Deep-Learning Based Automated Segmentation and Quantitative Volumetric Analysis of Orbital Muscle and Fat for Diagnosis of Thyroid Eye Disease

Adham M. Alkhadrawi,[1] Lisa Y. Lin,[2] Saul A. Langarica,[1] Kyungsu Kim,[1] Sierra K. Ha,[2] Nahyoung G. Lee,[2] and Synho Do[1,3,4]

[1]Department of Radiology, Lab of Medical Imaging and Computation, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States
[2]Department of Ophthalmology, Ophthalmic Plastic Surgery Service, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, United States,
[3]KU-Korea Institute of Science and Technology (KIST) Graduate School of Converging Science and Technology, Korea University, Seoul, Korea
[4]Kempner Institute, Harvard University, Boston, Massachusetts, United States

Correspondence: Synho Do, Kempner Institute, Harvard University, 125 Nashua St., Suite 2210, Boston, MA 02114, USA; sdo@mgh.harvard.edu.

AMA and LYL contributed equally to this work.

Received: December 1, 2023
Accepted: April 2, 2024
Published: May 2, 2024

Citation: Alkhadrawi AM, Lin LY, Langarica SA, et al. Deep-learning based automated segmentation and quantitative volumetric analysis of orbital muscle and fat for diagnosis of thyroid eye disease. *Invest Ophthalmol Vis Sci.* 2024;65(5):6. https://doi.org/10.1167/iovs.65.5.6

Purpose. Thyroid eye disease (TED) is characterized by proliferation of orbital tissues and complicated by compressive optic neuropathy (CON). This study aims to utilize a deep-learning (DL)-based automated segmentation model to segment orbital muscle and fat volumes on computed tomography (CT) images and provide quantitative volumetric data and a machine learning (ML)-based classifier to distinguish between TED and TED with CON.

Methods. Subjects with TED who underwent clinical evaluation and orbital CT imaging were included. Patients with clinical features of CON were classified as having severe TED, and those without were classified as having mild TED. Normal subjects were used for controls. A U-Net DL-model was used for automatic segmentation of orbital muscle and fat volumes from orbital CTs, and ensemble of Random Forest Classifiers were used for volumetric analysis of muscle and fat.

Results. Two hundred eighty-one subjects were included in this study. Automatic segmentation of orbital tissues was performed. Dice coefficient was recorded to be 0.902 and 0.921 for muscle and fat volumes, respectively. Muscle volumes among normal, mild, and severe TED were found to be statistically different. A classification model utilizing volume data and limited patient data had an accuracy of 0.838 and an area under the curve (AUC) of 0.929 in predicting normal, mild TED, and severe TED.

Conclusions. DL-based automated segmentation of orbital images for patients with TED was found to be accurate and efficient. An ML-based classification model using volumetrics and metadata led to high diagnostic accuracy in distinguishing TED and TED with CON. By enabling rapid and precise volumetric assessment, this may be a useful tool in future clinical studies.

Keywords: thyroid eye disease (TED), deep learning (DL), medical image segmentation

Thyroid eye disease (TED) is a heterogeneous autoimmune condition that can have a varied presentation. TED can be disfiguring, and early detection can be critical to avoid permanent vision loss. Due to aberrant signaling and overactivation of thyroid stimulating hormone receptor and insulin-like growth factor-1 pathways in TED, there is abnormal growth and uncontrolled proliferation of orbital tissues.[1–4] TED signs and symptoms, such as extraocular motility limitation and proptosis, are consequences to changes in the orbital fat and muscles, but diagnosis may be delayed due to its varied and nonspecific presenting symptoms. When the expansion of these tissues is severe, there may be vision-threatening complications, like compressive optic neuropathy (CON). Management and treatment strategies aim to alleviate symptoms, manage inflammation and disfigurement, and, in severe cases, prevent or address CON. However, identification of patients at risk for serious sequelae to achieve urgent triage to an oculoplastic surgeon is critical.

Volumetric segmentation is frequently utilized in radiotherapy for assessment of targeted organs and anatomic structures throughout the body. Automated segmentation of radiographic images has been a rapidly expanding field in the last decade to reduce the time and variability associated with manual segmentation utilizing techniques based on multi-atlas algorithms. Recently, deep learning (DL) has broadened the applicability of auto-segmentation and has allowed for broader generalization of new data.[5] For TED,

the quantification of orbital muscle and fat volumes in TED have been explored and volumetric based studies have helped better understand disease pathophysiology, corresponding clinical features, and response to therapies.[6–24] However, many of these studies depend on manual segmentation for demarcation of orbital tissues of interest, which is time and labor intensive.[11,13,14,25] Alternatively, they may involve the utilization of commercial or open-source software packages, many of which are programed based on normal orbital scans and may have less generalizability to patients with TED[6,8,12,16,18–21] or may still require some manual segmentation. Orbital imaging of TED poses a unique challenge for segmentation and volumetric analysis due to a relatively small target organ, the irregular expansion of the orbital muscles and fat, and the apical crowding within the bony orbit.

This study aims to introduce a DL-based fully automated approach for the segmentation of orbital muscle and fat, and subsequent quantitative volume measurement using orbital computed tomography (CT) scans in patients with TED and TED with CON. This study also aims to utilize this volumetric data in a DL-based classification model and integrates patient-specific data in conjunction with the image data for optimized classification.

Methods

Study Design

A retrospective cohort study was performed at Massachusetts Eye and Ear (MEE), a tertiary ophthalmic institution, over a 12-year period (August 2011 to 2023). The Massachusetts General Brigham (MGB) institutional review board approved this retrospective study, and the written informed consent was waived. 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.

Subject identification methods were previously described.[26] Subjects were included if they were 18 years of age or older with a dedicated CT scan of the orbit who had also undergone a clinical examination by an oculoplastic surgeon within a 3-month period of the scan. 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).

A multi-faceted approach to the DL algorithm was constructed, with an overview of the study methodology presented in Figure 1. First, the region of interest (ROI) was extracted (see Fig. 1, stage 1) from the orbital CTs, with details on methodology to follow. DL-based segmentation was then performed to segment muscle and fat volumes of each orbit (see Fig. 1, stage 2). Calculation of orbital muscle and orbital fat volumes were then performed. Then, the volumetric data was processed in synergy with patient metadata through a DL-based classifier to predict the presence of normal, mild TED, or severe TED (see Fig. 1, stage 3).

