RESEARCH Open Access

Comparative analysis of deep learning architectures for thyroid eye disease detection using facial photographs

Amirhossein Aghajani[1], Mohammad Taher Rajabi[1], Seyed Mohsen Rafizadeh[1], Amin Zand[1], Majid Rezaei[2], Mohammad Shojaeinia[3] and Elham Rahmanikhah[1*]

Abstract

Purpose To compare two artificial intelligence (AI) models, residual neural networks ResNet-50 and ResNet-101, for screening thyroid eye disease (TED) using frontal face photographs, and to test these models under clinical conditions.

Methods A total of 1601 face photographs were obtained. These photographs were preprocessed by cropping to a region centered around the eyes. For the deep learning process, photographs from 643 TED patients and 643 healthy individuals were used for training the ResNet models. Additionally, 81 photographs of TED patients and 74 of normal subjects were used as the validation dataset. Finally, 80 TED cases and 80 healthy subjects comprised the test dataset. For application tests under clinical conditions, data from 25 TED patients and 25 healthy individuals were utilized to evaluate the non-inferiority of the AI models, with general ophthalmologists and fellowships as the control group.

Results In the test set verification of the ResNet-50 AI model, the area under the receiver operating characteristic (ROC) curve (AUC), accuracy, sensitivity, and specificity were 0.94, 0.88, 0.64, and 0.92, respectively. For the ResNet-101 AI model, these metrics were 0.93, 0.84, 0.76, and 0.92, respectively. In the application tests under clinical conditions, to evaluate the non-inferiority of the ResNet-50 AI model, the AUC, accuracy, sensitivity, and specificity were 0.82, 0.82, 0.88, and 0.76, respectively. For the ResNet-101 AI model, these metrics were 0.91, 0.84, 0.92, and 0.76, respectively, with no statistically significant differences between the two models for any of the metrics (all p-values > 0.05).

Conclusions Face image-based TED screening using ResNet-50 and ResNet-101 AI models shows acceptable accuracy, sensitivity, and specificity for distinguishing TED from healthy subjects.

Keywords Artificial intelligence, Residual neural network, Thyroid eye disease, Face photography

*Correspondence:
Elham Rahmanikhah
elhamrahmanikhah@gmail.com
[1]Department of Oculo-Facial Plastic and Reconstructive Surgery, Farabi Eye Hospital, Tehran University of Medical Sciences, Qazvin Square, Tehran, Iran
[2]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[3]Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Introduction

Thyroid eye disease (TED) is a common orbital disease that can be sight-threatening and significantly impact the quality of life, particularly in moderate to severe cases [1, 2]. The diagnosis of TED relies on specific clinical and paraclinical criteria [3]. The initial step in diagnosing TED is a thorough ophthalmological examination, with exophthalmos and eyelid retraction being two key diagnostic signs. Active TED cases may also present with signs such as eyelid swelling, chemosis, and conjunctival injection [4].

The use of facial images for screening TED can potentially reduce the time required for patient visits and referrals. This technique has certain limitations, such as the inability to assess all clinical signs or perform specific examinations. Nonetheless, it facilitates timely diagnosis and management, thereby reducing the risk of unfavorable disease progression in TED [3–5]. This is particularly relevant in developing countries where access to expert ophthalmologists or oculoplastic surgeons may be limited [6]. Given the relatively high prevalence of TED and its controllable natural course, it is both logical and necessary to develop simple, low-cost screening tools for the general population [4, 6].

Computer-aided facial diagnosis systems offer an automated, rapid, and non-invasive method for screening and diagnosing diseases [7]. The application of artificial intelligence (AI), particularly deep learning technology, has significantly improved the diagnostic accuracy for several diseases, including eye diseases [7, 8]. Recent reports have demonstrated the use of AI tools for diagnosing TED based on facial images, showing acceptable accuracy, sensitivity, and specificity [9]. These diagnostic methods vary. Some focus on detecting specific ocular and periocular signs of TED, while others use heat-map analysis to examine facial image pixels [10, 11].

In this study, we compare two residual neural networks (ResNet-50 and ResNet-101), which have superior performance in image recognition tasks compared to traditional convolutional neural networks [12, 13]. Furthermore, previous studies showed although the deeper networks may have stronger learning ability, but as the networks get deeper, degradation issues begin to arise [14–16]. To resolve this problem, ResNet models were introduced residual connections. We conducted application tests under clinical conditions to verify the robustness and reliability of our AI models.

Methods

Data acquisition and labeling

Photographic and clinical data were collected from the medical records of patients evaluated by experienced orbital specialists from March 2020 to February 2024 at the oculoplastic department of Farabi Eye Hospital in Tehran, Iran. The study adhered to the principles of the Declaration of Helsinki, and written informed consent was obtained from each study participant. Approval for the study protocols was obtained from the ethics committee of Tehran University of Medical Sciences, Tehran, Iran (Ethics Code: IR.TUMS.MEDICINE.REC.1403.024).

