HAL open science

Utilisation de techniques d'intelligence artificielle dans le cadre de l'étude des tiques et de la maladie de Lyme

Sk Imran Hossain, Yann Frendo, Valerie Poux, Isabelle Lebert, Delphine Martineau, Olivier Lesens, Engelbert Mephu Nguifo, Jocelyn de Goër de Herve

To cite this version:

Sk Imran Hossain, Yann Frendo, Valerie Poux, Isabelle Lebert, Delphine Martineau, et al.. Utilisation de techniques d'intelligence artificielle dans le cadre de l'étude des tiques et de la maladie de Lyme. Journées INRAE systèmes d'information pour les données agro-environnementales, INRAE, Oct 2022, Clermont-Ferrand, France. hal-04087679v2

HAL Id: hal-04087679

https://hal.science/hal-04087679v2

Submitted on 3 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

EMScan, une application mobile pour l'assistance au diagnostic des formes précoces de la maladie de Lyme

Sk Imran Hossain*, Jocelyn de Goër de Herve**,***, Yann Frendo*,**
Delphine Martineau****, Isabelle Lebert**,***, Olivier Lesens****,♯
Engelbert Mephu Nguifo*

*Université Clermont Auvergne, CNRS, ENSMSE, LIMOS,
F-63000 Clermont-Ferrand, France
**Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA,
63122 Saint-Genès-Champanelle, France
***Université de Lyon, INRAE, VetAgro Sup, UMR EPIA,
F-69280 Marcy l'Etoile, France
****Infectious and Tropical Diseases Department, CHU Clermont-Ferrand, France
♯UMR CNRS 6023, Laboratoire Microorganismes: Génome Environnement (LMGE),
Université Clermont Auvergne, Clermont-Ferrand, France

Résumé. La maladie de Lyme est une maladie causée par la bactérie borrelia burgdorferi, transmise à l'homme lors d'une piqûre de tique infectée. Elle touche chaque année de nombreuses personnes en Europe et en Amérique. Le symptôme précoce le plus courant de cette maladie est une éruption cutanée appelée érythème migrant. EMScan est une application mobile d'aide au diagnostic précoce de la maladie de Lyme. EMScan établit un prédiagnostic, à partir de l'analyse d'image de l'éruption cutanée et d'un questionnaire de contexte d'apparition de celle-ci. L'analyse d'image est réalisée par un réseau neuronal convolutif entraîné via des méthodes d'apprentissage par transfert. L'analyse du questionnaire est réalisée par un modèle statistique, développé à partir de l'expertise d'un panel de médecins experts du domaine. L'application EMScan est en cours de certification en tant que Dispositif Médical et devrait être disponible auprès des professionnels de santé et du grand public, durant l'année 2023.

1 Introduction

La maladie de Lyme est une maladie causée par la bactérie borrelia burgdorferi, transmise à l'homme lors d'une piqûre de tique infectée. Au stade précoce, elle se manifeste le plus souvent par des lésions cutanées appelées érythème migrant (EM) (Hossain et al. (2022b). Aussi, un meilleur diagnostic permet généralement, via une antibiothérapie appropriée, de prévenir le passage aux formes tardives plus sévères. Ces dernières années, les techniques d'apprentissage profond ont facilité la création de solutions permettant le développement d'outils d'aide au diagnostic des lésions cutanées. Les réseaux neuronaux convolutifs (CNN), ont notamment été utilisés dans différentes études pour diagnostiquer des lésions cancéreuses et d'autres maladies dermatologie courantes, à partir d'images (Liu et al. (2020)), avec l'ambition de pouvoir

