الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N°Réf:……

Centre Universitaire

AbdElhafid Boussouf Mila

Institut des Sciences et Technologie **Département de Mathématiques et Informatique**

Mémoire préparé en vue de l'obtention du diplôme de Master

En : **Informatique**

Spécialité : Intelligence Artificiel et ces Applications (I2A)

Thème

Amélioration et Extraction d'Informations à partir de Documents Anciens Dégradés .

Préparé par:

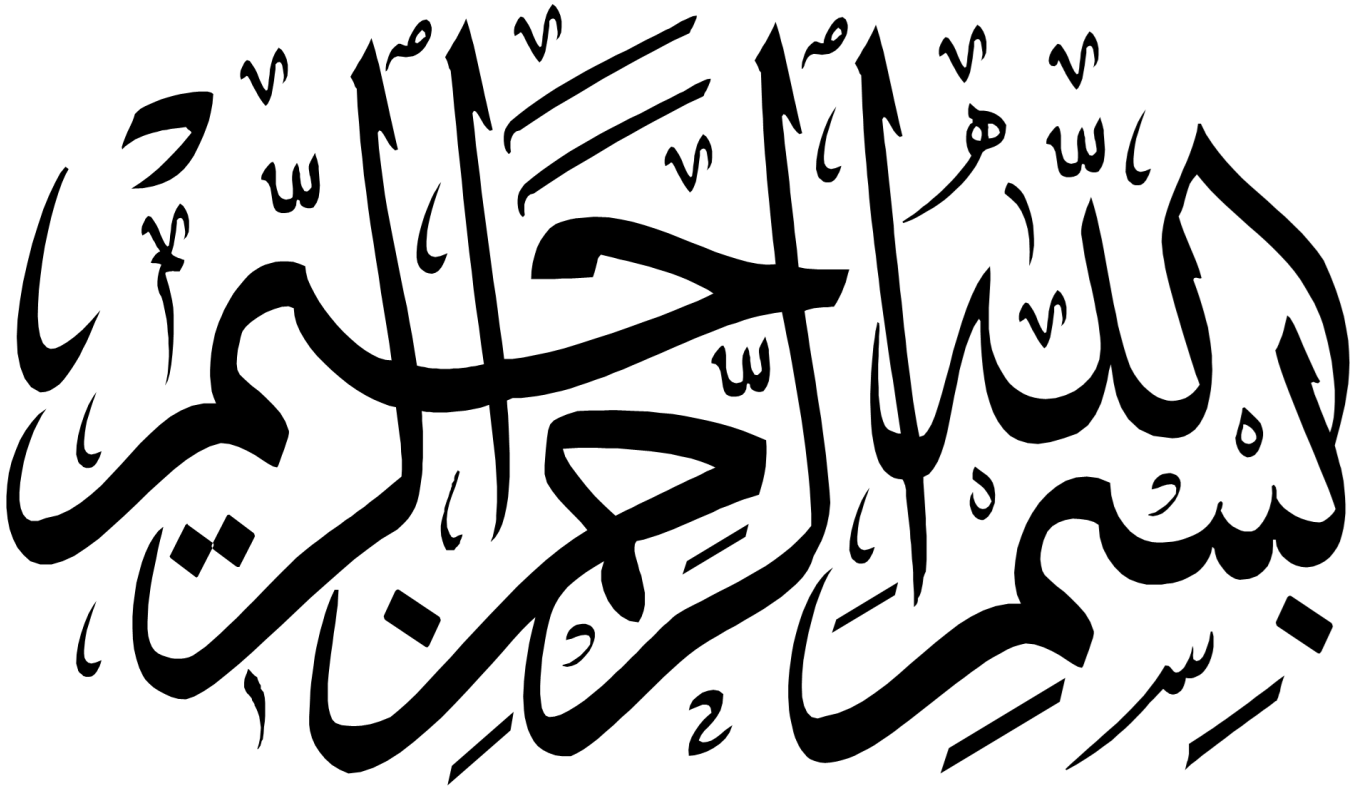
Zouaghi Yassamine

Ferhat Dounya

Soutenue devant le jury:

Président	**AOUAG Mouna**	**Grade MCB**
Examinateur	**BEKHOUCHE Maamar**	**Grade MAA**
Encadrant	**HADJI Atmane**	**Grade MAA**

Année Universitaire:2024/2025


بسم الله الرحمن الرحيم

Remerciement

Nous remercions, en premier lieu, ALLAH, pour la volonté, la force, la santé et la patience qu'Il nous a accordées afin de réaliser ce travail.

Nous tenons à adresser nos plus chaleureux remerciements à notre promoteur, Monsieur HADJI A., pour sa constante présence, sa confiance, sa patience ainsi que pour ses remarques pertinentes, ses conseils précieux et sa disponibilité tout au long de la réalisation de ce projet.

Nous exprimons également notre profonde gratitude aux membres du jury pour l'honneur qu'ils nous ont fait en acceptant d'évaluer notre travail.

Nous remercions sincèrement l'ensemble du corps enseignant et administratif du département Informatique pour leur soutien, leur disponibilité et leur gentillesse. Nos remerciements les plus sincères s'adressent également à nos chers parents, pour leurs sacrifices constants, leur soutien indéfectible et leurs encouragements sans faille.

Enfin, nous tenons à remercier toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce modeste travail.

Dédicace

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

وَءَاخِرُ دَعۡوَىٰهُمۡ أَنِ ٱلۡحَمۡدُ لِلَّهِ رَبِّ ٱلۡعَٰلَمِينَ ١٠

Tout d'abord, je tiens à remercier DIEU

De m 'avoir donné la force et le courage de mener , à bien ce modeste travail .

Je tiens à dédier cet humble travail à :

A ma tendre mère « MUNIRA » et mon très cher père « SEDDIK ».

Vous êtes la source de tout amour, de toute force et de toute bonté dans ma vie.

Je vous aime infiniment.

A ma sœur « IBTISSAM ».

A mes frère : « ABDOU » , « AMER », « DAOUD » .

A ma chére : « CHAYMA ».

À mes petits oiseaux préférés : « BYLASSAN, YAZEN et ELINE ».

Vous êtes la douceur de mon cœur . Je vous aime infiniment.

À ma partenaire de travail : « YASSAMINE ».

A mes meilleurs amis :

« SAMIRA , MARWA , MARIA , ROUMIASSA » .

Et enfin, un merci sincère à moi-même.

Pour mon courage, ma persévérance et tous les efforts fournis tout au long de ce parcours académique.

Tout ceux qui m'aiment et que j'aime.

DOUNYA

Dédicace

ٱلْحَمْدُ لِلَّهِ ٱلَّذِي عَلَّمَ بِٱلْقَلَمِ، عَلَّمَ ٱلْإِنسَٰنَ مَا لَمْ يَعْلَمْ

J'offre ce mémoire avec tout mon amour et ma gratitude à mes chers parents < Amar et Halima >, les piliers de ma vie, qui ont cru en moi même dans mes moments de doute,
et m'ont toujours soutenu avec une patience et un courage indescriptibles.

À ma chère grand-mère, à mes frères et sœurs < Farah, Amira et Mohamed >,
sources d'inspiration et de force dans les moments les plus difficiles.
À mes petits cœurs : < Mayassine, Sirine et Soujoud >.

À toute mes amies et ma partenaire de travail : < Dounya >,
et à toute ma famille < Zouaghi >.
À mes professeurs émérites, qui ont été une lumière éclairant mon chemin par leur savoir et leurs conseils.

Ce travail n'est pas qu'un simple assemblage de mots, mais un fragment de mon âme.
Il contient la fatigue des longues nuits, le murmure des prières, et les larmes de l'espérance.
Je l'offre à tous ceux qui ont cru en moi,
et m'ont poussé à croire en moi-même.
Et à la vie qui m'a mise à l'épreuve, pour mieux me réécrire.

