Causality-Inspired Graph Neural Network for Interpretable Strabismus Subtype Classification

Jiawen Zheng[1] *, Li Luo[2] *, Jiafan Zhuang[1] *, Peiwei Wei[3], Lihao Zhong[1], Xiaoling Xie[2], Jinming Guo[2], Meng Xie[4], Xiaoli Kang[4], Jie Cen[4], Lingyan Dong[4], Ce Zheng[4] †, and Zhun Fan[5] †

[1] Shantou University, Shantou, China
{23jwzheng, jfzhuang}@stu.edu.cn
[2] Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong, shantou, China
luoli1208@163.com
[3] MedicaI College, Shantou University, Shantou, China
[4] Department of Ophthalmology, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
zhengce@xinhuamed.com.cn
[5] Shenzhen Institute for Advanced Study, UESTC, Shenzhen, China
fanzhun@uestc.edu.cn

Abstract. Despite advances in deep learning, current automated methods for strabismus classification face two key challenges: limited interpretability and a lack of focus on strabismus subtypes. These issues undermine clinical trust, hinder practical adoption, and limit personalized treatment. To address this, we propose a Causality-Inspired Graph Neural Network (CI-GNN) framework that identifies causally related visual features from eye regions and constructs a graph structure for robust prediction, moving beyond reliance on raw image pixels. This causality-driven design enhances both interpretability and clinical relevance by providing more transparent diagnostic outcomes. We also establish a representative benchmark for strabismus subtype classification, focusing on deviation direction and horizontal angle variation (e.g., A/V-pattern). Experiments show that our method achieves state-of-the-art accuracy—89.8% and 88.1% on the two subtype tasks, respectively. Furthermore, by incorporating the SHapley explanation technique, CI-GNN offers clinician-friendly diagnostic evidence. Leveraging sparse causal features, the framework requires only 0.0003 GFLOPs, making it highly efficient and suitable for edge deployment. Overall, this work demonstrates the potential of integrating causal knowledge with GNNs to significantly enhance the performance, efficiency, and interpretability of strabismus diagnosis, offering promising directions for intelligent medical applications.

Keywords: Strabismus Subtype Classification · Causality · Graph Neural Network.

[1] * Co-first authors, these authors contributed equally to this work.
[2] † Co-corresponding author.

1 Introduction

Strabismus is a condition characterized by abnormal eye alignment, preventing both eyes from focusing on the same target simultaneously, resulting in the inability of the visual axes to intersect at the point of fixation. Approximately 3% of the global population is affected, making strabismus one of the most common visual disorders in children [1, 2]. Without timely diagnosis and intervention, strabismus can lead to significant binocular vision impairment. Early diagnosis and surgical correction, especially during critical periods of visual development, can effectively restore proper alignment and prevent long-term visual problems. Therefore, early screening and accurate diagnosis are essential[3–5].

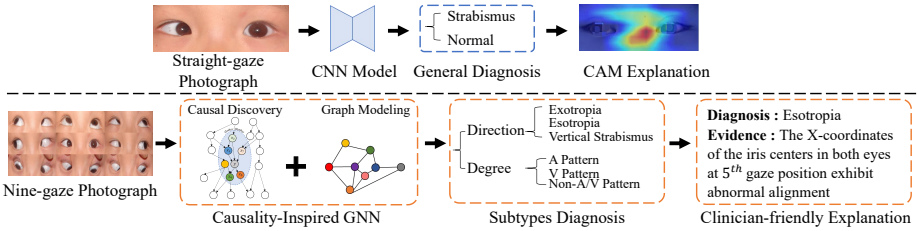

Fig. 1. Comparison of Previous Works and Our Approach. The upper section of the figure summarizes existing methods, while the lower section illustrates the approach proposed in this work, which addresses two major challenges: limited interpretability and the lack of subtype classification.

Traditionally, strabismus diagnosis and accurate subtype classification rely on specalized ophthalmic instruments, which are primarily available in tertiary hospitals. Recent advances in deep learning (DL) have shown promise in improving accessibility for strabismus diagnosis. However, while DL-based methods provide high diagnostic accuracy and adaptability, their "black-box" nature hinders interpretability and limits clinician trust[6–9]. Moreover, existing systems typically offer binary classification[10, 11], detecting only the presence of strabismus, which fails to capture fine-grained subtypes (*e.g.*, direction or angle changes), limiting their utility in personalized treatment planning and surgical strategy optimization[12–15].

