People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Setif 1 University Ferhat ABBAS

Master's Final Thesis submitted in partial fulfillment for the award of the Master's Degree in Computer Science
Academic program in Computer Science
Field: Computer Science
Speciality: Foundations and Engineering of Information and Imaging

Explainable Artificial Intelligence (XAI) for Enhancing Healthcare Diagnosis

Supervised by:
Dr. DRIF Ahlem
Associate Professor
Faculty of Sciences

Presented by:
Maaoui Anfel
Dib Amira

Academic year
2025–2026

Contents

1 Explainable AI in Healthcare: Fundamental Theories and Literature Review . . 15
1.1 Introduction 15
1.2 Making AI understandable to end users 15
1.2.1 From "Black Box" to "(Translucent) Glass Box" 15
1.2.2 Explainability: Transparent or Post-Hoc 15
1.2.3 Collaboration between Humans and AI 16
1.3 Why XAI is crucial for healthcare diagnosis 17
1.4 What is Easily Interpretable 18
1.5 Interpretability Metrics 18
1.6 Explainability vs. Interpretability 19
1.7 Model Transparency: White Box vs. Black Box 19
1.8 Classification of Explainable AI Models 20
1.8.1 Interpretation Types 21
1.8.2 Explanation Scopes 21
1.8.3 Model Specificity 22
1.8.4 Explanation Forms 23
1.9 XAI Models 24
1.9.1 Intrinsic Explainability 24
1.9.2 Post-hoc Explainability 24
1.10 Overview of Explainable AI Methods 26
1.10.1 Saliency 26
1.10.2 Class Activation Mapping 26
1.10.3 Occlusion Sensitivity 27
1.10.4 Testing with Concept Activation Vectors 27
1.10.5 Triplet Networks 28

1.10.6 Prototypes 28
1.10.7 Trainable Attention 29
1.10.8 Shapley Additive Explanations 29
1.10.9 Local Interpretable Model-Agnostic Explanations 30
1.10.10 Image Captioning 31
1.10.11 Recent XAI Methods 31
1.11 Clinical Applications of XAI 33
1.11.1 Radiology and Radiomics 33
1.11.2 Oncology 34
1.11.3 Cardiology 35
1.11.4 Other Pathology 36
1.11.5 Benefits of XAI in Clinical Practice 37
1.12 Ethical, Legal and Regulatory Considerations 38
1.12.1 Ethical Principles 38
1.13 Conclusion 40

2 A Hybrid Explainable ResNet50 Feature Extractors with Transformer Encoder for Pneumonia Diagnosis from Chest X-ray Images 42
2.1 Introduction 42
2.2 Methodology 43
2.2.1 Overview of Pneumoni 43
2.3 Data Description 45
2.4 Data Preprocessing and Augmentation 46
2.4.1 Data Preprocessing 47
2.4.2 Data Augmentation 47
2.4.3 Data Balancing for Imbalanced Classes 48
2.5 Baseline Architecture 49
2.5.1 Convolutional Neural Networks for Pneumonia Detection 49
2.5.2 ResNet50 50
2.5.3 Squeeze-and-Excitation Networks 50
2.5.4 Vision Transformers (ViT): Training Strategies 51
2.5.4.1 Training ViT from Scratch 51
2.5.4.2 Pretrained ViT with Fine-Tuning 51
2.5.5 Our Contribution (Hybrid ResNet50+ViT Model) 52
2.6 Conclusion 54

3 **Experiments and Results** 56
3.1 Introduction 56
3.2 Experimental Setup 56
3.2.1 Environment and Tools 57
3.2.1.1 Hardware Environment Settings 57
3.2.1.2 Software Environment Settings 57
3.3 Hyperparameters Settings 58
3.4 Baseline Performance 60
3.5 Our Model Results 61
3.5.1 Performance Metrics 62
3.5.2 Confusion Matrix 62
3.5.3 Training and Validation Curves 63
3.5.4 Discussion of Results 63
3.6 Comparative Analysis 64
3.6.1 Visualizing the Performance Differences 64
3.6.2 Discussion of Results and Medical Importance 65
3.7 Explainability and Feature Visualization Results 65
3.7.1 Feature Map Visualization 66
3.7.2 Classification Head Analysis 68
3.7.3 Grad-CAM Visual Explanations 68
3.8 Conclusion 70
Bibliography 72

