LABORATORY INVESTIGATION

journal homepage: https://laboratoryinvestigation.org/

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Research Article

The Future of Surgical Diagnostics: Artificial Intelligence-Enhanced Detection of Ganglion Cells for Hirschsprung Disease

Derya Demir[a], Kutsev Bengisu Ozyoruk[b,*], Yasin Durusoy[c], Ezgi Cinar[d], Gurdeniz Serin[a], Kayhan Basak[e], Emre Cagatay Kose[f], Malik Ergin[g], Murat Sezak[a], G. Evren Keles[h], Sergulen Dervisoglu[i], Basak Doganavsargil Yakut[a], Yavuz Nuri Ertas[j,k], Feras Alaqad[c,*], Mehmet Turan[c,*]

[a] Department of Pathology, Ege University Faculty of Medicine, Izmir, Turkey; [b] Artificial Intelligence Resource, Molecular Imaging Branch, National Cancer Institute, Bethesda, Maryland; [c] Department of Computer Engineering, Bogazici University, Istanbul, Turkey; [d] Department of Pathology, Bakırcay University Cigli Training and Research Hospital, Izmir, Turkey; [e] Department of Pathology, Saglık Bilimleri University, Kartal Dr. Lutfi Kırdar City Hospital, Istanbul, Turkey; [f] Deciphex, Dublin City University, Dublin, Ireland; [g] Department of Pathology, Dr. Behcet Uz Pediatrics and Surgery Training and Research Hospital, Izmir, Turkey; [h] Virasoft Corporation, New York, New York; [i] Department of Pathology, Medipol Mega University Hospital, Istanbul, Turkey; [j] Department of Biomedical Engineering, Erciyes University, Kayseri, Turkey; [k] Department of Technical Sciences, Western Caspian University, Baku, Azerbaijan

ARTICLE INFO

Article history:
Received 22 April 2024
Revised 9 November 2024
Accepted 14 November 2024
Available online 20 November 2024

Keywords:
artificial intelligence
Hirschsprung disease
frozen
surgical pathology

ABSTRACT

Hirschsprung disease, a congenital disease characterized by the absence of ganglion cells, presents significant surgical challenges. Addressing a critical gap in intraoperative diagnostics, we introduce transformative artificial intelligence approach that significantly enhances the detection of ganglion cells in frozen sections. The data set comprises 366 frozen and 302 formalin-fixed-paraffin-embedded hematoxylin and eosin–stained slides obtained from 164 patients from 3 centers. The ganglion cells were annotated on the whole-slide images (WSIs) using bounding boxes. Tissue regions within WSIs were segmented and split into patches of 2000 × 2000 pixels. A deep learning pipeline utilizing ResNet-50 model for feature extraction and gradient-weighted class activation mapping algorithm to generate heatmaps for ganglion cell localization was employed. The binary classification performance of the model was evaluated on independent test cohorts. In the multireader study, 10 pathologists assessed 50 frozen WSIs, with 25 slides containing ganglion cells, and 25 slides without. In the first phase of the study, pathologists evaluated the slides as a routine practice. After a 2-week washout period, pathologists re-evaluated the same WSIs along with the 4 patches with the highest probability of containing ganglion cells. The proposed deep learning approach achieved an accuracy of 91.3%, 92.8%, and 90.1% in detecting ganglion cells within WSIs in the test data set obtained from centers. In the reader study, on average, the pathologists' diagnostic accuracy increased from 77% to 85.8% with the model's heatmap support, whereas the diagnosis time decreased from an average of 139.7 to 70.5 seconds. Notably, when applied in real-world settings with a group of pathologists, our model's integration brought about substantial improvement in diagnosis precision and reduced the time required for diagnoses by half. This notable advance in artificial intelligence–driven diagnostics not only sets a new standard for

These authors contributed equally as co-first authors: Derya Demir and Kutsev Bengisu Ozyoruk.

These authors contributed equally as co-senior authors: Mehmet Turan, Feras Alaqad, and Kutsev Bengisu Ozyoruk.

* Corresponding authors.

E-mail addresses: kutsev.ozyoruk@nih.gov (K.B. Ozyoruk), ferasdc18@gmail.com (F. Alaqad), mehmet.turan@boun.edu.tr (M. Turan).

https://doi.org/10.1016/j.labinv.2024.102189

surgical decision making in Hirschsprung disease but also creates opportunities for its wider implementation in various clinical settings, highlighting its pivotal role in enhancing the efficacy and accuracy of frozen sections analyses.

Introduction

Hirschsprung disease (HD) is a rare congenital disorder primarily affecting newborns and pediatric patients.[1] It is characterized by the absence of ganglion cells in the Meissner and Auerbach plexuses of the distal colon, leading to impaired bowel innervation and motility. This condition can result in functional bowel obstruction, presenting as megacolon, chronic constipation, recurring colitis, or a life-threatening risk of perforation.[2] Biopsy and histologic examination of the affected colon segment are imperative for HD, with a specific focus on identifying the presence or absence of ganglion cells in the neural plexuses.[3]

During HD surgery, the intraoperative assessment of frozen sections (FSs) is a crucial step in determining the extent of the diseased bowel segment for resection. Although this process involves rapid freezing and staining of the tissue, the primary challenge lies in the accurate detection of ganglion cells within these sections. Inaccuracies in identifying these cells can arise due to various factors such as the pathologist's experience, intra/interobserver variability, and the inherent morphologic diversity of ganglion cells. These difficulties are compounded by the time-sensitive nature of intraoperative diagnostics and the comparatively lower resolution of FSs as opposed to formalin-fixed paraffin-embedded (FFPE) sections.

The diagnostic difficulties encountered in Hirschsprung surgery have significant implications, as they may lead to 2 major types of surgical errors: under-resection and over-resection.[4-6] Under-resection often demands follow-up surgeries, escalating the risk of additional health issues. In contrast, over-resection can lead to a spectrum of chronic health problems due to the resultant shortened colon, such as compromised bowel function, nutritional deficits, incontinence, and an elevated risk of bowel obstruction and small intestine bacterial overgrowth. This highlights the necessity for enhanced intraoperative techniques to precisely identify ganglion cells, especially in circumstances where specialized expertise may not be readily accessible.[1,7]

To overcome these diagnostic challenges, especially in less-experienced settings, our study proposes a deep learning model that provides a robust and real-time solution for accurately identifying ganglion cells in FSs. Our approach significantly enhances the precision of surgical decision making in Hirschsprung operations, thereby substantially reducing the risk of postoperative complications that often arise from misdiagnosis. Our artificial intelligence (AI)-driven methodology introduces a novel solution in pediatric surgical diagnostics, offering a scalable and reliable tool that can be adapted to a variety of clinical settings and types of frozen examinations. This advancement is poised to facilitate broader improvements in patient outcomes and increase health care efficiency.