To address these challenges, we propose the Causality-Inspired Graph Neural Network (CI-GNN) framework, which is compatible with mainstream causal discovery methods and is both scalable and adaptable, as shown in Fig. 1. By incorporating causal relationships, our approach enhances interpretability and offers a more clinically relevant model. The core innovation lies in automatically identifying and analyzing causally-related visual features from raw image pixels, which are then used to construct a clinician-interpretable, graph-like representation. Notably, these features align with strabismus clinical guidelines [16], confirming their biological relevance based on our experimental results. Leveraging

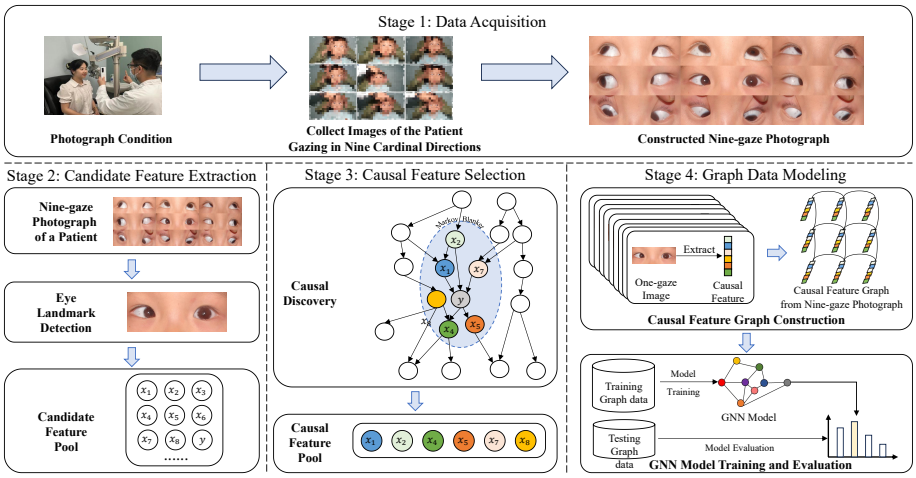

Fig. 2. Overview of our CI-GNN framework. The framework consists of four stages, including data acquisition, candidate feature extraction, causal feature selection, and graph data modeling.

these key representations, we apply a graph neural network for robust modeling and prediction, while providing clinician-friendly diagnostic evidence by highlighting abnormal visual features. This enables more intuitive and interpretable diagnoses than traditional DL models.

To advance strabismus subtype classification, we develop a benchmark targeting key subtypes: strabismus deviation direction (*e.g.*, esotropia, exotropia, vertical) and horizontal angle variations between upgaze and downgaze (*e.g.*, A-pattern, V-pattern, non-AV-pattern). Extensive experiments show that our framework outperforms state-of-the-art models in both accuracy and interpretability.

By leveraging causally-related graph features instead of raw image pixels, our method reduces computational cost to just 0.0003 GFLOPs—approximately $1/100,000$ of traditional CNNs—making it ideal for edge deployment(*e.g.*, Raspberry Pi), in resource-constrained settings.

This work demonstrates the potential of integrating causal knowledge with graph neural networks to significantly enhance the performance, efficiency, and interpretability of strabismus subtype classification, offering valuable insights for intelligent medical systems.

2 Method

2.1 Framework

We propose a Causality-Inspired Graph Neural Network (CI-GNN) framework for interpretable strabismus subtype classification, as illustrated in Fig. 2. In the

data acquisition stage, patients are instructed to gaze in nine specific directions, following the Hirschberg test [17], to capture abnormal eye movement patterns crucial for subtype classification. Facial landmarks are then detected to construct a nine-gaze photograph in a bottom-to-top, left-to-right order for structured representation.

The core innovation of our framework lies in extracting high-level candidate features from raw image pixels and incorporating a causal discovery mechanism to identify critical diagnostic features (Stages 2 and 3 in Fig. 2). The identified causal features are then structured into a graph representation based on the nine-gaze topology. Finally, a graph neural network (GNN) is employed for data modeling and prediction, ensuring both diagnostic accuracy and interpretability (Stage 4 in Fig.2).