Abstract AN

The integration of Deep Learning into clinical workflows improves automated disease detection, but the opaque “black-box” nature of these models limits medical trust. This thesis addresses this challenge by proposing an explainable diagnostic framework for pneumonia detection using chest X-rays. We develop a hybrid architecture combining a ResNet50 backbone for local feature extraction with a Vision Transformer (ViT) encoder to capture global contextual dependencies across lung regions. To ensure transparency, the framework integrates post-hoc Explainable AI (XAI) techniques, specifically Grad-CAM heatmaps and feature map visualizations. Implemented via PyTorch, our hybrid model achieves high diagnostic accuracy while ensuring predictions are based on true pulmonary opacities rather than artifacts. This work bridges the gap between high predictive performance and clinical interpretability.

Keywords: Deep Learning, Explainable AI (XAI), ResNet50, Vision Transformer (ViT), Pneumonia Detection, Grad-CAM.

Abstract FN

L'intégration de l'apprentissage profond (Deep Learning) dans les flux de travail cliniques améliore la détection automatisée des maladies. Cependant, la nature opaque de ces modèles, souvent considérés comme des « boîtes noires », limite la confiance des professionnels de santé dans leurs prédictions. Ce mémoire vise à répondre à ce défi en proposant un cadre diagnostique explicable pour la détection de la pneumonie à partir d'images radiographiques thoraciques (Chest X-rays).

Nous développons une architecture hybride combinant un réseau ResNet50 pour l'extraction des caractéristiques locales et un encodeur Vision Transformer (ViT) afin de capturer les dépendances contextuelles globales entre les différentes régions pulmonaires. Afin d'assurer la transparence du processus de décision, le cadre proposé intègre des techniques d'Intelligence Artificielle Explicable (XAI), notamment les cartes thermiques Grad-CAM et la visualisation des cartes de caractéristiques (Feature Maps).

Implémenté à l'aide de PyTorch, notre modèle hybride atteint une haute précision diagnostique tout en garantissant que les prédictions reposent sur de véritables opacités pulmonaires plutôt que sur des artefacts ou des informations non pertinentes. Ce travail contribue ainsi à réduire l'écart entre les performances prédictives élevées et l'interprétabilité clinique.

- **Mots-clés** : Apprentissage profond, Intelligence Artificielle Explicable (XAI), ResNet50, Vision Transformer (ViT), Détection de la pneumonie, Grad-CAM.

Abstract AR

يُسهم دمج التعلّم العميق (Deep Learning) في سير العمل السريري في تحسين الكشف الآلي عن الأمراض. ومع ذلك، فإن الطبيعة غير الشفافة لهذه النماذج، التي غالبًا ما تُعتبر «صناديق سوداء»، تحدّ من ثقة المتخصصين في الرعاية الصحية بنتائجها وتنبؤاتها. يهدف هذا البحث إلى معالجة هذا التحدي من خلال اقتراح إطار تشخيصي قابل للتفسير للكشف عن الالتهاب الرئوي اعتمادًا على صور الأشعة السينية للصدر .نقوم بتطوير بنية هجينة تجمع بين شبكة ResNet50لاستخراج السمات المحلية، ومُرمّز Vision Transformer (ViT)لالتقاط العلاقات السياقية الشاملة بين مختلف مناطق الرئتين. ولضمان شفافية عملية اتخاذ القرار، يدمج الإطار المقترح تقنيات الذكاء الاصطناعي القابل للتفسير(XAI) ، ولا سيما خرائط الحرارة Grad-CAMوتصوير خرائط السمات Feature). .(Mapsتم تنفيذ النموذج المقترح باستخدام مكتبة PyTorch، وقد حقق دقة تشخيصية عالية مع ضمان استناد التنبؤات إلى عتامات رئوية حقيقية بدلًا من الاعتماد على الشوائب أو المعلومات غير ذات الصلة. ويسهم هذا العمل في تقليص الفجوة بين الأداء التنبؤي المرتفع وقابلية التفسير السريري.

الكلمات المفتاحية :التعلّم العميق، الذكاء الاصطناعي القابل للتفسير(XAI) ، ResNet50، محول الرؤية (ViT)، الكشف عن الالتهاب الرئوي، .Grad-CAM.