ABSTRACT

Cele

Celem tego badania było (i) opracowanie możliwego do zinterpretowania systemu głębokiego uczenia się, o wysokiej dokładności, do klasyfikowania zmian w wątrobie w MRI wzmocnionym kontrastem, z przejrzystością umożliwiającą lekarzom uzasadnienie swoich decyzji oraz (ii) zatwierdzenie tego systemu przez porównanie jego wyników diagnostycznych z wynikami uzyskanymi przez radiologów.

Metody

Badanie objęło 296 pacjentów z 494 zmianami chorobowymi wątroby w sześciu kategoriach. Zmiany zostały zidentyfikowane za pomocą wielofazowego MRI i podzielone na zestawy treningowe (n=434) i testowe (n=60). Ustalone techniki powiększania obrazu zostały wykorzystane w celu zwiększenia liczby próbek szkoleniowych do 43 400. Ten zestaw szkoleniowy był wkładem do stworzonej na zamówienie sieci neuronów konwulsyjnych (CNN), składającej się z trzech warstw konwulsyjnych z powiązanymi prostymi jednostkami liniowymi, dwóch maksymalnych warstw zbiorczych oraz dwóch w pełni połączonych warstw. Do szkoleń modelarskich został użyty optymalizator Adama. Dodatkowo, do każdej zmiany chorobowej przypisano do podzbioru każdej klasy zmiany i zastosowano algorytm post hoc do wnioskowania o obecności tych cech w zestawie testowym na podstawie wzorów aktywacji (wytrenowanego) modelu CNN. Walidacja CNN została przeprowadzona poprzez porównanie wyników diagnostycznych CNN z wynikami dwóch radiologów posiadających certyfikat płytowy. Zostało to przeprowadzone przez Monte Carlo w ramach walidacji krzyżowej, a wyniki CNN na identycznym, niewidocznym zestawie testowym zostały porównane z wynikami radiologów. Wygenerowane zostały mapy cech wyróżniające regiony na oryginalnym obrazie, które odpowiadały poszczególnym cechom. Następnie do każdej zidentyfikowanej cechy przypisano ocenę istotności, oznaczającą względne znaczenie danej cechy dla przewidywanej klasyfikacji zmiany.

Wyniki

Interpretowalny system głębokiego uczenia się (DL) wykazał 92% czułości (Sn), 98% specyficzności (Sp) i 92% dokładności. Wydajność zestawu testowego w pojedynczym badaniu wykazała średnio 90% Sn i 98% Sp w sześciu typach zmian, w porównaniu do średnio 82,5% Sn i 96,5% Sp dla radiologów. Radiolodzy uzyskali Sn na poziomie 60%-70% do klasyfikacji raka wątrobowokomórkowego, natomiast system DL uzyskał Sn na poziomie 90%. Dla

szczególnego przypadku klasyfikacji HCC CNN uzyskał obszar charakterystyki pracy odbiornika pod krzywą 0,992. Czas obliczeniowy na jedno uszkodzenie wynosił 5,6 milisekundy.

Dodatnia wartość predykcyjna i Sn w identyfikacji prawidłowych cech radiologicznych występujących w każdej badanej zmianie wynosiły odpowiednio 76,5% i 82,9%, podczas gdy 12% zmian było źle sklasyfikowanych; te źle sklasyfikowane zmiany częściej prowadziły do błędnej identyfikacji cech niż prawidłowo sklasyfikowane (60,4% vs 85,6%). Oryginalne woksle obrazowe przyczyniające się do każdej funkcji obrazowania były spójne z wygenerowanymi mapami cech, a w każdej klasie najbardziej znaczące kryteria obrazowania były odzwierciedlone przez ich odpowiednie oceny istotności cech.

Wniosek

W pracy przedstawiono rozwój "interpretowalnego" prototypu systemu głębokiego uczenia się, którego dokładność przewyższa dokładność radiologów w klasyfikowaniu zmian w wątrobie na MRI wzmocnionym kontrastem, przy jednoczesnym oświetleniu procesu podejmowania decyzji przez algorytm. Przedstawiony interpretacyjny system DL wykazuje potencjał jako narzędzie wspomagające podejmowanie decyzji w diagnostyce zmian chorobowych w wątrobie; jednakże wpływ kliniczny narzędzia wspomagającego podejmowanie decyzji musi być zweryfikowany w badaniu prospektywnym, zanim będzie można rozważyć jego włączenie do praktyki klinicznej.

ABSTRACT

Objectives

The purpose of this study was (i) to develop an interpretable deep learning system, of high accuracy, for classifying hepatic lesions in contrast-enhanced MRI, with a transparency that allows justification of its decisions to physicians and (ii) to validate this system by comparison of its diagnostic performance with that of radiologists.

Methods

This study included 296 patients with 494 hepatic lesions in six categories. Lesions were identified by multiphasic MRI and divided into training (n=434) and test (n=60) sets. Established image augmentation techniques were used to increase the number of training samples to 43,400. This training set was input to a custom-made convolutional neural network (CNN), consisting of three convolutional layers with associated rectified linear units, two maximum pooling layers, and two fully connected layers. An Adam optimizer was used for model training. Additionally, up to four key imaging features per lesion were assigned to a subset of each lesion class and a post-hoc algorithm was used to infer the presence of these features in a test set on the basis of activation patterns of the (trained) CNN model. Validation of the CNN was performed by comparing the diagnostic performance of the CNN with that of two board-certified radiologists. This was carried out by Monte Carlo cross-validation, and the CNN's performance on an identical unseen test set was compared with that of the radiologists. Feature maps highlighting regions in the original image that corresponded to particular features were generated. A relevance score was then assigned to each feature identified, denoting the relative importance of the feature for the predicted lesion classification.

Results

The interpretable deep learning (DL) system demonstrated a 92% sensitivity (Sn), a 98% specificity (Sp), and a 92% accuracy. Test set performance in a single run showed an average 90% Sn and 98% Sp across the six lesion types, compared with an average 82.5% Sn and 96.5% Sp for radiologists, respectively. Radiologists achieved an Sn of 60%–70% for classifying hepatocellular carcinoma, while the DL system achieved an Sn of 90%. For the specific case of HCC classification the CNN achieved a receiver operating characteristic area under the curve of 0.992. Computation time per lesion was 5.6 milliseconds.

The positive predictive value and the Sn in identifying the correct radiological features present in each test lesion were 76.5% and 82.9%, respectively, while 12% of the lesions were

misclassified; these misclassified lesions led more often to wrongly identified features than the correctly classified ones did (60.4% vs. 85.6%). Original image voxels contributing to each imaging feature were consistent with the feature maps generated, and in each class the most prominent imaging criteria were reflected by their respective feature relevance scores.

Conclusion

This study presents the development of an "interpretable" deep learning system prototype, the accuracy of which exceeds that of radiologists in classifying hepatic lesions on contrast-enhanced MRI, while illuminating the algorithm's decision-making. The interpretable DL system presented demonstrates potential as a decision-support tool in liver lesion diagnosis; however, the clinical impact of the decision-support tool needs to be validated in a prospective study before the tool can be considered for integration into clinical practice.