Cross-Modal Attention for Multimodal Information Fusion: A Novel Approach to Attention Deficit Hyperactivity Disorder Detection

Christian Nash
Intelligent Sensing and Communications Research Group
Newcastle University, UK
Email: c.nash@newcastle.ac.uk

Rajesh Nair
Adult ADHD Services
Cumbria, Northumberland, Tyne and Wear, NHS Foundation Trust
Email: rajesh.nair@cntw.nhs.uk

Syed Mohsen Naqvi
Intelligent Sensing and Communications Research Group
Newcastle University, UK
Email: mohsen.naqvi@newcastle.ac.uk

***Abstract*—This paper presents a novel method for differentiating Attention Deficit Hyperactivity Disorder subjects from control participants by multimodal data fusion, including video observations and questionnaire responses. By exploiting the well known Video Vision Transformer model, we analyse the video modality to identify the complex spatial-temporal information of ADHD symptoms. Simultaneously, a Multi-Layer Perceptron model is applied to evaluate structured questionnaire data by capturing key cognitive and emotional indicators of the ADHD symptoms. To fuse the two modalities, a cross-modal attention mechanism assigns adaptive weights to each feature based on its classification relevance. The targeted weighting significantly refines the proposed model's decision-making capability by concentrating on the most critical elements of the aggregated information. For training and testing, our novel Multimodal ADHD dataset recorded under the Intelligent Sensing ADHD Trial in collaboration with Cumbria, Northumberland, Tyne and Wear NHS Foundation Trust UK is evaluated. The proposed model, ADViQ-AL achieves a 98.18% classification accuracy, 97.83% sensitivity, and 98.53% specificity in classifying ADHD and control groups.**

***Index Terms*—Attention Deficit Hyperactivity Disorder, Deep Learning, Machine Learning, Mental Health, Multimodal**

I. Introduction

The World Health Organisation (WHO) reports a 5.2% global prevalence of Attention Deficit Hyperactivity Disorder (ADHD) across all ages [1], a figure likely underestimated, particularly in adults due to diagnostic challenges and stigma. Systematic reviews reveal adult ADHD rates at 2.58% for persistent cases and 6.76% for symptomatic adults [2], indicating widespread underdiagnosis. Such underrecognition, especially given the Diagnostic Statistical Manual's (DSM-v) focus on childhood symptoms [3], provides the motivation for our ADHD symptoms detection system in adults. Further motivation for this is that undiagnosed ADHD carries significant socioeconomic costs. Research shows an undiagnosed cost of $66.8 billion in unemployment and $28.8 billion in productivity losses [4].

Recent surveys have shown that the vast majority of research in detecting and classifying ADHD has been done using Magnetic Resonance Imaging (MRI) data and Electroencephalography (EEG) signals [5] [6]. A very small amount of work has been done in the video domain, largely due to limited availability to the data [7] [8] [9]. This further provides motivation for the proposed research direction. Previous work carried out by the authors exploited a Long Short Term Memory (LSTM) network that tracked facial landmarks. This approach classified ADHD subjects with an accuracy of $88.24\% \pm 3.93\%$ [10].

ADHD classification using multimodal data is difficult to conduct due to the scarcity in the data.To the best of our knowledge, Qureshi et al. exploit sub-modalities from functional and structural MRI's to classify ADHD [11]. Using statistical methods and an extreme learning machine algorithm, a multi-class classification accuracy of 76.19% was achieved. Furthermore, Luo et al. utilised multiple sub-modalities of functional and structural MRI's to classify ADHD [12]. The results showed there were key differences in brain activity between controls and ADHD subjects. Using an ensemble learning technique with a support vector machine backbone, a classification accuracy of 81.6% was achieved. Kautzky et al. integrated genetic data and positron emission tomography imaging data within the serotenergic system to distinguish ADHD subjects [13]. Research showed it was 82% effective in distinguishing a control to an ADHD subject. Lohani and Rana sought to classify ADHD using structural MRI data (cortical thickness features) with personal characteristic data [14]. Using a radial-based support vector machine, a classification accuracy of 75% was achieved.