Data Preprocessing

Patient Metadata Collection and Categorization. Patient demographics, clinical history, smoking status, ancillary testing, and oculoplastic surgeon's clinical examination

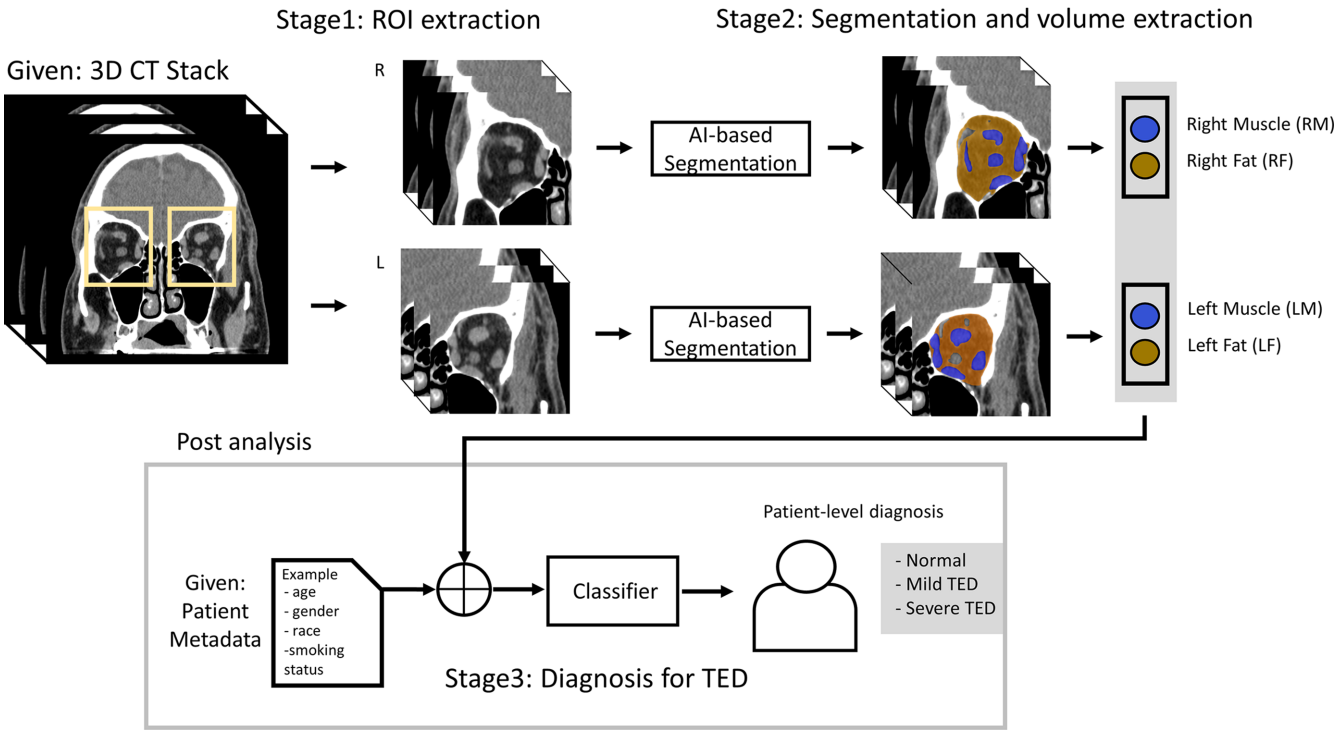

Figure 1. Overview of the proposed DL volumetric analysis methodology. In stage 1, the input whole CT scan is automatically cropped to extract right and left orbits. In stage 2, DL-based automated segmentation of the muscle and fat volumes in each orbit is performed and the volumes of each are subsequently calculated. In stage 3, the volumetric data is integrated with various patient metadata into an ML-based classifier which is used to diagnosis the patient with normal, mild TED, or severe TED.

Investigative Ophthalmology & Visual Science

closest in time to the CT were abstracted from the electronic medical record. Based on retrospective chart review, patients were categorized into three groups: normal patients, patients with mild TED (no optic neuropathy), and patients with severe TED (with optic neuropathy), as previously described.[26] Patients with severe TED were considered to have any signs of compressive optic neuropathy, including any one of the following: dyschromatopsia, Humphrey visual field (VF) changes in patterns consistent with TED, relative afferent pupillary defect, and/or optic nerve head changes.[27] Optic nerve head changes were identified on direct fundoscopic examination and including optic, and documented if there was any blurring of disc margins, disc edema, or disc pallor. These clinical categories were deemed the ground truth for model training.

Computed Tomography Data (CT Scans) Preprocessing. Orbital CTs from selected subjects were obtained. The slice thickness of the scans was 0.625 mm, and the pixel dimensions were 512 × 512 pixels. To maximize the clarity of the orbit, the image volume was clipped to 1000 pixels window level and 350 pixels window width. After this windowing process, each individual orbit (right and left orbit) was cropped to a 140 × 140 × 140-pixel volume. The right and left orbits of each CT scan were cropped by a fixed-size ROI. The orbital structures were not affected by the slight shifts in the ROI positions because all CT scans had the same pixel dimensions (512 × 512 pixels). This proof-of-concept model did not create a general cropping method that could handle different CT scan sizes and offsets. The cropped volumes of a subset of images were then manually segmented by experienced radiologists utilizing FIJI software and Labkit.[28,29] The manual segmentations of these orbits were performed to establish the ground truth for subsequent model training. The total number of manually segmented slices was 4200, which was split into three sets: training (3360), validation (420), and testing (420). The obtained segmentation masks of orbital muscles and fat were stored separately for model training. The dataset was divided into training, validation, and test sets, with the allocation ratios being 80%, 10%, and 10%, respectively.

Model Development

Orbital Volume Segmentation Model. A 2D U-Net[30] model was utilized for segmentation tasks. The adapted U-Net architecture featured an encoder-decoder design, with the encoder module utilizing a pretrained VGG16 model[31] for feature extraction from the processed CT volumes. The decoder module comprised an equivalent number of layers. During the training process, a learning rate of 1e-4 and a batch size of 4 were applied over 100 epochs, utilizing the cross-entropy loss function. To enhance dataset diversity and enhance the model's overall performance, normalization and resizing transforms were implemented as augmentation techniques during training. The following augmentation techniques were used: (a) rotation (limit = 50 degrees, probability = 1.0), (b) horizontal flip (probability = 0.5), and (c) vertical flip (probability = 0.5). Slices were then resized to 160 × 160 pixels for the purpose of training as the library "Segmentation Models Pytorch" (https://github.com/qubvel/segmentation_models.pytorch) required image height and width to be multiples of 32. The Dice Similarity Coefficient (DSC) was used as a metric to assess the model's accuracy, calculated through the following formula (Equation 1: $(A, B) = \frac{2(A \cap B)}{(A+B)}$). The output of the segmentation model includes both a 3-dimensional (3D) volumetric model as well as quantifications of the muscle volumes and fat volumes for each orbit. To assess the performance of the segmentation models, the model was applied to the test set data. A Monte Carlo five-fold cross-validation with data randomly shuffled was performed. The averaged DSC was then calculated.