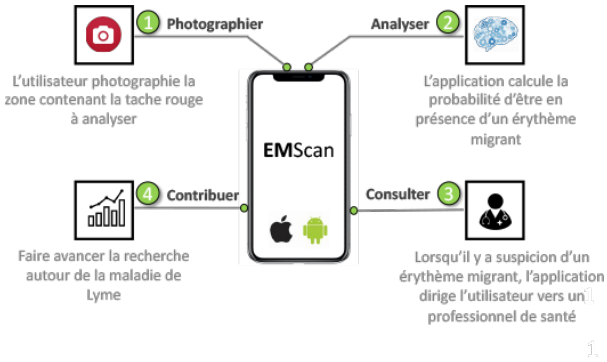

FIG. 1 – *Principe de fonctionnement de l'application EMScan.*

assister les professionnels de santé. Cependant, malgré de nombreux exemples d'utilisation de l'IA dans ce domaine, il n'existe que peu d'études traitant de la détection de la maladie de Lyme basées à partir d'images de lésions cutanées, via l'utilisation de CNN. Ceci peut être attribué au manque de jeux de données d'images d'EM publics, en raison de problèmes de confidentialité liés aux données médicales. Aussi les études présentes dans la littérature utilisent soit un ensemble de données privées, soit un très petit nombre d'images accessibles au public (Burlina et al. (2020); Čuk et al. (2014). Aucun jeu de données étiqueté public contenant des images d'EM n'étant disponible, nous avons dû en constituer un à partir d'images en provenance d'Internet et du Centre Hospitalier Universitaire de Clermont Ferrand et d'autres hôpitaux Français. Elles ont été classées par des dermatologues experts. Ce jeu de données peut être mis à la disposition des chercheurs sous réserve d'un accord de confidentialité. Au cours de ce travail, nous avons étudié de manière approfondie, l'efficacité des CNN pour l'aide au diagnostic des EM. En complément du premier jeu de donnée, nous avons utilisé un jeu de données contenant des lésions cutanées communes, appelé "Human Against Machine with 10000 training photos (HAM10000)", pour entraîner différent CNN, en complément de méthodes d'apprentissage par transfert, à partir de modèles pré-entraînés avec le jeu de données ImageNet (Tschandl et al. (2018)). Les résultats montrent que l'utilisation de HAM10000 apporte un réel bénéfice. Les travaux existants basent leur analyse, uniquement sur des images d'EM, sans prendre en compte les métadonnées associées. Cependant, les médecins estiment que les métadonnées des patients doivent également être prises en compte pour un diagnostic plus précis (Hossain et al. (2022b); Burlina et al. (2020). Cependant, la constitution d'un jeu de données d'images avec les métadonnées associées étant un processus long à réaliser, afin d'apporter une réponse à ce problème, nous avons fait appel à un panel de quinze médecins experts, qui ont tout d'abord été chargés du classement des images puis de définir les critères (avis d'experts) qui ont permis le développement d'une méthode de calcul de score de probabilité, à partir d'un questionnaire. L'application EMScan a été créée dans le but d'apporter une aide au diagnostic précoce de la maladie de Lyme. Elle est destinée à être utilisée par des professionnels de santé et le grand public et a comme objectif secondaire de faire avancer la recherche sur le diagnostic de la maladie de Lyme, via des méthodes d'intelligence artificielle. Pour établir un pré-diagnostic elle se base à la fois sur l'image d'une lésion et les données