Two groups of participants were enrolled in this study: (1) patients diagnosed with TED (including mild, moderate-to-severe, and sight-threatening cases) according to their medical records and (2) a control group of healthy subjects with no evidence of TED. The definition and severity of TED was based on the criteria of the European Group on Graves' Orbitopathy (EUGOGO), as described in detail elsewhere [3]. The diagnosis of TED was based on typical ocular signs, such as eyelid retraction, proptosis, restrictive strabismus, eyelid erythema or swelling, chemosis, or compressive optic neuropathy, in combination with immune-related thyroid dysfunction and radiographic evidence from orbital computed tomography (CT) imaging [17]. TED severity was categorized as mild, with lid retraction < 2 mm, mild soft-tissue involvement, proptosis < 3 mm, and no or intermittent diplopia. Moderate-to-severe, characterized by lid retraction of ≥ 2 mm, moderate-to-severe soft-tissue involvement, proptosis of ≥ 3 mm, or constant diplopia. Sight-threatening, defined by the presence of compressive optic neuropathy or corneal breakdown [3].

Patients with incomplete medical records, equivocal diagnoses, non-TED orbital conditions, or Graves' endocrine abnormalities with no evidence of orbitopathy were excluded. Photographs taken at each patient's initial consultation were collected and screened. These photographs were taken using a Canon EOS 7D digital single-lens reflex (DSLR) camera (Canon, Inc., Tokyo, Japan), ensuring that the images included both eyes, eyelids, canthi, brows, forehead, temples, glabella, and nasal dorsum. Images were stored as uncompressed tagged image format files (TIFF). Participants were excluded if the image quality was insufficient, if the photos were poorly focused, or if one or both eyelids were closed. For each participant, a single front-facing photograph with the patient's gaze in the primary position was selected. Each image was preprocessed by cropping to a region centered around the eyes, extending from above the eyebrows to below the lower eyelids, including part of the nasal dorsum and both temples, to exclude irrelevant areas and speed up training. This cropping preprocessing was performed manually by one investigator (E.R.). The images were then scaled to 280 × 460 pixels.

Deep learning of residual neural networks

Two neural networks, ResNet-50 (a residual neural network with 50 layers) and ResNet-101 (a residual neural network with 101 layers), were trained to diagnose the presence or absence of TED based on the training dataset, which included 643 photographs of TED patients and 643 of healthy individuals. The validation dataset (81 TED and 74 normal) was used to determine when to stop training and prevent overfitting. In the test setting, 80 normal and 80 TED photographs from the test dataset were used to evaluate the models.

Deep neural networks are capable of automatically learning features at various levels from images. In this study, we utilized residual neural networks with 50 layers (ResNet-50) and 101 layers (ResNet-101), implemented in Python programming language using PyTorch [14]. The models were pretrained on ImageNet, a dataset of general images designed for object recognition tasks, and then fine-tuned on our training and validation datasets for the specific task of classifying images as TED or healthy subjects. The Adam algorithm was employed to adjust the network weights while minimizing cross-entropy loss [18]. The learning rate was set to 0.0001, and the images were processed in batches of 32. Before being input into the network, the intensities of each image were normalized by subtracting the mean value and then dividing by the standard deviation. The training process used epochs as units, meaning all images in the training set were processed once per epoch. At the end of each epoch, the internal validation set was used to test the model, generating validation-accuracy and validation-loss values. Training stopped if the validation-loss did not decrease for 50 consecutive epochs. The epoch with the lowest loss on the validation set was selected as the final model, which was then used to predict the independent test set (Fig. 1).

Application tests

To evaluate the performance of our AI models, an application test was designed involving 50 cases (25 TED and 25 normal), assessed by the AI models (experimental groups), and by three general ophthalmologists and three

