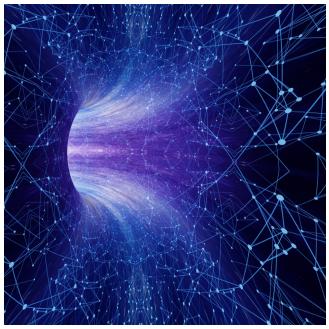
HealthManagement
Promoting Management and Leadership

Automated Labelling of Radiology Reports with LLMs



Accurate labelling of medical imaging data remains one of the most demanding aspects of training artificial intelligence models for diagnostic purposes. Traditionally, this task relies on expert annotations, a process that is both time-consuming and resource-intensive. Radiology reports often contain rich diagnostic insights that could be repurposed for data annotation. However, their free-text nature presents a challenge for automated interpretation and use. A new approach utilises large language models (LLMs) to generate labels from these reports. A recent study explored this method by training convolutional neural networks (CNNs) to detect ankle fractures based on LLM-generated labels, assessing whether such automation can deliver results comparable to those achieved through manual annotation.

Label Generation Using Open-Source Language Models
The research team employed the open-source model Mixtral-8×7B-Instruct-v0.1 to process radiology reports written in German. The LLM was tasked with producing binary classifications indicating whether or not an ankle fracture was present, using both zero-shot and few-shot prompting methods. After testing thirty-one different prompts, the final configuration yielded an accuracy of 92% on a dedicated test dataset. This optimal prompt significantly improved sensitivity while preserving a high level of specificity. To create a comprehensive dataset, the selected prompt was used to label 15,896 ankle X-ray images. This dataset included both positive and negative cases, extracted from the radiological archives of a hospital in Berlin. A validation procedure was carried out on 500 reports manually annotated by radiologists, which confirmed the high accuracy of the LLM-generated labels.

Must Read: ***Understanding Artificial Neural Networks in Modern Healthcare***

CNN Development and Performance Evaluation
With the labelled data in place, a CNN was developed using the DenseNet121 architecture. The training utilised data augmentation techniques such as random rotations and cropping to improve the model's generalisability. All images were standardised to a size of 640×640 pixels. Balanced datasets were ensured by under-sampling non-fracture images in the validation and test groups. The model was trained across 200 epochs using a university cluster with high-performance computing capabilities. The best-performing version of the model achieved an accuracy of 89.5% and a receiver operating characteristic area under the curve (ROC-AUC) of 0.926 on the test set. Sensitivity and specificity were also high, demonstrating the model's ability to detect fractures reliably. A sensitivity analysis excluding manually validated reports from the training data showed only marginal differences in performance, indicating the robustness of the overall approach.

Scalability and Implications for Medical AI Development
The study's methodology demonstrates a practical and efficient way to reduce the human effort involved in medical data annotation. LLMs provided accurate, scalable label generation from unstructured text, avoiding the limitations of traditional rule-based or machine learning approaches that require prior annotation. This approach is especially appealing due to its adaptability; LLMs can be redirected to new tasks through simple prompt modifications, eliminating the need for retraining. The use of an open-source model further enhances feasibility, as it avoids privacy concerns and operational costs associated with commercial LLMs. Additionally, the modest computational requirements of the Mixtral-8×7B model make it suitable for use in settings with limited technical infrastructure. These attributes position this approach as a promising tool for wider clinical AI adoption, allowing institutions to harness existing data more effectively and economically.

This study showcased the potential of large language models in automating the labelling of radiology reports for use in image-based AI applications. By training a CNN to detect ankle fractures using LLM-generated labels, the researchers demonstrated performance comparable to manually labelled datasets. The process significantly reduces the time and effort needed for annotation, which is a major bottleneck in AI development. With strong test results and an open-source foundation, this method offers a practical solution for expanding the capabilities of clinical AI systems. The approach provides a pathway for integrating unstructured textual data into diagnostic tools and has broad implications for improving scalability and efficiency in healthcare machine learning.

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.

Source: **Academic Radiology**

Image Credit: **iStock**

Published on : Thu, 24 Apr 2025

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.