We applied a post-processing step to resize the predicted masks back to their original 140 × 140 pixel dimensions to ensure accurate volume calculation. The muscle and fat volumes were calculated for each orbit by multiplying voxel volume (d_x,d_y, andd_z) by the predicted mask area (n_{mask}) for each slice in the volume (Equation 2: $V_{3D} = n_{mask} \cdot (d_x \cdot d_y \cdot d_z)$) and then compared for each classification (normal, mild TED, and severe TED) as well as subclassification based on male and female gender. Statistical comparison on outcomes was performed with 1-way ANOVA testing.

TED Classification Model. After the volumes were automatically segmented, an ML-based model was developed for classification of the images. The classification model utilized a Random Forest model design to classify among the three groups (normal, mild TED, and severe TED). The classification approach is represented in Figure 2. The input for the model was the predicted orbital muscle and orbital fat volumes obtained from the segmentation model. The input was then run through the ML-based classifier, which predicts the probability of each image belonging to each class in a binary fashion (i.e. normal versus not normal,

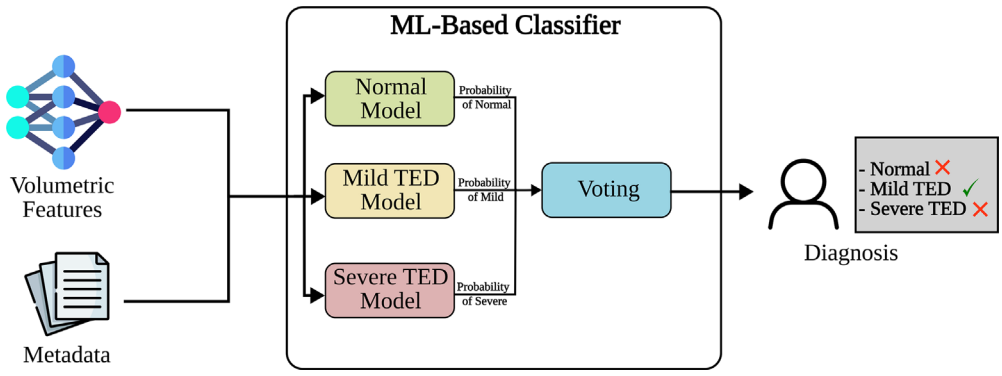

Figure 2. TED classification methodology. Three binary models, one for each class, was implemented. Each model outputs the estimated probability that the instance belongs to the corresponding class, and then the voting module decides the final class of the input instance based on these probabilities, ultimately predicting the diagnosis.

mild TED versus not mild TED, and severe TED versus not severe TED). Finally, the output of the three binary models is combined into a voting scheme, which predicts the final classification of each case (normal versus mild TED versus severe TED).

This classification model was also re-run in synergy with several different combinations of patient metadata, following similar model structure as demonstrated in Figure 2. The aim was to identify the minimal set of input variables to achieve the highest accuracy of prediction and overall performance. The initial classification model utilizing only volumetric data was called model 1 (M1). The second model (M2) only applied features from the patient's metadata, including demographic and clinical examination data as measured by an oculoplastic surgeon. The third model (M3) utilized all features from M1 and M2. The fourth model (M4) used volumetrics and some patient metadata that could be easily assessed without an oculoplastic examination (patient age, gender, race, and smoking status). Supplementary Table S1 lists all input variables for the four models. The accuracy and area under the curve (AUC) for all four model combinations was calculated, with outcomes for each binary model (i.e. normal versus not normal, mild TED versus not mild TED, and severe TED versus not Severe TED) and the main model (normal versus mild TED versus severe TED).

In order to pursue explainable artificial intelligence (AI) outcomes, feature importance of the Random Forest classifiers was also performed. The features analyzed include age, gender, race, smoking status, orbital muscle volumes, orbital fat volumes, and total orbital volumes. The impurity importance or the Gini importance[32] of all the input variables was calculated. Partial Dependence Plots (PDPs)[33] were also used to inspect the independent effects of all the input variables on the decisions of the classifiers.

Results

Two hundred eighty-one (281) patients, totaling 562 individual orbits, were included in this study. The average age of the patient was 55 years (range of 18-94 years), with 73% female patients, and 64% White patients. Table 1 demonstrates patient demographics.

Orbital Volume Segmentation Results and Analysis

The orbital fat and muscle volume segmentation model underwent training and was tested on 10% of the data. Representative examples of the DL-based auto-segmentation model results are shown in Figure 3 and the Supplementary Figures S1–S3. The model successfully took an input image of the CT orbit, identified, and segmented the muscle and fat within the orbit, generated a 3D volumetric model, and calculated the respective volumes of the muscle and fat of each orbit for patients with normal, mild TED, and severe TED.

The performance was cross-validated on the remaining 10% of the data using a 5-fold cross-validation with random shuffling of the data. The DSC of the muscle volumes had an average of 0.902 (0.882, 0.913, 0.906, 0.896, and 0.912 for each round of cross-validation split, respectively). The DSC of the fat volumes had an average of 0.921 (0.941, 0.922, 0.904, 0.924, and 0.908 for each round of cross-validation split, respectively).

The volumes of the orbital fat and muscle were then calculated based on the test dataset (28 patients). The outcomes are presented in Table 2, and subanalyzed for both male and female patients. The average orbital muscle volume per male patient was 3.21 ± 0.95 cm^3, 4.34 ± 1.39 cm^3, and 5.42 ± 0.83 cm^3 for normal, mild TED, and severe TED, respectively. This was found to be statistically significant (P value < 0.0003). The average orbital muscle volume per female patient was 3.50 ± 0.47 cm^3, 4.46 ± 1.29 cm^3, and 5.94 ± 2.19 cm^3 for normal, mild TED, and severe TED, respectively. This was also found to be statistically significant (P value < 0.0003). Furthermore, the average orbital fat volume per male patient among the three groups was found to be statistically significant (P value < 0.003), but there was no statistically significant difference in female patients.

TED Classification Results and Analysis

After the volumes of the muscle and fat of the orbit were calculated for all the cases, the TED classification model was used to classify each case into one of the three categories, namely, normal versus mild TED versus severe TED, using the composition of the three binary models (see Fig. 2). One notable advantage of training the binary models independently was that the strong class imbalance present in the data could be easily mitigated. For example, when training the binary model mild versus not mild, the not mild class included both the normal and severe cases, and for the mild class, only an equivalent number of cases were considered. The same procedure was followed to train the other binary models.

The data were divided into training, validation, and testing with corresponding ratios of 0.6, 0.2, and 0.2, respectively. The training and validation sets were used for hyperparameter optimization of the model, and the hyperparameters of the best performing model were used to classify the cases in the test set. The classification was performed on fat and muscle volumes calculated for both eyes as an average of the respective volumes for each patient.