Our objective is to evaluate the effectiveness of this deep learning–based decision support system in assisting pathologists with intraoperative diagnostics by automatically detecting ganglion cells within hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) of frozen tissue sections (Fig. 1).

Materials and Methods

Study Design

We proposed a deep learning–based decision support tool to enhance the accuracy and efficiency of the intraoperative diagnosis of HD. The tool identifies the regions within WSIs where ganglion cell appearance probability is the highest. The model is trained on a data set of 431 WSIs from Ege University Hospital (EUH), which is the mixture of both FFPE- and FS-WSIs. To cope with the scarcity of high-quality frozen WSIs for HD, we build the training data set by combining the FFPE and frozen images. The data sets were annotated by 3 pathologists (E.C., G.S., and D.D.) experienced in HD by tagging the ganglion cells in the WSIs using bounding boxes that encompassed only the regions containing the ganglion cells.

The model's intraoperative performance was tested exclusively on FS-WSIs. The generalizability of the model was validated using 2 independent data sets from different hospitals. By directing pathologists' attention to specific regions in WSIs, the model minimizes the requirement for reviewing the entire images. In this study, diagnoses on FSs were confirmed by examining the corresponding permanent sections. Pathologists conducted a detailed review of the FFPE sections to verify the presence or absence of ganglion cells identified in the FSs. Furthermore, a comprehensive analysis was conducted to correlate the clinical history and patient records with the diagnoses, ensuring the accuracy and relevance of the findings. This novel approach holds significant potential for enhancing HD diagnosis and management.

Data Collection

A meticulous data collection process was conducted to construct a comprehensive data set for training, validating, and testing our model. A total of 431 WSIs were acquired from the EUH, comprising 215 slides that were FFPE sections and 216 slides that were FSs. The WSIs were subjected to careful examination (expanded upon in the data collection section), and the presence of ganglion cells in the FSs were documented. These observations were compared against the "gold standard" FFPE sections.

To ensure the generalizability of our model, independent cohorts from 2 hospitals were incorporated. The first cohort, referred to as the Dr. Behcet Uz Hospital (BUH) cohort, consists of 197 WSIs obtained from BUH, of which 98 slides were FFPE sections, and the remaining 99 slides were FSs. The second cohort, referred to as the Medipol University Hospital (MUH) cohort, comprised 39 frozen WSIs acquired from MUH. The inclusion of these diverse cohorts facilitated the capture of variations in biopsy protocols, slide preparation techniques, staining mechanisms, and scanner vendors, thereby encompassing a wide spectrum of HD cases. Notably, to minimize potential confounding factors associated with daily variability in section staining and quality, archived sections were utilized instead of generating new ones.

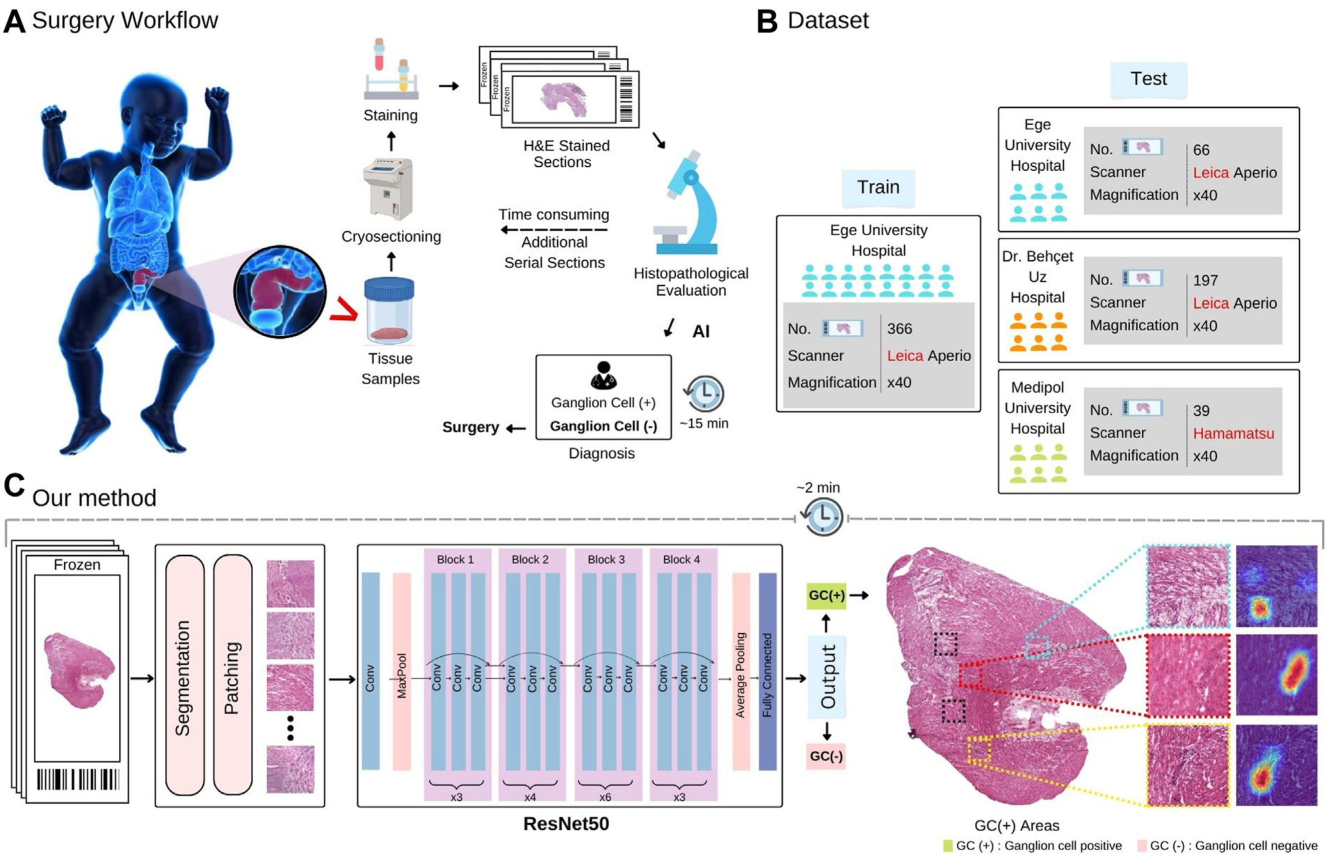

Figure 1.
Workflow overview of the proposed diagnostic decision support tool. (A) This diagram provides an overview of the general workflow for the proposed diagnostic decision support tool and how it integrates into the routine assessment of surgically excised specimens for histopathological evaluation. (B) The data set has been divided into training and test sets, with a specific emphasis on ensuring diversity by including data from 3 distinct centers. Additionally, the data set encompasses the use of various scanning devices, highlighting the tool's adaptability and resilience when dealing with different imaging equipment. (C) Our ganglion cell detection tool, which produces the ganglion cell observation probability heatmap in 2 minutes, is integrated into the frozen sectioning process as a last step. Our tool operates by taking digitized, high-resolution whole-slide frozen-section histology images as input. After a thorough segmentation process, the system produces a set of 2000 × 2000 mini-patches for each frozen-section- whole-slide images, and the ganglion cell probability heatmaps are generated to guide the pathologists. This strategic approach significantly enhances the speed and precision of intraoperative diagnostic processes.

