AMERICAN ACADEMY OF OPHTHALMOLOGY®

A Deep Learning Model for Screening Computed Tomography Imaging for Thyroid Eye Disease and Compressive Optic Neuropathy

Lisa Y. Lin, MD, MBE,[1,] Paul Zhou, MD, MS,[2,*] Min Shi, PhD,[3] Jonathan E. Lu, MD,[1] Soomin Jeon, PhD,[4] Doyun Kim, PhD,[5] Josephine M. Liu,[6] Mengyu Wang, PhD,[3] Synho Do, MS, PhD,[6] Nahyoung Grace Lee, MD[1]*

Purpose: Thyroid eye disease (TED) is an autoimmune condition with an array of clinical manifestations, which can be complicated by compressive optic neuropathy. It is important to identify patients with TED early to ensure close monitoring and treatment to prevent potential permanent disability or vision loss. Deep learning artificial intelligence (AI) algorithms have been utilized in ophthalmology and in other fields of medicine to detect disease. This study aims to introduce a deep learning model to evaluate orbital computed tomography (CT) images for the presence of TED and potential compressive optic neuropathy.

Design: Retrospective review and deep learning algorithm modeling.

Subjects: Patients with TED with dedicated orbital CT scans and with an examination by an oculoplastic surgeon over a 10-year period at a single academic institution. Patients with no TED and normal CTs were used as normal controls. Those with other diagnoses, such as tumors or other inflammatory processes, were excluded.

Methods: Orbital CTs were preprocessed and adopted for the Visual Geometry Group-16 network to distinguish patients with no TED, mild TED, and severe TED with compressive optic neuropathy. The primary model included training and testing of all 3 conditions. Binary model performance was also evaluated. An oculoplastic surgeon was also similarly tested with single and serial images for comparison.

Main Outcome Measures: Accuracy of deep learning model discernment of region of interest for CT scans to distinguish TED versus normal control, as well as TED with clinical signs of optic neuropathy.

Results: A total of 1187 photos from 141 patients were used to develop the AI model. The primary model trained on patients with no TED, mild TED, and severe TED had 89.5% accuracy (area under the curve: range, 0.96–0.99) in distinguishing patients with these clinical categories. In comparison, testing of an oculoplastic surgeon in these 3 categories showed decreased accuracy (70.0% accuracy in serial image testing).

Conclusions: The deep learning model developed in the study can accurately detect TED and further detect TED with clinical signs of optic neuropathy based on orbital CT. The model proved superior compared with human expert grading. With further optimization and validation, this TED deep learning model could help guide frontline health care providers in the detection of TED and help stratify the urgency of a referral to an oculoplastic surgeon and endocrinologist.

Financial Disclosure(s): The authors have no proprietary or commercial interest in any materials discussed in this article. *Ophthalmology Science 2024;4:100412 © 2023 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Thyroid eye disease (TED) is an autoimmune condition that, if delayed in diagnosis and management, can result in significant morbidity including pain, diplopia, disfigurement, and vision loss. Timely diagnosis, therefore, is paramount in preventing permanent sequelae. However, the diagnosis of TED is not always straightforward, as there exists a wide range of clinical manifestations, requiring a variety of clinical and functional tests interpreted by a specialist to diagnose and categorize the severity of disease. As a result, many patients present with late or severe stages of the disease.

Machine learning algorithms have been used in medicine for efficient automated screening. These artificial intelligence (AI) models are based on recognizing patterns in various visual data sets including fundus photography,[1–3] external photography of the periocular structures,[4] magnetic resonance imaging (MRI),[5] and computed tomography (CT). Artificial intelligence has also been incorporated in diagnosing orbital diseases, such as blowout fractures, orbital cavernous venous malformations, and TED.[6–9] Computed tomography scans of the face and head are

https://doi.org/10.1016/j.xops.2023.100412
ISSN 2666-9145/23

commonly ordered by an array of nonophthalmology providers and performed for a wide variety of indications, and therefore represent an opportunity for TED detection.

Various neural network-based methods have shown early promise for TED screening and detection.[5,10] However, there are limitations of the existing literature that prevent generalizability to patients in the United States and ability to scale the model for widespread use. Additionally, important clinical parameters such as compressive optic neuropathy were not included in these models.

To aid in earlier and more accurate detection of TED, the authors hereby seek to develop an accurate, efficient deep learning model to evaluate CT scans for the presence of TED. The ideal model will be able to triage the detected disease by severity, and particularly assess for the presence of compressive optic neuropathy.

Methods

A retrospective cohort study at Massachusetts Eye and Ear, a tertiary ophthalmic institution, over a 10-year period (August 2011 to 2021) was performed. The Massachusetts General Brigham institutional review board approved this retrospective study and waived the required written informed consent. The study was conducted following the ethical standards outlined in the Declaration of Helsinki and conducted in compliance with the Health Insurance Portability and Accountability Act.

Subjects were patients aged $>$ 18 years with a CT scan of the orbit who had a complete examination by an oculoplastic surgeon within 3 months of the scan being performed. Patients with a clinical diagnosis of TED were included. Patients with no TED or other orbital conditions who underwent CT orbits were included as normal controls. Patients were excluded if they had another orbital diagnosis, such as orbital tumors, fractures, or other inflammatory processes, and patients with any prior orbital surgery (e.g., orbital decompressions) were excluded. Demographics, clinical history, clinical examination from the oculoplastic surgeons' examination that was closest in time to the CT, and ancillary testing, including automated perimetry and laboratory testing, were abstracted from the electronic medical record.