The Adult ADHD Self-Report Scale (ASRS) is a key instrument in ADHD diagnosis, developed through collaboration between the WHO and the Workgroup on Adult ADHD. It was designed to screen adult ADHD symptoms in line with DSM-IV criteria, aiding early detection and management [15]. Figure 1 shows the first six questions in the ASRS. In total the ASRS is comprised of eighteen questions with the first six questions have the biggest predictive power of ADHD. If four or more responses to these six questions appear in the shaded boxes, then the patients are experiencing symptoms highly consistent with ADHD in adults.

In machine learning, the ASRS provides a standardized

Adult ADHD Self-Report Scale (ASRS-v1.1) Symptom Checklist

Patient Name		Today's Date					
Please answer the questions below, rating yourself on each of the criteria shown using the scale on the right side of the page. As you answer each question, place an X in the box that best describes how you have felt and conducted yourself over the past 6 months. Please give this completed checklist to your healthcare professional to discuss during today's appointment.			Never	Rarely	Sometimes	Often	Very Often
1. How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?							
2. How often do you have difficulty getting things in order when you have to do a task that requires organization?							
3. How often do you have problems remembering appointments or obligations?							
4. When you have a task that requires a lot of thought, how often do you avoid or delay getting started?							
5. How often do you fidget or squirm with your hands or feet when you have to sit down for a long time?							
6. How often do you feel overly active and compelled to do things, like you were driven by a motor?							

Fig. 1: The first six questions a participant would have to answer on the ASRS. The participant can only answer once per question. We encode the responses 0 - 4 (Never - Very often).

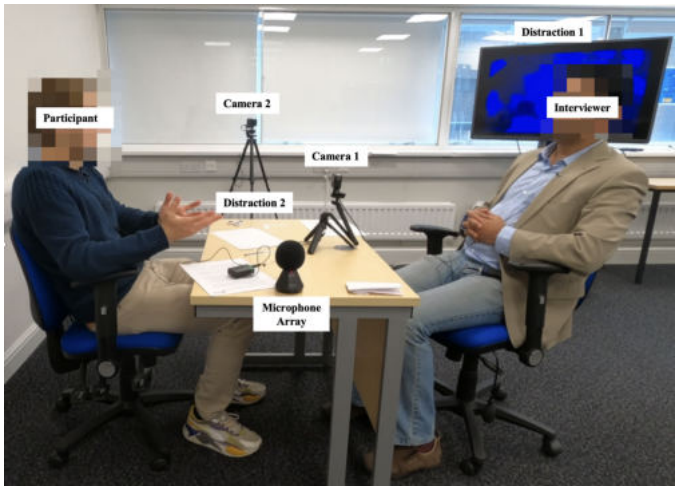

Fig. 2: The experimental design throughout recording the real multimodal ADHD dataset.

dataset crucial for training models, integrating well with multimodal data including RGB video for symptoms analysis. While the ASRS is a cost-effective and accessible tool, its efficacy is contingent upon the respondents' honesty and self-awareness. The potential for underreporting or overreporting symptoms necessitates sophisticated machine learning algorithms capable of discerning genuine patterns amidst data variability and bias.

The pivotal paper "Attention Is All You Need" has been instrumental in shaping the current landscape of machine learning by introducing the Transformer model, which centres around the novel concept of attention mechanisms [16]. Unlike previous models that processed data sequentially, Transformers compute outputs based on the entire input sequence simultaneously, allowing for a more nuanced and comprehensive understanding of data relationships. The core innovation of the Transformer is its reliance on self-attention – a mechanism that assesses the importance of different parts of the input data relative to each other. This approach enables the model to capture intricate dependencies without being constrained by the position or order of the data elements, a significant advancement over traditional sequence-based models like RNNs and LSTMs.

Identifying ADHD through visual cues alone presents challenges due to the variability of symptoms among individuals. The most observable symptoms, which are crucial for detection, encompass hyperactive behaviors such as; hand fidgeting, an inability to remain seated, undue physical activity, leg tapping, incessant talking, and averted gaze from assigned tasks [3]. In our research, we utilize only a front-facing camera to discern these hyperactive indicators. Leveraging the Intelligent Sensing ADHD Trial dataset, the proposed work has demonstrated a classification accuracy of 98.18%.