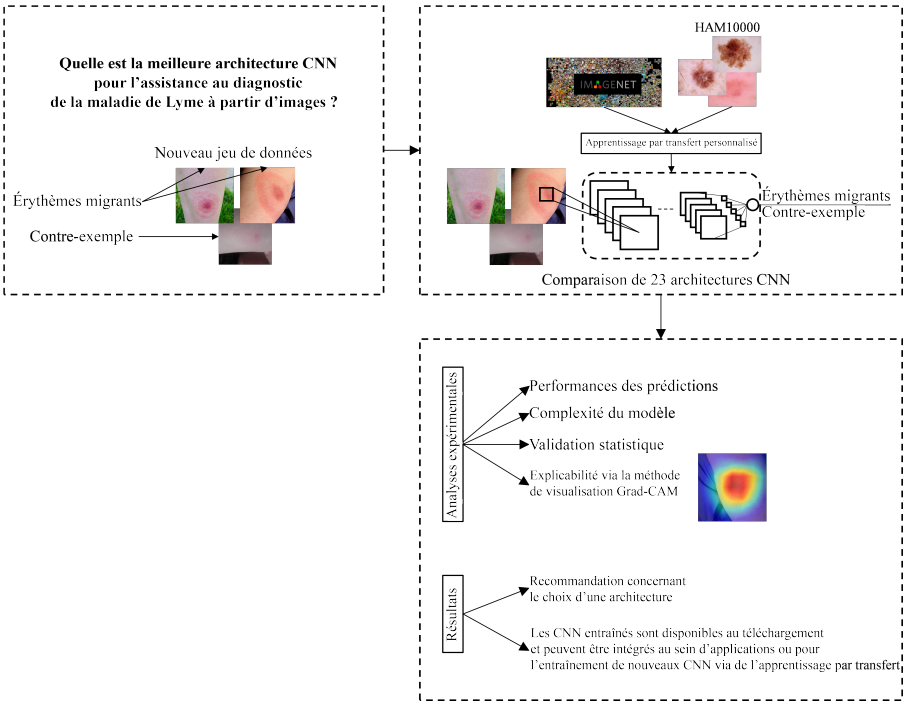

FIG. 2 – *Étude de l'efficacité des CNN pour le diagnostic de la maladie de Lyme à partir d'images (Hossain et al. (2022b)).*

relatives au patient. Le principe de fonctionnement de l'application est présenté figure 1.

2 Matériels and méthodes

Les sous-sections suivantes décrivent la préparation du jeu de données, les phases d'apprentissage des différents modèle CNN, le processus de recueil des avis d'experts et le principe de fonctionnementde l'application EMScan.

2.1 Préparation du jeu de données d'entraînement

Aucun jeu de données contenant des images d'EM labélisées n'étant disponible, nous avons dû en constituer un à partir d'images en provenance d'Internet, du Centre Hospitalier Universitaire de Clermont Ferrand et d'autres hôpitaux Français. Dans un premier temps, les images dupliquées ont été supprimées à l'aide d'un outil basé sur des techniques de hachage perceptuel d'images. Les images inappropriées ont été supprimées par les experts. Au final, 1672 images ont été utilisées durant cette étude. Puis ces images ont été répartis au sein de deux groupes par les experts. Ainsi, 866 photos ont été assignées à la classe « EM », tandis que 806 images ont

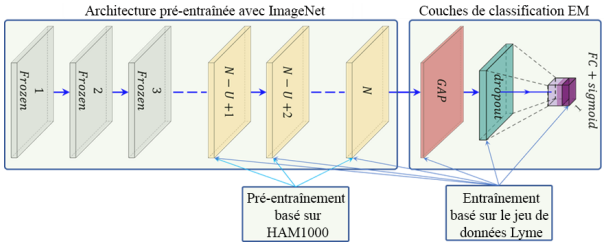

FIG. 3 – *Apprentissage par transfert d'un CNN pour le diagnostic de la maladie de Lyme. U représente le nombre de couches utilisées durant l'étape d'ajustement fin des paramètres et N est le nombre de couches pré-entraînées via ImageNet. GAP est l'abréviation de Global Average Polling (Hossain et al. (2022b)).*

été assignées à la classe « Contre-exemple ». Ce jeu de données a été publié au sein de l'article Hossain et al. (2022b) [1].