Based on retrospective chart review, patients were categorized into 3 groups: normal, mild TED (no optic neuropathy), and severe TED (with optic neuropathy). Each orbit was categorized separately due to the asymmetric nature of TED. The orbits were categorized based on clinical diagnosis, as measured and recorded in the chart by the oculoplastic surgeon, in conjunction with laboratory results and ancillary testing. Currently, the gold standard diagnosis of TED is based on clinical examination and therefore this was used for categorization. As TED is a heterogenous disease, the distinguishing factor between mild and severe TED was the presence of optic neuropathy. Mild TED was defined as an orbit with clinical signs of TED but without any clinical signs of optic neuropathy. Severe TED was defined as orbits with clinical signs of TED concurrent with signs of compressive optic neuropathy, including any one of the following: dyschromatopsia, Humphrey visual field changes in patterns consistent with TED,[11] relative afferent pupillary defect, and optic nerve head changes. These clinical categories were deemed the ground truth for deep learning model training.

Data Preprocessing

To assess the deep learning model's functionality in screening for thyroid eye disease and compressive optic neuropathy based on a single discrete image, the unprocessed images were obtained by manually cropping around the orbital region of interest. Right and left orbits were analyzed individually. The TED data set was then preprocessed for optimization in the deep learning model. The original 2-dimensional images were varied sizes, where the image width varied from 102 to 428 pixels and the image height varied from 75 to 410 pixels (Fig 1A). To accommodate all image samples and fit the input of the adopted deep learning model in this work, the images were resized to 256 $\times$ 256 and further transformed to the 3-channel Red Green Blue image (Fig 1B). The 2-dimensional region of interest images were split into 3 predetermined categories (normal, mild TED, and severe TED).

Model Training and Evaluation

The deep learning method for TED status prediction was trained based on the Visual Geometry Group (VGG)-16 network,[11] which classified the TED samples into 4 different experimental settings: (1) VGG-16 trained for normal, mild, and severe samples; (2) VGG-16 trained only for normal and mild samples; (3) VGG-16 trained only for normal and severe samples; and (4) VGG-16 trained only for mild and severe samples. For each experiment, 80% of the data were randomly sampled for training the model and the remaining 20% of image samples were held for testing the model prediction performance.

As shown in Figure 1C, the model took a 3-channel Red Green Blue image of size 256 $\times$ 256 $\times$ 3 as input, which subsequently went through 5 convolutional blocks to extract features. The first and second blocks each contained 2 convolutional layers followed by a max-pooling layer which aims to reduce the size of feature map. The next 3 convolutional blocks each contained 3 convolutional layers in a sequence followed by a max-pooling layer. The feature maps output from the fifth convolutional block was finally flattened in a feature vector which was used for TED status prediction after a densely connected layer together with a dropout layer. Note that the dense layer could output a 3-dimensional vector or a single scalar value, depending on the experimental settings. For example, when training the model to classify normal, mild, and severe samples all together (i.e., the 3-class prediction setting), the final output was a 3-dimensional vector which can be transformed to the respective class probabilities through a SoftMax function transformation. However, when training the model to classify 2 classes (i.e., the binary-class prediction setting, e.g., normal vs. mild, normal vs. severe, and mild vs. severe), the final output was a single scalar to indicate the binary prediction probability through a sigmoid function transformation.

To stabilize and enhance the model training, the pretrained VGG-16 model was trained on the ImageNet data set to initialize the model.[12] The model on the TED data set was fine-tuned with a total number of 100 epochs and a mini-batch size of 8. The Root Mean Square Propagation (RMSProp)[13] was adopted to optimize a categorical cross-entropy loss for the 3-class prediction setting,[14] and a binary cross-entropy loss for the binary-class prediction.[15] For all model training settings, the learning rate was set as 0.00001.

For each category, the deep learning model screens and decides based on the salient features of the input image. Figure 2 illustrates examples with the Grad-CAM (Gradient-weighted Class Activation Mapping) technique which demonstrates the salient features learned by the adopted deep learning model.

Comparison to Human Expert

To assess the model's abilities against human graders, a subset of images across the 3 categories were randomly selected from the data set. The images were randomized and presented to an oculoplastic