Table 1. Patient Demographics

		Normal	Mild TED	Severe TED	Total
Total	*N* participants (*N* orbits)	49 (98)	196 (392)	36 (72)	281 (562)
Age, y	Median (range)	56 (18-94)	55 (18-89)	62 (28-90)	55 (18-94)
Gender, *n* (%)	Male	19 (39%)	46 (23%)	10 (48%)	75 (27%)
	Female	30 (61%)	150 (77%)	26 (72%)	206 (73%)
Race, *n* (%)	White	33 (67%)	124 (63%)	24 (67%)	181 (64%)
	Asian	1 (2%)	23 (12%)	3 (8%)	27 (10%)
	Black	7 (14%)	17 (9%)	3 (8%)	27 (10%)
	Other	8 (16%)	32 (16%)	6 (17%)	46 (16%)

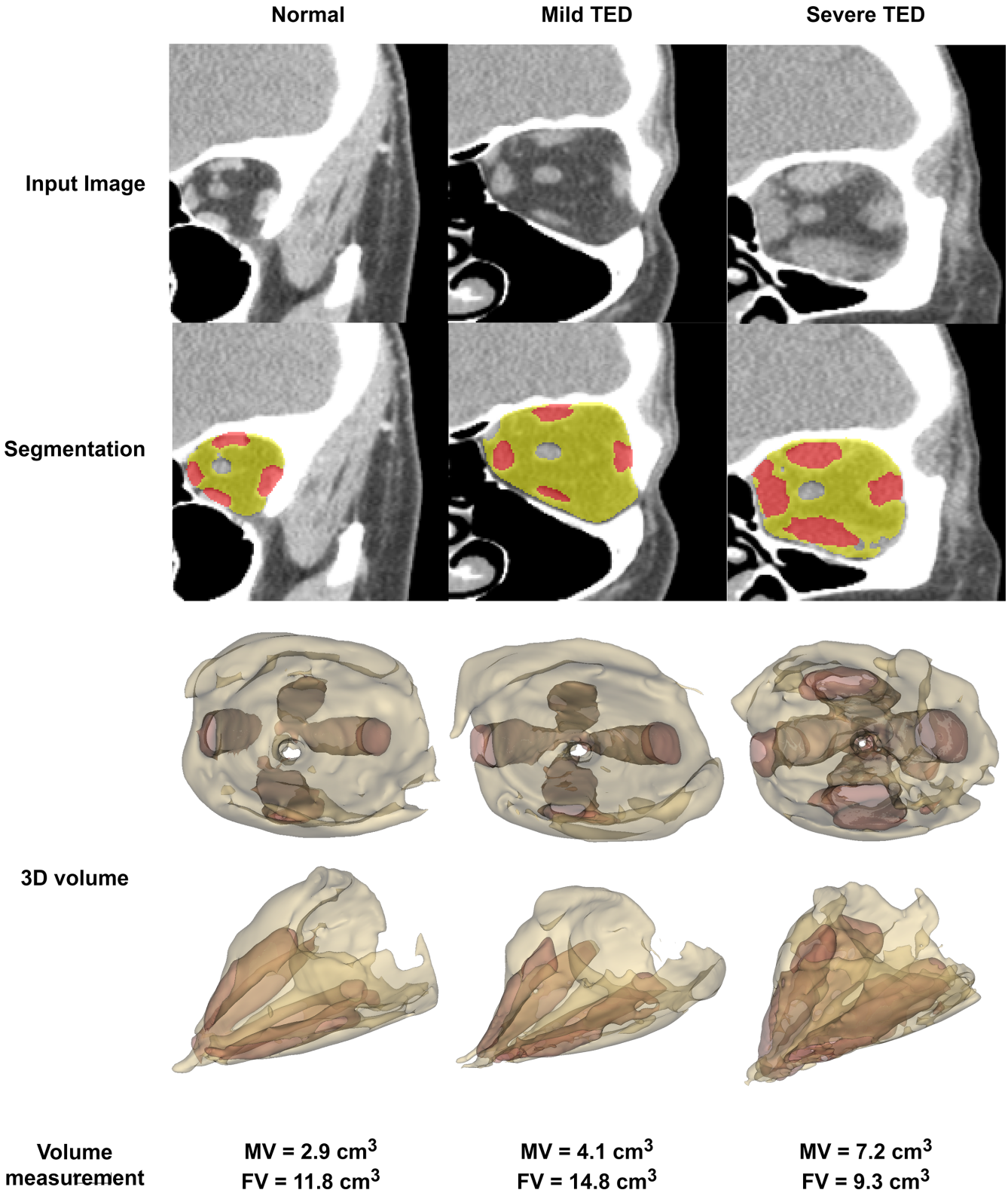

Figure 3. The segmentation of three randomly chosen orbits belonging to each category (normal, mild TED, and severe TED). The *top row* displays the input image into the model. The *middle row* displays the DL-based segmentation. The *yellow overlay* indicates fat segmentation and red overlay indicates muscle segmentation. The *bottom row* displays the 3D models created representing the muscle and fat volumes for each orbit. The model then calculates the resulting muscle volume (MV) and fat volume (FV) for each orbit.

The accuracy and AUC over the test set of each variation of the model are as follows. The M1 model (volume data only) had an accuracy of 0.720 and an AUC of 0.874 in predicting the main outcome between normal versus mild TED versus severe TED. The M2 model (patient metadata only) had an accuracy of 0.795 and an AUC of 0.831

Table 2. Orbital Muscle and Fat Volumes of Patients

Gender	Orbital Tissue	Normal (CM^3) ± SD	Mild TED (CM^3) ± SD	Severe TED (CM^3) ± SD	*P* Value
Male	Muscle	3.21 ± 0.95	4.34 ± 1.39	5.42 ± 0.83	**<0.0003**
	Fat	12.42 ± 3.8	15.7 ± 3.06	14.68 ± 1.97	**<0.003**
Female	Muscle	3.50 ± 0.47	4.46 ± 1.29	5.94 ± 2.19	**<0.0003**
	Fat	14.69 ± 2.82	15.01 ± 2.57	14.84 ± 2.78	0.83
Total	Muscle	3.38 ± 0.72	4.43 ± 1.32	5.8 ± 1.94	**0.0001**
	Fat	13.76 ± 3.44	15.18 ± 2.71	14.79 ± 2.59	**0.02**

SD, standard deviation of mean.
The *P* value was calculated by 1-way ANOVA.
Values represent the mean of the population of patients. Each patient is represented by the average of volumes of 2 orbits.
P values presented in bold face indicate statistical significance.