2.1.1 Développement de méthodes CNN via un apprentissage par transfert

Nous avons effectué une analyse comparative entre vingt-trois architectures CNN issues de la littérature que nous avons entraîné à partir de notre jeu de données. Puis les résultats obtenus ont été synthétisés en termes de mesures de performance prédictive, de complexité de calcul et de tests de statistique, comme le montre la figure 2. Durant cette étude comparative, nous avons opté pour l'utilisation de méthodes d'apprentissage par transfert, le nombre d'images présentes dans notre jeu de données n'étant pas suffisant pour pouvoir entraîner des CNN à partir de zéro. Ainsi, nous avons utilisé un réseau CNN au préalablement entraîné avec le jeu de données ImageNet, auquel nous avons ensuite ajouté une couche GAP (Global Average Pooling), une couche dropout et une couche entièrement connectée avec une fonction d'activation de type sigmoïde permettant de réaliser une classification binaire, comme le montre la figure 3. Cependant, les étapes d'ajustement fin des paramètres sur toutes les architectures CNN, après l'entraînement de la couche de classification, à partir de notre jeu de données n'ont pas donné de bons résultats face aux méthodes d'ajustement fin partiels des paramètres de plusieurs couches en bout d'architecture CNN, tout en gardant le reste des couches gelées. Aussi, pour chacune des architectures étudiées, nous avons déterminé empiriquement le nombre de couches à affiner à partir du nombre de couches pré-entraînées via ImageNet. D'après les résultats des expériences, le pré-entraînement des couches non gelées à partir du jeu de données HAM10000 a permis d'améliorer les performances de nos architectures CNN. Notre étude a confirmé que les CNN utilisant les méthodes d'apprentissage par transfert personnalisé, décrit ci-dessus, ont un fort potentiel pour les applications d'assistance au diagnostic de la maladie de Lyme. Pour des résultats détaillés, les lecteurs sont invités à consulter notre étude publiée Hossain et al. (2022b). En résumé, un graphique à bulles représentant le taux d'exactitude du modèle en fonction des opérations en virgule flottante (FLOP) est présenté figure 4. Le nombre de paramètres pour chaque modèle est représenté par la taille de chaque bulle. Ce diagramme

1. Les personnes intéressées peuvent contacter l'équipe du projet à l'adresse dappem-project@inrae.fr

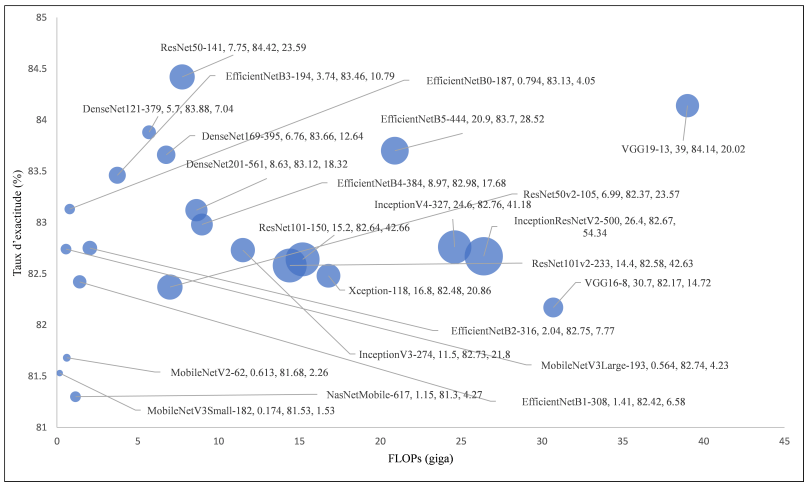

FIG. 4 – *Comparaison des taux d'exactitude des différents modèles CNN entraînés, par rapport aux nombre d'opérations à virgule flottante (FLOPs) nécessaires pour leurs interrogations. Le nombre après le trait d'union du nom du modèle correspond au nombre de couches non gelées durant les phases d'apprentissage par transfert. Les trois valeurs suivantes, représentent respectivement les FLOPs, le taux d'exactitude et les paramètres des modèles. La taille de chaque bulle représente le nombre de paramètres du modèle, mesuré en millions d'unités. (Hossain et al. (2022b)).*

peut être utilisé pour choisir les modèles en fonction de leur complexité et de leur taux d'exactitude. La figure montre que EfficientNetB0-187 est une bonne solution pour être intégré au sein d'une application mobiles ayant des ressources de calcul limitée, avec toutefois un taux d'exactitude raisonnable. De plus, EfficientNetB0-187 a également obtenu de bons résultats durant les tests de visualisation Grad-CAM. (Selvaraju et al. (2017). Enfin, le modèle RestNet50-141 peut aussi être utilisé pour obtenir le meilleur taux d'exactitude si les ressources ne sont pas un critère limitant.

