

# SmartWoodID

## Smart classification of Congolese timbers: deep learning techniques for enforcing forest conservation

Contract - B2 / 202 / P2 / SmartWoodID

### RÉSUMÉ

#### Contexte

Une part importante du commerce du bois demeure illégale, et l'exploitation forestière illégale constitue le crime le plus lucratif contre la biodiversité. ONU Environnement estime que l'exploitation forestière illégale et le commerce du bois qui y est associé représentent entre 50 et 152 milliards de dollars par an. L'exploitation forestière illégale comporte un risque élevé de dommages irréversibles aux écosystèmes liés à l'exploitation d'espèces très recherchées, parfois protégées. Des réglementations sur le bois sont déjà en vigueur (CITES, FLEGT, RBUE), mais leur mise en œuvre et leur application constituent un défi. Actuellement, la Belgique est perçue négativement comme une « plaque tournante du commerce illégal du bois ». 27,5 % des importations totales de produits de bois tropicaux primaires de l'UE-28 transitent par la Belgique (principalement via le port d'Anvers). L'identification du bois est un processus clé de l'application de la loi, qui doit vérifier si l'expédition correspond aux produits mentionnés sur les documents d'accompagnement. C'est pourquoi il existe une demande croissante d'outils d'identification du bois utilisables par les forces de l'ordre (e.g. les douanes).

Le xylarium de Tervuren est la collection gouvernementale belge d'échantillons de bois. Il s'agit d'un élément du patrimoine scientifique fédéral de renommée internationale, hébergé par le Musée royal de l'Afrique centrale, qui comprend du matériel de référence sur 13 000 espèces botaniques différentes. L'une des fonctions actuelles croissantes de la collection est de soutenir la recherche médico-légale par la vérification de l'identité d'une espèce. La technique la plus courante d'identification du bois est l'évaluation anatomique du bois. Les méthodes d'apprentissage automatique devraient faciliter le processus d'identification du bois pour les non-spécialistes. Les essences de bois présentent en effet des caractéristiques à différents grossissements microscopiques. Cependant, certaines de ces caractéristiques sont très variables, ce qui entrave le développement de clés d'identification dichotomiques classiques utilisables par les non-spécialistes. De plus, de nombreuses caractéristiques observées à la surface du bois doivent être considérées comme des artefacts (fissures, traces de dommages mécaniques, attaques de champignons et d'insectes) et ne sont pas toujours faciles à distinguer des caractéristiques diagnostiques pour un œil non averti. Le xylarium de Tervuren offre l'assemblage le plus complet de matériel de référence pour le développement de nouvelles approches d'identification du bois.

Le projet visait à automatiser une partie du processus d'identification du bois en appliquant des techniques d'intelligence artificielle à l'analyse d'images anatomiques d'essences de bois de la République démocratique du Congo. La flore arborée d'Afrique centrale comprend 3 013 espèces, dont 27 appartiennent à la classe 1 des essences commerciales de la RDC et sont actuellement exploitées et commercialisées de manière intensive, 20 à la classe 2 (ayant une valeur commerciale potentielle importante), 44 à la classe 3 (considérées comme valorisées) et 879 à la classe 4 (valeur commerciale encore inconnue). Le projet a utilisé des échantillons de xylarium de toutes les espèces des quatre classes et a tiré parti de la puissance des approches modernes d'apprentissage profond. Le projet s'est appuyé sur des descriptions anatomiques du bois réalisées par des experts, qui ont servi de données d'apprentissage annotées pour développer le logiciel. Ce projet était unique en raison du grand nombre d'essences africaines, de l'application de l'apprentissage profond et d'une base de données de descriptions standardisées disponibles en ligne. Dans un premier lot de travaux, des annotations d'experts sur des images microscopiques et mésoscopiques de surfaces transversales de 1 000 essences de bois congolaises ont été réalisées. Dans le cadre du lot de travail 2, nous avons développé un pipeline de traitement d'images pour l'annotation semi-automatisée de coupes de bois microscopiques et mésoscopiques. Dans le lot de travail 3, nous nous sommes concentrés sur la création d'une interface utilisateur intuitive.

## Objectifs

Le projet SmartwoodID vise à améliorer le processus d'analyse des essences de bois de la République démocratique du Congo (RDC), un vaste pays riche en essences où l'identification est une tâche complexe, entraînant régulièrement des infractions à la loi. L'ambition générale est d'explorer les possibilités d'automatisation de l'identification des essences de bois par l'apprentissage automatique, en utilisant de vastes ensembles de données de caractéristiques macroscopiques et des scans optiques haute résolution des surfaces de bois de bout.