II. Methods

A. Multimodal ADHD Dataset

The Intelligent Sensing ADHD Trial (ISAT) is a novel multimodal dataset (audio, video, history/questionnaire, keyboard tracking) [17]. It contains 22 participants, 10 subjects with ADHD, and 12 controls. Our ADHD subjects have been provided by the Cumbria, Northumberland, Tyne and Wear (CNTW) National Healthcare Services (NHS) Foundation Trust. The CNTW NHS Foundation Trust is one of the largest mental health NHS Trusts in the UK. For validation purposes, every participant in the dataset performed the Adult ADHD Self-Report Scale (ASRS-v1.1) symptoms checklist [15].

In the creation of our novel multimodal ADHD dataset depicted in Figure 2, we utilized three GoPro cameras recording at a 3840×2160 (4K) resolution and 30 fps. To ensure clear audio capture with minimal background noise, all participants were equipped with Lavelier microphones. We further enhanced environmental sound capture through a quartet of microphones positioned on the table, aimed at detecting ambient distractions like sudden loud noises. The setup also included a fidget spinner and a notepad to serve as potential distractions, alongside a TV monitor cycling through various wallpapers and emitting diverse sounds to simulate real-life distractions.

The data collection protocol commenced with an interview segment where participants answered 21 questions from the Diagnostic Interview for ADHD in Adults (DIVA) [18], followed by the administration of the CANTAB [19] computerized tasks assessing various cognitive functions. Subsequent to these tasks, participants engaged in a response action task, signaling responses via hand raises to specific auditory cues. The session concluded with the viewing of two videos designed to elicit differing levels of engagement. Each session varied in length from 60 to 90 minutes, tailored to individual participants. For our analysis, we focused exclusively on data captured by the front-facing camera.

B. Video Pre-Processing

Given that the original videos in our unique multimodal dataset were recorded in 4K resolution, it was crucial to downscale the video dimensions to manage computational efficiency

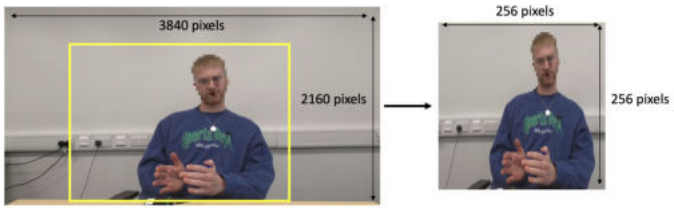

Fig. 3: The participant's movement is tracked across 150 frames at a recording speed of 30 fps. A bounding box is created around the region where the participant moves throughout the 5-second duration. Subsequently, the video is cropped to this specific area and resized to 256×256 pixels.

while retaining sufficient detail to observe behavioral nuances. Consequently, for the RGB video classification model, the 5-second video clips were resized to 256×256 pixels. A significant challenge we faced was the static background present in these interview-style recordings, which could introduce a considerable amount of redundant information into our analysis. To address this, we employed a background subtraction strategy. Utilizing the Mask-RCNN model [20], we were able to identify and track the human subject within each 5-second clip, generating a bounding box that followed the participant's movements frame by frame. This bounding box was then used to isolate the subject from the static background, after which the extracted segment was resized to 256×256 pixels, as illustrated in Figure 3. This procedure ensured that the focal point of each video clip was consistently the participant, thereby eliminating extraneous data and enhancing the efficiency and focus of our model.

C. ASRS Questionnaire Generation

In the proposed work, we faced a dataset challenge. Each of the 18 participants, evenly divided between ADHD subjects and controls, had only one set of ASRS questionnaire responses against 4,050 video clips. To overcome this and deepen our dataset, we generated artificial questionnaire responses based on the ASRS's 0 (Never) - 4 (Very Often) encoding scale. This approach commenced with an in-depth analysis of the existing responses, aiming to identify and encapsulate the nuanced response patterns characteristic of each group. The WHO analysed the effectiveness of the ASRS in 2005 [15]. In the report, average responses per question for each respective group are provided. Using this information, the random response per question had a probability that is reflected in the research. This approach allowed us to create nuanced, probabilistically informed artificial responses, introducing controlled variability to mirror real-world response diversity and enhance our model's generalisation capability while addressing overfitting risks.