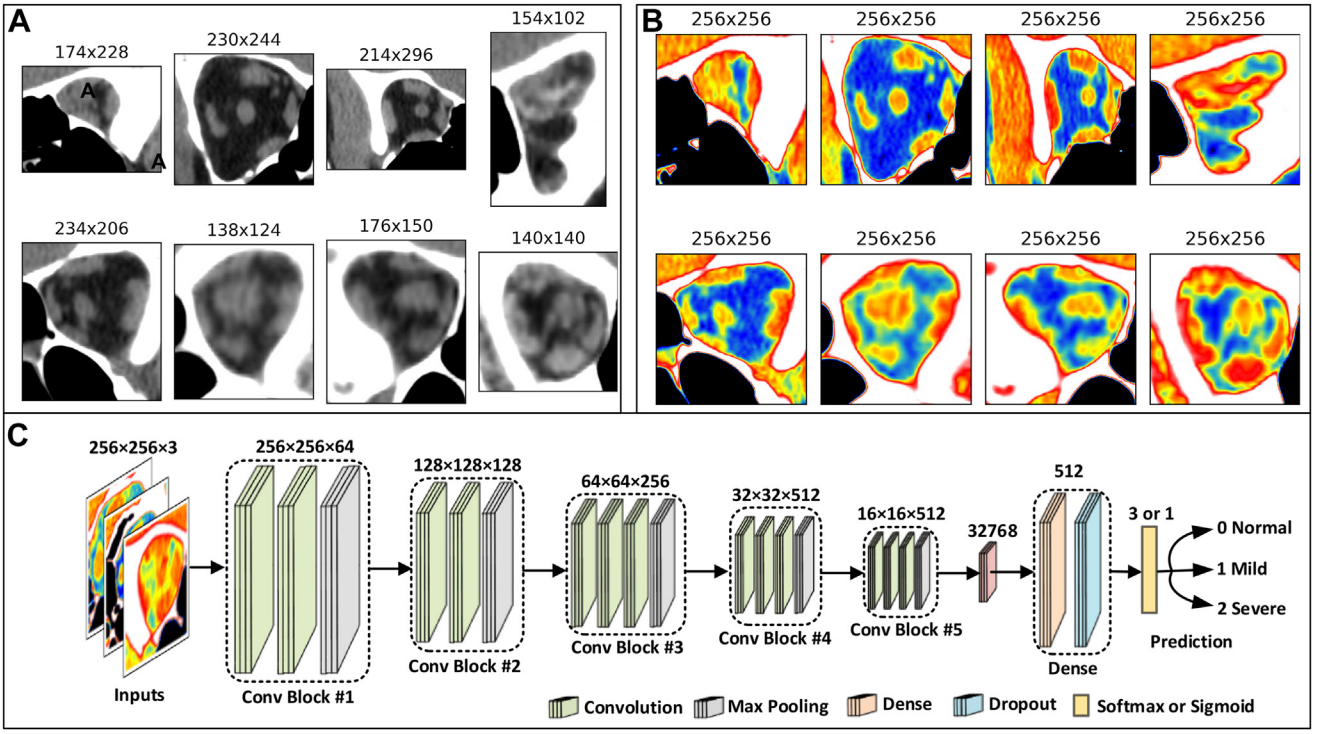

Figure 1. Data preprocessing and Visual Geometry Group (VGG)-16 model training. **A,** A sample of original 2-dimensional orbital computed tomography (CT) images of orbital CTs is demonstrated here, with varying image width and height. **B,** The images were resized to 256 × 256 and further transformed to the 3-channel Red Green Blue image. **C,** The adopted deep learning framework for thyroid eye disease status prediction based on the VGG-16 network. The size of the feature map at each layer is shown on the top of each block, e.g., for the convolutional block # 1, 256 × 256 × 64 means the size of feature map is 256 × 256 and the adopted number of convolutional filters is 64.

surgeon (N.G.L.) who was masked to the clinical information for grading. The grader (N.G.L.) was tasked with categorizing the images into normal, mild TED, and severe TED. The grader was tested on discrete snapshot images, directly obtained from the image data set, and serial cross sectional images, to simulate the real clinical assessment process. Performance metrics were calculated based on the number of eyes correctly identified as belonging to the normal control, mild TED group, or the severe TED group.

Results

One hundred forty-one patients were identified who met the criteria for inclusion in the study. Thirty-one patients were included in the normal control group and a total of 123 patients were included in the mild and severe TED groups. The median age of all participants was 71 years (range: 25−95 years). One hundred four (74%) patients were female, and 75% of patients identified as White. Patient demographics are presented in Table 1. A total of 1187 images were included in the data set: 373 images from 31 patients (32 eyes) were controls, 459 images from 83 patients (84 eyes) with mild TED, and 355 images from 40 patients (76 eyes) with severe TED. Thirteen patients were included in both the mild and severe TED groups as they demonstrated features of mild TED in 1 eye and severe TED with compressive optic neuropathy in the other.

The deep learning model was trained and tested with samples from all 3 groups (normal, mild TED, and severe TED). The training data set (approximately 80% of total images) consisted of 296 normal images, 375 mild TED images, and 278 severe TED images. The testing data set (238 images, approximately 20% of total images) consisted of 77, 84, and 77 images from the normal control, mild TED, and severe TED groups, respectively.

The primary model experiment was performed by training the model on all 3 categories of images: normal, mild TED, and severe TED. This model achieved an overall accuracy of 89.5% (confusion matrix presented in Fig 3A). The receiver operating characteristic curve demonstrates the area under the curve (AUC) performance of the model at 3 classification thresholds (Fig 3B), and ranged from 0.96 to 0.99. The model misclassified 3 severe TED images as mild TED, but no severe cases were classified as normal (Fig 3C).

Binary modeling was performed with normal and mild TED images. The training data set consisted of 289 images from the normal group and 376 images from the mild group, and the testing data set had 84 normal images and 84 mild images. Visual Geometry Group-16 achieved an overall accuracy of 93.4% in determining the normal images and mild TED CT images with an AUC of 0.982 (Fig 4A1, 2). The model misclassified 9 mild TED images as normal.

Binary modeling for normal and severe TED groups only had a training data set with 297 images from the normal group and 285 images from the severe TED group, and a