YASSAMINE

Résumé

Les documents anciens constituent une source précieuse de mémoire historique et culturelle, et servent également de pièces officielles pour prouver les propriétés privées et publiques. Cependant, avec le temps, ils subissent des dégradations physiques (jaunissement du papier, encre effacée, moisissures) qui rendent leur lecture difficile. La restauration numérique et l'extraction d'informations fiables à partir de ces documents sont donc essentielles pour leur préservation et leur mise à disposition.

L'intelligence artificielle (IA), notamment les techniques de vision par ordinateur et de traitement automatique du langage naturel (TALN), offre des solutions innovantes. Grâce aux réseaux de neurones et à la reconnaissance optique de caractères (OCR), il est possible de restaurer visuellement les documents et d'en extraire le texte avec précision.

Cependant, les méthodes classiques d'OCR ne s'adaptent pas facilement aux écritures anciennes ou manuscrites. Pour pallier ces difficultés, nous proposons une approche hybride combinant restauration d'images, OCR multilingue et analyse linguistique assistée par machine learning.

Cette recherche vise à concevoir un système intelligent permettant de restaurer des documents anciens, améliorer la reconnaissance des caractères et extraire des informations pertinentes (noms, dates, lieux). Le système développé est modulaire, flexible et prend en charge plusieurs formats (image, PDF). Il utilise des techniques avancées comme CLAHE pour le contraste, PaddleOCR et EasyOCR pour la reconnaissance de texte, ainsi que des outils de post-traitement linguistique.

Les tests effectués sur divers documents anciens en arabe, français et anglais ont montré une amélioration notable de la qualité visuelle (jusqu'à 20 %) et une extraction textuelle précise (taux moyen supérieur à 75 %).

Mots-clés : Documents anciens, OCR, IA, Restauration d'images, Extraction d'information

Abstract

Ancient documents represent a valuable source of historical and cultural memory, and also serve as official records for verifying private and public ownership. However, over time, they undergo physical degradation such as paper yellowing, faded ink, and mold, which makes reading them difficult. Digital restoration and reliable information extraction are therefore essential for their preservation and accessibility.

Artificial Intelligence (AI), particularly through computer vision and natural language processing (NLP) techniques, offers innovative solutions. Thanks to neural networks and Optical Character Recognition (OCR), it is possible to visually restore documents and extract text with precision.

However, traditional OCR methods do not easily adapt to ancient or handwritten scripts. To overcome these difficulties, we propose a hybrid approach combining image restoration, multilingual OCR, and linguistic analysis assisted by machine learning.

This research aims to design an intelligent system capable of restoring old documents, improving character recognition, and extracting relevant information (names, dates, places). The developed system is modular, flexible, and supports multiple formats (image, PDF). It uses advanced techniques such as CLAHE for contrast enhancement, PaddleOCR and EasyOCR for text recognition, and linguistic post-processing tools.

Tests conducted on various ancient documents in Arabic, French, and English have shown a significant improvement in visual quality (up to 20%) and precise text extraction (average rate exceeding 75%).

Keywords : Anciens documents, OCR, AI, Image restauration, Information extraction

الملخص

تشكل الوثائق القديمة مصدرًا ثمينًا للذاكرة التاريخية والثقافية ووثائق رسمية للملكيات الفردية والعامة، ومع مرور الزمن، تتعرض هذه الوثائق للتدهور و التآكل المادي مثل اصفرار الورق وزوال الحبر والفطريات، مما يجعل قراءتها صعبة. ومن هنا تظهر أهمية الترميم الرقمي واستخراج المعلومات الموثوقة من هذه الوثائق لتسهيل حفظها وجعلها في متناول الباحثين والمؤرخين .

يقدم الذكاء الاصطناعي(IA) ، وخاصةً تقنيات الرؤية الآلية ومعالجة اللغة الطبيعية (TALN) ، حلولًا مبتكرة. بفضل الشبكات العصبية والتعرف على الحروف الضوئية(OCR) ، يمكن ترميم الوثائق بصريًا واستخراج النصوص بدقة عالية.

لكن لا تتناسب طرق OCR التقليدية بسهولة مع الخطوط القديمة أو الكتابة اليدوية. وللتغلب على هذه الصعوبات، نقترح منهجية هجينة تجمع بين تحسين الصور و OCR متعدد اللغات والتحليل اللغوي المعزز بالتعلم الآلي.

يرمي هذا البحث إلى تصميم نظام ذكي قادر على ترميم الوثائق القديمة وتحسين التعرف على الحروف واستخراج المعلومات ذات الصلة (أسماء، تواريخ، أماكن). النظام المطور قابل للتوسيع ومرن ويتعامل مع عدة تنسيقات (صور، PDF) كما يستخدم تقنيات متقدمة مثل CLAHE لتحسين التباين، و PaddleOCR و EasyOCR للاستخراج النصي، بالإضافة إلى أدوات للمعالجة اللغوية الثانوية.

أظهرت الاختبارات التي أجريت على مختلف الوثائق القديمة باللغات العربية والفرنسية والإنجليزية تحسنًا ملحوظًا في الجودة البصرية (حتى 20٪) واستخراجًا دقيقًا للنصوص (معدل يزيد عن 75٪)

الكلمات المفتاحية :الوثائق القديمة، OCR، الذكاء الاصطناعي، تحسين الصور، استخراج المعلومات.

Table des Matière

Remerciement I
Dédicace II
Dédicace III
Résumé IV
Abstract V
الملخص VI
Table des Matières 7
Liste des Figures X
Liste des Tableaux XI
Liste des sigles et acronymes XII
Introduction générale 1

Chapitre 01: Restauration et Reconnaissance des Documents Anciens Dégradé

1.1 Introduction 4
1.2 Dégradation et numérisation des documents anciens 4
1.2.1 Définition de la dégradation des documents anciens 4
1.2.2 Définition La numérisation des documents anciens 4
1.2.3 Problèmes de conservation des documents papiers et leur impact 5
1.2.4 Archives et numérisation 5
1.2.4.1 Numérisation des matériaux d'archives classiques 6
1.2.4.2 Numérisation des matériaux d'archives microfilmés 6
1.2.4.3 Disposition et préservation des documents d'archives produits à l'ère de l'informatique 7
1.3 Techniques de restauration d'images 7
1.3.1 Modèles basés sur l'IA 7
1.3.1.1 Réseaux antagonistes génératifs (Gans) 7
1.3.1.2 Super résolution d'image 8
1.3.1.3 Filtres adaptatifs 10
1.4 Reconnaissance optique de caractères (OCR) 10
1.4.1 Définition de l'OCR 10
1.4.2 Défis spécifiques liés aux documents anciens 11
1.4.3 Approches modernes d'OCR 12

1.4.3.1 Modèle d'OCR basé sur les Transformers 12
1.4.3.2 Modèles CRNN 12
1.4.4 Modèles OCR disponibles 13
1.4.4.1 MMOCR 13
1.4.4.2 EasyOCR 13
1.4.4.3 PaddleOCR 13
1.4.4.4 Keras OCR 14
1.4.4.5 TesseractOCR 14
1.4.5 Comparaison entre bibliothèques OCR 14
1.5 Métriques d'évaluation de la qualité restaurée 16
1.5.1 Erreur quadratique moyenne (MSE) 16
1.5.2 Rapport crête signal sur bruit (PSNR) 16
1.5.3 Rapport signal sur bruit (SNR) 17
1.6 Conclusion 17