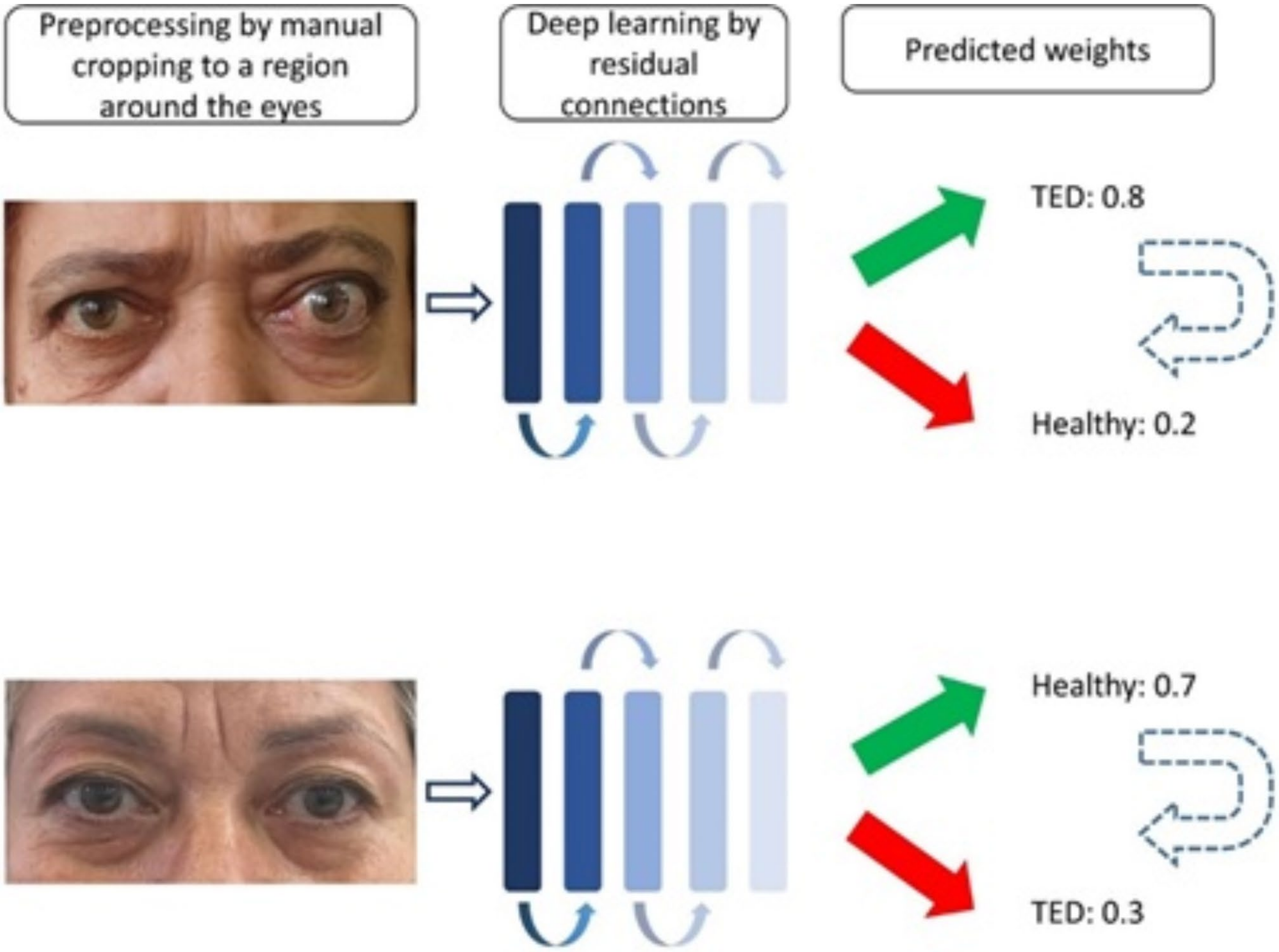
Preprocessing by manual cropping to a region around the eyes
Deep learning by residual connections
Predicted weights
TED: 0.8
Healthy: 0.2
Healthy: 0.7
TED: 0.3

Fig. 1 The schematics illustrate the training process for the residual neural networks. The baseline photographs were manually cropped around the eye region, as described in detail. The neural networks then automatically learn features at different levels (each represented by a rectangle) from the images. Residual connections (curved arrows) were used to mitigate degradation issues. After global average pooling of the features from different levels, the model predicted a weight for a binary thyroid eye disease (TED) or healthy diagnosis

Table 1 Demographics and clinical characteristics among the TED and healthy groups

Variable	TED (*n* = 643)	Healthy (*n* = 643)	*P*
Age (years), mean ± SD (range)	50.64 ± 12.01 (23–78)	48.92 ± 12.53 (20–75)	0.917
Gender, n (%)	Female: 332 (51.6) Male: 311 (48.4)	Female: 412 (64.1) Male: 231 (35.9)	0.192
Severity of TED, n (%)	Mild: 81 (12.6) Moderate-to-severe: 476 (74.0) Sight-threatening: 86 (13.4)		

TED: Thyroid eye disease, n: number

fellowships of oculoplastic surgery (control groups). The aim was to demonstrate the non-inferiority of the AI models compared to general ophthalmologists and fellowships in distinguishing TED from healthy subjects based on facial photographs. With a non-inferiority margin of 10% and a standard deviation of 10%, and considering a one-sided type I error of 5.0% and 80% power, the sample size was estimated to be 50 face photos, considering a 95% confidence interval [19].

Interpretability of the models

To address the inherent black-box nature of deep learning models, we applied the occlusion sensitivity method on test set images used in the application tests to demonstrate model interpretability [20, 21]. Briefly, this method systematically occludes various regions of an input image and observes the effect on the model's output. By assessing the changes in the model's predictions when different parts of the image are occluded, this approach provides insights into the importance of various image regions. The method generates heatmaps that visualize which areas of the image contributed most to the model's prediction for a specific class [20, 22].

Statistical analysis

Statistical analysis was performed using SPSS software version 24 (Chicago, IL). The demographic characteristics of the TED and healthy groups were compared using a *t*-Test for age and a chi-square test for gender. A general linear model was applied to compare accuracy, sensitivity, and specificity metrics among the experimental and control subgroups during the application test. Pairwise comparisons with Bonferroni correction were used for further analysis. A p-value of < 0.05 was considered statistically significant.

Table 2 Performance of TED ResNet-50 screening model on the training set, validation set, and testing set

	AUC	Accuracy	Sensitivity	Specificity
Training Set	0.94	0.87	0.87	0.88
Validation Set	0.96	0.89	0.91	0.86
Testing Set	0.94	0.88 (CI: 0.83–0.93)	0.64 (CI: 0.57–0.71)	0.92 (CI: 0.88–0.96)

TED: thyroid eye disease, ResNet: residual neural network, AUC: area under the receiver operating characteristic curve, CI: 95% confidence interval

Results

Demographics and clinical characteristics

A total of 643 frontal face photographs were included in both the TED and healthy groups. The demographic data are summarized in Table 1. The mean age in the TED group was 50.64 ± 12.01 years, compared to 48.92 ± 12.53 years in the healthy group, with no statistically significant difference ($P = 0.917$). Additionally, no significant difference was observed in gender distribution between the TED group (51.6% female) and the healthy group (64.1% female) ($P = 0.192$). Regarding TED severity, 81 cases (12.6%) were classified as mild, 476 cases (74.0%) as moderate to severe, and 86 cases (13.4%) as sight-threatening.

