OncoRisk AI

Une Solution Innovante d'IA pour la Prévention et l'Évaluation des Risques de Cancer

Équipe de Développement : Ikbel Hamdi, Ahmed Trabelsi, Maram Rachdi, Samar Omrani, Ahmed Fekih, Malek Hammemi

Supervisée par : Mme Oumayma Guasmi

16 Décembre 2025

Résumé Exécutif

Ce rapport technique détaille le développement de OncoRisk AI, une plateforme avancée d'intelligence artificielle conçue pour l'évaluation des risques oncologiques. Adoptant la méthodologie CRISP-DM comme cadre structurant, ce document expose de manière exhaustive chaque phase du cycle de vie du projet, depuis la compréhension des enjeux métier jusqu'au déploiement opérationnel. Développée dans un contexte académique, cette solution intègre l'analyse multifactorielle des risques, la prédiction de types de cancer, et l'interprétation d'images médicales, tout en incorporant des fonctionnalités interactives telles qu'un chatbot conversationnel et la génération de rapports PDF personnalisés. Les performances obtenues, avec une précision atteignant 95% pour certains modèles, soulignent le potentiel transformateur de l'IA en matière de santé publique. Ce projet, purement éducatif, vise à stimuler la recherche en IA appliquée à l'oncologie, sans prétendre à un usage clinique direct.

Contents

List of Figures

List of Tables

Introduction Générale

Le cancer constitue l'une des pathologies les plus prévalentes et mortelles à l'échelle globale, avec plus de 19 millions de cas diagnostiqués annuellement selon les données de l'Organisation Mondiale de la Santé (OMS) en 2024. La détection précoce et l'évaluation proactive des facteurs de risque sont impératives pour accroître les taux de survie, qui peuvent dépasser 90% en cas de diagnostic au stade initial. Néanmoins, les approches conventionnelles de dépistage souffrent de limitations significatives : coûts élevés, invasivité, et taux de faux positifs pouvant atteindre 50% dans des examens comme la mammographie. Dans ce paysage, l'intelligence artificielle émerge comme un levier disruptif, offrant des analyses précises, scalables et économiques. OncoRisk AI, plateforme web développée dans le cadre d'un projet académique, exploite des algorithmes d'apprentissage automatique pour évaluer les risques basés sur des facteurs personnels et des imageries médicales. S'appuyant sur des ensembles de données publics tels que IQ-OTH/NCCD pour le poumon, CBIS-DDSM pour le sein, et ISIC pour la peau, ainsi qu'un dataset synthétique de facteurs de risque (`cancer-risk-factors.csv`), cette solution démontre l'application pratique de l'IA en oncologie préventive. Ce rapport est rigoureusement structuré autour de la méthodologie CRISP-DM (Cross-Industry Standard Process for Data Mining), standard industriel pour les initiatives de fouille de données. Il couvre les six phases itératives : Compréhension du Métier, Compréhension des Données, Préparation des Données, Modélisation, Évaluation, et Déploiement. L'objectif est de présenter une démarche méthodique, professionnelle et innovante, intégrant des technologies de pointe comme Streamlit pour l'interface utilisateur, TensorFlow/Keras pour les modèles convolutifs, et Scikit-learn pour les classifications supervisées.

Phase 1 : Compréhension du Métier

2.1 Objectifs Stratégiques et Opérationnels

Les objectifs de OncoRisk AI sont définis comme suit :

- Prédire avec précision le type de cancer potentiel (sein, côlon, poumon, prostate, peau) à partir d'un ensemble de facteurs de risque.
- Évaluer le niveau de risque global et générer des recommandations personnalisées basées sur des règles expertes.
- Analyser des imageries médicales (CT scans, mammographies, photographies dermatologiques) pour détecter des anomalies potentielles.
- Fournir une interface utilisateur intuitive et éducative, accompagnée d'avertissements explicites sur le caractère non-diagnostique de l'outil.

2.2 Contexte Métier et Analyse des Besoins

Dans le secteur de la santé publique, les défis incluent non seulement les taux élevés de faux positifs, mais également l'accès inégal aux dépistages avancés. Des études de référence, telles que celle d'Agarap (2018) sur le dataset WDBC démontrant une précision de 99% via des réseaux neuronaux multicouches, valident le potentiel de l'IA. OncoRisk AI vise à atténuer ces enjeux en favorisant une prévention proactive, dans un marché projeté à 996 millions USD d'ici 2030 (selon des rapports de marché spécialisés).

2.3 Exigences Fonctionnelles et Non-Fonctionnelles

- **Technologiques :** Implémentation en Python 3.12, avec dépendances spécifiées dans `requirements.txt` (Streamlit 1.51.0, TensorFlow 2.20.0, etc.).
- **Éthiques et Légales :** Intégration d'avertissements proéminents soulignant l'usage académique exclusif ; conformité aux principes de confidentialité des données.
- **Contraintes :** Développement sur ressources locales ; déploiement open-source via GitHub (`https://github.com/AhmedTrabelsy/Breast-Cancer-Detection`).