Chapitre 02: Extraction d'Informations et Archivage Numérique Structuré

2.1 Introduction 19
2.2 Extraction d'informations à partir de textes anciens 19
2.2.1 Reconnaissance d'entités nommées historiques 19
2.2.2 Définition du Traitement Automatique du Langage Naturel (TALN) 21
2.2.3 Définition de l'apprentissage automatique 21
2.2.4 Approches hybrides combinant TALN et apprentissage automatique 22
2.3 Préservation Numérique et archivage 23
2.3.1 Formats modernes d'archivage numérique 23
2.4 Évaluation des performances d'extraction d'information 24
2.4.1 Métriques classiques d'évaluation (précision, rappel, F1-score) 25
2.4.2 Évaluation humaine vs automatique 25
2.5 Conclusion 26

Chapitre 03: Approche Proposée : Conception et mise en œuvre

3.1 Introduction 28
3.2 Notre Approche Proposée 28
3.2.1 Vision Globale 28

3.2.2 Objectifs de l'Approche 28
3.2.3 Innovation de l'Approche 28
3.3 Architecture du Modèle Proposé 29
3.4 Étapes de l'approche proposée 30
3.4.1 Étape 1 : Réception et validation des fichiers....30
3.4.2 Étape 2 : Prétraitement et Amélioration d'Image....31
3.4.2.1 CLAHE (Contrast Limited Adaptive Histogram Equalization)....31
3.4.2.2 Filtre bilatéral....31
3.4.3 Étape 3 : Traitement Différencié par Type de Document 32
3.4.3.1 EasyOCR....32
3.4.3.2 PaddleOCR 33
3.4.4 Étape 4 : Évaluation de la qualité 34
3.4.5 Étape 5 : Post-traitement et formatage 35
3.4.6 Étape 6 : Génération et export des résultats 36
3.5 Conclusion....37

Chapitre 04 : Réalisation et Évaluation des performances

4.1 Introduction 39
4.2 Environnement de travail....39
4.2.1 Matériel utilisé....39
4.2.2 Outils de développement 39
4.2.2.1 Google Colab 39
4.2.2.2 NGROK: 40
4.3 Présentation de l'application 41
4.4 Evaluation de l'Efficacité de Notre Approche Proposée....45
4.5 Evaluation basé sur les métriques de performance....47
4.6 Conclusion....49

Conclusion générale 50
Bibliographies....52

Liste des Figures

Figure 1. 1: Exemple de numérisation d'une image dégradée.....5
Figure 1. 2: Exemple d'architecture de modèle de réseau antagoniste génératif.....8
Figure 1. 3: Les différentes étapes pour augmenter la résolution 9
Figure 1. 4: Comparaison des bibliothèques OCR (Français, Anglais, Arabe) 16
Figure 2. 1:TALN avec ses trois domaines fondamentaux 21
Figure 2. 2: Les deux phases essentielles d'apprentissage automatique.....22
Figure 3. 1: Architecture générale du système proposé 29
Figure 3. 2: Architecture de Réception et Validation des Fichiers 30
Figure 3. 3: Architecture de Prétraitement et Amélioration d'Image.....32
Figure 3. 4: Architecture de Traitement Différencié par Type Image 33
Figure 3. 5: Architecture de Traitement Différencié par Type PDF 34
Figure 3. 6: Architecture d' Évaluation de la Qualité.....35
Figure 3. 7: Architecture de Post-traitement et Formatage 36
Figure 3. 8: Architecture de Génération et Export des Résultats.....37
Figure 4. 1: Google Colab capture d'écran.....40
Figure 4. 2: Capture d'écran Ngrok.....40
Figure 4. 3: Interface principale de notre application 42
Figure 4. 4: Exemple de résultat Image/Français.....43
Figure 4. 5: Exemple de résultat Image/Anglais.....43
Figure 4. 6: Exemple de résultat PDF/Arabe 44
Figure 4. 7: Exemple sur les textes manuscrits.....45
Figure 4. 8: Capture d'un graphique comparatif.....47

Liste des Tableaux

Tableau 1. 1: Comparatif structuré des bibliothèques OCR dans plusieurs critéres.....14
Tableau 1. 2: Comparaison des bibliothèques OCR dans différentes langues 15
Tableau 2. 1: Tableau de contingence.....25
Tableau 2. 2: comparatif entre Évaluation humaine et Évaluation automatique 26
Tableau 4. 1: Les statistiques de clarté sur plusieurs documents arabes 46
Tableau 4. 2: Les statistiques de clarté sur plusieurs documents anglais 46
Tableau 4. 3: Les statistiques de clarté sur plusieurs documents français 47
Tableau 4. 4: Evaluation basé sur les métriques de performance des documents arabes 48
Tableau 4. 5: Evaluation basé sur les métriques de performance des documents anglais 48
Tableau 4. 6: Evaluation basé sur les métriques de performance des documents français.....49

Liste des sigles et acronymes

OCR	Reconnaissance Optique de Caractères (Optical Character Recognition)
MSE	Erreur Quadratique Moyenne (MeanSquaredError)
SNR	Rapport Signal sur Bruit (Signal-to-Noise Ratio)
PSNR	Rapport Crête Signal sur Bruit (Peak Signal-to-Noise Ratio)
GANs	Réseaux Antagonistes Génératifs (GenerativeAdversarial Networks)
CLAHE	Égalisation de l'histogramme adaptative limitée au contraste (Contrast Limited Adaptive HistogramEqualization)
NER	Reconnaissance d'Entités Nommées (NamedEntity Recognition)
CPU	Unité Centrale de Traitement (Central Processing Unit)
GPU	Unité de Traitement Graphique (GraphicsProcessing Unit)
RAM	Mémoire d'Accès Aléatoire (Random Access Memory)
API	Interface de Programmation (Application Programming Interface)
PDF	Format de Document Portable (Portable Document Format)
IA	Intelligence Artificielle
MMOCR	Multi-Modal Optical Character Recognition
Keras OCR	Framework basé sur TensorFlow pour l'OCR
PaddleOCR	Toolkit d'OCR développé par Baidu
Tesseract OCR	Moteur OCR open-source développé par Google
SIMD	Single Instruction Multiple Data
RNN	Réseaux de Neurones Récurrents (Recurrent Neural Networks)
CRNN	Réseaux de Neurones Convolutifs et Récurrents (Convolutional Recurrent Neural Networks)
LSTM	Mémoire à Long Terme (Long Short-Term Memory)
CNN	Réseaux de Neurones Convolutifs (Convolutional Neural Networks)
NLP	Traitement du Langage Naturel (Natural Language Processing)
F1-score	Mesure de précision d'un test statistique
JSON	Notation Objet JavaScript (JavaScript Object Notation)
RGB	Rouge, Vert, Bleu (format de couleur)
PNG	Portable Network Graphics
JPG / JPEG	Joint Photographic Experts Group

Introduction générale

Depuis des siècles, les documents anciens sont des vecteurs de transmission des savoirs, de la mémoire collective, des repères juridiques, économiques et culturels. Ils sont de plus en plus variés en matière de supports (actes notariés, registres de cadastre, lettres manuscrites ou encore décrets officiels) favorisant ainsi le témoignage de la vie administrative, de la vie sociale et de la vie patrimoniale des sociétés passées.

La numérisation, la conservation et l'exploitation de ces documents à des fins de recherche et de valorisation historique sont l'objet d'un intérêt croissant, du fait de l'évolution des outils numériques. Cependant, la plupart de ces documents ont subi au cours du temps des dégradations importantes : papier jauni, encre effacée, moisissures, déchirures et/ou détériorations provoquées lors d'une conservation inadéquate. Toutes ces dégradations compromettent leur lisibilité, ce qui complique l'analyse manuelle et l'interprétation automatique. Dès lors, la restauration numérique et l'extraction d'information utile constituent des opérations centrales pour faciliter la conservation et l'accessibilité de ces archives.

