al*. Management of Thyroid Eye Disease: A Consensus Statement by the American Thyroid Association and the European Thyroid Association. *Thyroid* 2022;32:1439–70.
7 Perros P, Žarković M, Pearce SH, *et al*. Inter-observer Variability of Clinical Activity Score: Assessments in Patients With Thyroid Eye Disease. *Am J Ophthalmol* 2023;252:94–100.
8 Dickinson AJ, Perros P. Controversies in the clinical evaluation of active thyroid-associated orbitopathy: use of a detailed protocol with comparative photographs for objective assessment. *Clin Endocrinol (Oxf)* 2001;55:283–303.
9 Moon JH, Shin K, Lee GM, *et al*. Machine learning-assisted system using digital facial images to predict the clinical activity score in thyroid-associated orbitopathy. *Sci Rep* 2022;12:22085.
10 Shin K, Choung H, Lee MJ, *et al*. A Preliminary Evaluation of the Diagnostic Performance of a Smartphone-Based Machine Learning-Assisted System for Evaluation of Clinical Activity Score in Digital Images of Thyroid-Associated Orbitopathy. *Thyroid* 2024;34:744–52.
11 Wiersinga WM, Perros P, Kahaly GJ, *et al*. Clinical assessment of patients with Graves' orbitopathy: the European Group on Graves' Orbitopathy recommendations to generalists, specialists and clinical researchers. *Eur J Endocrinol* 2006;155:387–9.
12 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 770–8. Las Vegas, NV, USA.
13 ISO/IEC JTC 1/SC 37 Biometrics. ISO/IEC TR 29794-5:2010 information technology — Biometric sample quality — Part 5: face image data: international organization for standardization. 2010.
14 Park J, Yoon JS, Kim N, *et al*. Deep Learning-Driven Exophthalmometry through Facial Photographs in Thyroid Eye Disease. *Ophthalmol Sci* 2025;5:100791.
15 Yi C, Niu G, Zhang Y, *et al*. Advances in artificial intelligence in thyroid-associated ophthalmopathy. *Front Endocrinol (Lausanne)* 2024;15:1356055.

16 Song X, Liu Z, Li L, *et al*. Artificial intelligence CT screening model for thyroid-associated ophthalmopathy and tests under clinical conditions. *Int J Comput Assist Radiol Surg* 2021;16:323–30.
17 Zhang Y, Rao J, Wu X, *et al*. Automatic measurement of exophthalmos based orbital CT images using deep learning. *Front Cell Dev Biol* 2023;11:1135959.
18 Fu R, Bandos A, Leader JK, *et al*. Artificial Intelligence Automation of Proptosis Measurement: An Indicator for Pediatric Orbital Abscess Surgery. *Ophthalmol Ther* 2023;12:2479–91.
19 Lin C, Song X, Li L, *et al*. Detection of active and inactive phases of thyroid-associated ophthalmopathy using deep convolutional neural network. *BMC Ophthalmol* 2021;21:39.
20 Li Y, Ma J, Xiao J, *et al*. Use of extreme gradient boosting, light gradient boosting machine, and deep neural networks to evaluate the activity stage of extraocular muscles in thyroid-associated ophthalmopathy. *Graefes Arch Clin Exp Ophthalmol* 2024;262:203–10.
21 Lee J, Lee S, Lee WJ, *et al*. Neural network application for assessing thyroid-associated orbitopathy activity using orbital computed tomography. *Sci Rep* 2023;13:13018.
22 Wang Y, Wang H, Li L, *et al*. Novel observational study protocol to develop a prediction model that identifies patients with Graves' ophthalmopathy insensitive to intravenous glucocorticoids pulse therapy. *BMJ Open* 2021;11:e053173.
23 Karlin J, Gai L, LaPierre N, *et al*. Ensemble neural network model for detecting thyroid eye disease using external photographs. *Br J Ophthalmol* 2023;107:1722–9.
24 Huang X, Ju L, Li J, *et al*. An Intelligent Diagnostic System for Thyroid-Associated Ophthalmopathy Based on Facial Images. *Front Med* 2022;9:920716.