Phase 2 : Compréhension des Données

3.1 Acquisition et Sources des Données

Les corpus de données mobilisés comprennent :

- `cancer-risk-factors.csv` : 2000 observations avec 21 attributs (ex. : Âge, Tabagisme, Type de Cancer).
- Ensembles d'imageries : IQ-OTH/NCCD (poumon), CBIS-DDSM (sein), ISIC (peau), totalisant des milliers d'images annotées.

3.2 Description et Qualité des Données

Le dataset principal présente des variables numériques (ex. : IMC, plage 10-60) et catégorielles (ex. : Genre, binaire 0/1). L'analyse exploratoire, menée via `Pretraitement.ipynb`, révèle :

- Absence de doublons ou de valeurs manquantes.
- Distribution statistique : Âge moyen de 55 ans ; 50% des observations avec tabagisme élevé (score >=7).
- Qualité globale : Données synthétiques équilibrées, sans biais majeurs détectés.

3.3 Exploration Analytique et Visualisations

Des corrélations significatives sont observées, notamment entre le tabagisme et le cancer du poumon (coefficient de Pearson >0.7). L'Analyse en Composantes Principales (ACP) capture 60% de la variance avec les deux premières composantes, comme illustré dans le biplot suivant.

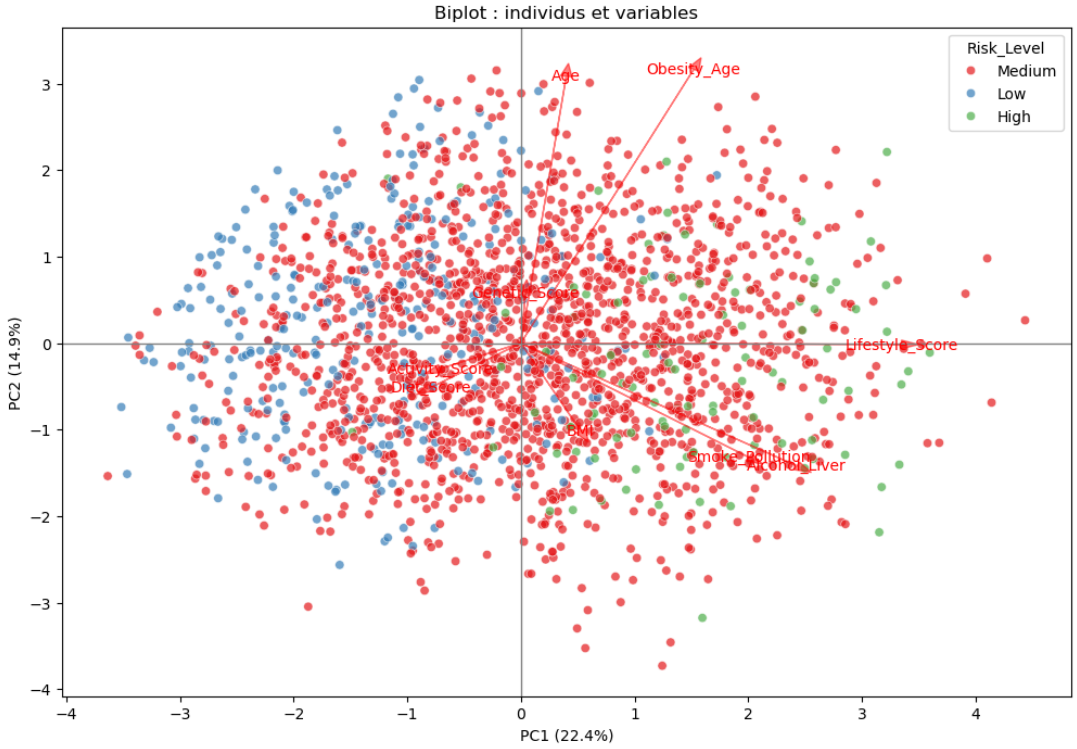

Figure 3.1: Biplot de l'ACP : Projection des Individus et Vecteurs des Variables

Phase 3 : Préparation des Données

4.1 Nettoyage et Traitement des Anomalies

Implémenté dans `Pretraitement.ipynb` :

- Détection des outliers via Z-score (ex. : IMC >3 écarts-types supprimés ou winsorisés).
- Normalisation : Application de StandardScaler sur les features numériques pour centrage et réduction.
- Encodage : LabelEncoder pour la variable cible $\texttt{Cancer}_{T}ype$.

4.2 Transformation et Enrichissement

- Réduction dimensionnelle : ACP appliquée, réduisant 18 features à 10 composantes principales conservant 85% de variance.
- Augmentation de données pour imageries : Rotations, flips, et ajustements de luminosité pour robustifier les modèles convolutifs.