L'intelligence artificielle (IA) et les récents progrès en vision par ordinateur et en traitement automatique du langage (TAL) proposent de nouvelles perspectives dans ce domaine. À la suite des avancées dans les modèles de réseaux de neurones, il est désormais possible d'améliorer visuellement les images dégradées, mais aussi de transcrire automatiquement le contenu textuel de documents papier grâce à de nouvelles méthodes de reconnaissance optique de caractères (OCR). Une fois ce texte extrait, il est possible d'en déduire et d'organiser les informations historiques, telles que les noms de lieux, de personnes ou de dates.

La reconnaissance et l'exploitation des entités permet non seulement d'offrir une lisibilité et une possibilité de consultation aux documents, mais encore de construire des bases de connaissance historiques présentant un potentiel d'utilisations diversifiées, qu'il s'agisse de recherche scientifique, de valorisation patrimoniale, d'analyse historique ou, bien sûr, de gestion des archives publiques.

Néanmoins, ces traitements sont confrontés à plusieurs difficultés. Les modèles classiques de reconnaissance optique de caractères ne sont pas adaptés aux anciennes polices ou aux styles d'écriture manuscrite et l'extraction d'informations exactes reste peu efficace face aux variations linguistiques, aux formes de phrase anciennes, voire disons archaïques, ou au faible niveau de numérisation des images. Pour surmonter ces difficultés, il serait nécessaire de recourir à une approche hybride qui conjugue traitements d'images, modèles linguistiques (ou de machine learning), techniques de DeepLearning, et méthodes d'extraction d'informations.

Cette recherche vise principalement à concevoir une méthode automatisée et intelligente permettant la restauration visuelle des documents anciens, l'amélioration de la reconnaissance des caractères qu'ils contiennent et l'extraction fiable des informations pertinentes qu'ils présentent. Ce processus repose sur l'articulation d'une approche combinant IA pour la restauration d'image, adaptation d'un OCR aux spécificités historiques et algorithmes TALN et d'apprentissage automatique pour l'analyse du contenu textuel.

Pour répondre à ces enjeux, notre mémoire est structurée en quatre chapitres :

- Dans le premier chapitre, nous abordons les défis liés à la dégradation des documents anciens, les techniques de restauration basées sur l'intelligence artificielle ainsi que les approches modernes de reconnaissance optique de caractères (OCR).
- Dans le deuxième chapitre, nous explorons les méthodes d'extraction d'informations à partir des textes restaurés, les modèles hybrides TALN/apprentissage automatique, ainsi que les formats numériques d'archivage adaptés à une conservation structurée et pérenne.
- Dans le troisième chapitre, nous présentons notre méthodologie proposée pour un système hybride multilingue basé sur l'OCR (Reconnaissance Optique de Caractères), visant à améliorer la qualité des images et à extraire du texte avec une grande précision, en mettant l'accent sur les techniques de prétraitement d'image, l'évaluation quantitative et la gestion du contexte linguistique.
- Dans le quatrième chapitre, nous présentons l'expérimentation et les résultats pratiques qui démontrent l'efficacité de la solution proposée, en la testant sur une variété de documents historiques en arabe, français et anglais.

Chapitre 01:

Restauration et Reconnaissance des Documents Anciens Dégradé

1.1 Introduction

La préservation et l'exploitation des documents anciens jouent un rôle crucial dans la valorisation du patrimoine historique, culturel et scientifique. Cependant, ces documents sont souvent fortement dégradés par le temps, l'environnement ou la manipulation humaine, rendant leur accès difficile et leur contenu peu exploitable. Face à ces défis, les avancées technologiques, notamment dans les domaines de la numérisation, de la restauration d'images et de la reconnaissance optique de caractères (OCR), offrent des solutions prometteuses.

Ce chapitre présente les fondements et les techniques modernes utilisées pour restaurer et reconnaître automatiquement le contenu des documents anciens dégradés. Il commence par définir les concepts de dégradation physique et de numérisation des archives, tout en abordant les problématiques de conservation. Ensuite, il explore les méthodes de restauration d'images à l'aide de modèles d'intelligence artificielle et les critères d'évaluation de la qualité des images restaurées. Enfin, il se focalise sur la technologie OCR, en décrivant ses principes, les défis propres aux documents historiques, les approches modernes, ainsi qu'une comparaison des principales bibliothèques OCR disponibles aujourd'hui.

1.2 Dégradation et numérisation des documents anciens

1.2.1 Définition de la dégradation des documents anciens

La notion de dégradation des images tient une place centrale dans le présent mémoire. Il convient donc de la définir. On appelle dégradation tous les effets indésirables cumulés nuisant à la lisibilité, au traitement ou à la conservation des images. Les dégradations ont plusieurs origines et l'accumulation des défauts rend difficile leurs séparations. Les dégradations ont des effets plus destructeurs sur les images de documents que sur toutes autres images dites «naturelles » [1].

1.2.2 Définition de La numérisation des documents anciens

Est un enjeu important pour les services d'archives, les bibliothèques, les historiens et les chercheurs en sciences littéraires pour les possibilités de manipulation, de visualisation et de recherche d'information qui en découlent. La numérisation physique : scannérisation (ou digitalisation), consiste à créer une image du document (un tableau de points ou de pixels1) à l'aide d'une caméra numérique ou d'un scanneur. Une haute résolution est souvent nécessaire (de 300 à 600 dpi) pour restituer les éléments les plus fins de l'écriture et des graphismes. L'image obtenue est en couleur, en niveaux de gris (un nombre fini de teintes de gris) ou bitonale (noir et blanc) suivant les possibilités du capteur et les choix de numérisation. La question du format de sauvegarde (ou stockage), dépend de l'application visée et de la taille du support de conservation.

Les possibilités de choix étant de conserver l'image brute ou de la comprimer avec ou sans perte d'information [2].

La figure ci-dessous représente un exemple de numérisation une image dégradée

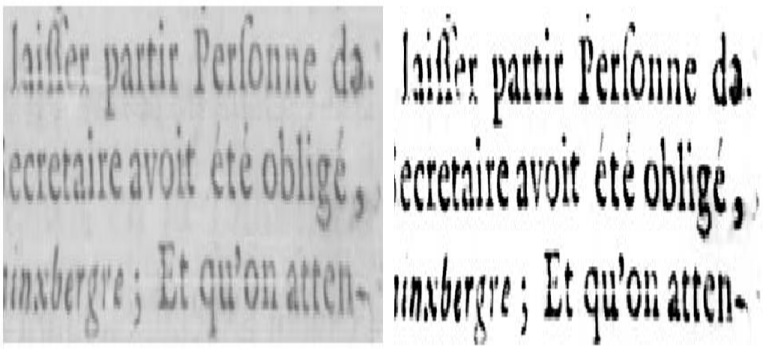


Figure 1. 1: Exemple de numérisation d'une image dégradée

1.2.3 Problèmes de conservation des documents papiers et leur impact

- L'encollage à l'alun et à la colophane, qui, au contact de l'humidité, produit de l'acide sulfurique, provoquant ainsi la dégradation, le jaunissement et la fragilisation du papier.
- Les impuretés présentes dans la pâte à papier, telles que le cuivre et le fer, qui accélèrent l'oxydation et la décomposition de la cellulose, entraînant un changement de couleur et une perte de résistance du papier.
- Les conditions de stockage inappropriées, notamment des températures et un taux d'humidité élevés, qui accélèrent les réactions chimiques responsables de la détérioration du papier. En effet, une température plus basse ralentit ces réactions et peut doubler la durée de vie du papier à chaque réduction de 10 °C.
- L'exposition à la lumière et aux polluants atmosphériques, comme le dioxyde de soufre, l'ozone, les oxydes d'azote et le sulfure d'hydrogène, qui contribuent à la détérioration du papier.
- La présence de lignine dans la pâte de bois, une substance organique qui, en se décomposant, produit des acides responsables du brunissement et de la fragilisation du papier.
- L'usage précoce du chlore pour le blanchiment, qui n'était pas entièrement éliminé après traitement, causant la dégradation de nombreux ouvrages des XVIIIe et XIXe siècles [3].