2.2 Candidate Feature Extraction

Interpretable model behavior requires extracting high-level features from raw image pixels, such as the iris and orbital centers. To achieve this, we perform candidate feature extraction guided by clinical guidelines and ophthalmologists' observations, effectively integrating expert knowledge.

Since diagnosing deviation direction and angle changes involves different information, we construct separate candidate feature pools for each task. Table 1 presents the features used for deviation direction diagnosis. This transformation converts raw image data into clinician-interpretable features, establishing the basis for explainable diagnosis.

Table 1. Candidate Features for Strabismus Direction Classification

Candidate Features	**Meaning**
(x^L_{iris}, y^L_{iris}), (x^R_{iris}, y^R_{iris})	Coordinates of the left and right iris centers
(x^L_{eye}, y^L_{eye}), (x^R_{eye}, y^R_{eye})	Coordinates of the left and right orbital centers
(x^L_{ls}, y^L_{ls}), (x^R_{ls}, y^R_{ls})	Coordinates of the left and right eye light spots
d_{ls}	Distance between the left and right eye light spots
d_{irls}	Distance between the centers of the left and right irises
$\theta_{deviation}$	Cosine angle between the left and right eye light spots

2.3 Causal Feature Selection

The constructed candidate feature pool contains both continuous variables (*e.g.*, coordinate points) and discrete variables (*e.g.*, strabismus diagnostic categories such as direction and angle). To extract essential diagnostic features, we employ

a score-based causal reasoning approach for mixed-type data, inspired by the HCM algorithm [18].

First, we apply the PC [19] algorithm and MRCIT [18] for skeleton learning, followed by a greedy algorithm for causal directed acyclic graph (DAG) construction. The process initializes an empty DAG and iteratively adds edges based on the maximum score gain.

The initial score for the j-th candidate feature from the candidate feature pool is defined as:

$$S(\emptyset, X_j) = \frac{1}{n} \sum_{i=1}^{n} \log \left(P_r(X_j = x_{i,j}) \right), \tag{1}$$

where X_j represents the j-th feature, and $x_{i,j}$ is its value for the i-th observation. $P_r(X_j = x_{i,j})$ denotes the empirical probability, estimated via frequency for discrete variables and kernel density estimation (KDE)[20] for continuous variables.

At each iteration, the gain from adding a potential edge $l \to j$ is computed as:

$$S_j(P_j^G \cup \{l\}; X_j). \tag{2}$$

The edge yielding the highest gain is added to the DAG, ensuring the most informative causal relationships for strabismus diagnosis. The learned DAG is subsequently pruned using MRCIT to test conditional independence among parent-child pairs, eliminating redundant edges. The final causal DAG encapsulates the most significant diagnostic dependencies.

To refine feature selection, we extract the Markov blanket[21] of the target variable y (strabismus diagnosis outcome), comprising its direct causes, direct effects, and variables that render y conditionally independent from the rest. This ensures a compact yet informative causal feature set, enhancing both interpretability and efficiency while mitigating overfitting and aligning with clinical diagnostic reasoning.

2.4 Graph Data Modeling

We define the graph structure for strabismus diagnosis as:

$$G = (\nu, \varepsilon, \chi), \tag{3}$$

where $\nu = \{v_1, v_2, \ldots, v_9\}$ denotes the nine nodes corresponding to gaze positions in the nine-gaze photograph. The graph is fully connected, i.e., $\varepsilon \subseteq \nu \times \nu$. The node feature matrix $X \subseteq \mathbb{R}^{N \times d}$ contains d-dimensional feature vectors $X_i \subseteq \mathbb{R}^d$ for each node v_i where d is the number of selected causal variables for classification. We use a Graph Convolutional Network(GCN)[22] to extract features by propagating information through the graph. At each layer, node features are updated using a normalized adjacency matrix and a learnable weight matrix W. The graph convolution is defined as:

$$H^{(l+1)} = \sigma(\hat{A} H^{(l)} W^{(l)}), \tag{4}$$

Table 2. Discovered Causal Features for Strabismus Subtype Classification

Deviation Direction	**Meaning**
(x^L_{iris}, y^L_{iris}), (x^R_{iris}, y^R_{iris})	Coordinates of the left and right iris centers
x^L_{ls}, x^R_{ls}	The x coordinates of the left and right eye light spots
x^L_{iris}	The x coordinates of the left iris centers
Deviation Angle	**Meaning**
θ^L_{eye} θ^R_{eye}	The cosine angle between the iris and orbital centers of the eyes
$\Delta\theta_2$, $\Delta\theta_7$ $\Delta\theta_8$, $\Delta\theta_9$	The angle difference $\Delta\theta_i$ between the eyes at the i^{th} region in nine-gaze photograph
$\Delta\theta_{1-7}$, $\Delta\theta_{2-8}$,$\Delta\theta_{3-9}$	The angle difference $\Delta\theta_{i-j}$ between the i^{th} and j^{th}

where $H^{(l)}$ is the node feature matrix at layer l, $\hat{A} = D^{-1/2}AD^{-1/2}$ is the normalized adjacency matrix, A is the adjacency matrix, D is the degree matrix, and σ is the ReLU[23] activation function. $W^{(l)}$ is the learnable weight matrix at layer l. This graph-based formulation captures spatial dependencies among gaze positions, improving both classification performance and interpretability.

3 Experiment and Results

3.1 Datasets

Existing strabismus datasets[24] often suffer from limited sample sizes and coarse-grained labels. To address this, we constructed a large-scale clinical dataset comprising 1,075 real cases for strabismus subtype classification. Each patient contributed nine facial images, each corresponding to a specific gaze direction. The dataset includes six label categories: A-pattern, V-pattern, non-AV-pattern, esotropia, exotropia, and vertical strabismus. We split the dataset into training, validation, and test sets in an 8:1:1 ratio.

3.2 Biological Interpretation of Selected Causal Features

The causal features selected by our method, as listed in Table 2, align with clinical guidelines [16], reinforcing their interpretability and biological relevance in strabismus diagnosis.

For direction classification, including esotropia (inward eye tilt), exotropia (outward eye tilt), and vertical strabismus (upward or downward misalignment), the variables (x^L_{iris}, y^L_{iris}), (x^R_{iris}, y^R_{iris}), and x^L_{ls}, x^R_{ls} capture key positional data of the irises and eye light spots, directly reflecting the horizontal and vertical misalignment of the eyes.

Table 3. Performance comparison with main-stream models

Performance Comparison					
Model	VGG-16	ResNet50	ViT	Swin ViT	Ours
Acc_{ang}	0.738	0.794	0.869	0.850	**0.881**
Acc_{dir}	0.878	0.869	0.850	0.850	**0.898**
Cross-Environment Validation					
Model	VGG-16	ResNet50	ViT	Swin ViT	Ours
Acc_{ang}	0.655	0.689	0.689	0.689	**0.825**
Acc_{dir}	0.689	0.613	0.638	0.428	**0.750**

For angle variation, A-pattern is characterized by a greater downward gaze angle than the upward gaze angle, while V-pattern exhibits the opposite trend. The variables θ_{eye}^{L} and θ_{eye}^{R}, representing the cosine angles between the iris and orbital centers, capture eye alignment. Additionally, $\Delta\theta_2$, $\Delta\theta_7$, $\Delta\theta_8$, and $\Delta\theta_9$ quantify interocular misalignments in specific gaze regions, while $\Delta\theta_{1-7}$, $\Delta\theta_{2-8}$, and $\Delta\theta_{3-9}$ capture angular differences between upward and downward gaze positions, facilitating A/V-pattern recognition.

3.3 Results

Performance Comparison We evaluated our method against state-of-the-art models, including VGG [25], ResNet [26],ViT [27], and SwinViT [28], following standard image classification pipelines, ensuring fair comparisons with standardized data augmentation and training settings. Classification performance was assessed using Acc_{dir} and Acc_{deg} for strabismus deviation direction and deviation angle classification, respectively. As shown in Table 3, our method consistently outperforms existing models.

For generalizability, we conducted cross-environment validation on 130 strabismus patients from a major medical institution (2016–2019). Our framework demonstrated robustness across diverse imaging conditions, demographics, and equipment, ensuring reliability in clinical deployment.

Leveraging causal feature selection, our method drastically reduces computational cost. Compared to ViT's 22.08 GFLOPs, our approach requires only 0.0003 GFLOPs—$1/100,000$ of ViT's cost—making it highly efficient and suitable for resource-limited medical environments.