Le projet avait trois objectifs de recherche :

- **Objectif de recherche 1** : Construction d'une base de données de scans à plat des surfaces de bois de bout et des caractéristiques macroscopiques des essences de bois de la République démocratique du Congo.
- **Objectif de recherche 2** : Développement de clés de classification illustrées pour l'identification du bois du Congo.
- **Objectif de recherche 3** : Développement d'une clé de classification illustrée assistée par l'apprentissage automatique.

## Conclusions

La base de données d'images SmartWoodID offre de nouvelles possibilités pour le développement de systèmes d'identification basés sur la reconnaissance des caractéristiques anatomiques diagnostiques du bois. Cette base de données est unique car elle couvre un grand nombre d'espèces d'arbres et de taxons inférieurs africains, dont la structure macroscopique est visualisée et décrite. La Collection de Bois de Tervuren fournit ces données grâce à son héritage de plus d'un siècle de collecte de matériel de référence fiable. Au total, 56 % de toutes les espèces d'arbres et de taxons inférieurs de RDC sont actuellement disponibles dans la Collection de Bois de Tervuren. La première version de la base de données d'images SmartWoodID présentée ici comprend un ensemble de 954 essences de bois et taxons inférieurs présents dans les forêts de RDC. La base de données se concentre sur les caractéristiques anatomiques macroscopiques observables sur un scan haute résolution de la surface du bois de bout. La base de données prend en compte les irrégularités et la variabilité naturelle, en utilisant plusieurs spécimens présentant de grandes surfaces de bois de bout. Cela en fait une base de données de référence robuste pour la recherche sur le bois en général et permettra le développement d'outils pour aider les forces de l'ordre à lutter contre l'exploitation forestière illégale.

L'étude des clés d'identification (WP2) met en évidence les limites inhérentes aux 31 caractéristiques transversales macroscopiques accessibles et définies par des experts pour l'identification taxonomique d'une grande diversité d'essences de bois, telles que celles du bassin du Congo. Si la précision de la classification s'améliore aux rangs taxonomiques supérieurs, les prédictions à l'échelle du genre et de la famille restent limitées en raison du chevauchement des caractéristiques anatomiques entre les taxons. Néanmoins, les caractéristiques transversales macroscopiques conservent une valeur diagnostique lorsqu'elles sont appliquées à des périmètres taxonomiques plus restreints. La discrimination réussie des espèces de *Pterocarpus*, autrefois considérées comme indiscernables sans méthodes de laboratoire, démontre que des caractéristiques anatomiques facilement observables peuvent permettre une identification à l'échelle de l'espèce sur le terrain. L'intégration de l'imagerie haute résolution à grande surface et de jeux de données multi-spécimens permet d'améliorer encore la précision diagnostique. Ces approches capturent plus efficacement la variabilité anatomique intraspécifique que les méthodes conventionnelles à échantillon unique et permettent l'extraction d'informations anatomiques quantitatives à une échelle plus fine. L'intégration de ces améliorations offre des pistes prometteuses pour accroître la résolution taxonomique et la fiabilité de la classification. Cependant, les contraintes pratiques sur le terrain, telles que les contraintes de temps et les équipements limités, nécessitent des stratégies alternatives pour une identification fiable du bois in situ. La poursuite des progrès dépendra de l'évolution des systèmes d'identification basés sur les CV, notamment les CNN, capables de traiter directement les images macroscopiques pour une classification précise et rapide. Ces modèles offrent un fort potentiel pour des applications de vérification du bois évolutives, efficaces et prêtes à l'emploi.

Les résultats du WP3 montrent que les CNN, comme l'architecture Xception appliquée, peuvent extraire avec succès des caractéristiques permettant de classer des fragments d'images de coupes transversales poncées afin de classer différents bois au niveau du genre et de distinguer le bois sans anomalie du bois anormal. Les performances sur les données de test variaient selon les genres, certains bénéficiant d'un entraînement sur des images sans anomalie, tandis que pour