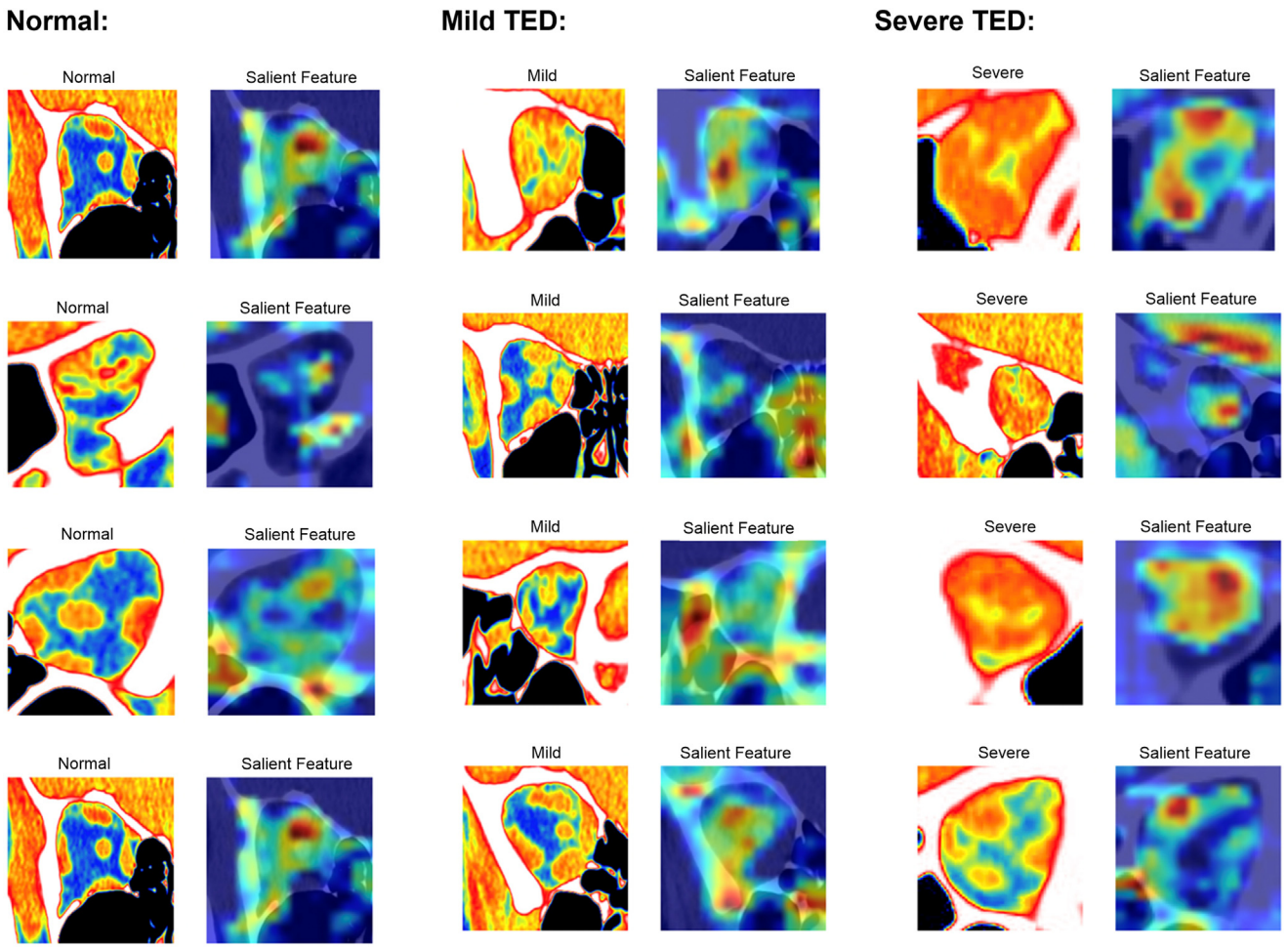

Figure 2. Salient feature maps. Feature importance map of representative salient features learned by the adopted deep learning model for all 3 groups (normal [left panel], mild thyroid eye disease [TED] [middle panel], and severe TED [right panel]).

testing data set with 76 images from the normal group and 70 images from the severe TED group. The overall accuracy was 97.9% in distinguishing between normal and severe TED, and the AUC was determined to be 0.993 (Fig 4B1, 2). The model misclassified 2 severe TED images as normal.

Lastly, the binary model was trained on mild TED and severe TED only. The training data set included 366 images from the mild TED group and 285 images from the severe TED group, and the testing data set included 93 images from the mild group and 70 images from the severe group. The

Table 1. Patient Demographics

	Groups	All Participants	Normal	Mild TED	Severe TED
Demographic	**Total, N**	**141 (1187 images)**	**31 (373 images)**	**83* (459 images)**	**40* (355 images)**
Age (yrs)	Median (range)	71 (25–95)	61 (31–91)	74 (25–90)	65 (32–95)
Sex, n (%)	Female	104 (74)	22 (71)	65 (78)	27 (68)
	Male	37 (26)	9 (29)	18 (22)	13 (32)
Race, n (%)	White	115 (81)	23 (74)	63 (76)	31 (78)
	Asian	11 (8)	2 (6)	8 (10)	3 (8)
	Black	8 (6)	3 (10)	6 (7)	1 (2)
	Others	7 (5)	3 (10)	6 (10)	5 (12)

TED = thyroid eye disease.
*Thirteen patients were included in both the mild and severe TED groups because 1 eye demonstrated features of compressive optic neuropathy (severe TED) and the other eye had mild TED.

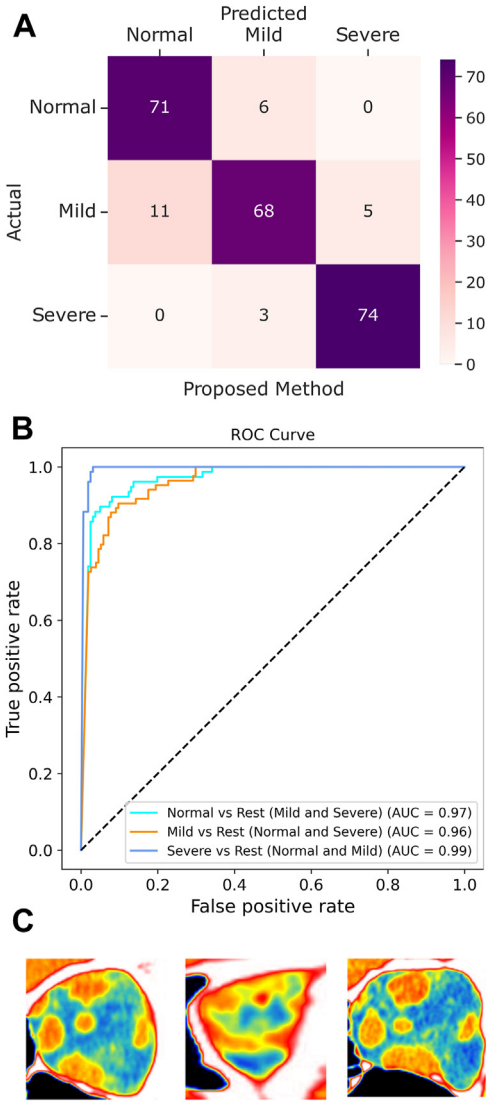

Figure 3. Results of primary model. **A,** The results of the testing of the primary model trained on all 3 categories are shown, with the confusion matrix indicating an accuracy of 89.5%. **B,** The receiver operating characteristic (ROC) curve demonstrates the 3 curves in a category vs. rest model (normal vs. mild and severe; mild vs. normal and severe; and severe vs. normal and mild) with an area under the curve (AUC) ranging from 0.96 to 0.99. **C,** The 3 images of severe thyroid eye disease (TED) that the model misclassified as mild TED.