Ablation Study In this subsection, we conduct experiment to reveal the effectiveness of our proposed method.

Effect of Proposed Components Our proposed CI-GNN framework comprises three key components: candidate feature extraction (CFE), causal feature se-

Table 4. Ablation study on the proposed components

CFE	CFS	GDM	Acc_{ang}	Acc_{dir}
			0.729	0.729
✓			0.835	0.231
✓	✓		0.817	0.815
✓	✓	✓	**0.881**	**0.898**

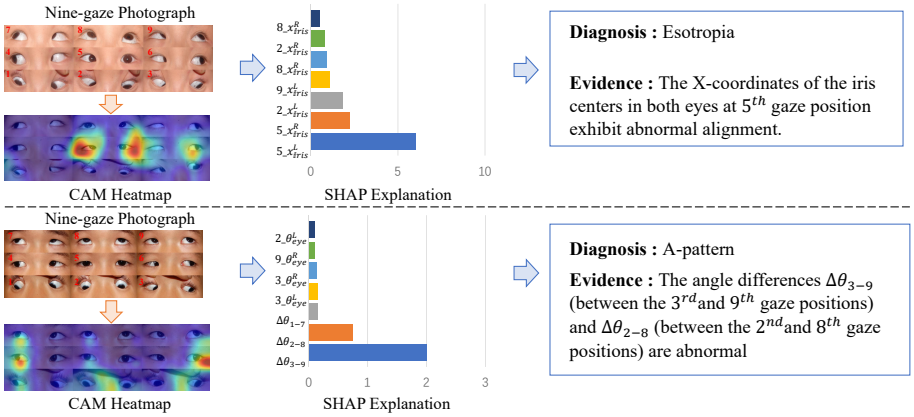

Fig. 3. Illustration of Interpretable Diagnosis. Compared to CAM, our method generates diagnostic evidence that is more easily interpretable by clinicians.

lection (CFS), and graph data modeling (GDM). To evaluate their contributions, we conducted an ablation study by systematically removing each module: (1) bypassing CFE by using raw image data for prediction, (2) omitting CFS by utilizing the entire candidate feature pool, and (3) replacing GDM with a standard MLP structure. In Table 4, each component contributes significantly to performance improvement, demonstrating its effectiveness. Notably, the full model achieves the highest overall accuracy, highlighting the synergy among the proposed modules.

Interpretability While our method achieves the highest accuracy in strabismus subtype diagnosis, its key advantage lies in providing clinically interpretable diagnostic evidence. To illustrate this, we present interpretability analyses for two representative cases and compare them with CAM [29] heatmaps generated by ResNet, as shown in Fig.3. Leveraging Shapley-based explanation techniques, our approach precisely localizes critical eye features associated with the patient's condition, enhancing clinicians' understanding and validation of the model's decision-making process. In contrast, CAM heatmaps highlight broad, less specific regions, making it challenging for clinicians to interpret the underlying diagnostic reasoning.

4 Conclusion

We proposed a Causality-Inspired Graph Neural Network (CI-GNN) framework, for interpretable strabismus subtype classification. By integrating causal discovery with graph neural networks, our approach enhances both diagnostic accuracy and interpretability. Experimental results demonstrate its superiority over existing methods, achieving state-of-the-art performance while significantly reducing computational costs. Future work will explore its extension to broader ophthalmic disorders and real-time clinical applications.

Acknowledgments. This work was supported by the National Science and Technology Major Project(2021ZD0111502), the Science and Technology Planning Project of Guangdong Province of China(2025A1515010800), Hospital Funded Clinical Research, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (21XJMR02), the National Natural Science Foundation of China(62176147, 62406186, 62476163, 62441612), the STU Scientific Research Foundation for Talent(NTF22030),