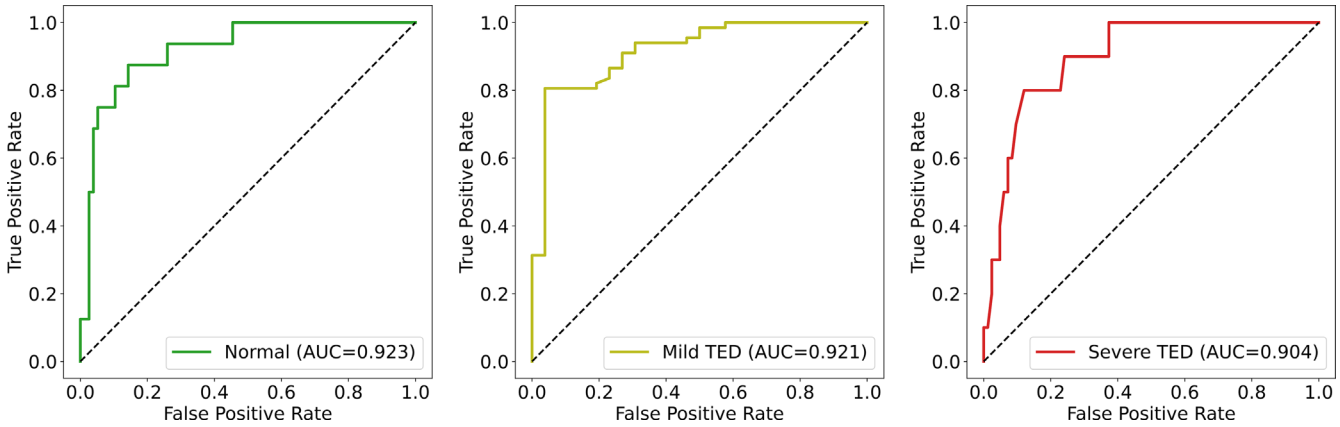

Figure 4. ROC curves for the binary models that compose the M4 model (volume data and limited patient data).

in predicting the main outcome. The M3 model (M1 and M2 data combined) had an accuracy of 0.828 and an AUC of 0.903 in predicting the main outcome. The M4 model (volume data and limited patient data) had the highest accuracy of 0.838 and an AUC of 0.929 in predicting the main outcome. The receiver operator curve (ROC) for the three binary models in M4 is presented in Figure 4.

Random Forest models were performed to determine feature importance for each classification and are presented in Figure 5. The orbital muscle volume was the most important feature (0.40 feature importance) for classification between normal healthy patients and patients with severe TED. For the mild TED cases, the most important feature was orbital fat volume, followed closely by orbital muscle volume, total orbital volume, and patient age.

Further analysis on feature importance demonstrated that muscle volume alone can nearly completely separate normal and severe cases of TED. When the muscle volume was approximately equal to or greater than 4.5 cm^3, nearly all cases were identified as severe TED (upper plot, Fig. 6). When subanalyzed by gender, the muscle volume boundary axis between normal and severe TED cases was different for male and female patients compared to boundary axis determined in all cases (see lower plots, Fig. 6). Additional analysis for important features and metadata was not able

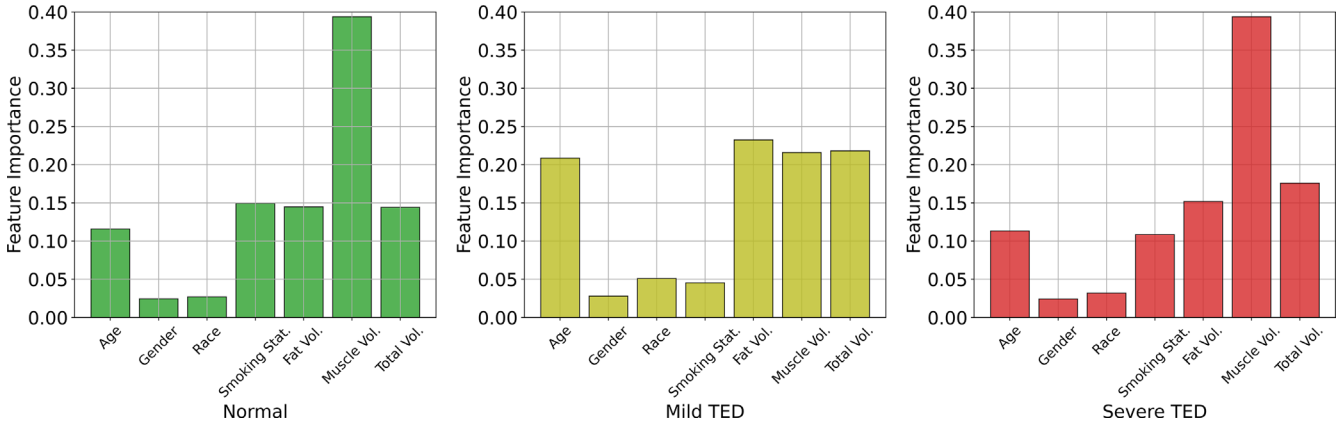

Figure 5. Feature (Gini) importance of the Random Forest classifiers for patients with normal, mild TED, and severe TED.

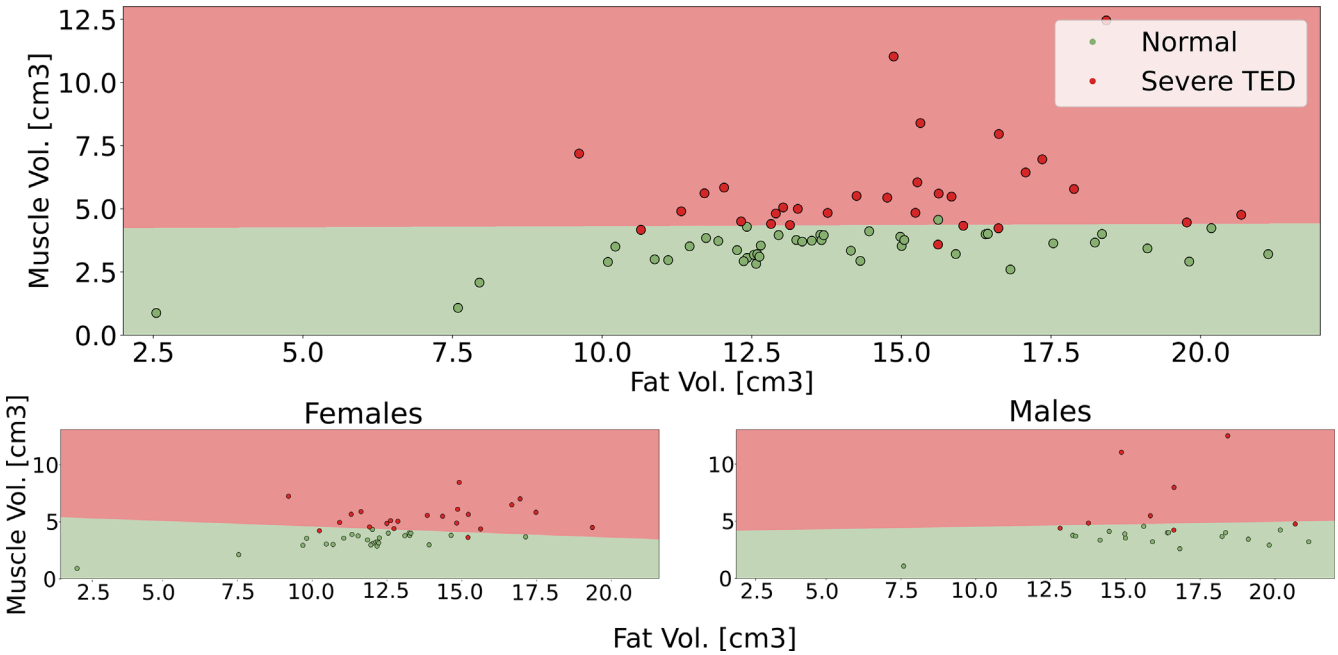

Figure 6. Muscle volume as the most important feature to differentiate between normal and severe TED. The upper plot demonstrates that muscle volume can linearly separate normal and severe cases in general. The lower left and right plots demonstrate the separation by muscle volume when differentiating between female and male patients, respectively.