overall accuracy was 93.8% in distinguishing mild TED and severe TED, and the AUC was determined to be 0.985 (Fig 4C1, 2). The model misclassified 7 severe TED images as mild TED. A bootstrapped *t* test was used to compare the performance between the different models. The primary model's accuracy was found to be significantly greater than that of the 3 binary models ($P < 0.001$).

For comparison with human graders, a subset of 373 images (31.4%) from 90 orbits across the 3 groups were randomly selected for screening. The accuracy of the oculoplastic surgeon on discrete single images was 43.8%. Subsequent evaluation mimicking realistic clinical assessment with serial region of interest images per orbit had a prediction accuracy of 70.0% (Fig 5). In serial image testing, the oculoplastic surgeon incorrectly identified 2 normal patients as mild TED and 1 normal patient as severe TED. Notably, the oculoplastic surgeon never misdiagnosed normal controls as mild or severe TED.

Discussion

This study presents a deep learning model that can accurately distinguish orbital CTs from patients with normal, mild TED, and severe TED with optic neuropathy. The primary model can distinguish between all 3 categories with 89.5% accuracy and AUC of 0.96 to 0.99. Binary testing showed improved accuracy, with normal versus mild TED yielding 93.4% accuracy (AUC, 0.982); normal versus severe TED yielding 97.9% accuracy (AUC, 0.993), and mild versus severe TED yielding 94% accuracy (AUC, 0.985). It was also shown to be more accurate in distinguishing all 3 categories compared with a human expert review of imaging.

Deep learning models for TED have demonstrated early promise with various strategies for radiographic images.[5,10] These have included novel neural networks as well as those built on existing pretrained machine learning algorithms. One such study by Lee et al[10] compared their own neural network model with GoogleNet Inception, Deep Residual Learning (ResNet), VGG, and oculoplastic surgeons, with successful demonstration of the potential effectiveness of AI. Variations of ResNet and VGG methods have also been applied to TED patients who underwent MRI with success in distinguishing active versus inactive phases of disease.[5] Some models employ inputting multiple pieces of clinical data, visual fields, and orbital imaging.[16] Other studies have tried to delineate specific radiographic hallmarks of TED, usually focusing on the extraocular muscles.[17,18] These and similar studies evaluating distinct radiographic characteristics of TED may also aid future machine learning algorithms and increase yield. Visual Geometry Group-16 has been widely used in medical imaging. It has a relatively simple architecture, making the model easier to understand, implement, and modify. It can effectively learn and generalize complex patterns, which is crucial in medical imaging when the model needs to recognize subtle and complex patterns within the images. Visual Geometry Group-16 is also beneficial for smaller data sets, and a larger model has a higher risk of overfitting small data sets.[19] Therefore, VGG-16 was used for this model development.

Limitations in the existing literature includes models having difficulty distinguishing normal controls from patients with mild TED, the need for large numbers of images per patient, significant manual/clinical input, and selection bias due regional data sets usually from tertiary academic hospitals and challenges with obtaining "normal" CT scans for control patients. Additionally, there is a lack of AI for radiographic detection of TED studies performed in the United States. This may limit the generalizability of the previously developed AI models to patients in the United