The EUH data set was divided into 3 groups: training (70%), validation (15%), and holdout testing (15%). We ensured an equitable representation of ganglion cell—positive and —negative annotations across all parts. The independent test cohorts from BUH and MUH were completely used to assess the model's adaptability to the high data variability present across different hospitals. To ensure a robust evaluation of our deep-learning (DL) model, we employed a randomization process to split the slides into training and test sets.

Data Preprocessing

In this study, a thorough quality control process was implemented to assess the integrity and reliability of the WSIs included in the data set. Each WSI underwent meticulous scrutiny to identify any potential issues related to staining procedures or scanner artifacts. WSIs that exhibited problems or abnormalities were promptly identified and eliminated from the data set to maintain the overall quality and reliability of the data set.

We used the tissue segmentation pipeline from the clustering-constrained attention multiple instance learning (CLAM)[8] method for WSIs. The method allowed us to capture tissue regions and extract patches of 2000 × 2000 pixels. These patches were subsequently resized to 512 × 512 pixels using the NumPy[9] resize function, which does not use any interpolation method.

Our training data set consists of positive and negative classes. Twenty thousand patches were sampled from WSIs without ganglion cells, forming the negative class. The positive class was formed by extracting and resizing 4000 patches, which were extracted from areas annotated by pathologists as containing ganglion cells. To balance the classes, we applied a data augmentation method that randomly altered the ganglion cells' position within patches, thereby increasing the positive patch count to 20,000. This strategy resulted in robust patch images showcasing ganglion cells amidst diverse background tissues.

Data Set Labeling

The labeling process of the data set involved the annotation of WSIs by a panel of 3 highly experienced pathologists specializing in pediatric gastrointestinal pathology. Each pathologist independently examined the WSIs to identify and annotate the ganglion cells. To ensure the accuracy and consistency of the annotations, consensus meetings were conducted to discuss and resolve any discrepancies among the pathologists. In our study, class 1 refers to slides that contain ganglion cells, whereas class 0 refers to slides that do not contain ganglion cells.

Model Learning

We utilized ResNet-50[10] architecture pretrained on the ImageNet data set[11] for feature extraction. Residual networks are widely recognized for their effectiveness in addressing the vanishing gradient problem and are widely used for feature extraction in the literature. The architecture includes residual blocks consisting of multiple convolutional layers, batch normalization, activation functions, and skip connections, facilitating the learning of intricate image representations.

The weights of the pretrained model were retained, and the entire network was fine-tuned to adapt it specifically for the task of ganglion cell detection in HD. Stochastic gradient descent optimization with a learning rate of 0.001 and a momentum of 0.85 was employed to optimize the learning process. Two Nvidia Tesla A100 GPUs were employed for training. The model was trained for 50 epochs, utilizing a batch size of 16. The binary cross-entropy loss function was used to measure the discrepancy between the predicted probabilities and the ground truth labels.

Furthermore, to enhance our model's interpretability, we used the gradient-weighted class activation mapping (Grad-CAM)[12] algorithm, which generates a heatmap by computing the gradient of the predicted class score with respect to the feature maps of the final layer in the model. This algorithm identifies the most salient regions of an input image for prediction. We used Grad-CAM to locate regions with ganglion cells and displayed a heatmap of these regions to the pathologist. This facilitated a targeted examination of ganglion cells and an informed interpretation of the model predictions.

Model Selection

We conducted a comprehensive comparison of state-of-the-art deep learning models, including MobileNetV2,[13] MobileNetV3,[14] ResNet101,[10] ResNet18,[10] ResNet-50,[10] Swin Transformer,[15] Deep ViT,[16] EfficientViT,[17] and CrossViT,[18] to determine the optimal model for our data set (Supplementary Fig. S1). The same training parameter set was shared across all models.

Our results indicated that ResNet-50 achieved superior performance over the vision transformers and other models in our comparison. Despite certain models, such as Swin Trans former, exhibiting higher accuracy on the EUH data set, their performance on the independent cohort was comparatively lower.

Considering the tradeoff between inference time and accuracy, ResNet-50 emerged as the optimal model. This finding highlights the superiority of ResNet-50 in the context of HD diagnosis compared with a diverse range of deep learning models. By selecting ResNet-50 as the preferred model, we ensured efficient and reliable diagnostic capabilities, enhancing the overall diagnostic accuracy and clinical utility of the developed deep learning model.

Reader Study

We conducted a reader study with 10 pathologists to evaluate the influence of our AI model on their decision making for detecting ganglion cells in FSs. The study involved 2 phases with a 2-week washout period, where pathologists were presented with 50 FS-WSIs in each phase. Within these images, 25 slides contained ganglion cells, whereas the other 25 slides did not include ganglion cells. In the first phase, pathologists were provided with the complete WSIs and were allowed to navigate the entire slide at different magnifications, simulating the conventional diagnostic practice. They evaluated the FS slides as they would in their routine practice, without any AI assistance. The pathologists recorded their diagnostic decisions, classifying each slide as "positive" (indicating the presence of ganglion cells), "negative" (indicating the absence of ganglion cells), or "uncertain" (suggesting the need for further examination using FFPE sections). In the second phase, after at least a 2-week washout period, the same WSIs were presented to the pathologists along with 2 patches per slide, which had the highest probability of containing ganglion cells as predicted by the AI model. The mean washout period was 3.06 weeks, with an SD of ±0.89 weeks. These patches were accompanied by corresponding heatmaps to guide the pathologists' attention to regions of interest. The pathologists reevaluated the slides, and their diagnostic decisions, and decision times were recorded.

Results

Evaluation of Model Performance

The model achieved state-of-the-art performance on various data sets. On the holdout EUH data set, consisting of 660 positive patches (class 1) and 640 negative patches (class 0), it achieved an accuracy of 91.3% and an F1 score of 91.3% for class 0, whereas it achieved an F1 score of 91.2% for class 1. On the BUH data set, which comprised 1523 positive patches (class 1) and 1285 negative patches (class 0), the model displayed an accuracy of 92.8% and an F1 score of 92.3% for class 0, along with a 93.2% F1 score for class 1. Similarly, on the MUH data set, consisting of 810 positive patches (class 1) and 1140 negative patches (class 0), the model exhibited an accuracy of 90.1% and an F1 score of 91.6% for class 0, with an 87.7% F1 score for class 1. These results were obtained at the patch level, indicating the model's ability to accurately detect ganglion cells and distinguish between different classes (Fig. 2).