d'autres genres, comme *Cynometra*, un rappel plus élevé a été observé pour le modèle entraîné sur des images anormales. L'analyse Grad-CAM a révélé la préférence du modèle pour les régions des fragments présentant des tissus anatomiques du bois non masqués, soulignant l'importance d'une anatomie claire du bois dans l'entraînement des CNN pour l'identification du bois. Cela pourrait permettre aux CNN de capturer plus efficacement les schémas diagnostiques, ce qui conduirait à une meilleure discrimination entre les bois, même lorsqu'ils sont appliqués à des spécimens anormaux sur le terrain. L'inclusion de fragments anormaux a eu un impact limité, mais a légèrement amélioré les performances sur ces fragments. Les résultats suggèrent donc que les CNN (comme Xception) affichent les meilleures performances en matière de classification du bois lorsqu'ils sont entraînés sur des images exemptes d'anomalies, ce qui rend cette approche très efficace pour développer des modèles d'identification du bois basés sur le CV, destinés à être déployés sur le terrain. Cela démontre le potentiel de l'apprentissage profond pour l'identification automatisée des genres de bois. Les résultats soulignent que, si tous les modèles de classification capturent des schémas sous-jacents similaires, les CNN surpassent les clés d'identification sur les 31 caractéristiques macroscopiques transversales IAWA grâce à leur capacité à extraire des caractéristiques d'image discriminantes sans recourir à des descripteurs prédéfinis. Les tendances de performance selon les différentes portées taxonomiques soulignent l'importance de la diversité des données d'entraînement, car les modèles entraînés sur des ensembles de données plus larges présentent de meilleures capacités de généralisation que ceux entraînés exclusivement sur des bois commerciaux.

Au-delà des CNN de classification multiclasse standard, les approches CNN de ré-identification d'objets offrent des alternatives intéressantes, notamment dans les contextes médico-légaux où l'identification peut être moins critique que l'exclusion de certains bois. L'approche de vérification binaire affiche d'excellentes performances à cet égard, bien que son efficacité soit limitée par les limitations de classement et les exigences de calcul. La ré-identification par intégration, bien qu'efficace sur le plan informatique, est moins performante dans cette étude, ce qui suggère que des stratégies d'exploration et des fonctions de perte améliorées pourraient améliorer sa fiabilité. De plus, la ré-identification d'objets produit des informations sur la similarité avec des spécimens de référence spécifiques, plutôt que de produire une prédiction directe des classes (par exemple, des genres dans cette étude), fournissant ainsi des informations précieuses aux chercheurs en criminalistique.

Les résultats obtenus avec une clé intégrée utilisant le CV et les 31 caractéristiques définies par des experts démontrent que, si l'intégration de caractéristiques anatomiques macroscopiques définies par des experts peut apporter des améliorations modérées aux prédictions de genre basées sur les CNN, ces avantages dépendent fortement des genres spécifiques et de la profondeur du reclassement appliqué. Les résultats globaux confirment que les modèles CNN encodent déjà à eux seuls des informations taxonomiques substantielles, probablement grâce à leur entraînement sur des comparaisons diagnostiques complexes qui vont au-delà des descripteurs anatomiques traditionnels. Il est crucial de noter que le reclassement parmi les deux à cinq premières prédictions CNN offre les gains de performance les plus constants en termes d'exactitude, de précision et de rappel, en particulier au seuil des trois premières. Au-delà, les performances diminuent en raison d'erreurs de classification introduites par une sur-accentuation des caractéristiques anatomiques faibles ou trompeuses. Ce résultat souligne l'utilité limitée, mais

stratégique, de l'anatomie macroscopique transversale du bois pour affiner les identifications. Du point de vue de l'application sur le terrain, notamment dans le contexte de la vérification du bois de première ligne, les implications sont doubles. Premièrement, le modèle CNN offre une méthode rapide et accessible d'identification au niveau du genre, déjà performante dans la plupart des cas. Deuxièmement, les méthodes d'affinement, telles que le reclassement, doivent être appliquées de manière sélective, car une utilisation indiscriminée, notamment sur les taxons protégés, peut réduire le rappel, augmentant ainsi le risque de négliger des bois prioritaires comme le Khaya.

Les recherches futures devraient s'attaquer aux limites des modèles basés sur l'intégration en utilisant les technologies de pointe actuelles afin de proposer une approche performante pour l'identification automatisée du bois. Ces modèles devraient être explorés dans des études hybrides intégrant plusieurs modalités de données diagnostiques (images macroscopiques, anatomie microscopique du bois et empreintes chimiques) afin d'améliorer la précision de la classification et de permettre une identification avec une résolution taxonomique plus précise. Alors que les efforts mondiaux pour lutter contre l'exploitation forestière illégale et faire respecter les réglementations en matière de commerce durable s'intensifient, l'identification du bois par l'IA sera essentielle pour renforcer les capacités d'analyse forensique et garantir une gestion responsable des ressources.

## Mots-clés

- Anatomie du bois
- Identification du bois
- Évaluation anatomique macroscopique du bois
- Coupe transversale
- Identification du bois par vision artificielle
- Object re-identification
- Gradient-weighted class activation mapping