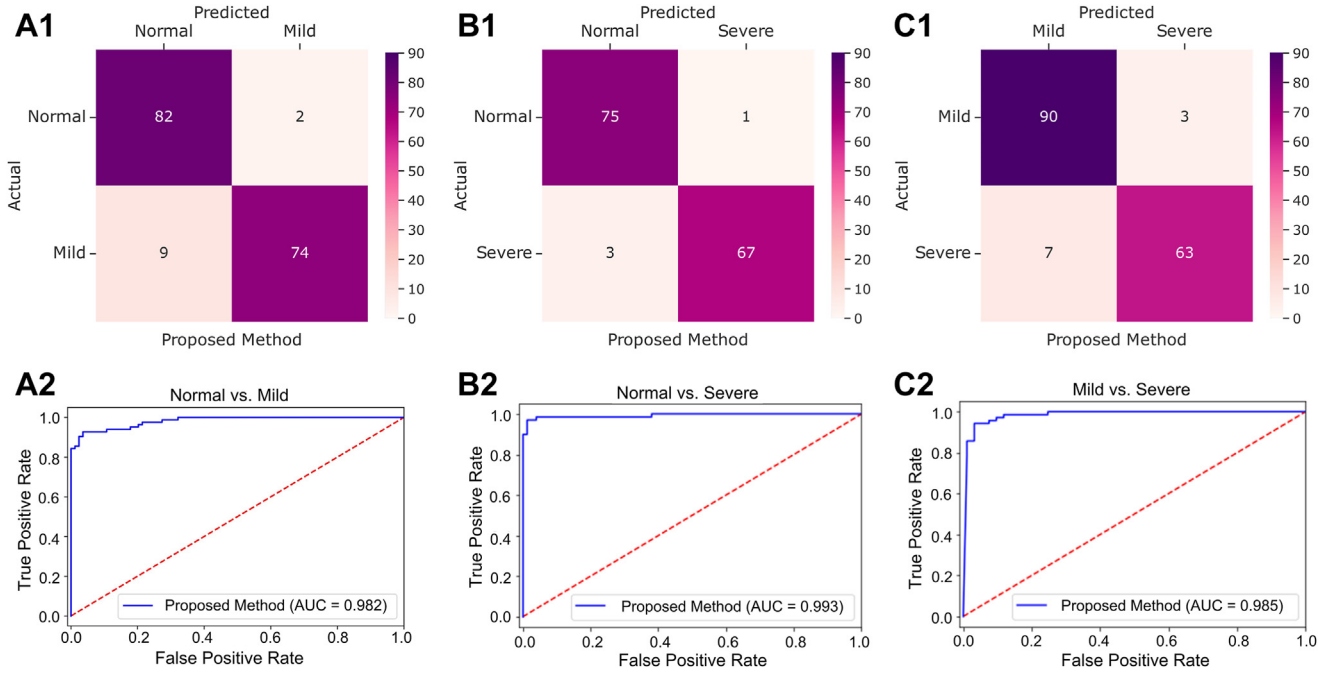

Figure 4. Results of binary model. The results of the binary models are shown, with the confusion matrixes presented in the top panels and the receiver operating characteristic in the bottom panels. The model trained on normal vs. mild thyroid eye disease (TED) is shown on the left panel (**A1, 2**), normal vs. severe TED in the middle panel (**B1, 2**), and mild vs. severe TED in the right panel (**C1, 2**, with accuracies of 93.4%, 97.9%, and 93.8% respectively, and areas under the curve (AUCs) of 0.982, 0.993, and 0.985 respectively.

States due to potential demographic and equipment differences.

This study demonstrated great accuracy in distinguishing between normal scans, mild TED with no optic neuropathy, and severe TED with optic neuropathy on orbital CTs (AUC, 0.96–0.99). This is the only model in the literature, to the authors' knowledge, that can distinguish between these 3 categories. Binary testing also performed well, with normal versus mild (accuracy, 0.982; AUC, 93.4), normal versus severe (accuracy, 0.993; AUC, 97.9), and mild versus severe (accuracy, 0.985; AUC, 93.8). Two models in the literature that used CT scans to distinguish normal CTs from TED reported AUCs of 0.919[8] and 0.946,[18] but did not include analysis on the severity of the disease. One model performed binary testing utilizing CTs to distinguish normal versus mild and normal versus severe TED, with AUC of 0.826 and 0.930 and accuracy of 0.895 and 0.979, respectively.[10] Another model utilized MRI images rather than CT and distinguished active versus inactive TED (accuracy: 0.863 and 0.855, AUC, 0.922).[5] Overall, compared with other studies in the literature, this model is more accurate and also can accurately distinguish normal, mild TED with no optic neuropathy from severe TED with optic neuropathy, which is critically important for triaging urgent referrals to oculoplastic surgeons.[5,8,10,18]

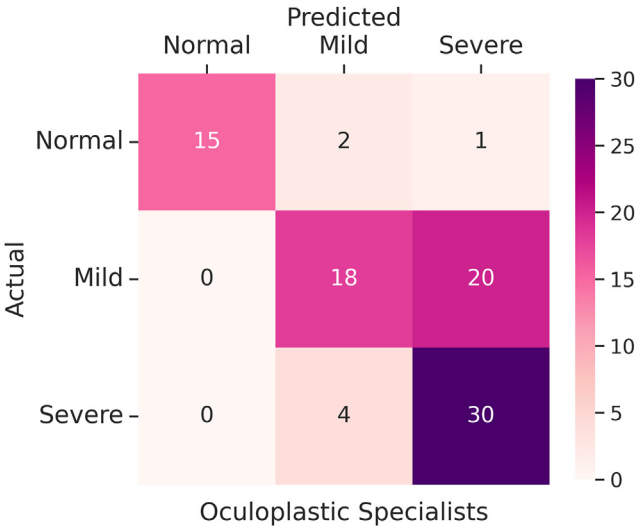

Figure 5. Results of serial image testing of an oculoplastic surgeon. The results of serial image testing of an oculoplastic surgeon demonstrate an overall accuracy of 0.7. There were no mild or severe cases of thyroid eye disease that were mistaken as normal.

Additionally, unlike some models which utilize MRIs,[5,20] this screening model utilizes orbital CTs which is an imaging modality that is more accessible and less expensive. One model uses external photographs as the imaging modality,[4,21] but orbital imaging is ideal for training AI compared with other data sets because there is less background noise, leading to greater potential detection accuracy.[6] However, CTs do have a cost and expose the patient to ionizing radiation which may limit its widespread use as a screening tool. However, by detecting TED, and in particular TED with optic neuropathy, early, expedited referrals may allow patients

to initiate lifestyle modifications and potential pharmacologic or surgical treatments sooner, potentially decreasing morbidity and irreversible vision loss.

The primary model had an accuracy of 89.5%, which is slightly lower than the accuracy of the binary models. However, the primary model never underclassified a severe TED patient with optic neuropathy as a normal scan, and only misclassified 3 severe TED patients as mild TED. The binary models had greater overall accuracy compared with the primary model, but had several underclassifications (9 mild as normal; 2 severe as normal; and 7 severe as mild). An ideal screening model that can be clinically applied should have both high sensitivity and specificity, but also importantly never underclassify conditions to ensure a clinically grave condition is not missed. Although the binary models are more accurate, the primary model is more clinically impactful as it appeared to never miss a severe TED case in the testing phase. Interestingly, the human expert grader, although less accurate overall than the models, never underclassified a mild or severe TED as a normal scan.