ResNet-50 AI model

After 30 epochs of training, the ResNet-50 TED AI screening model achieved an accuracy of 0.88 on the test set, with sensitivity and specificity of 0.64 and 0.92, respectively. Additionally, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were used to evaluate the model's performance. The algorithm achieved an AUC of 0.94, 0.96, and 0.94 on the training, validation, and test sets, respectively. The performance of the model on the three datasets (training, validation, and test) is summarized in Table 2.

ResNet-101 AI model

After 100 epochs of training, the ResNet-101 TED AI screening model achieved an accuracy of 0.84 on the test set, with sensitivity and specificity of 0.76 and 0.92, respectively. The ROC curve and AUC were also employed to evaluate the model's performance. The algorithm achieved an AUC of 0.93, 0.96, and 0.93 on the training, validation, and test sets, respectively. The performance of the model on the three datasets is summarized in Table 3.

Application tests

In the evaluation of 50 face images (25 TED and 25 healthy individuals), the general ophthalmologist group had false positive and negative rates of 9.3% and 13.3%, respectively. The fellowship group had false positive and negative rates of 5.3% and 10.7%, respectively. For the ResNet-50 AI model, the false positive rate was 24.0%, and the false negative rate was 12.0%. For the ResNet-101

Table 3 Performance of TED ResNet-101 screening model on the training set, validation set, and testing set

	AUC	Accuracy	Sensitivity	Specificity
Training Set	0.93	0.86	0.86	0.86
Validation Set	0.96	0.86	0.78	0.95
Testing Set	0.93	0.84 (CI: 0.78–0.90)	0.76 (CI: 0.69–0.83)	0.92 (CI: 0.88–0.96)

TED: thyroid eye disease, ResNet: residual neural network, AUC: area under the receiver operating characteristic curve, CI: 95% confidence interval

Table 4 Performance of TED screening using face photographs in experimental (ResNet-50 and ResNet-101 AI models) and control (general ophthalmologists and fellowships) groups

	ResNet-50	ResNet-101	General ophthalmologists	Fellowships	*P**
Accuracy (%)	82.0	84.0	88.7	92.0	0.413
Sensitivity (%)	88.0	92.0	86.7	89.3	0.947
Specificity (%)	76.0	76.0	90.7	94.7	0.052

TED: thyroid eye disease, ResNet: residual neural network, AI: artificial intelligence

* general linear model with pairwise Bonferroni corrections

AI model, the false positive and negative rates were 24.0% and 8.0%, respectively. The accuracy rates were 88.7% and 92.0% in the general ophthalmologist and fellowship groups (control groups), respectively. For the ResNet-50 and ResNet-101 AI models, the accuracy rates were 82.0% and 84.0%, respectively. According to the general linear model with pairwise Bonferroni corrections, no statistically significant differences were observed in accuracy, sensitivity, or specificity metrics between the ResNet-50 AI model, ResNet-101 AI model, general ophthalmologists, or fellowship groups (all p-values > 0.05, Table 4). Confusion matrices showing breakdown by class are shown in Fig. 2. The algorithm achieved an AUC of 0.82, and 0.91 on the application test for ResNet-50 and ResNet-101 models, respectively (Fig. 3).

The occlusion sensitivity method was used to identify which areas of the facial photographs were most important for TED detection by the AI models. According to the heatmap analysis, the pixels corresponding to the periocular regions were most strongly associated with TED detection by the models (Fig. 4).

Discussion

Our study demonstrated the high diagnostic performance of the AI models in identifying features of TED relevant to disease assessment. We showed that face image-based TED screening using both ResNet-50 and ResNet-101 AI models achieved acceptable accuracy, sensitivity, and specificity for distinguishing TED from healthy subjects, when compared to the control subgroups (general ophthalmologists and fellowships). Furthermore, we found no statistically significant differences in the performance of the two tested AI models.

Various AI models have been developed for TED screening, with some based on face photographs and others on paraclinical diagnostic tools like orbital CT imaging [9]. Song et al. showed that ResNet-18 is an accurate model (87%) with considerable sensitivity (88%) and specificity (85%) for screening TED patients using orbital