Our model analyzes a complete slide within an average duration of 2 minutes, which is obtained by averaging the processing time over 100 WSIs. The model extracts proposed ganglion cell patches from the WSI and displays them with associated heatmaps (Fig. 2), which enables the pathologist to make a more accurate decision in a much shorter time frame, as shown in the reader study results in Figure 3.

Model Interpretability

We demonstrated that providing heatmaps to pathologists enhanced the accuracy of the diagnosis, as corroborated by the reader study results in Figures 2 and 3.

We analyzed positive patches with ganglion cells and observed that our model assigned high attention scores only to the regions with ganglion cells. Conversely, in negative patches without ganglion cells, the heatmaps were dispersed. These results indicate that our model effectively learned to focus solely on ganglion cells for prediction, as illustrated in Figures 2 and 3 (positive and negative, respectively).

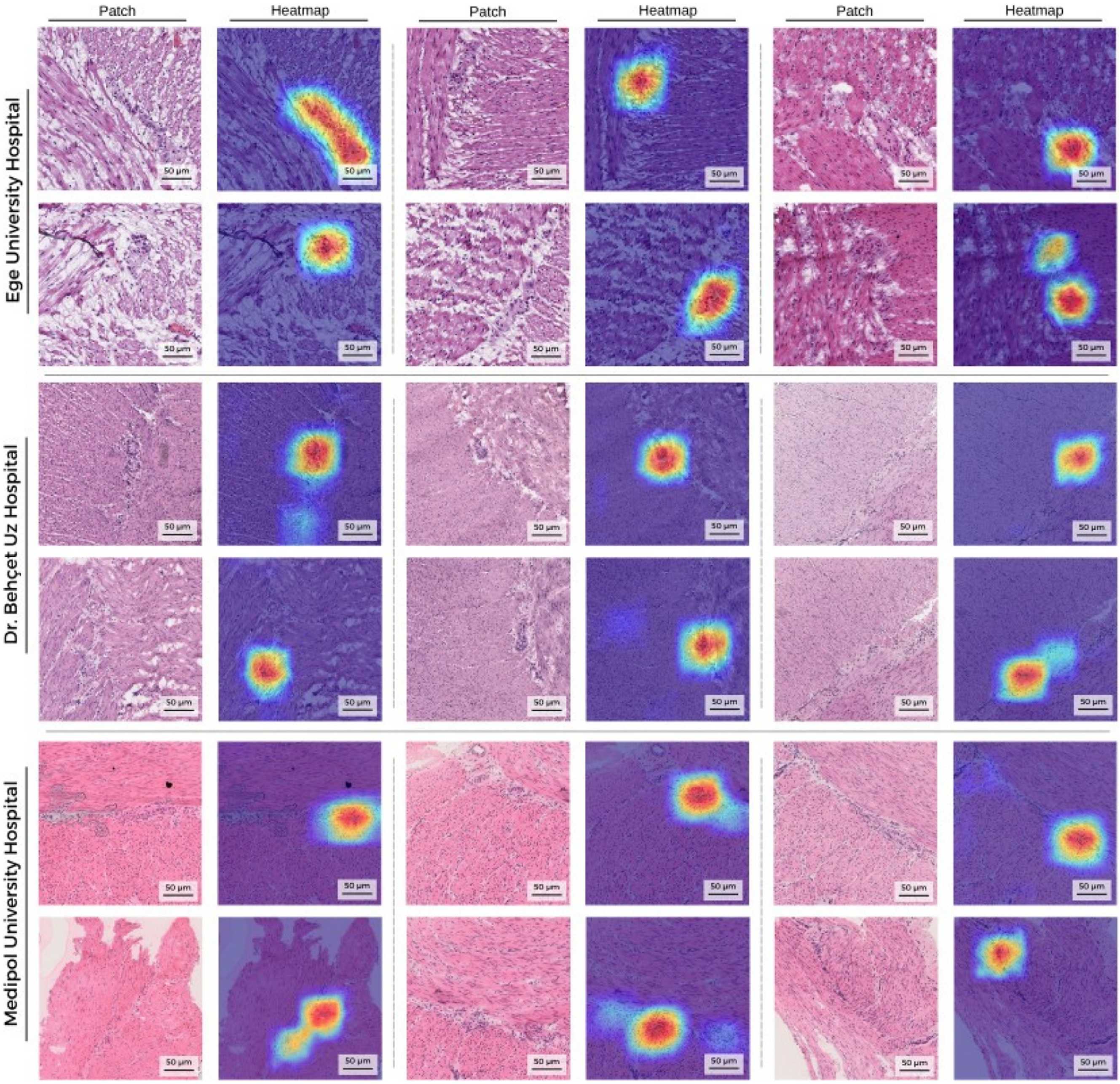

Figure 2.
Patch-level heatmaps. The figure illustrates patch-level heatmaps that highlight the regions with the highest likelihood of containing ganglion cells within 512 × 512 patches. These heatmaps serve as a visual representation of the artificial intelligence model's attention, revealing the specific areas it considers most indicative of ganglion cell presence in frozen sections. Heatmaps show the highest probability region for ganglion cells on each 512 × 512 patches. Distinct variations are observed in sections of patches obtained from different centers, notably in patches from Ege University Hospital where freezing artifacts are pronounced, and in patches from Dr. Behcet Uz Hospital where significant deviations in staining are evident. Freezing artifacts in patches from Ege University Hospital and variations in staining in patches from Dr. Behcet Uz Hospital complicate the optical evaluation of sections, making it challenging to ascertain the presence of potential neural plexuses and whether they contain ganglion cells. Heatmaps in patch groups from all 3 centers highlight areas of possible ganglion cell presence within neural plexuses, thereby directing attention to these areas and facilitating the detection of ganglion cells. Importantly, it is worth noting that the performance of the heatmaps in representing ganglion cells across all 3 data sets from different centers was successful despite the stain, scanner, and frozen sections preparation technique differences.