Post hoc review of the model-generated salient features maps was notable for the model's ability to generally identify key features that oculoplastic surgeons and radiologists review, such as the muscles and optic nerves. However, some salient feature maps were found to highlight other areas such as neighboring sinuses, which may account for some of the inaccuracies of the models. Post hoc review of the misclassified scans did not identify a clear pattern for misclassifications. It is difficult to know what the model may focus on in an image down to a pixel level that a human grader may or may not be attuned to. Future directions could include refining the model's ability to detect specific components without the orbit to analyze or having the model auto-segment the orbit and isolate features such as the muscles, optic nerve, or orbital fat to improve overall accuracy.

The current gold standard of an objective TED diagnosis is clinical, with a combination of clinical examination, laboratory testing, ancillary testing, and orbital imaging aiding in diagnosis. This model was successful in diagnosing TED and its severity based on previously defined clinical categories using imaging review alone without any clinical information. Human expert review of imaging alone was less accurate, particularly in distinguishing between mild TED with no optic neuropathy and severe TED with optic neuropathy. Although this should not supplant clinical examinations, this model holds promise for utilization in radiographic reviews of imaging and could be integrated with radiology to both diagnose incidental TED if a patient is undergoing neuroimaging for another reason, or to aide in diagnosis and triage for non−ophthalmology-trained clinicians. Patients with an unknown diagnosis may present to a primary care or emergency department with their eye symptoms, eye pain, or headache. Non−ophthalmology-trained clinicians may overlook clinical features of TED and obtain a CT orbit for further evaluation. Additionally, unfamiliarity with TED and its disease process can lead non-ophthalmology clinicians to be uncertain with the tempo of referral to ophthalmology. Although it would be costly and inefficient to routinely screen patients with thyroid problems with a CT, oculoplastic surgeons are limited in the United States and globally. Although oculoplastic examination is likely a more comprehensive and directed approach to caring for patients with TED, this CT model has potential to be utilized in settings where an oculoplastics evaluation is difficult to access.

The strength of this study is the inclusion of a large cohort of patients with paired clinical data that allows for correlation of CTs to clinical examination and the presence of compressive optic neuropathy. However, there are several limitations to consider. The study cohort was biased toward abnormal CTs. Healthy patients without any clinical abnormalities do not routinely undergo orbital CTs. Those with prior trauma, tumors, or other forms of orbital inflammation were also excluded, and therefore the number of radiographically normal control scans was limited. However, although the total number of normal patients included was low, the total number of image slices used was comparable in number to the other categories (373 images slices for normal; 459 for mild; 355 for severe). Additionally, the scans may not represent a truly normal distribution of the population as there was likely a higher index of suspicion for an orbital process if the patient was undergoing a scan in the first place. However, it does represent clinical practice, where a CT may be obtained for vague orbital pain or proptosis to evaluate for TED, but ultimately results as normal. Additionally, the study did not include patients with enlarged extraocular muscles due to other etiologies. This limits the predictive value of this model in clinical practice, as a clinical examination is still warranted to correlate with the AI model findings. Future studies include expanding the model to distinguish TED from other TED mimickers and differential diagnoses, as well as integrating other clinical information, such as patient demographics, laboratory data, reported symptoms, or clinical examination findings to optimize the model's accuracy. This study is also limited due to sample bias, as an ideal model would be trained on a broad and diverse demographic and validated across different cohorts and regions as orbital anatomy varies across ethnicities. Due to the location of the hospital in the New England area, the demographic is skewed toward White. However, other AI models utilizing orbital images for detection of TED that have been developed thus far have been predominantly based on Asian populations (Japan,[18] South Korea,[10] and China[5,8]). This model helps to expand and diversify the models that are already present in the literature.

Future directions for this particular research, and health care-related machine learning models in general, include expanding this system into an explainable machine learning model. Explainable machine learning models provide human-interpretable explanations for the model decisions.[22,23] Although this model provided salient feature maps, it was not always clear that what the model was emphasizing was what the human expert deemed to be clinically relevant. A model that can be more widely implanted clinically should be both accurate but also provide prediction uncertainty with transparent prediction-uncertainty metrics, and thus explainable machine learning models are critical for health care implentation.[22] Future

directions may also include auto-segmentation of images which may provide higher quality and more consistently formatted images for model development and testing.

Thyroid eye disease is a disfiguring, and potentially vision-threatening, condition. Timely diagnosis and referral to oculoplastics for further treatment and management is critical. Our AI model presents an accurate prediction algorithm that can effectively distinguish between normal patients and those with TED. Furthermore, it can distinguish patients with TED who have clinical signs of compressive optic neuropathy. With further optimization and validation, this model can be used to help guide frontline providers in the detection of TED and stratify the urgency of referrals to oculoplastic surgeons and endocrinologists.

Footnotes and Disclosures

Originally received: May 23, 2023.
Final revision: October 7, 2023.
Accepted: October 9, 2023.
Available online: October 13, 2023. Manuscript no. XOPS-D-23-00110R2.

[1] Department of Ophthalmology, Ophthalmic Plastic Surgery Service, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts.

[2] Department of Ophthalmology, Gavin Herbert Eye Institute, University of California Irvine, Irvine, California.

[3] Harvard Ophthalmology AI Lab, Schepens Eye Research Institute of Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts.

[4] Department of Information Sciences and Mathematics, Dong-A University, Busan, Republic of Korea.

[5] Data Science, Athenahealth, Watertown, Massachusetts.

[6] Department of Radiology, Lab of Medical Imaging and Computation, Massachusetts General Brigham and Harvard Medical School, Boston, Massachusetts.

*L.Y.L. and P.Z. contributed equally to this work.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have no proprietary or commercial interest in any materials discussed in this article.

Presented at the Women in Ophthalmology Annual Meeting. August 25-28, 2022, Monterey, California; American Society of Ophthalmic Plastic & Reconstructive Surgery (ASOPRS) Annual Fall Meeting, September 29–30, 2022, Chicago, Illinois; American Academy of Ophthalmology (AAO) Annual Meeting. September 30 to October 3, 2022, Chicago, Illinois; Orbit Society International Meeting. December 2–3, 2022, Seoul, Korea.