2.2 Utilisation d'avis d'experts en l'absence de métadonnées patients

Étant donné que nous ne disposions pas des métadonnées correspondantes aux patients pour chacune des images de notre jeu de données, nous avons fait appel à des avis d'experts. Les questionnaires ainsi que l'attribution de poids pour chacune des questions, ont été préparés en recueillant les données de quinze dermatologues experts. Étant donné qu'il est difficile pour les médecins de fournir des estimations des paramètres d'une distribution de probabilité ou des estimations de probabilité pour un grand nombre de cas possibles, des valeurs de pondération ont été attribué aux différents symptômes dans un intervalle de -1 à +3 (la valeur plus élevée représentant une plus grande contribution du symptôme à la possibilité de la maladie). Nous avons résumé chacun des cas sous la forme la somme des poids moyens, avec une normali-

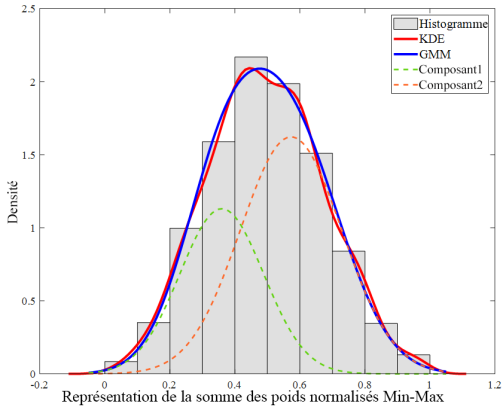

FIG. 5 – *Représentation des trois approches utilisées pour l'analyse des données d'experts. GMM et KDE signifient respectivement modèle de mélange gaussien et estimation par noyau (Hossain et al. (2022a).*

sation min-max. Sur la base de tous les cas possibles, nous avons proposé aux experts, trois approches pour convertir la somme des poids en un score de probabilité :

1. Probabilité cumulée des sommes des poids normalisés basée sur une estimation par noyau ;
2. Probabilité cumulée d'une somme de poids normalisée basée sur l'estimation de la densité d'un modèle de mélange gaussien à deux composantes ;
3. Probabilité à posteriori d'une somme de poids normalisée par rapport à la deuxième composante du mélange gaussien (en supposant que la deuxième composante représente la sous-population atteinte).

Ces différentes approches sont présentées dans la figure 5. Les experts ont opté pour la seconde approche. Le modèle obtenu a été validé en utilisant un arbre de décision et un approche d'analyse formelle des concepts. Nous avons également proposé un moyen de combiner le score de probabilité du classificateur d'images CNN avec la probabilité des données des patients, en veillant à ce que la probabilité basée sur les données des patients ait une influence majeure sur le score final. Les détails de cette méthode sont disponibles au sein de l'article Hossain et al. (2022a).

2.3 Principales étapes d'utilisation de l'application EMScan

La figure 6 La figure montre le principe de fonctionnement de l'application mobile EMScan. Tout d'abord, l'utilisateur prend une photo de la lésion cutanée à l'aide de l'appareil photo du mobile. Ensuite, l'application détecte et recadre automatiquement la lésion cutanée, qui peut également être ajustée manuellement par l'utilisateur. Troisièmement, les données du patient relatives à la lésion cutanée sont recueillies au travers d'un questionnaire comportant 15 questions. Quatrièmement, l'image de la lésion est analysée par le CNN et les données du

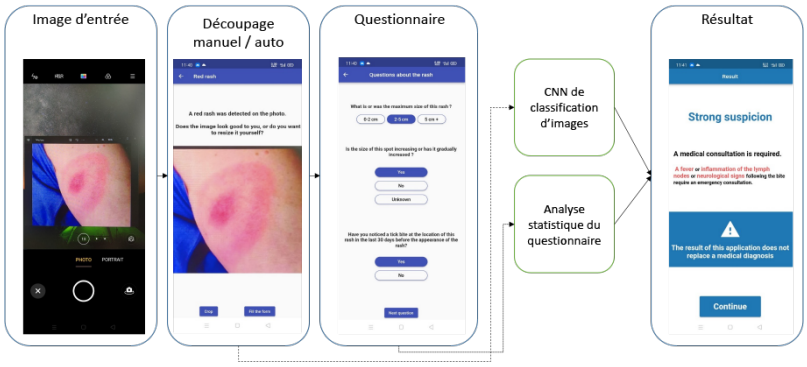

FIG. 6 – *Enchaînement des principales étapes d'analyse de l'application EMScan.*

