

Volume de données nécessaires pour entraîner un modèle IA de diagnostic de la thyroïde

Ce document présente les estimations du nombre de patients ou d'exemples nécessaires pour entraîner différents modèles d'intelligence artificielle appliqués au diagnostic des pathologies thyroïdiennes.

1. Modèle Machine Learning (données biologiques et cliniques)

- Données requises : résultats de TSH, T3, T4, symptômes, sexe, âge, antécédents.
- Nombre minimum recommandé : 300 à 500 patients.
- Nombre idéal : 1000+ patients.
- Ces modèles sont robustes même avec des jeux de données modérés, à condition que les variables soient informatives et bien structurées.

2. Modèle Deep Learning (imagerie échographique)

- Données requises : images échographiques annotées (nodule, goitre, normal, etc.).
- Nombre minimum recommandé : 1000 images.
- Nombre idéal : 5000+ images.
- Pour démarrer plus vite, on peut utiliser des techniques de transfert d'apprentissage avec des modèles préentraînés.

3. Modèle combiné (biologie + clinique + image)

- Données requises : chaque patient doit avoir des données biologiques, cliniques et une échographie annotée.
- Nombre minimum recommandé : 500 patients complets.
- Nombre idéal : 2000+ patients.
- Cette approche permet une modélisation plus réaliste et complète du raisonnement médical, mais nécessite plus de données alignées.

4. Recommandation pour Djibouti

Étapes proposées :

1. Commencer par un modèle simple basé sur des données tabulaires : viser la collecte de 500 dossiers complets.
 2. Lancer un projet pilote de collecte d'images échographiques annotées : objectif 1000+ images.
 3. Regrouper les cas complets (bio + imagerie) pour préparer un modèle combiné.
- Un plan de collecte progressif et réaliste permettra de construire une base solide pour le développement d'un outil intelligent fiable et acceptable par les praticiens.