However, generating artificial responses poses challenges regarding data authenticity and potential bias, given their synthetic nature. These issues necessitate careful consideration to maintain the integrity and applicability of the proposed approach.

D. Classification

1) Video Vision Transformer: We exploit Video Vision Transformer (ViVIT) for video data processing to harness the spatial and temporal richness of video clips, crucial for differentiating ADHD subjects from controls [21]. ViVIT, distinct from conventional CNNs, employs a self-attention mechanism, allowing a holistic analysis of video frames over time, which is vital for capturing the intricate behaviours in our dataset. This method contrasts with CNNs that prioritise local spatial features, whereas ViVIT's architecture assesses the entire frame sequence, providing a deeper insight into content and enabling the detection of nuanced behavioral patterns associated with ADHD.

The adoption of ViVIT requires video preprocessing to meet model specifications, including segmenting and normalizing video clips. Post-processing, ViVIT generates feature vectors that encapsulate the videos' semantic content, offering a detailed portrayal of actions and contexts within the clips. These vectors are then integrated with questionnaire data using a cross-modal attention mechanism, enhancing the classification's accuracy. The choice of ViVIT enriches our model's interpretative capacity, ensuring a detailed and context-aware analysis that aligns with our objective of discerning between ADHD and control participants effectively.

2) Multi-layer Perceptron: In our approach, we utilize a Multi-Layer Perceptron (MLP) to analyze questionnaire data, effectively handling the structured numerical information it provides. The MLP's straightforward architecture is ideal for transforming questionnaire responses into numerical vectors, which are then used to classify participants as either ADHD or control group members.

The process involves numerically encoding responses to feed into the MLP, which comprises an input layer, several hidden layers, and an output layer. The hidden layers enable the MLP to discern complex relationships within the data, influencing the final classification output.

Despite its simplicity, the MLP adeptly identifies nuanced patterns in questionnaire responses, distinguishing between ADHD and control participants. This capability demonstrates the MLP's effectiveness in extracting meaningful insights from seemingly straightforward data, demonstrating its value in our multimodal classification framework.

3) Attention Mechanism for Fusion: In the proposed work, we leverage cross-modal attention for feature fusion, aimed at enhancing the classification accuracy between ADHD subjects and control groups. This approach harnesses the strengths of two distinct data modalities: RGB video data processed by the ViVIT model and nuanced questionnaire data interpreted through a Multi-Layer Perceptron (MLP).

The essential part of our method lies in the application of a cross-modal attention mechanism that intelligently integrates the heterogeneous features extracted from both modalities. The mechanism is designed to dynamically adjust the influence of each feature based on its relevance to the classification task. This enables a more nuanced and context-aware decision-making process. By assigning variable weights to features

from the video and questionnaire data, the model adeptly focuses on the most informative aspects, enhancing both the robustness and interpretability of the classification.

One of the significant strengths of this methodology is its adaptability, allowing the model to prioritise features that are most indicative of ADHD, irrespective of their modality. This adaptability not only improves classification accuracy but also offers insights into the complex interplay between behavioural cues and self-reported symptoms, potentially shedding light on novel ADHD markers.

Saying this, the approach is not without challenges. The implementation of a cross-modal attention mechanism introduces additional complexity to the model architecture, demanding sophisticated training strategies and potentially increasing computational requirements. Moreover, the efficacy of the attention-based fusion relies heavily on the quality and representativeness of the data from both modalities, highlighting the importance of comprehensive dataset curation.

Despite these challenges, the advantages of employing cross-modal attention for feature fusion in ADHD classification are compelling. This methodology not only sets a new standard for the accuracy and interpretability of machine learning models in mental health diagnostics but also paves the way for future advancements in multimodal data integration for a broad range of clinical applications.