Reader Study

Figure 4 illustrates the outcomes of the study, revealing notable improvements in both accuracy and diagnosis time when utilizing the AI model. On average, the pathologists' accuracy increased from 77% in phase 1 to 85.8% in phase 2 (Fig. 3). This enhancement indicates that the AI model's predictions, along with the provided patches and heatmaps, effectively aided the pathologists in making accurate diagnoses (Fig. 2). Furthermore, the incorporation of the AI model resulted in a substantial reduction in diagnosis time. The average diagnosis time decreased from 139.7 seconds in phase 1 to 70.5 seconds in phase 2 (Fig. 3). This considerable timesaving highlights the efficiency and effectiveness of the AI model in expediting the diagnostic process. To

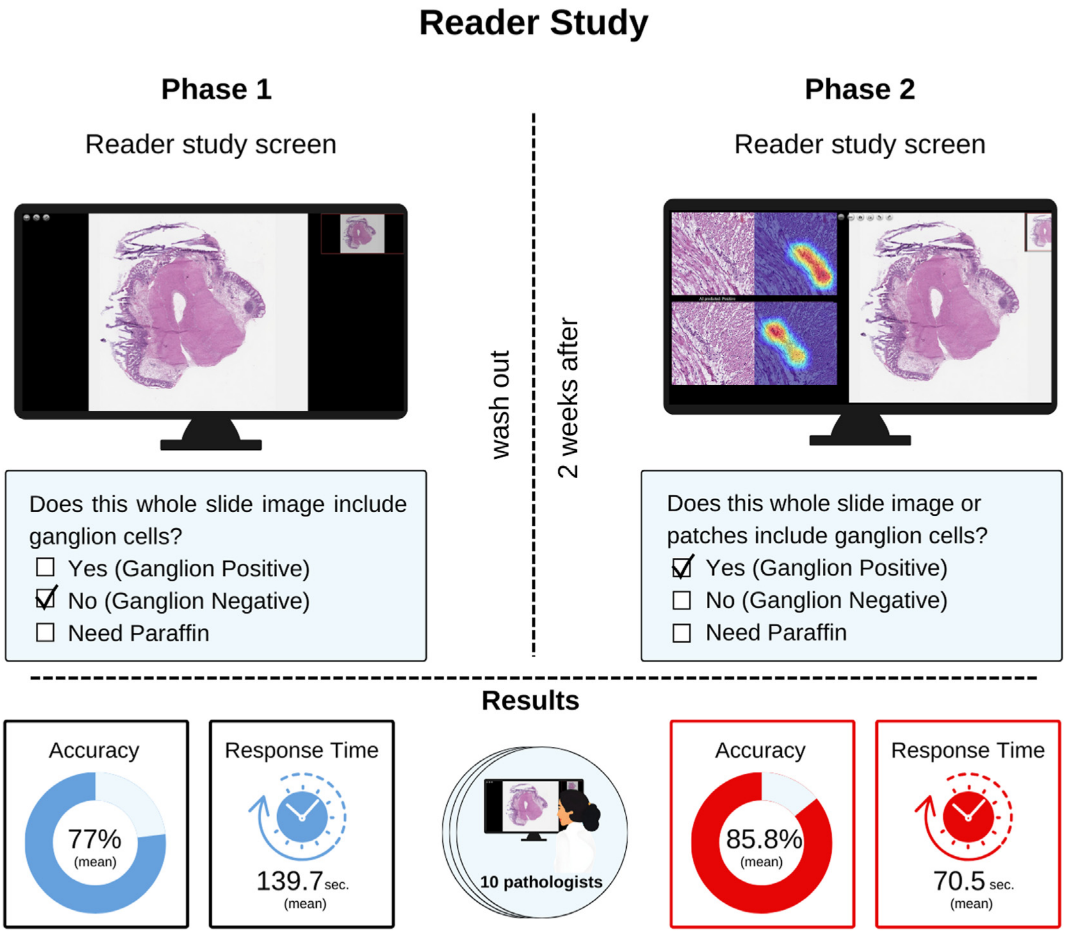

Figure 3.

Reader study results. In each phase of the study, 10 pathologists evaluated 50 frozen-section whole-slide images, with 25 slides containing ganglion cells and 25 slides without ganglion cells. In the first phase, pathologists navigated the entire whole-slide images at different magnifications, simulating routine practice without artificial intelligence assistance. In the second phase, the same whole-slide images were re-evaluated with the artificial intelligence model's heatmap support, highlighting regions with the highest probability of containing ganglion cells. On average, the pathologists' diagnostic accuracy increased from 77% to 85.8% after model's heatmap support, whereas the diagnosis time decreased from 139.7 to 70.5 seconds.

assess the accuracy achieved using the diagnostic support tool compared with the conventional pipeline, a Wilcoxon signed-rank test was conducted. The analysis yielded a test statistic of 0.0 with a *P* value of <.01, indicating that there is a statistically significant difference in the accuracy of ganglion cell detection between the 2 groups. This finding supports the notion that diagnostic support tools provide sufficient clinical information to determine the presence of ganglion cells. Apart from these, we conducted inter-rater agreement analysis for 5 pathologists who were experts in HD and 5 general pathologists in phases 1 and 2. The mean Cohen kappa for the HD pathologists was 0.54 and 0.68 in phases 1 and 2, respectively. The mean Cohen kappa for the general pathologists was 0.31 and 0.58 in phases 1 and 2, respectively. For HD expert pathologists, the mean kappa values for phases 1 and 2 were compared using a paired *t* test. The Wilcoxon signed-rank statistic was 0.0 with a *P* value of .0019. Similarly, for general pathologists, the Wilcoxon signed-rank statistic was 7.0 with a *P* value of .037. Because the *P* values were less than the significance level of .05, we rejected the null hypothesis. This suggests a statistically significant difference in mean kappa values between phases 1 and 2 for both HD and general pathologists. The readers were categorized into 2 groups: general and HD pathologists. For general pathologists, the AUC scores were 0.85, 0.90, 0.92, 0.60, and 0.82. Their sensitivity scores were 1.00, 0.88, 0.84, 0.88, and 0.92 and specificity scores were 0.76, 0.92, 1.00, 0.56, and 0.84. For HD pathologists, the AUC scores were 0.85, 0.91, 0.94, 0.90, and 0.94. Sensitivity scores for HD pathologists were 0.96, 1.00, 0.88, 0.88, and 0.96, and their specificity scores were 0.80, 0.84, 1.00, 0.96, and 0.92.

The proposed deep learning approach achieved an accuracy of 91.3%, 92.8%, and 90.1% in detecting ganglion cells within WSIs from the test data sets obtained from 3 centers. In the reader study, on average, the pathologists' diagnostic accuracy increased from 77% to 85.8% with the model's heatmap support. The accuracy estimate included the observer's scoring for all 50 slides, both with and without ganglion cells, ensuring a comprehensive assessment of the model's performance. The diagnosis time decreased from an average of 139.7 to 70.5 seconds. For a detailed comparison between the DL model's recommendations and the pathologists' final choices, including percentages of adherence and deviations, refer to Supplementary Table S1. It is important to note that the DL model did not output a "need-paraffin" decision. It always provided a definitive recommendation of either "positive" or "negative." This feature ensured that the model consistently offered actionable insights, further streamlining the diagnostic process and reducing ambiguity during intraoperative decision making.