1.2.4 Archives et numérisation

Ce changement technologique dans le domaine des technologies de l'information a influencé le travail des archives, posant ainsi d'importants défis technologiques face à la documentation produite sur la base de cette technologie. Les obligations qui incombent aux archives en général sont variées et immédiates : gérer le matériel d'archives existant ainsi que celui formé de manière numérique à l'ère de l'informatique et d'Internet, organiser, classer, préserver et mettre ces documents à disposition des utilisateurs grâce aux technologies de l'information, sans compromettre

ni leur identité ni leur validité [4].

Cela nécessite que l'archivage soit réalisé de trois manières :

1.2.4.1 Numérisation des matériaux d'archives classiques

Pour rendre les documents d'archives existants accessibles aux utilisateurs et compatibles avec les technologies de l'information (TI) de l'ère de l'informatique et d'Internet, il est nécessaire de numériser les documents classiques (documents écrits sur papier) afin de répondre aux besoins pratiques d'utilisation.« La documentation sur papier est encore très présente dans la création, le traitement et l'échange d'informations. En raison de son volume important, ce type de documentation devient de plus en plus difficile d'accès et occupe un grand espace ».

Le processus de numérisation a été rendu possible grâce aux technologies modernes, notamment l'invention du scanner et des logiciels permettant de stocker les documents numérisés avec leur apparence d'origine. « L'archive électronique représente actuellement la méthode la plus moderne d'archivage des documents. Cette méthode repose sur l'utilisation du scanner, qui permet de convertir les documents (photos) en format électronique sur un ordinateur où les images des documents sont stockées » .Cela demande bien sûr du temps et des ressources, mais c'est une nécessité imposée par l'ère de l'informatique et de la numérisation de tous les services vitaux, y compris le besoin d'utiliser la documentation archivistique [5].

1.2.4.2 Numérisation des matériaux d'archives microfilmés

Dès l'apparition des outils électroniques techniques, les techniques de microfilmage ont été développées comme une nouvelle méthode de préservation des documents d'archives en réduisant leur volume et en les rendant plus facilement accessibles. Avec le développement accéléré des technologies de l'information et de l'informatique, la conversion des archives microfilmées en documents numériques est devenue indispensable, sans que personne ne ressente de nostalgie pour cette technique. « L'archivage mixte représente un processus technico-technologique permettant de protéger les informations archivistiques grâce à une méthode combinée de microfilmage et de numérisation. Il implique l'acceptation et l'application de toutes les bonnes qualités du microfilm tout en tirant parti des avantages de la numérisation.

C'est une méthode fiable et économiquement viable pour protéger les documents d'archives et d'autres éléments du patrimoine culturel écrit ». Les programmes informatiques, grâce aux éléments de liaison et au transfert d'informations depuis les appareils de microfilmage vers les bases de données numériques, ont permis cette transition, facilitant ainsi le processus de conversion et réduisant le temps nécessaire [6].

1.2.4.3 Disposition et préservation des documents d'archives produits à l'ère de l'informatique

Avec l'installation des ordinateurs personnels comme outils de travail dans l'administration publique, les institutions publiques et privées, la production d'archives numériques est devenue une réalité, nécessitant des solutions pour leur classification, leur stockage et leur gestion. « le support des documents n'est en aucun cas le papier, mais un support électronique, à condition qu'il réponde aux exigences de caractère et de crédibilité prévues par la loi. Ainsi, pour sensibiliser à l'importance de la préservation des documents, les professionnels ont adopté de nouvelles méthodes, notamment l'archivage électronique des documents » [7].

1.3 Techniques de restauration d'images

Les techniques de restauration d'images sont essentielles pour améliorer la qualité visuelle des images corrompues ou dégradées. Voici quelques-unes des techniques et concepts associés:

1.3.1 Modèles basés sur l'IA

1.3.1.1 Réseaux antagonistes génératifs (Gans)

Les réseaux antagonistes génératifs (GAN) sont une architecture révolutionnaire dans le domaine de l'intelligence artificielle [8]. En 2014, ils représentent une technologie très prometteuse en apprentissage automatique. Les GAN utilisent des méthodes d'apprentissage profond, comme les réseaux de neurones convolutifs, pour la modélisation générative, permettant de découvrir et d'apprendre automatiquement les modèles présents dans les données d'entrée. Ils reposent sur un concept ingénieux de deux sous-modèles: le générateur, formé pour produire de nouveaux exemples, et le discriminateur, qui cherche à distinguer les exemples réels des exemples générés [9].

La figure ci-dessous représente l'architecture de modèle de réseau antagoniste génératif :

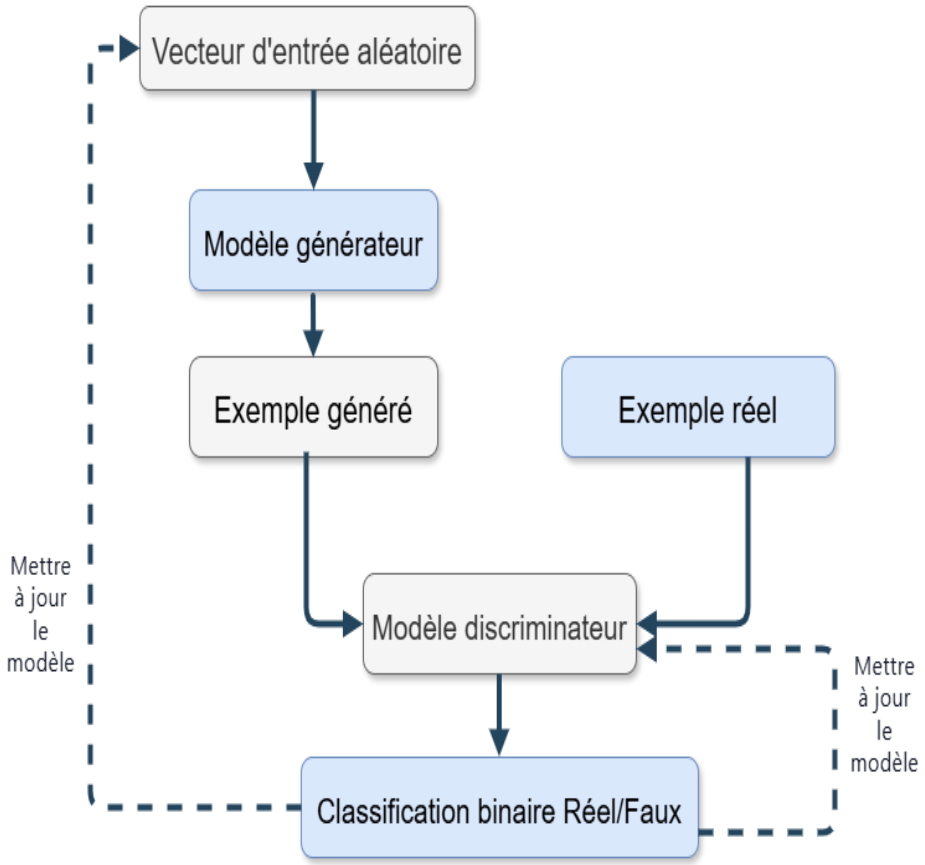


Figure 1. 2: Exemple d'architecture de modèle de réseau antagoniste génératif

1.3.1.2 Super résolution d'image

La super résolution d'image (SRI) est une technologie de pointe qui utilise des algorithmes avancés et des techniques d'apprentissage automatique pour produire des images à haute résolution à partir d'entrées à faible résolution. Le principe de la SRI est de générer une image haute résolution en déduisant les détails et les informations manquants d'images basse résolution. Pour ce faire, l'image basse résolution est analysée et des modèles d'apprentissage automatique sont utilisés pour estimer les informations manquantes. Ces modèles sont entraînés sur un vaste ensemble de données d'images à haute résolution, ce qui leur permet de prédire avec précision les informations manquantes dans l'image à basse résolution [10].