III. Results

Leave-one-out cross-validation was implemented during training to increase the reliability of the results. Implementing this increases the overall training times because every participant has their own respective model trained. Each iteration of training included 75 epochs with a learning rate of 0.001. Early stopping was implemented to reduce training times and to further mitigate overfitting. We then choose the best overall model for testing on the holdout set, where the preference is high performance and low computational complexity. The dataset used contains 4,050 samples per participant. To ensure a balanced dataset, we used 18 participants, 9 controls, and 9 ADHD subjects. During training and validation, 14 participants (7 ADHD/7 control) are used. The remaining 4 participants were left out, as their respective clips will be used as the hold-out dataset. The goal of the holdout set is to classify ADHD, therefore, we felt that 8,100 samples are enough for a holdout dataset size. This was done to ensure that there was no cross-over with the training of the models. An additional benefit is that the effects of overfitting are mitigated which increases the validity of the results. We recognise that scarcity in medical data is a common problem. Saying this, we feel we have mitigated the effects of overfitting to the best of our ability while also training on a dataset size of 56,700 samples and a testing size of 16,200 samples. The proposed approach was trained and tested on a system that contains a 13^{th} Gen Intel i7 processor, NVIDIA RTX3090 GPU, 64GB RAM with Ubuntu running CUDA 11.

We propose ADViQ-AL (Attention-Driven Video and Questionnaire ADHD Learning). Figure 4 shows the predictive

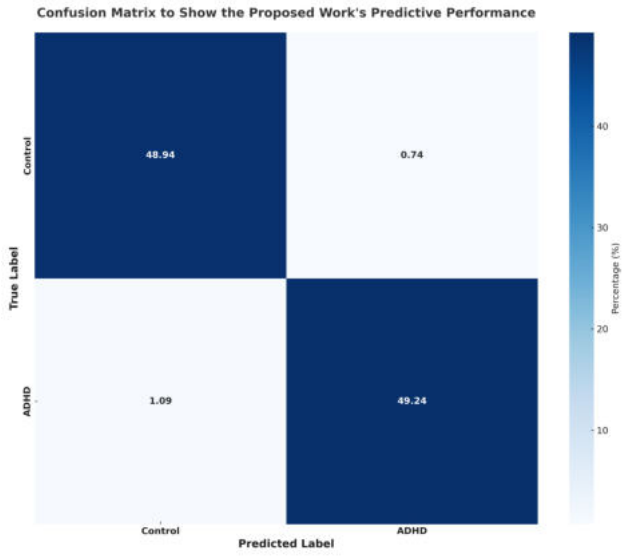

Fig. 4: Confusion matrix illustrating the predictive performance of the proposed work in classifying ADHD subjects from control participants. The matrix quantifies the models accuracy in terms of percentage.

accuracy of ADViQ-AL, using RGB video and questionnaire data through our advanced multimodal approach. The model demonstrates a robust capability in accurately identifying ADHD cases, as highlighted by a high proportion of true positives. This significant percentage reflects ADViQ-AL's sensitivity/recall for the ADHD class, illustrating its effectiveness in capturing ADHD instances, which is crucial for reliable ADHD detection.

In contrast, ADViQ-AL exhibits a low percentage of false negatives, indicating its efficiency in rarely misclassifying ADHD instances as control. This performance is key in reducing Type 2 errors, essential in medical diagnostics where overlooking ADHD can have profound implications. Additionally, the model's low false positive rate emphasises its high specificity, illustrating its success in precisely identifying control cases, thereby minimising Type 1 errors and avoiding incorrect ADHD labels for control instances. We believe that in certain clips for the ADHD subjects, there is very minimal movement or the movement could be something as trivial as scratching the head. This is behaviour that is also typical of the controls which could explain the small amount of false negatives we are witnessing. The high percentage of true negatives further validates the model's dependability in accurately recognizing control participants, reinforcing its fair performance across both ADHD and control classifications. This equilibrium suggests that the model is adept at discerning between ADHD and control classes without introducing bias towards either. This is a testament to the effectiveness of our multimodal approach and attention mechanism in enhancing predictive performance.

Overall, the percentage-based breakdown of the confusion matrix not only highlights the model's strengths in sensitivity

TABLE I: Performance metric comparison of the proposed model with and without the attention mechanism. All results shown is the performance on the hold-out dataset and shown in percentages.