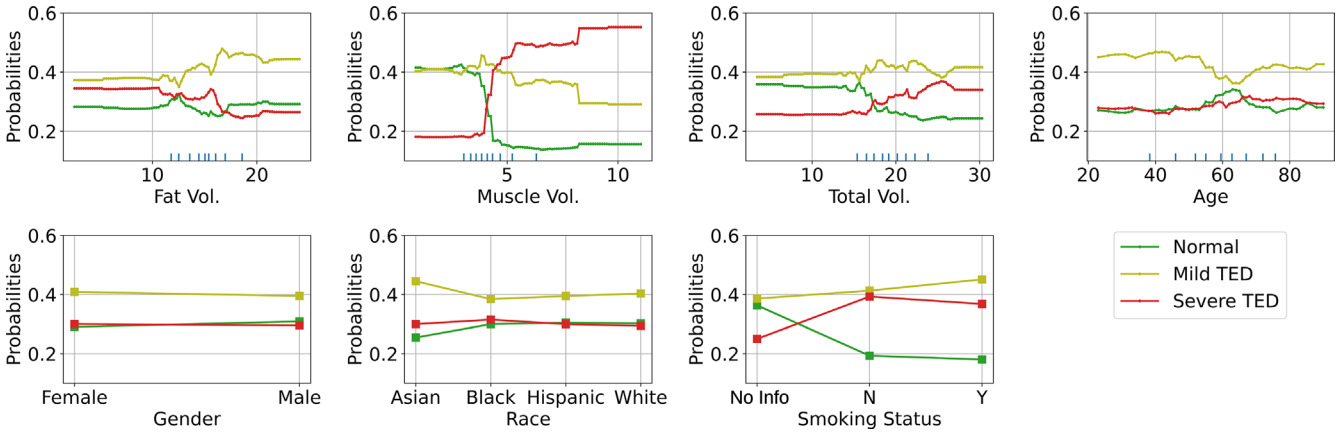

Figure 7. Partial dependence plot of the Random Forest classifier.

to identify a unique variable that had the same potential for separating mild TED cases.

PDP were generated to determine the independent effects of all input variables on the decisions of the classifiers. Figure 7 demonstrates the PDP for each variable, with the range of the variable (either numerical or categorical) on the x-axis and the probability of classification of normal versus mild TED versus severe TED on the y-axis. Muscle volume is again demonstrated as the most important feature to distinguish between normal and severe cases. As muscle volume increases, the probability of severe cases increases dramatically, and the probability of normal cases decreases. Fat volume also demonstrated importance in the PDP, with the probability of mild cases increasing with increased fat volumes. Race and age did not appear to have an effect on classification. This finding was supported by the analysis of variance (ANOVA) test for muscle and fat volumes in normal subjects among different races. The results showed that muscle and fat volumes did not differ significantly among different race groups ($P = 0.7$ and $P = 0.85$ for muscle and fat, respectively).

Discussion

This study presents a DL model that can automatically segment the orbital muscle and fat volumes based on orbital CTs. The volumetric data analysis found orbital muscle volumes were significantly different with increasing volumes among normal, mild TED, and severe TED. Additionally, this study utilizes an ML-based classification model using a combination of volumetric data and patient metadata. The M4 model (volumetric data + limited patient data) was found to be the most accurate in distinguishing among all three categories (normal versus mild TED versus severe TED).

The orbital volume segmentation model was able to segment the orbital muscle and fat volumes accurately and efficiently. When the model is used, it can obtain the volumetric data from the orbital CTs rapidly, with both segmentation and quantitative volume results in approximately 30 seconds or less. It was also found to be highly accurate in cross-validation, with an average DSC of 0.902 and 0.921 for muscle and fat volumes, respectively. The slightly higher accuracy for the fat volume segmentation may be attributed to the lower density of fat tissue in Hounsfield Units compared to orbital muscle or globe, which may allow for the model to distinguish the fat more easily from surrounding tissues. The volumetric outcomes were subanalyzed by gender. The average orbital muscle volumes for normal, mild TED, and severe TED, respectively, were found to be 3.21 ± 0.95, 4.34 ± 1.39, and 5.42 ± 0.83 cm^3 for male patients and 3.50 ± 0.47, 4.46 ± 1.29, and 5.94 ± 2.19 cm^3 for female patients, similar to values determined in other studies.[12,22]

There was significant variation in muscle volumes across the three categories ($P < 0.0003$), with a notable progressive increase in volumes from normal healthy patients to patients with mild and severe TED, in alignment with muscle volume changes in other studies.[7,9,20] In contrast, the average orbital fat volumes were statistically significant for normal, mild TED, and severe TED for male patients ($P < 0.003$), but not for female patients. The fat volumes also did not progressively increase, and the patients with mild TED trended toward having higher fat volumes than patients with severe TED.

There is variability in orbital fat and muscle volume in the literature for "normal" orbits, likely driven by variations in age, gender, and ethnicity.[18,34–37] Other investigations of orbital muscle and fat volumes in patients with TED specifically have both a wide range of methodology and results. Some investigations on orbital volumes were based on manual segmentations of orbital imaging.[11,13,14,22,25] In 1982, Feldon and Weiner provided some of the first quantitative volumetric data based on manual tracings of orbital tissues on CT for eight patients with TED.[22] They similarly found a regular increase in extraocular muscle volume with worsening severity of TED. In 1982, a more automated approach was developed utilizing region-growing and automatic tracing algorithms as early attempts to measure orbital tissue volumes on CT for patients with TED.[7] However, this program still required moderate additional effort and manual data processing. They found comparable measurements in healthy individuals to this study. Commercial and open-source software packages provide tools for manual or semi-automated segmentation of 3D images that have aided volumetric studies in TED.[8,12,18–20] Some of these studies have explored correlation of volumetric data to clinical data, such as correlation between the orbital muscle volumes to vertical strabismus,[8] or the predictive potential of orbital muscle volume in diagnosing CON, identifying the medial rectus muscle volume as the most robust predictor.[12] Other studies utilized magnetic resonance imaging (MRI) to assess orbital volumes.[6,17,24,34,38–40] One study utilized a semi-automatic approach for segmenting orbital muscles in TED using MRIs and found a substantial reduction in time, from 20 minutes for the manual segmentation to 7 minutes.[17] Although MRIs are higher resolution, they are more costly and time intensive compared to CTs. For TED diagnosis and management, CTs are usually sufficient and generally the imaging modality of choice. Therefore, this study utilized CTs to improve generalizability and provide a more real-world accessible tool.