La super-résolution d'image est une technique utilisée pour améliorer la résolution d'une image, ce qui permet d'obtenir une image de meilleure qualité et plus détaillée. Le processus implique l'utilisation d'algorithmes qui analysent l'image et prédisent les données manquantes pour augmenter la résolution. Ce processus comporte différentes étapes :

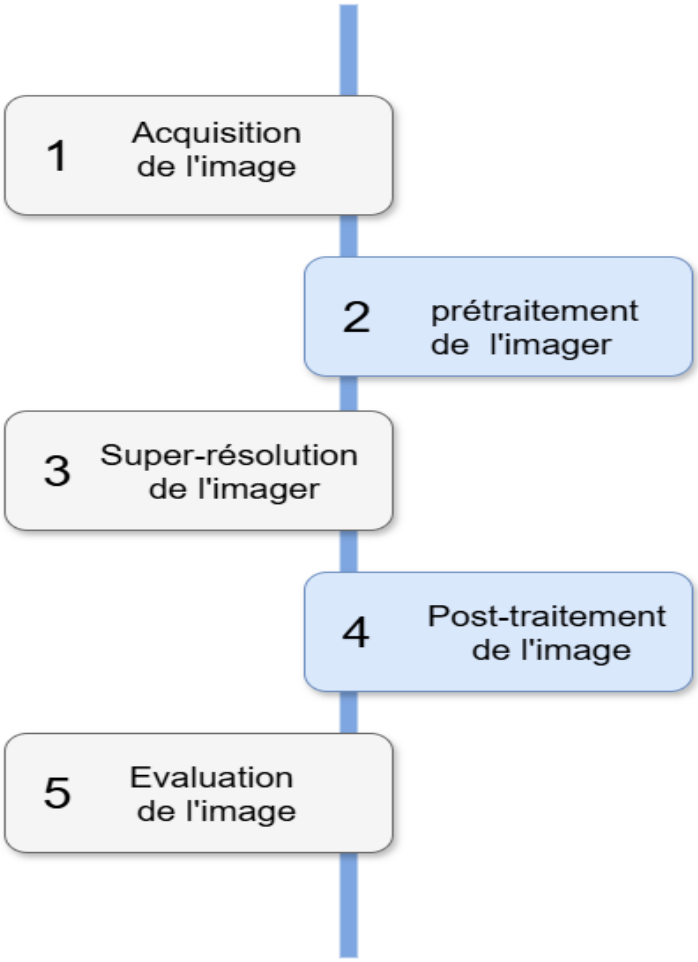


Figure 1.3: Différentes étapes pour augmenter la résolution

Acquisition de l'image : elle consiste à capturer une image à l'aide d'un appareil photo ou d'un scanner. L'image est généralement de faible résolution en raison des limitations de l'appareil d'imagerie ou d'un sous-échantillonnage pour le stockage ou la transmission [11].

Prétraitement de l'image : ce stade consiste à traiter l'image à faible résolution pour le processus de super-résolution. Cette étape peut consister à supprimer le bruit ou le flou de l'image, à recadrer l'image ou à aligner plusieurs images [11].

Super-résolution de l'image : dans cette étape, un modèle et/ou un algorithme est utilisé pour augmenter la résolution de l'image. Différentes méthodes permettent de le faire, telles que l'interpolation, la reconstruction ou les méthodes basées sur l'apprentissage profond [12].

Post-traitement de l'image : ce stade implique d'affiner l'image de sortie en supprimant les artefacts ou en améliorant la qualité de l'image. Cela peut inclure le flou, le débruitage ou la compensation des couleurs [11].

Évaluation de l'image : Ce processus permet l'évaluation de la qualité de l'image de sortie par rapport à l'image haute résolution d'origine, si cette dernière est déjà connue. Cela peut se faire à l'aide de mesures spécifiques telles que le PSNR, le SSIM ou le VIF [12].

1.3.1.3 Filtres adaptatifs

Dans ce type de filtrage, le calcul de la nouvelle valeur du pixel central est obtenu par un calcul qui exclut toute relation linéaire entre le pixel et son voisinage. Les principaux filtres non linéaires sont :

Filtre d'ordre : Il sélectionne parmi le voisinage un ou plusieurs niveaux de gris et les combine d'une certaine façon, dans ce type de filtre on note :

Filtre médian : Le filtrage par la médiane est le filtre d'ordre le plus connu. Il permet d'adoucir l'image en remplaçant la valeur du pixel central d'une fenêtre par une valeur représentative de son voisinage. Contrairement aux filtres linéaires classiques qui utilisent une moyenne, ce filtre retient la valeur médiane, c'est-à-dire celle qui, après avoir trié les valeurs des pixels de la fenêtre en ordre croissant, se situe exactement au centre de cette distribution. Cette méthode est particulièrement efficace pour éliminer le bruit impulsionnel (bruit de type sel et poivre) tout en préservant les contours des objets dans l'image.

Filtre de position : Dans ce type de filtre, le voisinage le plus homogène est sélectionné, et son niveau de gris moyen est affecté au pixel central [13].

1.4 Reconnaissance optique de caractères (OCR)

1.4.1 Définition de l'OCR

La reconnaissance optique de caractères (OCR) est une technologie qui extrait automatiquement les données pour convertir rapidement les images de texte dans un format lisible par les machines.

Parfois appelé « reconnaissance de texte », l'OCR permet d'extraire et de réaffecter les données à partir de documents numérisés, d'images prises avec un appareil photo et de fichiers

PDF contenant uniquement des images. Le logiciel OCR distingue les lettres figurant dans les images, forme des mots, puis des phrases pour permettre l'accès au contenu d'origine, ainsi que sa modification. Il élimine les tâches de saisie manuelle répétitives.

Les systèmes OCR associent matériel et logiciel pour convertir les documents imprimés sur supports physiques en texte lisible par les machines. Le matériel, tel qu'un scanner optique ou une carte de circuit imprimé spécialisée, copie ou lit le texte, puis le logiciel se charge du traitement avancé.

Les logiciels OCR peuvent exploiter l'intelligence artificielle (IA) pour appliquer des méthodes de reconnaissance intelligente de caractères (ICR) plus avancées et identifier ainsi les langues ou l'écriture manuscrite. Les entreprises font souvent appel à l'OCR pour convertir les documents juridiques ou historiques imprimés au format PDF. Les utilisateurs peuvent ainsi

modifier, mettre en forme et interroger les documents comme s'ils avaient été créés avec un logiciel de traitement de texte [14].