patient sont analysées par le modèle statistique. Enfin, une estimation de la présence de la maladie avec des recommandations sont fournies à l'utilisateur, sur la base de l'analyse de l'image de la lésion cutanée et des données du patient. L'application fournit également à l'utilisateur des informations sur la maladie de Lyme. Enfin, les données annonymisées sont collectées à des fins de recherche, avec le consentement de l'utilisateur.

3 Conclusion

Nous avons développé une application mobile d'aide au diagnostic précoce de la maladie de Lyme. Le principal défi scientifique a été de faire face à la faible quantité de données disponibles. Cependant, ces difficultés ont pu être surmontées grâce à une méthode personnalisée d'apprentissage par transfert et au développement d'une méthode basée sur des avis d'experts. L'application EMScan est en cours de labélisation en tant que Dispositif Médical et sera par la suite mise à la disposition du grand public.

Utilisation et mise à disposition des données

Les données utilisées dans le cadre de ce projet de recherche sont disponibles en libre accès sur le site du projet DAPPEM (Développement d'une APPlication d'identification des Erythèmes Migrants à partir de photographies) [2].

Remerciements

Ce projet de recherche a été financé par le Fonds Européen de Développement Régional, projet DAPPEM-AV0021029. Le projet DAPPEM est coordonné par Pr. Olivier Lesens et réalisé dans le cadre de l'appel à propositions "Pack Ambition Researche" de la Région Auvergne-

2. https ://dappem.limos.fr

Rhône-Alpes, France. Ce travail a également été partiellement financé par la Mutualité Sociale Agricole (MSA), France.

Références

Burlina, P. M., N. J. Joshi, P. A. Mathew, W. Paul, A. W. Rebman, et J. N. Aucott (2020). Ai-based detection of erythema migrans and disambiguation against other skin lesions. *Computers in biology and medicine 125*, 103977.

Čuk, E., M. Gams, M. Možek, F. Strle, V. M. Čarman, et J. F. Tasič (2014). Supervised visual system for recognition of erythema migrans, an early skin manifestation of lyme borreliosis. *Strojniški vestnik-Journal of Mechanical Engineering 60*(2), 115–123.

Hossain, S. I., J. d. G. de Herve, D. Abrial, R. Emillion, I. Lebert, Y. Frendo, D. Martineau, O. Lesens, et E. M. Nguifo (2022a). Expert opinion elicitation for assisting deep learning based lyme disease classifier with patient data. *arXiv preprint arXiv :2208.14384 1*, 1.

Hossain, S. I., J. d. G. de Herve, M. S. Hassan, D. Martineau, E. Petrosyan, V. Corbin, J. Beytout, I. Lebert, J. Durand, I. Carravieri, et al. (2022b). Exploring convolutional neural networks with transfer learning for diagnosing lyme disease from skin lesion images. *Computer Methods and Programs in Biomedicine 215*, 106624.

Liu, Y., A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, et al. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature medicine 26*(6), 900–908.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, et D. Batra (2017). Grad-cam : Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Tschandl, P., C. Rosendahl, et H. Kittler (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data 5*(1), 1–9.

Summary

Lyme disease a chronic disease caused by a pathogen transmitted by infected ticks affects a lot of people in Europe and America. The most common early symptom of Lyme disease is a red skin rash called erythema migrans. EMScan is a mobile application to assist with early diagnosis of Lyme disease. EMScan does the diagnosis based on image analysis of the rash and also considers patient data in the context of the appearance of the red rash. The Image analysis is powered by a deep convolutional neural network utilizing custom transfer learning. The image analysis is assisted by another probability score calculated from related patient data based on a statistical model created by eliciting opinions from a panel of expert doctors. The EMScan application is in the process of being certified as a Medical Device and should be available to healthcare professionals and the general public during the year 2023.