With the expansion of DL based volumetric studies in radiology, the integration of DL techniques has been applied to the study of TED as well. Jiang et al. presented a study on DL-based FCN-8s network derived auto-segmentation model for CTs to determine the clinical target volumes for potential radiotherapy, a region which includes the soft tissues within the bony orbit posterior to the globe. They found the DL-based model was similar to manual contouring, but they did not segment orbital fat and muscles separately. Another study utilizing a DL-based technique using Semantic V-Net was developed for the automated segmentation and volume measurement of orbital muscles of 97 presumably predominantly Chinese patients, with an overall Intersection over Union (IOU) score of 0.8207.[23] This study provides similar automatic segmentation tools with training across a broader cohort of patients, with a DSC of 0.902.

The segmentation and volume extraction data were subsequently utilized to correlate radiographic findings with a clinical diagnosis. The M1 model utilized volume data alone and was moderately accurate at distinguishing among normal, mild TED, and severe TED (accuracy of 0.720 and AUC of 0.874). The M2 model with patient clinical data and no volumetric data demonstrated similar results (accuracy of 0.795 and AUC of 0.831), and the M3 model which combined all the data from M1 and M2 mode further improved the accuracy in distinguishing all three categories (accuracy of 0.828 and AUC of 0.903). However, the patient clinical data in M2 and M3 are clinical measurements obtained by the oculoplastic surgeon, and included features such as lagophthalmos, and inferior and superior scleral show. In order for a model to be generalizable and easily adoptable, it was felt that the clinical features that are integrated into the model should be easily assessed by a non-ophthalmology trained provider. Thus, M4 included volume data and limited patient data (age, gender, race, and smoking history). Interestingly, the pared down clinical data in synergy with the volumetric data was the most accurate model (accuracy of 0.838 and AUC of 0.929) in distinguishing the three categories. This model may be capturing the heterogeneity in which TED is manifested across different age, gender, and race, and thus improving the model accuracy. Given that the M4 model had the best overall performance and no reliance on ophthalmic expertise, this model was felt to be the best use in practice and all subsequent analysis was based on this model.

Additional analysis was performed to better understand the decision making in the ML model. Explainable DL models provide human-interpretable explanations for the model decisions.[41,42] Although it is not always clear what ML-based models emphasize, it is important to explore the features that the model emphasizes in order to provide transparency for the model, which is critical for healthcare implementation. On feature importance analysis, orbital muscle volume was determined to be the most important feature in distinguishing normal healthy patients and patients with severe TED, whereas orbital fat volume, followed closely by orbital muscle volume, total orbital volume, and age, was the most important feature for patients with mild TED.

The strengths of this study include its large cohort of patients, and that the volumetric DL model was trained on both normal healthy patients and patients with TED, which allows for higher accuracy in future TED studies. Volume analysis and segmentation in patients with TED is particularly challenging due to the small confined bony space and

crowding at the apex. By training the model on patients with TED rather than a normal cohort alone, the segmentation and volume analysis is more adaptable to the heterogeneity of TED orbital tissues on imaging and improvement in accuracy of measurements. This study was also performed on CTs rather than MRIs, which is more clinically applicable in the real-world. However, this study also has several limitations to consider. As this was a single-centered study, the patient demographic is skewed toward Caucasian patients and may make the model less generalizable in various patient demographics. Additionally, this model only included normal or TED CT scans, and excluded those with other orbital processes or prior surgeries. This again limits the model's generalizability and utility in implementation. In addition, the use of volume data obtained from the segmentation model to train the classification model might introduce some bias, however, we assume that the characteristics of the data differ significantly between the images and the calculated volumes, which we believed would mitigate the risk of bias. Last, the imaging analysis was performed retrospectively and subjects were not required to have fixation targets, which may lead to variation in axis of segmentation or contraction states of the muscle. However, this is more applicable to real-world utilization where subjects are routinely getting imaging. Future studies include applying the model to different patient cohorts, with a more diverse patient demographic, for further optimization and validation. Comparison of this model to the accuracy of human graders, including radiologists and oculoplastic surgeons, should also be explored.

In conclusion, this study demonstrates an accurate and efficient model that automatically segments and precisely calculates the orbital fat and muscle volumes in under thirty seconds. This provides high-quality, quantitative data in analyzing the orbital soft tissues for patients with TED in a rapid fashion. This volumetric model was then utilized to provide precise quantitative volumetric data to build a highly accurate classification model for predicting the presence of TED and TED with CON. The model also synergistically used patient clinical information to optimize the classification accuracy. With further validation, this model may serve as a platform for future clinical studies and real-time incorporation into clinical practice and triage. Overall, this tool may have potential to enhance patient care through supporting broader radiographic based clinical research.

Acknowledgments

The authors gratefully acknowledge the partial support of this project by the Massachusetts Lions Eye Research Fund, Inc.

Disclosure: **A.M. Alkhadrawi**, None; **L.Y. Lin**, None; **S.A. Langarica**, None; **K. Kim**, None; **S.K. Ha**, None; **N.G. Lee**, None; **S. Do**, None