1.4.2 Défis spécifiques liés aux documents anciens

Malgré les deux décennies de recherche sur la numérisation des documents, toujours il n'existe pas de solution fiable permettant de passer d'un document numérisé à une version en mode texte sans erreurs. Les SOCR permettent de détecter et transposer un mot à partir d'une image, mais ils laissent encore quelques imperfections pour par venir à une réédition du document ce qu'il peut être dû à différents problèmes. Parmi ces problèmes nous citons :

La qualité du document : un document télécopié ou photocopié plusieurs fois est plus difficile à traiter que la copie originale. L'écriture peut devenir plus mince ou au contraire plus épaisse, dégradée avec des parties du texte qui manquent ou des tâches qui apparaissent, des ouvertures ou des bouchages de boucles …

L'impression : un document composé est de meilleure qualité qu'un document dactylographié qui, à son tour, est plus clair qu'un texte issu d'une imprimante matricielle. Une imprimante à jet d'encre peut introduire des tâches d'encre et un étalement des caractères, une imprimante laser peut générer des lignes ou des fonds ...

La résolution :la qualité du contenu de l'image dépend principalement de sa résolution. La difficulté souvent rencontrée est de pouvoir adapter la résolution aux différentes tailles de caractères et épaisseurs de graphiques présents dans le document, ne nécessitant pas le même niveau de précision. Le mauvais choix de la résolution peut conduire à un manque d'information dans le cas de la réduction de la résolution et à l'apparition d'un bruit plus abondant dans le cas contraire.

La discrimination de la forme : selon le style de la fonte utilisée (son corps et sa graisse...), le caractère change du graphisme. Le nombre de formes est d'autant plus important que le nombre de styles d'écriture est élevé. De plus, plusieurs caractères présentent une forte ressemblance.

Le support de l'information : tel que le papier, joue également sur les performances de la reconnaissance par sa qualité : son grammage, sa granulation et sa couleur.

L'acquisition : La numérisation en temps réel introduit souvent des distorsions dans l'image. Dans le cas hors ligne la qualité du texte numérisé est un compromis entre les variations de la position (inclinaison, translation, rétrécissement...), la propreté de la vitre du dispositif de numérisation et sa résolution.

L'inclinaison : est une source d'erreur classique, relativement gênante pour les systèmes de reconnaissance qui utilisent l'horizontal comme référentiel de base pour l'extraction des lignes de texte et la modélisation de la forme des lettres. Elle est de plus en plus maîtrisée grâce à l'existence de logiciels de redressement appliqués systématiquement sur les documents à leur entrée.

Les variations des dimensions : le « pitch » qui représente la distance en millimètres entre

les centres de deux pixels adjacents, peut varier entre 10, 12 ou 16. Cette valeur indique la finesse de l'image et par conséquent plus cette distance est réduite, plus le point est net.

Nombre de scripteurs :la difficulté de reconnaissance croît avec ce nombre, divisant l'échelle en trois : mono, multi et omni-scripteurs. En multi-scripteur, le système doit s'adapter à l'écriture de plusieurs scripteurs, tandis qu'en omni-scripteur, le système doit être capable de généraliser son apprentissage à n'importe quel type d'écriture.

Taille du vocabulaire : les applications à vocabulaire limité (< 100 mots) se distinguent de celles à vocabulaire très étendu (> 10 000 mots). Il est évidente que dans le premier cas, la complexité est moindre, car la réduction du nombre limite l'encombrement mémoire et favorise l'utilisation des méthodes directes de reconnaissance et donc rapides, par balayage systématique de l'ensemble des mots du lexique [15].

1.4.3 Approches modernes d'OCR

1.4.3.1 Modèle d'OCR basé sur les Transformers

Le TrOCR est un modèle pré-entraîné qui est donc plus avantageux en termes de réduction du temps et des ressources nécessaires. Il offre une précision sur la prédiction améliorée grâce au mécanisme d'attention intégré dans son architecture de Transformer classique. Cependant, comme il a été entrainé sur des textes en langue anglaise, il est moins performant en langue française. Par ailleurs, l'efficacité de ce modèle repose sur la condition que chaque image soumise doit contenir uniquement une ligne de texte, nécessitant ainsi une segmentation initiale de l'image en plusieurs sous-images. Chacune de ces dernières doit inclure une unique ligne de texte pour garantir un traitement optimisé. Dans de nombreuses études précédentes [16].

1.4.3.2 Modèles CRNN

Les CRNNs (Convolutional Recurrent Neural Networks) combinent les CNNs et les RNNs, et sont conçus pour prédire une séquence d'étiquettes à partir des images. Une image d'entrée est d'abord divisée en trames de même taille appelées champs réceptifs, et la couche CNN extrait les caractéristiques de l'image de chaque trame en utilisant des couches de convolution et de max-pooling, tandis que les couches entièrement connectées sont supprimées. Les caractéristiques des trames sont ensuite utilisées comme entrées pour les couches LSTM bidirectionnelles. Les couches récurrentes prédisent une distribution d'étiquettes de caractères pour chaque trame dans la séquence de caractéristiques. L'idée de la prédiction de séquence d'étiquettes est similaire aux CRFs : l'étiquette prédite pour chaque trame dépend des étiquettes des trames avoisinantes. Les modèles basés sur les CRNNs surpassent les modèles basés sur la détection dans les cas où les textes sont plus densément répartis. Dans cette mémoire, notre système de reconnaissance optique de caractères

(OCR) utilise les CRNNs comme architecture de base [17].

1.4.4 Modèles OCR disponibles

Texte extrait d'une image Grâce aux capacités automatisées d'extraction et de stockage de données, la technologie OCR(reconnaissance optique de caractères) constitue une méthode commerciale utile qui permet d'économiser du temps, de l'argent et d'autres ressources. Les logiciels OCR récupèrent les informations à partir de documents numérisés, d'images prises par une caméra et d'images au format PDF. Nous pouvons accéder au contenu original et le modifier à l'aide de logiciels OCR, qui séparent les lettres dans l'image, les transforment en mots, puis convertissent ces mots en phrases. De plus, cela élimine le besoin de saisie manuelle des données.

Les logiciels OCR populaires tels que Tesseract, EasyOCR, Keras OCR, MMOCR et PaddleOCR peuvent tous être utilisés pour extraire du texte à partir d'images. Les cadres OCR largement utilisés pour la reconnaissance de texte dans les images incluent MMOCR, EasyOCR, PaddleOCR, Keras OCR et TesseractOCR. Bien que leurs méthodes et applications de base soient différentes les unes des autres, nous en donnerons une brève description ci-dessous [18].

1.4.4.1 MMOCR

Basé sur l'architecture PyTorch, MMOCR est une boîte à outils OCR open source créée par l'équipe OpenMMLab. Elle offre une variété de tâches OCR, notamment la détection de texte, la reconnaissance et l'analyse de mise en page. Des modèles et méthodes modernes tels que TextSnake, Proposal Attention Network (PAN) et Scene Text Attention Recognition (SAR) sont intégrés à MMOCR. Cela facilite une reconnaissance précise du texte dans des environnements complexes [19].

1.4.4.2 EasyOCR

EasyOCR est un package OCR open source qui propose une interface simple pour la reconnaissance de texte. Il repose sur le moteur Tesseract OCR et utilise des modèles d'apprentissage profond pour une extraction précise du texte. EasyOCR identifie d'abord les régions d'intérêt (ROIs) contenant du texte dans l'image, à l'aide d'une technique d'identification de texte. Plusieurs algorithmes peuvent être utilisés pour reconnaître le texte, notamment ceux basés sur les bords, les contours ou les réseaux neuronaux. Pour créer une solution conviviale et précise de reconnaissance de texte, EasyOCR combine la puissance de Tesseract OCR avec des modèles d'apprentissage profond. Il utilise la détection de texte, la reconnaissance basée sur le deeplearning, la modélisation linguistique et des méthodes de post-traitement. La bibliothèque simplifie la complexité de l'OCR et offre aux développeurs une API facile à utiliser pour intégrer cette fonctionnalité dans leurs projets[20].