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hatt, S.R., Leske, D.A., Kirgis, P.A., Bradley, E.A., Holmes, J.M.: The effects of strabismus on quality of life in adults. American journal of ophthalmology **144**(5), 643–647 (2007)
2. Ticho, B.H.: Strabismus. Pediatric Clinics **50**(1), 173–188 (2003)
3. Menon, V., Saha, J., Tandon, R., Mehta, M., Khokhar, S.: Study of the psychosocial aspects of strabismus (2002)
4. VanderVeen, D.K., Bremer, D.L., Fellows, R.R., Hardy, R.J., Neely, D.E., Palmer, E.A., Rogers, D.L., Tung, B., Good, W.V., for Retinopathy of Prematurity Cooperative Group, E.T., et al.: Prevalence and course of strabismus through age 6 years in participants of the early treatment for retinopathy of prematurity randomized trial. Journal of American Association for Pediatric Ophthalmology and Strabismus **15**(6), 536–540 (2011)
5. Nelson, B.A., Gunton, K.B., Lasker, J.N., Nelson, L.B., Drohan, L.A.: The psychosocial aspects of strabismus in teenagers and adults and the impact of surgical correction. Journal of American Association for Pediatric Ophthalmology and Strabismus **12**(1), 72–76 (2008)
6. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
7. Hatherley, J.J.: Limits of trust in medical ai. Journal of medical ethics **46**(7), 478–481 (2020)
8. Zohuri, B., Moghaddam, M.: Deep learning limitations and flaws. Mod. Approaches Mater. Sci **2**, 241–250 (2020)
9. Sparrow, R., Hatherley, J.: High hopes for "deep medicine"? ai, economics, and the future of care. The Hastings Center Report **50**(1), 14–17 (2020)
10. De Almeida, J.D.S., Silva, A.C., de Paiva, A.C., Teixeira, J.A.M.: Computational methodology for automatic detection of strabismus in digital images through hirschberg test. Computers in biology and medicine **42**(1), 135–146 (2012)

11. Gattass, Marcelo, Azevedo, Valente, Thales, Levi, Meireles, Teixeira, Jorge, and, A.: Automatic diagnosis of strabismus in digital videos through cover test. Computer Methods and Programs in Biomedicine: An International Journal Devoted to the Development, Implementation and Exchange of Computing Methodology and Software Systems in Biomedical Research and Medical Practice (2017)
12. Zheng, C., Yao, Q., Lu, J., Xie, X., Lin, S., Wang, Z., Wang, S., Fan, Z., Qiao, T.: Detection of referable horizontal strabismus in children's primary gaze photographs using deep learning. Translational vision science & technology **10**(1), 33–33 (2021)
13. Hamid, H.S., AlKindy, B., Abbas, A.H., Al-Kendi, W.B.: An intelligent strabismus detection method based on convolution neural network. TELKOMNIKA (Telecommunication Computing Electronics and Control) **20**(6), 1288–1296 (2022)
14. Zhang, G., Xu, W., Gong, H., Sun, L., Li, C., Chen, H., Xiang, D.: Multi-feature fusion-based strabismus detection for children. IET Image Processing **17**(5), 1590–1602 (2023)
15. de Oliveira Simoes, T., Souza, J.C., de Almeida, J.D.S., Silva, A.C., de Paiva, A.C.: Automatic ocular alignment evaluation for strabismus detection using u-net and resnet networks. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 239–244. IEEE (2019)
16. Griffin, J.R., Grisham, J.D.: Binocular anomalies: diagnosis and vision therapy. (No Title) (2002)
17. Miller, J.M., Mellinger, M., Greivenkemp, J., Simons, K.: Videographic hirschberg measurement of simulated strabismic deviations. Investigative ophthalmology & visual science **34**(11), 3220–3229 (1993)
18. Li, Y., Xia, R., Liu, C., Sun, L.: A hybrid causal structure learning algorithm for mixed-type data. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2022) (2022)
19. Colombo, D., Maathuis, M.H., et al.: Order-independent constraint-based causal structure learning. J. Mach. Learn. Res. **15**(1), 3741–3782 (2014)
20. Xie, Z., Yan, J.: Kernel density estimation of traffic accidents in a network space. Computers, environment and urban systems **32**(5), 396–406 (2008)
21. Pearl, J.: Probabilistic reasoning in expert systems (1988)
22. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
23. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 315–323. JMLR Workshop and Conference Proceedings (2011)
24. Chen, Z., Fu, H., Lo, W.L., Chi, Z.: Strabismus recognition using eye-tracking data and convolutional neural networks. Journal of healthcare engineering **2018**(1), 7692198 (2018)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings

of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)