HUMAN SUBJECTS: Human subjects were included in this study. The Massachusetts General Brigham (MGB) institutional review board approved this retrospective study, and the required written informed consent was waived by the MGB institutional review board. All research adhered to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Lee, Lin, Shi, Zhou, Wang, Do

Data collection: Lee, Shi, Zhou, Lu, Jeon, Wang, Do, Liu, Kim

Analysis and interpretation: Lee, Lin, Shi, Zhou, Lu, Jeon, Wang, Do

Obtained funding: N/A

Overall responsibility: Lee, Lin, Shi, Zhou, Do

Abbreviations and Acronyms:
AI = artificial intelligence; **AUC** = area under the curve; **CT** = computed tomography; **MRI** = magnetic resonance imaging; **TED** = thyroid eye disease; **VGG** = Visual Geometry Group.

Keywords:
Artificial intelligence, Compressive, Deep learning, Optic neuropathy, Thyroid eye disease.

Correspondence:
Nahyoung Grace Lee, Massachusetts Eye and Ear Infirmary, 243 Charles St., 10th floor – Eye Plastics, Boston, MA 02114. E-mail: grace_lee@meei.harvard.edu.

References

1. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223. https://doi.org/10.1001/jama.2017.18152.
2. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–175. https://doi.org/10.1136/bjophthalmol-2018-313173.
3. Reid JE, Eaton E. Artificial intelligence for pediatric ophthalmology. *Curr Opin Ophthalmol*. 2019;30:337–346. https://doi.org/10.1097/ICU.0000000000000593.
4. Karlin J, Gai L, LaPierre N, et al. Ensemble neural network model for detecting thyroid eye disease using external photographs. *Br J Ophthalmol*. 2022: bjophthalmol-321833. https://doi.org/10.1136/bjo-2022-321833.
5. 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. https://doi.org/10.1186/s12886-020-01783-5.
6. Bao XL, Sun YJ, Zhan X, Li GY. Orbital and eyelid diseases: the next breakthrough in artificial intelligence? *Front Cell Dev Biol*. 2022;10:1069248. https://doi.org/10.3389/fcell.2022.1069248.
7. Li L, Song X, Guo Y, et al. Deep convolutional neural networks for automatic detection of orbital blowout fractures. *J Craniofac Surg*. 2020;31:400–403. https://doi.org/10.1097/SCS.0000000000006069.
8. Song X, Liu Z, Li L, et al. Artificial intelligence CT screening model for thyroid-associated ophthalmopathy and tests under clinical conditions. *Int J CARS*. 2021;16:323–330. https://doi.org/10.1007/s11548-020-02281-1.
9. Han Q, Du L, Mo Y, Huang C, Yuan Q. Machine learning based non-enhanced CT radiomics for the identification of orbital cavernous venous malformations: an innovative tool. *J Craniofac Surg*. 2022;33:814–820. https://doi.org/10.1097/SCS.0000000000008446.
10. Lee J, Seo W, Park J, et al. Neural network-based method for diagnosis and severity assessment of Graves' orbitopathy

using orbital computed tomography. *Sci Rep*. 2022;12:12071. https://doi.org/10.1038/s41598-022-16217-z.

11. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Cornel University Library. April 10, 2015. doi:10.48550/arXiv.1409.1556
12. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. 2009. *IEEE Conference on Computer Vision and Pattern Recognition*. 2009:248−255. https://doi.org/10.1109/CVPR.2009.5206848.
13. Hinton G, Srivastava N, Swerskey K. Neural networks for machine Learning. Lecture 6a Overview of mini- batch gradient descent. https://docplayer.net/22723085-Neural-networks-for-machine-learning-lecture-6a-overview-of-mini-batch-gradient-descent-geoffrey-hinton-with-srivastava-kevin-swersky.html. Accessed December 28, 2022.
14. Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html. Accessed December 28, 2022.
15. Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for Image classification. *Int J Adv Trends Comput Sci Eng*. 2020;9. https://doi.org/10.30534/ijatcse/2020/175942020.
16. Salvi M, Dazzi D, Pellistri I, et al. Classification and prediction of the progression of thyroid-associated ophthalmopathy by an artificial neural network. *Ophthalmology*. 2002;109:1703−1708.
17. Yu B, Gong C, Ji YF, et al. Predictive parameters on CT scan for dysthyroid optic neuropathy. *Int J Ophthalmol*. 2020;13:1266−1271.
18. Hanai K, Tabuchi H, Nagasato D, et al. Automated detection of enlarged extraocular muscle in Graves' ophthalmopathy with computed tomography and deep neural network. *Sci Rep*. 2022;12:16036. https://doi.org/10.1038/s41598-022-20279-4.
19. Kaur T, Gandhi TK. Automated brain image classification based on VGG-16 and transfer learning. 2019. *International Conference on Information Technology (ICIT)*. 2019:94−98. https://doi.org/10.1109/ICIT48102.2019.00023.
20. Wu H, Luo B, Zhao Y, et al. Radiomics analysis of the optic nerve for detecting dysthyroid optic neuropathy, based on water-fat imaging. *Insights Imaging*. 2022;13:154. https://doi.org/10.1186/s13244-022-01292-7.
21. Huang X, Ju L, Li J, et al. An intelligent diagnostic system for thyroid-associated ophthalmopathy based on facial images. *Front Med (Lausanne)*. 2022;9:920716. https://doi.org/10.3389/fmed.2022.920716.
22. Chua M, Kim D, Choi J, et al. Tackling prediction uncertainty in machine learning for healthcare. *Nat Biomed Eng*. 2023;7:711−718.
23. Kim D, Chung J, Choi J, et al. Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. *Nat Commun*. 2022;13:1867. https://doi.org/10.1038/s41467-022-29437-8.