	Concatenation	ADViQ-AL
Accuracy	95.02	***98.18***
Precision	95.25	***98.52***
Sensitivity	94.68	***97.83***
Specificity	95.36	***98.53***
F1 Score	94.96	***98.17***

and specificity but also offers a comprehensive perspective on its balanced and nuanced performance in the context of ADHD classification. Such detailed insights are pivotal for gauging the model's clinical applicability and reliability, where accurate and balanced diagnostic capabilities are essential.

Table I demonstrates that the incorporation of an attention mechanism has enhanced the model's performance across all pertinent metrics when compared with a model employing late-fusion concatenation. ADViQ-AL exhibits a marked improvement in accuracy, rising to 98.18% from 95.02% observed in the late-fusion concatenation model. This enhancement is not merely a numerical increase but signifies a substantial advancement in the model's ability to discern between ADHD and control groups accurately. Such precision is paramount in clinical settings where the cost of misclassification can have implications on patient care and therapeutic direction. The precision metric, which ascends to 98.52% in ADViQ-AL, reveals the model's enhanced specificity in identifying true ADHD cases, minimizing the risk of false positives. In the realm of healthcare, where diagnostic accuracy is imperative, the elevated precision ensures that individuals are correctly identified as having ADHD, thereby facilitating timely and appropriate intervention.

Furthermore, the ADViQ-AL achieves a sensitivity of 97.83%, surpassing the late-fusion concatenation model's 94.68%. This increase is indicative of ADViQ-AL's strengthened ability to capture the majority of true ADHD instances, thereby reducing the likelihood of Type 2 errors. This is a critical consideration in medical diagnostics where failing to detect an ADHD case could delay or deny necessary treatment. Specificity, which stands at 98.53% with ADViQ-AL, highlights the model's capability to correctly negate false ADHD diagnoses, an essential factor in preventing unnecessary medical interventions for control subjects. Additionally, the F1 score, which harmonises precision and recall, reflects a balance at 98.17%, suggesting that ADViQ-AL achieves an optimal equilibrium between detecting ADHD cases and avoiding misclassifications.

The core advantage provided by ADViQ-AL stems from its capacity to dynamically allocate importance to different features from the RGB video and questionnaire data, facilitating a context-aware integration of information. This is a significant improvement from the late-fusion approach, which merely concatenates features without regard for their interrelatedness or the contextual implications of their combination. In essence, the employment of an attention mechanism within our multimodal framework not only amplifies the model's diagnostic accuracy but also enriches its interpretability, offering insights into the data-driven inferences it draws.

TABLE II: Comparison with the state-of-the-art multimodal methods for ADHD detection.

	Accuracy (%)	Validation	Age Group (years)
Structural and Functional MRI [11]	76.19	✓	> 18
Structural MRI Sub-modalities (>4) [12]	81.6	✓	> 18
Eye Movement + Facial Expression + 3D posture [22]	80.25	✗	Specified as children
Genetic + PET Imagery [13]	82.0	✓	> 18
Structural MRI + PC Question Data [14]	75.0	✓	18 - 40
ADViQ-AL	***98.18***	✓	***18 - 53***

Table II compares the proposed work with existing state-of-the-art work that exploits video data. The results show that ADViQ-AL improves from our own existing work, while outperforming the state-of-the-art performance. [22] is closely related to the proposed work, in terms of analysing facial behaviour, with the proposed work outperforming the method. [14] provides the closest comparison for multimodal structures. Both papers utilise a form of medical record for data with the difference being the type of imaging data. The proposed work suggests that the RGB video data is more effective in identifying ADHD as ADViQ-AL outperforms the suggested approach. Furthermore this comparison could provide insight into the usefulness of the medical records for classification. In [14], the records display age, weight and more personal information. Whereas in the proposed work, we use the ASRS which is specifically designed for detecting ADHD. This could be an interesting future work direction to see if the relationship between the medical record and the partner modality makes a difference.