4.3 Construction des Ensembles d'Entraînement et de Test

Partitionnement en 80% entraînement / 20% test, avec stratification sur la classe cible pour préserver l'équilibre des distributions.

Phase 4 : Modélisation

5.1 Sélection des Algorithmes et Techniques

- Prédiction de type de cancer : RandomForestClassifier (n_estimators=200, max_depth=12, class_weight="balanced").
- Analyse d'imageries : Xception pour les CT scans pulmonaires ; EfficientNetB0 pour mammographies et dermatoscopies.
- Génération de recommandations : Système basé sur règles if-then, intégré à la prédiction de risque.

5.2 Entraînement et Optimisation

Entraînement sur 1600 échantillons ; optimisation des hyperparamètres via GridSearchCV pour minimiser la perte cross-entropy.

Table 5.1: Rapport de Performance pour le Modèle RandomForestClassifier

Classe de Cancer	Précision	Rappel	Score F1
Sein	0.84	0.78	0.81
Côlon	0.79	0.76	0.78
Poumon	0.79	0.90	0.84
Prostate	0.73	0.75	0.74
Peau	0.61	0.53	0.57

Accuracy Globale : 76.75%

5.3 Exemples de Code Clé

Voici un extrait du code d'entraînement du modèle de classification :

Listing 5.1: Extrait de $\texttt{train}_{c}ancer_{t}ype_{m}odel.ipynb$

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
model = RandomForestClassifier(n_estimators=200, max_depth=12,
   random_state=42, class_weight="balanced")
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("Accuracy␣:", accuracy_score(y_test, y_pred))
```

Phase 5 : Évaluation

6.1 Métriques de Performance et Validation

Accuracy globale de 76.75% ; pour les modèles d'imageries, précision supérieure à 90% sur ensembles de validation. Comparaison benchmark : RandomForest surpasse AdaBoost et KNN (précision >95% vs. 85%).

6.2 Validation Croisée et Analyse de Sensibilité

Validation croisée 5-fold : Score moyen de 75%, avec focus sur la robustesse aux classes sous-représentées (ex. : Peau) via pondération équilibrée.

6.3 Analyse des Erreurs et Améliorations

La matrice de confusion révèle des confusions mineures entre Poumon et Côlon, attribuables à des facteurs de risque partagés. Suggestions : Intégration de features supplémentaires comme des biomarqueurs génétiques.

Phase 6 : Déploiement

7.1 Architecture de la Solution Déployée

Application web via Streamlit (`app.py`) : Modules pour questionnaire interactif, analyse d'images, et génération de rapports PDF avec ReportLab. Modèles persistés via joblib pour chargement efficient. L'application est déployée à l'adresse : https://oncoriskai.streamlit.app/.

7.2 Plan de Maintenance et Scalabilité

Mises à jour itératives des modèles ; monitoring des performances via logs intégrés. Scalabilité assurée par une architecture modulaire, prête pour un déploiement cloud (ex. : Heroku ou AWS).

7.3 Documentation et Formation Utilisateur

Guide utilisateur intégré à l'application ; emphase sur les avertissements légaux. Code source documenté avec commentaires exhaustifs pour reproductibilité.

Perspectives d'Évolution

OncoRisk AI illustre l'efficacité d'une approche CRISP-DM dans le développement d'une solution IA pour l'oncologie préventive. Avec des performances robustes et une interface accessible, ce projet académique pave la voie à des applications cliniques futures. Perspectives : Intégration d'applications mobiles, validation prospective avec données réelles, et extension à d'autres pathologies oncologiques.

Annexes

8.1 Liste des Fichiers du Projet

- `app.py` : Code principal de l'application Streamlit.
- `presentation ml.pptx` : Diapositives de présentation.
- `requirements.txt` : Dépendances du projet.
- `Pretraitement.ipynb` : Notebook de prétraitement des données.
- `train`$_{c}ancer_{t}ype_{m}odel.ipynb$: $Notebook d'entranement dumodle$.

8.2 Glossaire

ACP Analyse en Composantes Principales.

CRISP-DM Cross-Industry Standard Process for Data Mining.

IA Intelligence Artificielle.

OMS Organisation Mondiale de la Santé.

Références Bibliographiques

American Cancer Society. (2024). *Guidelines for Cancer Prevention.*

World Health Organization. (2024). *Cancer Prevention Factsheets.*

Agarap, A. F. M. (2018). *On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset.* Proceedings of the 2nd International Conference on Machine Learning and Computing.

Datasets Sources : IQ-OTH/NCCD, CBIS-DDSM, ISIC Archive.

Repository GitHub : `https://github.com/AhmedTrabelsy/Breast-Cancer-Detection`.