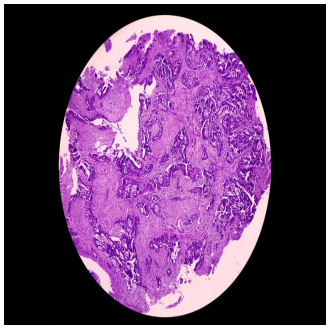
HealthManagement
Promoting Management and Leadership

Sustainable AI Benchmarking in Histopathology



Artificial Intelligence (AI) has revolutionised the field of histopathology, particularly through deep learning (DL) models that assist in medical diagnostics. However, while most research focuses on improving the diagnostic accuracy of these models, the environmental impact of their development and usage has been largely overlooked. The high computational demands of training and deploying DL models lead to substantial carbon dioxide equivalent (CO_2eq) emissions, especially in regions with non-renewable energy sources. To address this, a new benchmarking framework known as the Environmentally Sustainable Performance (ESPer) score has been introduced, combining diagnostic performance with carbon footprint measurements for a more holistic assessment.

Balancing Diagnostic Accuracy and Environmental Impact

The ESPer score was developed to assess both the diagnostic accuracy and carbon footprint of DL models in computational pathology. It evaluates a model's ability to classify complex medical cases while factoring in the CO_2eq emissions during training and inference phases. The study tested multiple models, including TransMIL, CLAM, InceptionV3, Vision Transformer (ViT) and Prov-GigaPath, using datasets for renal cell carcinoma (RCC) and kidney transplant pathology. While TransMIL and CLAM demonstrated high diagnostic accuracy with relatively low carbon footprints, other models like InceptionV3 and ViT had significantly higher emissions without a proportional increase in performance. This indicates that some models can maintain strong diagnostic results with a lower environmental cost.

The ESPer score provides a comprehensive comparison by normalising both performance and emissions, ensuring a fair assessment of each model’s sustainability. This metric is particularly important as the energy requirements for AI development can vary significantly depending on the model architecture and dataset size. By incorporating the ESPer score, researchers can make informed decisions on model selection that balance effectiveness and ecological responsibility.

CO_2 Emissions and the ESPer Metric

CO_2eq emissions were measured during the models' training and inference phases, highlighting significant disparities among them. TransMIL and CLAM emerged as the most efficient models, producing the least emissions while maintaining high diagnostic performance. Conversely, models like ViT and Prov-GigaPath produced much higher emissions despite comparable diagnostic outputs. The ESPer score normalises performance and CO_2eq emissions, allowing a direct comparison of models on both fronts. For instance, CLAM achieved a higher ESPer score due to its superior balance of accuracy and emissions, while models with lower diagnostic accuracy but higher emissions scored poorly. The score also factors in long-term emissions by considering the number of inferences made over time, providing a clearer picture of a model’s sustainability.

Moreover, the study emphasises the impact of energy grids and power sources on CO_2eq emissions. Models run in regions with higher proportions of renewable energy sources produced lower emissions, underlining the importance of location-specific evaluations when benchmarking sustainability.

Reduction Strategies for Sustainable AI Use

The study explored several strategies to reduce CO_2 emissions without compromising diagnostic accuracy. These included using larger image tiles with lower resolution and reducing the number of tiles processed per whole slide image (WSI). TransMIL, the model with the highest ESPer score, demonstrated that using fewer tiles or lower resolutions significantly reduced emissions while maintaining diagnostic performance. Additionally, early stopping techniques and selective data sampling further improved sustainability. These strategies emphasise that it is possible to develop powerful AI tools for histopathology without excessive environmental costs by prioritising efficiency during both the development and inference stages.

Further methods mentioned include model pruning and quantisation, which involve reducing the complexity of neural networks without a significant drop in performance. These approaches can be particularly beneficial for real-world deployments where energy efficiency and rapid inference times are critical.

The ESPer score provides a comprehensive approach to evaluating AI models in histopathology by integrating both performance and ecological sustainability. As AI continues to transform medical diagnostics, balancing accuracy with environmental responsibility is essential. By adopting metrics like ESPer and implementing reduction strategies, researchers can develop DL models that are not only diagnostically effective but also sustainable in the long term. The ESPer framework sets a precedent for environmentally conscious AI development, urging the medical and AI research communities to prioritise ecological sustainability alongside technological advancements.

Ultimately, ecologically sustainable benchmarking frameworks like ESPer encourage transparency and accountability in the development of medical AI tools. They empower researchers and healthcare institutions to make more informed decisions, ensuring that progress in medical technology aligns with global sustainability goals.

Source: **npj Digital Medicine**

Image Credit: **iStock**

Published on : Fri, 10 Jan 2025