The prevailing method in the limited existing literature for multimodal ADHD detection is MRI data. Structural MRI's and functional MRI's are two different modalities because they serve different purposes and provide distinct types of neurological information. The public ADHD-200 dataset is the driving factor for the research prevalence and the functionality of splitting MRI data into different sub-modalities. A big drawback of MRI data is that the machines to capture the images are expensive and require training to use. Users of these machines can be made to feel uncomfortable due to their claustrophobic nature. In contrast, our proposed work exploits only a stationary camera capable of capturing the subject's face and body, enabling the detection of hyperactivity symptoms associated with ADHD without wearable devices. Our approach demonstrates the feasibility of accurately identifying ADHD symptoms through a non-wearable, more accessible means. This simplicity paves the way for the integration of our system

into future smart devices through applications. Our innovative system is designed to recognize multiple ADHD symptoms.

IV. Conclusion

In conclusion, the proposed work introduced an approach to the classification of ADHD by employing a novel implementation of cross-modal attention fusion, integrating RGB video data and questionnaire responses. We capitalised on the intrinsic strengths of the ViVIT model and a Multi-Layer Perceptron to process these distinct data modalities, subsequently harnessing a cross-attention mechanism to dynamically synthesise the extracted features. This fusion strategy has proven to be instrumental in capturing the nuanced interplay between the visual cues from video data and the self-reported symptoms and behaviours from questionnaires, offering a comprehensive and contextually enriched analysis.

The results showed the high performance of the cross-modal attention mechanism, as shown by the performance metrics compared to the baseline late-fusion concatenation model. Specifically, ADViQ-AL achieved a high accuracy of 98.18%, a precision of 98.52%, and an F1 score of 98.17%. This indicates not only its robustness in classification accuracy but also its precision in distinguishing between ADHD and control groups. The sensitivity and specificity metrics further attest to ADViQ-AL's capability to correctly identify ADHD instances while minimizing false positives, a testament to its clinical utility. By effectively capturing the relationships between the video and questionnaire data, ADViQ-AL offers insights into the multifaceted nature of ADHD, potentially unveiling novel behavioural and symptomatic markers. Moreover, the compelling performance improvements highlighted in our findings not only validate the effectiveness of our proposed model but also emphasise the value of cross-modal attention in extracting meaningful, discriminative features from complex, multimodal datasets.

The results of ADViQ-AL suggest cross-modal attention is a powerful approach when understanding multimodal data. With recent work on our own novel Multimodal ADHD dataset including acoustic and text feature analysis [23], our future work will look to bridging the gap between the RGB video and audio data. With the aim being to further understand the subtle relationships between the different data modalities which can provide further insight into technological mental health diagnosis.

References

[1] W. H. Organization, "Attention Deficit Hyperactivity Disorder (ADHD)," 2019.

[2] P. Song, M. Zha, Q. Yang, Y. Zhang, X. Lie, and I. Rudan, "The Prevalence of Adult Attention-Deficit Hyperactivity Disorder: A Global Systematic Review and Meta-analysis," *Journal of Global Health*, vol. 11, pp. 1–9, 2021.

[3] A. P. Association, "Diagnostic and Statistical Manual of Mental Disorders," *Diagnostic and Statistical Manual of Mental Disorders*, 2013.

[4] J. Schein, A. A. Lenard, A. Childress, P. Gagnon-Sanschagrin, M. Davidson, F. Kinkead, M. Cloutier, A. Guérin, and P. Lefebvre, "Economic Burden of Attention-Deficit/Hyperactivity Disorder Among Adults in the United States: a Societal Perspective," *Journal of Managed Care and Specialty Pharmacy*, vol. 28, pp. 168–179, 2022.

[5] C. Nash, R. Nair, and S. Naqvi, "Machine Learning and ADHD Mental Health Detection - A Short Survey," *International Conference on Information Fusion (FUSION)*.

[6] C. Nash, R. Nair, and S. Naqvi, "Machine Learning in ADHD and Depression Mental Health Diagnosis: A Survey," *IEEE Access*, 2023.

[7] Y. Li, R. Nair, and M. Naqvi, "Video-Based Skeleton Data Analysis for ADHD Detection," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023, pp. 1–6.