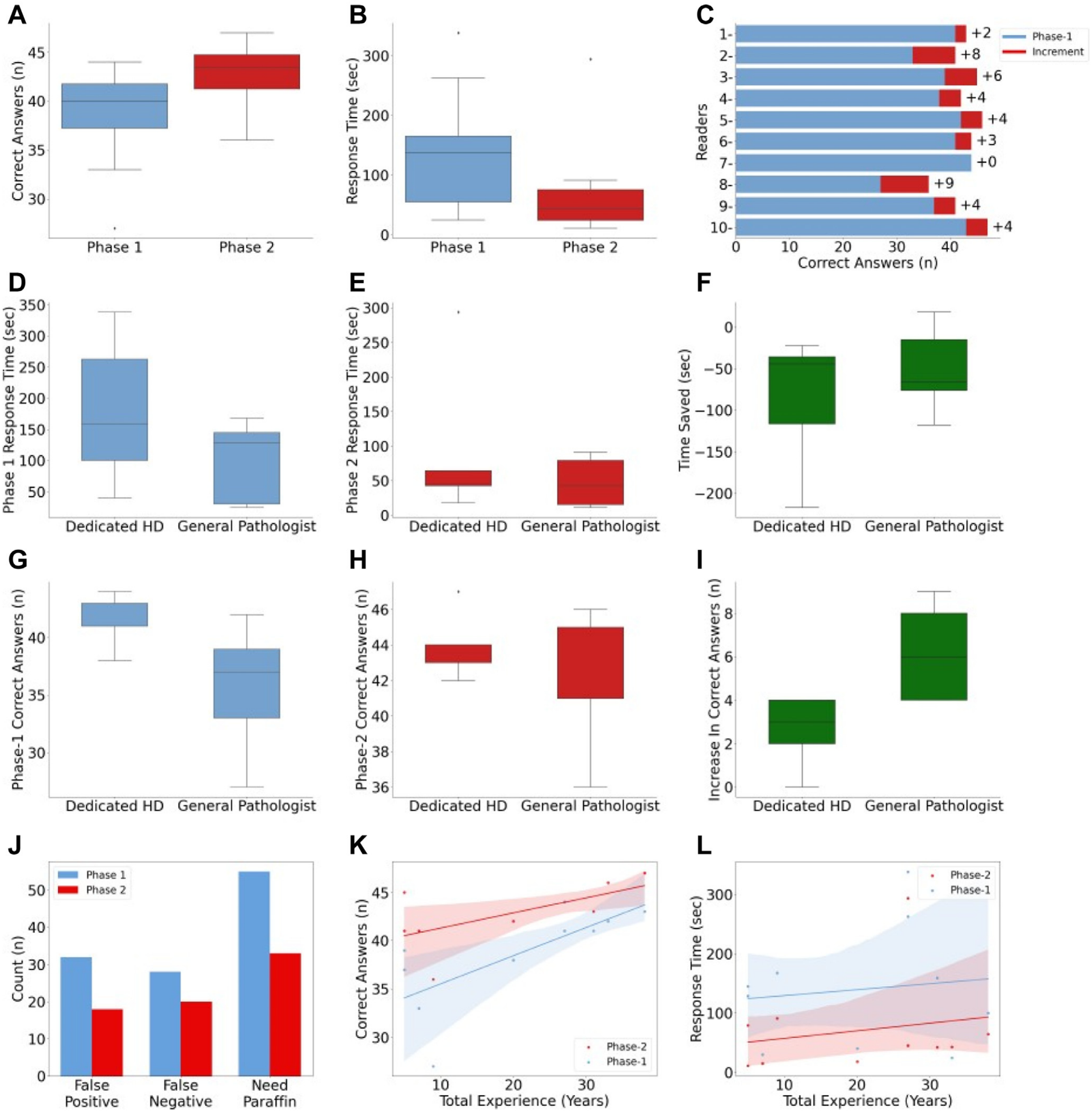

Figure 4.

Reader study detailed result. (A) Phase 2 clearly demonstrates a substantial increase in the correct answer rate, reflecting improved diagnostic accuracy. The integration of our artificial intelligence (AI) model significantly raised pathologists' diagnostic accuracy, ensuring more precise diagnoses. (B) Phase 2 reveals a marked reduction in response times, indicating enhanced efficiency during the diagnostic process. The utilization of our AI model resulted in notably shorter response times, streamlining and expediting the diagnostic procedure. (C) In phase 2, a conspicuous increase in the correct answers by 10 pathologists is illustrated in the red bars, suggesting improved diagnostic accuracy. The introduction of our AI model led to a significant boost in diagnostic accuracy, as evident in the responses from the pathologists. (D, E) During phase 1, dedicated pathologists had longer response times, whereas in phase 2, response times reduced notably in both groups. The study highlights a substantial decrease in response times in phase 2, with dedicated pathologists particularly benefiting from this efficiency improvement. (F) A noteworthy timesaving effect is observed in both groups, with general pathologists showing a more significant improvement. Our AI model substantially expedited the diagnostic process for all pathologists, with general pathologists benefiting most from this efficiency enhancement. (G) In phase 1, dedicated pathologists demonstrated a noticeable increase in the correct answer rate. The diagnostic efficiency was already enhanced in phase 1 for dedicated pathologists, laying the foundation for further improvements in phase 2. (H) Phase 2 highlights an increase in general pathologists' diagnostic accuracy. General pathologists witnessed a significant increase in diagnostic accuracy in phase 2, further supporting the role of the AI model in enhancing diagnostic precision. (I) Among general pathologists, the increase in correct answers is particularly striking. The improvement in diagnostic accuracy was notably pronounced among general pathologists, reaffirming the utility of the AI model in their diagnostic process. (J) In phase 1, both false positives and false negatives were higher, especially for false positives, whereas in phase 2, there was a significant reduction in the "need-paraffin" responses. The AI model contributed to a substantial decrease in the necessity for paraffin sections. (K) The increase in correct answers in phase 1 shows a more significant improvement. Phase 1 already witnessed improvements in diagnostic accuracy, setting the stage for further enhancement in phase 2. (I) Response times were reduced in both groups during phase 2, with dedicated and general pathologists experiencing a substantial decrease in diagnostic time. Both dedicated and general pathologists benefited from a substantial reduction in response times, highlighting the timesaving capabilities of the AI model.

Discussion

In this study, we introduced an intraoperative decision support tool designed to assist pathologists in diagnosing HD during surgery. The tool highlights the regions of interest that likely contain ganglion cells in WSIs, directing the pathologist's focus effectively. HD diagnosis is a tedious, time-consuming task that usually involves looking through 10 or more sections from the biopsy sample to conclude an intraoperative diagnosis, which presents a challenge for the time-sensitive HD diagnosis. Compounding on this is the fact that HD is a rare disease that is seen in ~1 in 5000 live births,[19] which means that unless the pathologists examining the slides are specialized in this area, they usually have little experience with HD diagnosis.[20] Our model aimed to aid in this regard by scanning the WSI of the FS in seconds and presenting the best candidate patches for a ganglion cell and the overall decision with WSI. Although there are studies supporting their findings with clinical validation by reader study results, the number of readers may not be sufficient for a strong claim.[21] In our reader study, all but one of the total 10 pathologists involved had significant decision time improvement that came with no performance penalty on the diagnostic accuracy. Besides, these readers deferred the diagnosis to FFPE sections significantly less. This could indicate that, especially in suboptimal settings where consultation with a specialist is unavailable or would take a long time, our model can expedite the diagnosis in this time-critical process.