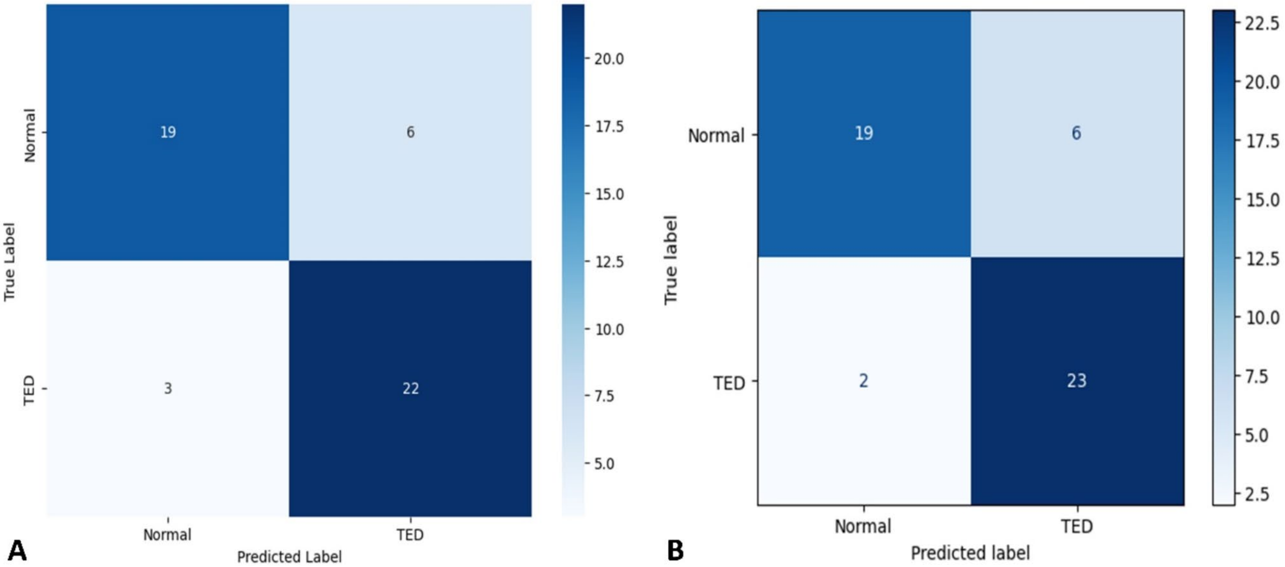

Fig. 2 Confusion matrices for the classification of Thyroid Eye Disease (TED) using the ResNet-50 (**A**) and ResNet-101 (**B**) models on a clinical application test set of 50 face images. The numbers indicate the counts for each actual and predicted class

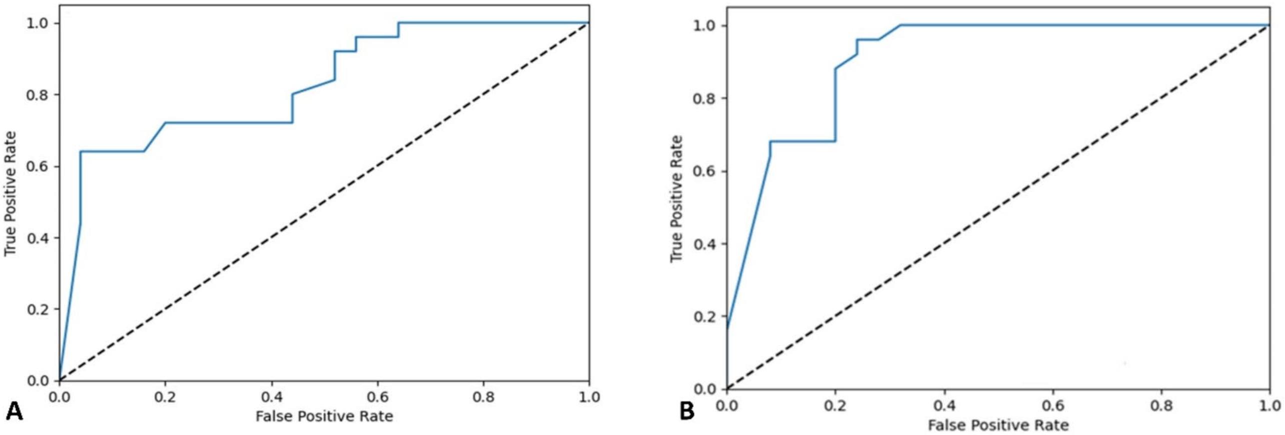

Fig. 3 Receiver operating characteristic (ROC) curve for application test of ResNet-50 (**A**) and ResNet-101 (**B**) models. The area under curve (AUC) was 0.82 for ResNet-50 and 0.91 for ResNet-101