References

1. Alessi DR, Andjelkovic M, Caudwell B, et al. Mechanism of activation of protein kinase B by insulin and IGF-1. *EMBO J*. 1996;15(23):6541–6551.
2. Crudden C, Song D, Cismas S, et al. Below the surface: IGF-1R therapeutic targeting and its endocytic journey. *Cells*. 2019;8(10):1223.
3. Crudden C, Shibano T, Song D. Inhibition of G protein coupled receptor Kinase 2 promotes unbiased downregulation of IGF1 receptor and restrains malignant cell growth. *Cancer Res*. 2021;81(2):501–514.
4. Worrall C, Suleymanova N, Crudden C. Unbalancing p53/Mdm2/IGF-1R axis by Mdm2 activation restrains the IGF-1- dependent invasive phenotype of skin melanoma. *Oncogene*. 2017;36(23):3274–3286.
5. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. *Semin Radiat Oncol*. 2019;29(3):185–197.
6. Paniagua L, Bande MF, Abalo-Lojo JM, Gonzalez F. Computer aided volumetric assessment of orbital structures in patients with Graves' orbitopathy: correlation with serum thyroid antiperoxidase antibodies and disease activity. *Int Ophthalmol*. 2023;43(9):3377–3384.
7. Forbes G, Gorman CA, Gehring D, Baker HL, Jr. Computer analysis of orbital fat and muscle volumes in Graves ophthalmopathy. *AJNR Am J Neuroradiol*. 1983;4(3):737–740.
8. Lee JY, Bae K, Park KA, Lyu IJ, Oh SY. Correlation between extraocular muscle size measured by computed tomography and the vertical angle of deviation in thyroid eye disease. *PLoS One*. 2016;11(1):e0148167.
9. Ma L, Hui S, Li Y, et al. Different characteristics of orbital soft tissue expansion in Graves orbitopathy: extraocular muscle expansion is correlated to disease activity while fat tissue volume with duration. *J Craniofac Surg*. 2022;33(8):2354–2359.
10. Bontzos G, Papadaki E, Mazonakis M, et al. Extraocular muscle volumetry for assessment of thyroid eye disease. *J Neuroophthalmol*. 2022;42(1):e274–e280.
11. Hallin ES, Feldon SE. Graves' ophthalmopathy: I. Simple CT estimates of extraocular muscle volume. *Br J Ophthalmol*. 1988;72(9):674–677.
12. Berger M, Matlach J, Pitz S, et al. Imaging of the medial rectus muscle predicts the development of optic neuropathy in thyroid eye disease. *Sci Rep*. 2022;12(1):6259.
13. Al-Bakri M, Rasmussen AK, Thomsen C, Toft PB. Orbital volumetry in Graves' orbitopathy: muscle and fat involvement in relation to dysthyroid optic neuropathy. *ISRN Ophthalmol*. 2014;2014:435276.
14. Pieroni Gonçalves AC, Silva LN, Gebrim EMMS, Matayoshi S, Monteiro MLR. Predicting dysthyroid optic neuropathy using computed tomography volumetric analyses of orbital structures. *Clinics (Sao Paulo)*. 2012;67(8):891–896.
15. Kim M, Chang JH, Lee NK. Quantitative analysis of extraocular muscle volume and exophthalmos reduction after radiation therapy to treat Graves' ophthalmopathy: a pilot study. *Eur J Ophthalmol*. 2021;31(2):340–345.
16. Law JJ, Mundy KM, Kupcha AC, et al. Correlation of automated computed tomography volumetric analysis metrics with motility disturbances in thyroid eye disease. *Ophthal Plast Reconstr Surg*. 2021;37(4):372–376.
17. Firbank MJ, Harrison RM, Williams ED, Coulthard A. Measuring extraocular muscle volume using dynamic contours. *Magn Reson Imaging*. 2001;19(2):257–265.
18. Regensburg NI, Kok PHB, Zonneveld FW, et al. A new and validated CT-based method for the calculation of orbital soft tissue volumes. *Invest Ophthalmol Vis Sci*. 2008;49(5):1758.
19. Shyu VBH, Hsu CE, Chen CH, Chen CT. 3D-assisted quantitative assessment of orbital volume using an open-source software platform in a Taiwanese population. *PLoS One*. 2015;10(3):e0119589.
20. Bao Y, Zhang Z, Li C, et al. Geometric and volumetric measurements of orbital structures in CT scans in thyroid eye disease classification. *Appl Sci (Basel)*. 2021;11(11):4873.
21. Yu B. Predictive parameters on CT scan for dysthyroid optic neuropathy. *Int J Ophthalmol*. 2020;13(8):1266–1271.

22. Feldon SE, Weiner JM. Clinical significance of extraocular muscle volumes in graves' ophthalmopathy: a quantitative computed tomography study. *Arch Ophthalmol*. 1982;100(8):1266.
23. Zhu F, Gao Z, Zhao C, et al. Semantic segmentation using deep learning to extract total extraocular muscles and optic nerve from orbital computed tomography images. *Optik (Stuttg)*. 2021;244(167551):167551.
24. Kaichi Y, Tanitame K, Terada H, et al. Thyroid-associated orbitopathy: quantitative evaluation of the orbital fat volume and edema using IDEAL-FSE. *Eur J Radiol Open*. 2019;6:182–186.
25. Weis E. Clinical and soft-tissue computed tomographic predictors of dysthyroid optic neuropathy. *Arch Ophthalmol*. 2011;129(10):1332.
26. Lin LY, Zhou P, Shi M, et al. A deep learning model for screening computed tomography imaging for thyroid eye disease and compressive optic neuropathy. *Ophthalmol Sci*. 2023;(100412):100412.
27. Freitag SK, Tanking T. A nomenclature to describe the sequence of visual field defects in progressive thyroid eye disease-compressive optic neuropathy (an American Ophthalmological Society thesis). *Am J Ophthalmol*. 2020;213:293–305.
28. Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods*. 2012;9(7):676–682.
29. Arzt M, Deschamps J, Schmied C, et al. LABKIT: labeling and segmentation toolkit for big image data. *Front Comput Sci*. 2022;4:10.
30. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. New York, NY: Springer International Publishing; 2015:234–241.
31. Liu S, Deng W. Very deep convolutional neural network based image classification using small training sample size. In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE; 2015, doi:10.1109/acpr.2015.7486599.
32. Nembrini S, König IR, Wright MN. The revival of the Gini importance? *Bioinformatics*. 2018;34(21):3711–3718.
33. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232.
34. Tian S, Nishida Y, Isberg B, Lennerstrand G. MRI measurements of normal extraocular muscles and other orbital structures. *Arbeitsphysiologie*. 2000;238(5):393–404.
35. Perros P, Crombie AL, Matthews JNS, Kendall-Taylor P. Age and gender influence the severity of thyroid-associated ophthalmopathy: a study of 101 patients attending a combined thyroid-eye clinic. *Clin Endocrinol (Oxf)*. 1993;38(4):367–372.
36. Krahe T, Schlolaut K, Poß T, Trier H, Lackner K. Computertomographische Volumetrie der Orbita bei endokriner Orbitopathie [article in German]. *Rofo*. 1989;151(11):597–601.
37. Kavoussi S, Giacometti J, Servat J, Levin F. The relationship between sex and symmetry in thyroid eye disease. *Clin Ophthalmol*. 2014;8:1295–1300.
38. Nishida Y, Tian S, Isberg B, Tallstedt L, Lennerstrand G. MRI measurements of orbital tissues in dysthyroid ophthalmopathy. *Arbeitsphysiologie*. 2001;239(11):824–831.
39. Keene KR, van Vught L, van de Velde NM, et al. The feasibility of quantitative MRI of extra-ocular muscles in myasthenia gravis and Graves' orbitopathy. *NMR Biomed*. 2021;34(1):e4407.
40. Song C, Luo Y, Huang W, et al. Extraocular muscle volume index at the orbital apex with optic neuritis: a combined parameter for diagnosis of dysthyroid optic neuropathy. *Eur Radiol*. 2023;33(12):9203–9212.
41. 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(1):1867.
42. Chua M, Kim D, Choi J, et al. Tackling prediction uncertainty in machine learning for healthcare. *Nat Biomed Eng*. 2022;7(6):711–718.

Investigative Ophthalmology & Visual Science