When pathologists find the signs in the FSs of the biopsy inconclusive, they may choose to defer the evaluation to the permanent sections. The discordance of FFPE and FS has a wide range in the literature, with reports as low as 3%,[22] and Maia[6] reported a discordance rate up to 33% in a heterogeneous group of specimens from different intestinal sites. The main factors responsible for this discrepancy are as follows: sampling from the transition zone, insufficient samples where all layers cannot be seen completely,[23] artifacts due to the FS,[24] immature ganglion cells, technical difficulties, and the lack of experience in HD. Misdiagnosis of HD in FS may lead to suboptimal surgical treatment and life-long disability for the patient.[6] We saw a significant decline in the deferral of cases to the FFPE sections without a penalty in diagnostic accuracy. This means that our model can potentially prevent secondary surgeries and life-long disabilities.

Variations in section preparation, staining technique, WSI scanner type, and FS artifacts,[25] introduce difficulty in obtaining AI performance for multicenter. Naturally, this problem is exacerbated when an AI model is trained on the data set from one laboratory and evaluated in another. On the contrary to Schilling et al's model,[26] one of the main strengths of our AI model is that it is trained on a large data set from a center with all of the aforementioned variations present. Combined with the fact that our model achieved good performance on holdout and independent test sets that consisted of the entire data sets from multiple centers meant that our model could generalize to unseen data and is robust to variations. To help even further with the generalizability, we trained on the WSIs of both FSs and FFPE tissue sections. We did not evaluate our model's capabilities on FFPE tissue WSIs, but we intend to explore this area further in future work.

Explainability is a crucial aspect of an AI tool that is designed to help and work alongside humans. For that reason, we utilized a convolutional neural network architecture that is commonly used in digital pathology, ResNet-50, and incorporated a Grad-CAM layer into our model to provide interpretable heatmaps that highlight the regions of interest for pathologists. Our reader study results confirm that this helps pathologists in both decision time and diagnostic accuracy. The heatmaps from the Grad-CAM layer also assisted in understanding if the AI model attended to the same areas' pathologists would attend to, indicating that the model learned useful representations. The heatmaps showed that the model generally attributed high importance to areas where neural plexuses are usually located and very low importance to all other areas. They were also low in the areas with HD-associated signs, such as hypertrophic muscularis mucosa or hypertrophic nerve trunks. These results indicate that our model could indeed learn clinically meaningful and robust representations.

Alongside the rarity of the disease itself, there are usually only a few ganglion cells in a given biopsy, depending on the size and composition of the tissue itself,[5,27] which results in very high specificity values in detection true negatives.[28] This meant a significantly sparse positive label representation in our data set, for which we used data augmentation techniques to compensate for. This method both made the model more robust to diverse representations of ganglion cells and served to balance the labels in our data set.

Our study has several limitations that should be addressed in future work. First, our data set was relatively small and imbalanced, as HD is a rare condition. Therefore, although great care was taken to ensure the highest possible diversity and quality of our data set, our model may not capture all the possible variations and nuances of ganglion cell morphology and distribution. To improve our model's performance and reliability, we need to collect more data from diverse sources and apply data augmentation techniques to increase the diversity of our training samples. Second, our model was only evaluated on FS-WSIs, which are typically used for intraoperative consultation. However, FS-WSIs have lower quality and resolution than FFPE-WSIs, which are used for definitive diagnosis. Therefore, our model may not be able to detect subtle or ambiguous features of ganglion cells that are more visible in FFPE-WSIs. To address this issue, we need to test our model on both FFPE and frozen WSIs and compare its performance with conventional histopathological methods. Third, our model was only designed to detect ganglion cells as a binary classification task. However, HD is a disorder with a complex set of differential diagnoses that involve various types of nerve abnormalities such as hypoganglionosis, hyperganglionosis, immature ganglia, and giant ganglia.[2] Therefore, our model's inability to distinguish between these different subtypes of nerve defects or provide more detailed information about their severity or extent may diminish its clinical applicability. To overcome this limitation, we need to extend our model to perform multiclass or multilabel classification tasks that can identify different types of nerve abnormalities. Fourth, because our model works on the patch level, it currently does not utilize the clues that can be gathered outside the immediately surrounding tissue. The hypertrophy of the muscularis mucosa, hypertrophy of the nerve trunks, the orientation of the tissue, the general quality of the staining, and the presence of FS artifacts are all clues for experienced pathologists to determine if a suspected cell is a ganglion cell or not. We believe this can be explored in a further study with a larger data set and an multiple-instance learning-like model approach to holistically evaluate the clues in the WSI from the eyes of a neural network. Although the traditional approach to FS analysis, which involves examining multiple levels, achieves high diagnostic accuracy, often approaching 100% in experienced hands, our study aimed to evaluate the potential of the AI model to assist in scenarios with constrained resources. Future studies should compare the AI-assisted approach to the traditional method of examining multiple levels to determine if the AI improves accuracy and speed within this comprehensive diagnostic process. This will provide a clearer understanding of the AI model's practical benefits in enhancing the accuracy and efficiency of HD diagnosis.

In conclusion, we present a deep learning model that can assist pathologists in detecting ganglion cells in FS-WSIs during Hirschsprung surgery. Our model has shown promising results on both internal and external data sets, as well as providing interpretable heatmaps that can facilitate diagnosis. Our study contributes to the advancement of AI applications in pathology by demonstrating how deep learning models can improve the accuracy and efficiency of intraoperative consultation for rare diseases such as HD.

Acknowledgments

Mehmet Turan is especially grateful to the Scientific and Technological Research Council of Turkey (TUBITAK) for 2232 International Outstanding Researcher Fellowship and TUBITAK ULAKBIM for TRUBA cluster and data storage service.

Author Contributions

D.D., G.S., K.B.O., E.C., and M.T. contributed to conceptualization. F.A., Y.D., M.T., Y.N.E., and K.B.O. contributed to experimental analysis. B.D., S.D., M.S., M.E., E.C., G.S., D.D., K.B.O., and M.T. contributed to methodology. F.A., K.B.O., Y.D., D.D., G.S., B.D., S.D., E.C.K., K.B., M.E., E.C., and M.T. performed investigation and analyzed the results. F.A., Y.D., K.B.O., D.D., K.B., and M.T. performed visualization. K.B.O., D.D., F.A., E.K., S.D., K.B., E.C.K., Y.N.E., M.T., and B.D. performed supervision. D.D., K.B.O., F.A., and M.T. contributed to writing—original draft. All authors contributed to writing—review and editing and critically reviewed and approved the final version of the manuscript.