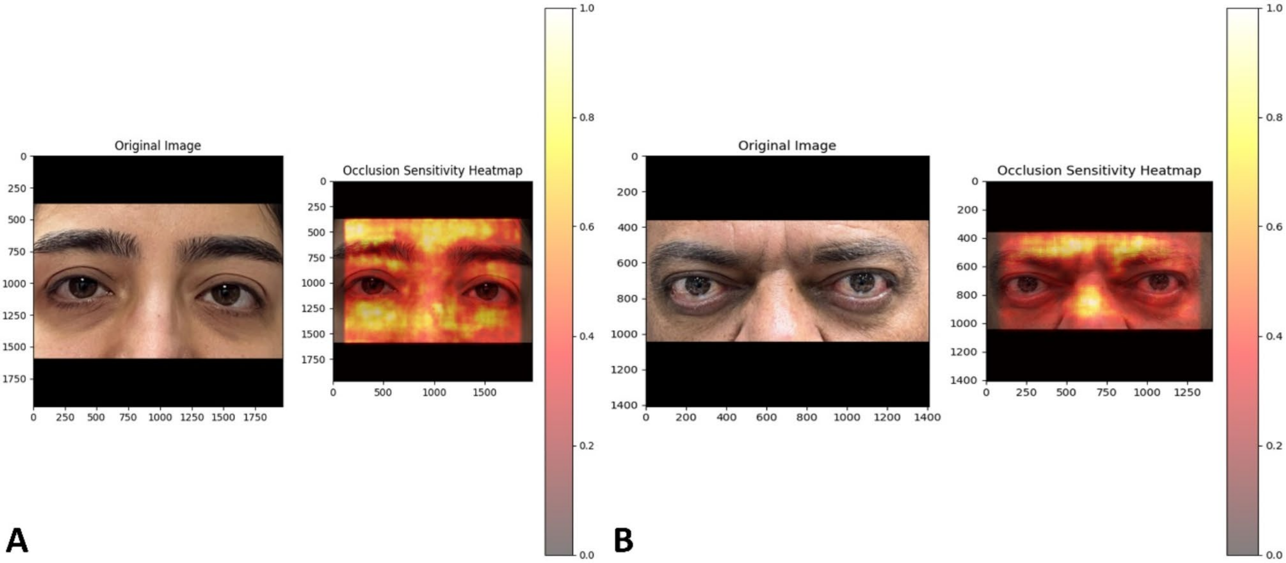

Fig. 4 Visual interpretability of cropped frontal face photographs using the occlusion sensitivity method. Two representative photos from the test dataset are shown. Cooler (yellow) colors indicate areas with little to no activation for TED detection. In a healthy case (**A**), the heatmap predominantly displays cooler (yellow) tones in the periocular region. In contrast, in a case with bilateral TED (**B**), these cooler (yellow) tones are notably reduced around the eyes, emphasizing the areas most relevant to TED detection

CT imaging [23]. However, CT imaging has limitations as a screening tool, including availability, exposure to ionizing radiation, high variability in imaging techniques and quality, and cost. In contrast, screening TED via face photographs is generally more available, low-cost, non-office-based, and potentially more generalizable through the development of self-evaluation applications as new tools in promoting public health. Furthermore, AI models have been used to predict postoperative facial appearance after orbital decompression surgery in TED patients, showing promising results [24].

Previous studies have used different criteria to diagnose TED with AI models based on facial images [9]. Karlin et al. used a gradient-weighted class activation mapping (Grad-CAM) AI model with heat-map analysis of face photographs for TED detection. This technique highlighted the pixels in the ocular and periocular areas that were most affected by the disease. Their model achieved a test set accuracy of 89.2%, specificity of 86.9%, and sensitivity of 93.4% [10]. Shao et al. developed an automatic system to measure eyelid position in TED cases by assessing margin to reflex distance (MRD) for both upper (MRD-1) and lower (MRD-2) lids [11]. They compared the repeated automatic (by their AI model) and manual (by experts) measurements of MRD1 and MRD2 in both TED and normal subjects. They found the intraclass correlation coefficients (ICCs) between repeated automatic measurements of MRDs were up to 0.998 ($P<0.001$), showing the high repeatability of their AI model. Huang et al. developed a diagnostic method based on modules evaluating eye location (Module I), ocular dyskinesis (Module II), and other TED signs (Module III), showing a mean AUC of 0.85, sensitivity of 80%, and specificity of 79% [25]. Images of three face positions and nine eye positions are included in their study. However, we used just a single front-face photograph for screening of TED for each subject in our study. In our study, TED was diagnosed by two AI models (ResNet-50 and ResNet-101) using facial images to detect several TED signs after deep learning. Regardless of the methods used for TED detection in our study and previous studies, the AUC, accuracy, specificity, and sensitivity of our test dataset were comparable with those reported in similar studies. In this study, we utilized deeper ResNet models compared to previous researches to enhance performance in detecting TED using facial photographs. ResNet architectures are widely recognized for their effectiveness in image recognition tasks, owing to their ability to train deep networks efficiently [13]. Additionally, the residual connections in ResNet models address concerns about degradation and overfitting, especially in deeper architectures like ResNet-50 and ResNet-101 compared to ResNet-18. This feature significantly mitigates validation challenges that are often encountered with traditional neural networks [14–16].

In the study by Karlin et al., compared to expert clinicians, the AI model demonstrated higher sensitivity (89% vs. 58%) but lower specificity (84% vs. 90%) in detecting TED using facial images [10]. However, in our study, the sensitivity, specificity, and accuracy rates were similar between the AI models and the control group (general ophthalmologists and fellowships). Discrepancies between our findings and previous literature may be due to differences in the AI models used, characteristics of the face photos of TED and healthy subjects, factors used for TED detection via AI, and the experience level of the control groups.

Deep convolutional neural networks can automatically learn features from images; the deeper the network, the stronger the learning ability. However, deeper networks face degradation and overfitting issues [14, 16, 26]. ResNet addresses these problems through residual connections [27]. In this study, we used ResNet-50 and ResNet-101 without significant errors in training, validation, or test sets. We found that the ResNet-50 model had a lower testing error than the ResNet-101 model (12% vs. 16%), with nearly similar train and validation errors. Furthermore, the AUC for ResNet-50 and ResNet-101 models were similar to each other (0.94 vs. 0.93) in our test dataset. Karlin et al. showed that a ResNet-18 ensemble model achieved a test set accuracy of 89.2% for distinguishing TED from non-TED in face photographs [10]. Huang et al. reported similar mean AUCs for detecting different TED signs between ResNet-50 and ResNet-101 (0.91 vs. 0.92) [25]. These negligible discrepancies can be explained by previously mentioned factors.