[8] P. V. Casal, F. L. Esposito, I. M. Martinez, A. Capdevila, M. S. Puig, N. de la Osa, L. Ezpeleta, A. P. I. Lluna, S. V. Faraone, J. A. Ramos-Quiroga, H. Super, and J. Canete, "Clinical Validation of Eye Vergence as an Objective Marker for Diagnosis of ADHD in Children," *Journal of Attention Disorders*, vol. 23, pp. 599–614, 2019.

[9] S. D. Silva, S. Dayarathna, G. Ariyarathne, D. Meedeniya, S. Jayarathna, and A. M. P. Michalek, "Computational Decision Support System for ADHD Identification," *International Journal of Automation and Computing*, vol. 18, pp. 233–255, 2021.

[10] C. Nash, R. Nair, and S. Naqvi, "ADHD Mental Health Symptoms Detection Based on Facial Landmark Tracking," *1st International Conference on Artificial Intelligence, Robotics, Signal and Image Processing (AIRoSIP)*, 2023.

[11] M.Qureshi, J. Oh, B. Min, H. J. Jo, and B. Lee, "Multi-modal, Multi-measure, and Multi-class Discrimination of ADHD with Hierarchical Feature Extraction and Extreme Learning Machine Using Structural and Functional Brain MRI," *Frontiers in Human Neuroscience*, vol. 11, 4 2017.

[12] Y. Luo, T. Alvarez, J. Halperin, and X. Li, "Multimodal Neuroimaging-based Prediction of Adult Outcomes in Childhood-onset ADHD Using Ensemble Learning Techniques," *NeuroImage: Clinical*, vol. 26, p. 102238, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158220300759

[13] A. Kautzky, T. Vanicek, C. Philippe, G. S. Kranz, W. Wadsak, M. Mitterhauser, A. Hartmann, A. Hahn, M. Hacker, D. Rujescu, S. Kasper, and R. Lanzenberger, "Machine Learning Classification of ADHD and HC by Multimodal Serotonergic Data," *Translational Psychiatry 2020 10:1*, vol. 10, pp. 1–9, 4 2020.

[14] D. Lohani and B. Rana, "ADHD Diagnosis Using Structural Brain MRI and Personal Characteristic Data with Machine Learning Framework," *Psychiatry Research: Neuroimaging*, vol. 334, p. 111689, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925492723000999

[15] R. C. Kessler, L. Adler, M. Ames, O. Demler, S. Faraone, E. V. A. Hiripi, M. J. Howes, R. Jin, K. Secnik, T. Spencer, T. B. Ustun, and E. E. Walters, "The World Health Organization Adult ADHD Self-Report Scale (ASRS): a Short Screening Scale For Use in the General Population," *Psychol. Med.*, vol. 35, no. 2, pp. 245–256, 2005.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 6000–6010.

[17] Y. Li, S. Li, C. Nash, S. Naqvi, and R. Nair, "24 Intelligent Sensing in ADHD Trial (ISAT) – Pilot Study," *Journal of Neurology, Neurosurgery Psychiatry*, vol. 94, p. e2, 12 2023. [Online]. Available: https://jnnp.bmj.com/content/94/12/e2.35https://jnnp.bmj.com/content/94/12/e2.35.abstract

[18] J. S. Kooij and M. Francken, "Diagnostic Interview for ADHD in Adults (DIVA)," 2010.

[19] C. Cognition, "Cambridge Cognition: CANTAB," https://www.cambridgecognition.com/cantab, 2022.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A Video Vision Transformer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6816–6826, 2021.

[22] Y. Zhang, M. Kong, T. Zhao, W. Hong, D. Xie, C. Wang, R. Yang, R. Li, and Q. Zhu, "Auxiliary Diagnostic System for ADHD in Children Based on AI Technology," *Frontiers of Information Technology and Electronic Engineering*, vol. 22, pp. 400–414, 2021.

[23] S. Li, R. Nair, and S. Naqvi, "Acoustic and Text Features Analysis for Adult ADHD Screening: A Data-Driven Approach Utilizing DIVA Interview," *IEEE Journal of Translational Engineering in Health and Medicine*, 2024.