Data Availability

All code was implemented in Python using PyTorch as the primary deep learning library. The complete pipeline for processing whole-slide images and training and evaluating deep learning models is available at repository in https://github.com/FerasAlaqad/Ganglion-Detector and can be used to reproduce the experiments of this paper.

Funding

Mehmet Turan was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) and the Turkish Directorate of Strategy and Budget, Turkey, under TAM project number 2007K12-873. Mehmet Turan also thanks TUBITAK UlakBim for its TRUBA High-Performance Computing Service.

Declaration of Competing Interest

None reported.

Ethics Approval and Consent to Participate

The Ege University Ethics Committee(s) approved the study with ID number 23-8T/25.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.labinv.2024.102189

References

1. Smith C, Ambartsumyan L, Kapur RP. Surgery, surgical pathology, and postoperative management of patients with Hirschsprung disease. *Pediatr Dev Pathol*. 2020;23(1):23–39. https://doi.org/10.1177/1093526619889436
2. Lotfollahzadeh S, Taherian M, Anand S. Hirschsprung Disease. In: *StatPearls*. StatPearls Publishing; 2023. Accessed January 7, 2024. http://www.ncbi.nlm.nih.gov/books/NBK562142/
3. Martucciello G. Hirschsprung's disease, one of the most difficult diagnoses in pediatric surgery: a review of the problems from clinical practice to the bench. *Eur J Pediatr Surg*. 2008;18(3):140–149. https://doi.org/10.1055/s-2008-1038625
4. Smith M, Chhabra S, Shukla R, et al. The transition zone in Hirschsprung's bowel contains abnormal hybrid ganglia with characteristics of extrinsic nerves. *J Cell Mol Med*. 2023;27(2):287–298. https://doi.org/10.1111/jcmm.17659
5. Kapur RP. Histology of the transition zone in Hirschsprung disease. *Am J Surg Pathol*. 2016;40(12):1637–1646. https://doi.org/10.1097/PAS.0000000000000711
6. Maia DM. The reliability of frozen-section diagnosis in the pathologic evaluation of Hirschsprung's disease. *Am J Surg Pathol*. 2000;24(12):1675–1677. https://doi.org/10.1097/00000478-200012000-00013
7. Meyrat BJ, Lesbros Y, Laurini RN. Assessment of the colon innervation with serial biopsies above the aganglionic zone before the pull-through procedure in Hirschsprung's disease. *Pediatr Surg Int*. 2001;17(2-3):129–135. https://doi.org/10.1007/s003830000507
8. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555–570. https://doi.org/10.1038/s41551-020-00682-w
9. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2
10. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:770–778. https://doi.org/10.1109/CVPR.2016.90
11. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2009:248–255. https://doi.org/10.1109/CVPR.2009.5206848
12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 618–626. https://doi.org/10.1109/ICCV.2017.74
13. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. *Preprint*. Posted online January 13, 2018. https://doi.org/10.48550/arXiv.1801.04381
14. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. *Preprint*. Posted online May 6, 2019. bioRxiv 1905.02244. https://doi.org/10.48550/arXiv.1905.02244
15. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021:9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
16. Zhou D, Kang B, Jin X, et al. DeepViT: towards deeper vision transformer. *Preprint*. Posted online March 22, 2021. bioRxiv 2103.11886v2. https://doi.org/10.48550/arXiv.2103.11886
17. Liu X, Peng H, Zheng N, Yang Y, Hu H, Yuan Y. EfficientViT: memory efficient vision transformer with cascaded group attention. Preprint. Posted online May 11 2023. bioRxiv 2305.07027. https://doi.org/10.48550/arXiv.2305.07027
18. Chen CFR, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021:347–356. https://doi.org/10.1109/ICCV48922.2021.00041
19. Butler Tjaden NE, Trainor PA. The developmental etiology and pathogenesis of Hirschsprung disease. *Transl Res*. 2013;162(1):1–15. https://doi.org/10.1016/j.trsl.2013.03.001
20. Alexandrescu S, Rosenberg H, Tatevian N. Role of calretinin immunohistochemical stain in evaluation of Hirschsprung disease: an institutional experience. *Int J Clin Exp Pathol*. 2013;6(12):2955–2961. PMID:24294384.
21. Greenberg A, Aizic A, Zubkov A, Borsekofsky S, Hagege RR, Hershkovitz D. Automatic ganglion cell detection for improving the efficiency and accuracy of hirschprung disease diagnosis. *Sci Rep*. 2021;11(1):3306. https://doi.org/10.1038/s41598-021-82869-y

22. Shayan K, Smith C, Langer JC. Reliability of intraoperative frozen sections in the management of Hirschsprung's disease. *J Pediatr Surg*. 2004;39(9): 1345–1348. https://doi.org/10.1016/j.jpedsurg.2004.05.009
23. Beltman L, Shirinskiy I, Donner N, et al. Determining the correct resection level in patients with Hirschsprung disease using contrast enema and full thickness biopsies: can the diagnostic accuracy be improved by examining submucosal nerve fiber thickness? *J Pediatr Surg*. 2023;58(8):1463–1470. https://doi.org/10.1016/j.jpedsurg.2022.08.019
24. Matsukuma K, Gui D, Saadai P. Hirschsprung disease for the practicing surgical pathologist. *Am J Clin Pathol*. 2023;159(3):228–241. https://doi.org/10.1093/ajcp/aqac141
25. Jaafar H. Intra-operative frozen section consultation: concepts, applications and limitations. *Malays J Med Sci*. 2006;13(1):4–12. PMID:22589584.
26. Schilling F, Geppert CE, Strehl J, et al. Digital pathology imaging and computer-aided diagnostics as a novel tool for standardization of evaluation of aganglionic megacolon (Hirschsprung disease) histopathology. *Cell Tissue Res*. 2019;375(2):371–381. https://doi.org/10.1007/s00441-018-2911-1
27. Kapur RP, Kennedy AJ. Histopathologic delineation of the transition zone in short-segment Hirschsprung disease. *Pediatr Dev Pathol*. 2013;16(4): 252–266. https://doi.org/10.2350/12-12-1282-OA.1
28. Kapur RP, Reed RC, Finn LS, Patterson K, Johanson J, Rutledge JC. Calretinin immunohistochemistry versus acetylcholinesterase histochemistry in the evaluation of suction rectal biopsies for Hirschsprung disease. *Pediatr Dev Pathol*. 2009;12(1):6–15. https://doi.org/10.2350/08-02-0424.1