In the application tests under clinical conditions, we showed that the ResNet-50 AI model, the accuracy, sensitivity, and specificity were 0.82, 0.88, and 0.76, respectively. For the ResNet-101 AI model, these metrics were 0.84, 0.92, and 0.76, respectively. Furthermore, the occlusion sensitivity method confirmed that the models accurately localized clinically relevant areas for TED detection in the photographs. As shown in the representative heatmaps, the model focused on the periocular region, which clinicians consider important for diagnosing TED based on facial photographs [3].

We observed no significant difference in the performance metrics of ResNet-50 and ResNet-101 for TED detection using face photographs. Therefore, increasing the number of layers and deepening the model (ResNet-101) does not negatively impact the performance of ResNet models in our study, even with the potential risk of overfitting. As previously mentioned, this phenomenon can be attributed to the residual connections inherent in ResNet architectures [14]. Karlin et al. found that their ResNet-18 ensemble model for screening TED based on face photographs achieved accuracy 0.86, sensitivity 0.89 and specificity 0.84 at application test step after completing deep learning processes [10].

In this study, we compared the performance of two AI models for TED screening to that of trained specialists using only frontal face photographs, without additional diagnostic tools like exophthalmometry or orbital imaging. While this basic screening method may have lower performance compared to comprehensive assessments, it offers a potential solution in regions with limited access to specialists and advanced diagnostic equipment. The future goal is to develop AI-based self-evaluation applications, promoting public health and facilitating earlier detection of TED, particularly in underserved areas.

This study had some limitations. First, we did not classify TED cases by disease severity or activity. Performing stratified analysis based on TED severity or activity would further clarify the performance of these AI models, particularly in distinguishing mild or non-active disease. However, as mentioned previously, the focus of our study was to introduce AI models for distinguishing TED from healthy subjects; with an emphasis on their potential as screening tools, especially in regions with limited access to ophthalmologists or paraclinical investigations. Second, we used just a single front-face photograph for screening of TED. Using images of nine eye positions may increase the performance of the AI models for screening of TED as a disease with various ocular diskynesis presentations. Third, the model's generalizability to other clinical centers was not evaluated. Future studies should test the accuracy and sensitivity of our ResNet TED AI screening models using new collections of TED patient images from different centers. Fourth, although our evaluated ResNet models had good accuracy with sufficient external validation, there is a gap between these models and real clinical challenges in diagnosing TED [28]. To address this, combining face photograph data with other examinations necessary for TED evaluation, such as exophthalmometry measurements, thyroid function tests, and orbital CT imaging in AI models, may improve the sensitivity and specificity of TED detection and diagnosis.

In conclusion, promising face image-based TED screening ResNet-50 and ResNet-101 AI models were established and passed application tests under clinical conditions. Both models had acceptable accuracy, sensitivity, and specificity for distinguishing TED from healthy subjects based on face photos. Therefore, applications for self-evaluation using these AI models could be developed as new tools in promoting public health.

Abbreviations

TED	Thyroid eye disease
AI	Artificial intelligence
ResNet	Residual neural network
ROC	Receiver operating characteristic curve
AUC	Area under the receiver operating characteristic curve
CT	Computed tomography
MRD	Margin to reflex distance

Acknowledgements

The authors are grateful to the research participants and all the hospital staff who took interest and helped out in the study.

Author contributions

A.A. and M.T.R. designed the study and supervised the project. A.A., M.T.R., S.M.R. and E.R.captured the photographs and collected the data. M.R. and M.S. conducted the deep learning modeling and analysis. A.A., E.R., and A.Z. wrote and revised the main manuscript text. All the authors read and approved the final manuscript.

Funding

This study was not supported by any funding.

Data availability

The datasets used during the current study are available on the Eye Research Center of the Farabi Eye Hospital, Tehran, Iran. The data is not available publicly due confidentiality. However, upon a reasonable request, the data can be obtained from the corresponding author.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Tehran University of Medical Sciences (ethics code: IR.TUMS.MEDICINE.REC.1403.024). The research adhered to the ethical guidelines stipulated in the Declaration of Helsinki. Prior to enrollment in the study, written informed consent was secured from all participants.

Consent for publication

Consent for publication was obtained from the participants.

Competing interests

The authors declare no competing interests.

Received: 11 September 2024 / Accepted: 17 March 2025
Published online: 01 April 2025

References

1. Bahn RS. Graves' ophthalmopathy. N Engl J Med. 2010;362(8):726–38. https://doi.org/10.1056/NEJMra0905750.
2. Bruscolini A, Sacchetti M, La Cava M, et al. Quality of life and neuropsychiatric disorders in patients with graves' orbitopathy: current concepts. Autoimmun Rev. 2018;17(7):639–43. https://doi.org/10.1016/j.autrev.2017.12.012.
3. Bartalena L, Baldeschi L, Boboridis K, et al. The 2016 European thyroid Association/European group on graves' orbitopathy guidelines for the management of graves' orbitopathy. Eur Thyroid J. 2016;5(1):9–26. https://doi.org/10.1159/000443828.
4. Szelog J, Swanson H, Sniegowski MC, Lyon DB. Thyroid eye disease. Mo Med. 2022;119(4):343–50. http://www.ncbi.nlm.nih.gov/pubmed/36118816.
5. Kahaly GJ, Grebe SKG, Lupo MA, McDonald N, Sipos JA. Graves' disease: diagnostic and therapeutic challenges (Multimedia Activity). Am J Med. 2011;124(6):S2–3. https://doi.org/10.1016/j.amjmed.2011.03.001.
6. Yu C, Ford R, Wester S, Shriver E. Update on thyroid eye disease: regional variations in prevalence, diagnosis, and management. Indian J Ophthalmol. 2022;70(7):2335. https://doi.org/10.4103/ijo.IJO_3217_21.
7. Attallah O. A deep learning-based diagnostic tool for identifying various diseases via facial images. Digit Heal. 2022;8:20552076221124432. https://doi.org/10.1177/20552076221124432.
8. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43. https://doi.org/10.1136/svn-2017-000101.
9. Chng CL, Zheng K, Kwee AK, et al. Application of artificial intelligence in the assessment of thyroid eye disease (TED) - a scoping review. Front Endocrinol (Lausanne). 2023;14. https://doi.org/10.3389/fendo.2023.1300196.
10. 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(11):1722–9. https://doi.org/10.1136/bjo-2022-321833.
11. Shao J, Huang X, Gao T, et al. Deep learning-based image analysis of eyelid morphology in thyroid-associated ophthalmopathy. Quant Imaging Med Surg. 2023;13(3):1592–604. https://doi.org/10.21037/qims-22-551.
12. Zhao X, Wang L, Zhang Y, Han X, Deveci M, Parmar M. A review of convolutional neural networks in computer vision. Artif Intell Rev. 2024;57(4):99. https://doi.org/10.1007/s10462-024-10721-6.
13. Wang H, Li K, Xu C. A New Generation of ResNet Model Based on Artificial Intelligence and Few Data Driven and Its Construction in Image Recognition Model. Ning X, ed. Comput Intell Neurosci. 2022;2022:1–10. https://doi.org/10.1155/2022/5976155
14. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In:, Recognition P. (CVPR). IEEE; 2016:770–778. https://doi.org/10.1109/CVPR.2016.90
15. Li H, Li J, Guan X, Liang B, Lai Y, Luo X. Research on Overfitting of Deep Learning. In:, Intelligence, Security. (CIS). IEEE; 2019:78–81. https://doi.org/10.1109/CIS.2019.00025
16. Lawrence S, Giles CL. Overfitting and neural networks: conjugate gradient and backpropagation. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium. IEEE; 2000:114–119 vol.1. https://doi.org/10.1109/IJCNN.2000.857823
17. Barrio-Barrio J, Sabater AL, Bonet-Farriol E, Velázquez-Villoria Á, Galofré JC. Graves' ophthalmopathy: VISA versus EUGOGO classification, assessment, and management. J Ophthalmol. 2015;2015:1–16. https://doi.org/10.1155/2015/249125.
18. Kingma DP, Ba J, Adam. A method for stochastic optimization. ArXiv Published Online 2017:14126980.
19. Flight L, Julious SA. Practical guide to sample size calculations: non-inferiority and equivalence trials. Pharm Stat. 2016;15(1):80–9. https://doi.org/10.1002/pst.1716.
20. Khodabandeh Z, Rabbani H, Bidabadi NS, Bonyani M, Kafieh R. Comprehensive evaluation of artificial intelligence models for diagnosis of multiple sclerosis using information from retinal layers multicenter OCT images. Published Online March. 2024;6. https://doi.org/10.1101/2024.03.05.24303789.

21. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. JAMA Ophthalmol. 2019;137(12):1353. https://doi.org/10.1001/jamaophthalmol.2019.3501.
22. Zhang Qshi, Zhu S. chun. Visual interpretability for deep learning: a survey. Front Inf Technol Electron Eng. 2018;19(1):27–39. https://doi.org/10.1631/FITEE.1700808
23. 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(2):323–30. https://doi.org/10.1007/s11548-020-02281-1.
24. Yoo TK, Choi JY, Kim HK. A generative adversarial network approach to predicting postoperative appearance after orbital decompression surgery for thyroid eye disease. Comput Biol Med. 2020;118:103628. https://doi.org/10.1016/j.compbiomed.2020.103628.
25. Huang X, Ju L, Li J, et al. An intelligent diagnostic system for Thyroid-Associated ophthalmopathy based on facial images. Front Med. 2022;9. https://doi.org/10.3389/fmed.2022.920716.
26. Habibollahi Najaf Abadi H, Modarres M. Predicting system degradation with a guided neural network approach. Sensors. 2023;23(14):6346. https://doi.org/10.3390/s23146346.
27. He F, Liu T, Tao D. Why ResNet works?? Residuals generalize. IEEE Trans Neural Networks Learn Syst. 2020;31(12):5349–62. https://doi.org/10.1109/TNNLS.2020.2966319.
28. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799. https://doi.org/10.1136/bmjopen-2016-012799.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.