1.4.4.3 PaddleOCR

Il s'agit d'un cadre OCR open source créé par l'entreprise technologique d'IAPaddlePaddle. Il

offre une gamme complète d'outils pour la détection et la reconnaissance detexte utilisant des modèles pré-entraînés. Avec un accent sur l'efficacité et la convivialité, PaddleOCR s'appuie sur la technologie d'apprentissage profond PaddlePaddle. Des réseaux neuronaux convolutifs (CNN) et des réseaux neuronaux récurrents (RNN) sont utilisés dans laconstruction de ces modèles[21].

1.4.4.4 Keras OCR

La bibliothèque Keras OCR utilise le framework de deeplearning Keras pour fournir des fonctionnalités OCR. Elle vise à offrir une interface simple et compréhensible pour les opérations de reconnaissance de texte. En général, les modèles OCR basés sur des architectures d'apprentissage profond comme les CNN ou les RNN sont combinés avec des méthodes de prétraitement telles que la binarisation d'image et la suppression de bruit dans Keras OCR. Les procédures spécifiques de Keras OCR peuvent varier selon le modèle ou la technique utilisée[22].

1.4.4.5 TesseractOCR

L'un des moteurs OCR les plus connus et les plus utilisés est Tesseract OCR. Il est performant en reconnaissance de texte et constitue un projet open source initialement développé par HP, aujourd'hui soutenu par Google. Tesseract combine des modèles de Markov cachés (HMM) et d'autres algorithmes d'apprentissage automatique, en plus des méthodes traditionnelles de vision par ordinateur pour reconnaître le texte. La version la plus récente, Tesseract 4, intègre des approches d'apprentissage profond afin d'améliorer les performances, contrairement aux versions antérieures qui utilisaient principalement des méthodes classiques. Les modèles d'apprentissage profond utilisés par Tesseract sont basés sur des réseaux de neurones LSTM (Long Short Term Memory)[23].

1.4.5 Comparaison entre bibliothèques OCR

Une brève explication de la manière de comparer et d'évaluer différentes bibliothèques OCR, y compris Tesseract OCR, MMOCR, Keras OCR, PaddleOCR et EasyOCR, est donnée dans le document. Celui-ci énumère des éléments importants tels que la précision, la prise en charge des langues, les modèles pré-entraînés, les options de personnalisation et d'entraînement, l'intégration et l'accessibilité via une API, la vitesse et les performances, la communauté et la documentation, les tarifs et les licences, ainsi que la disponibilité pour l'intégration et l'entraînement.

La précision de reconnaissance de cinq bibliothèques OCR — Tesseract...[24]

Tableau 1.1: Comparatif structuré des bibliothèques OCR dans plusieurs critères

Critére	Tesseract OCR	MMOCR	Keras OCR	Paddle OCR	EasyOCR
Précision	Bonne, dépend du prétraitement	Très bonne avec des images bien	Moyenne à bonne selon les cas	Excellente, surtout avec modèles	Bonne pour plusieurs langues

		structurées		multilingues	
Langues supportées	+100 langues	Plusieurs langues supportées (dépend des modèles)	Principalement anglais et chiffres	+80 langues, y compris français et arabe	+80 langues supportées
Facilité d'utilisation	Moyenne, nécessite des étapes manuelles	Utilisation avancée, nécessite configuration	Simple à utiliser avec Keras/Tensor Flow	Interface simple, bien documentée	Très simple à utiliser
Reconnaissance mise en page	Limitée	Bonne (surtout en combinaison avec des détecteurs de mise en page)	Limitée	Très bonne, inclut détection de structure	Moyenne
Formats de sortie	Texte brut (nécessite outils externes pour PDF)	JSON, texte	Texte ou données structurées	Texte, JSON, structure, PDF possible	Texte brut, prise en charge de PDF
Manuscrits	Très limité	Faible à moyenne	Limité	Moyenne, meilleure avec modèles personnalisés	Moyenne
Vitesse	Rapide pour petits volumes, lent pour documents volumineux	Optimisé, mais dépend des modèles utilisés	Relativement rapide	Très rapide, support GPU	Rapide
Coût	Gratuit et open-source	Gratuit (open-source, Apache 2.0)	Gratuit	Gratuit (open-source, licence Apache 2.0)	Gratuit (open-source, licence Apache 2.0)
Utilisation hors ligne	Oui	Oui	Oui	Oui	Oui

Tableau 1. 2: Comparaison des bibliothèques OCR dans différentes langues

Bibliothèque OCR	**Français**	**Anglais**	**Arabe**
Tesseract OCR	95 %	92 %	74 %
MMOCR	94 %	89 %	73 %
PaddleOCR	97 %	89 %	91 %
Keras OCR	90 %	72 %	71 %

EasyOCR	89 %	71 %	67 %

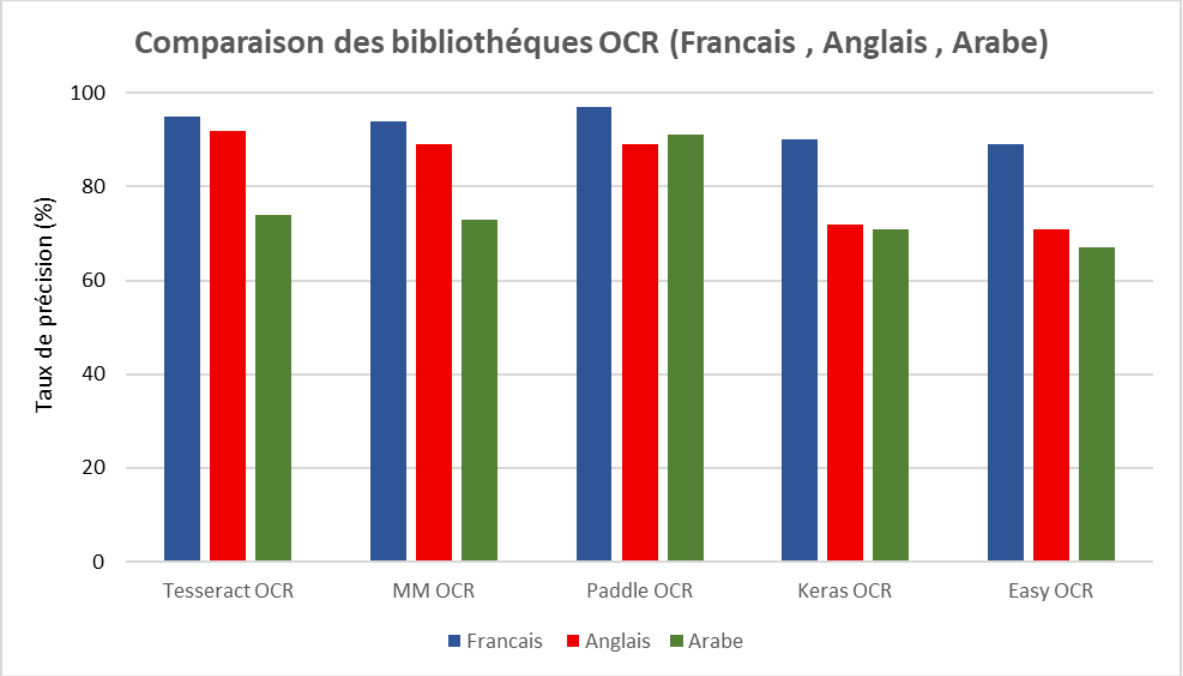


Figure 1.4: Comparaison des bibliothèques OCR (Français, Anglais, Arabe)

1.5 Métriques d'évaluation de la qualité restaurée

1.5.1 Erreur quadratique moyenne (MSE)

L'image dégradée Î est toujours comparée à l'originale I pour déterminer son rapport de ressemblance. Ce critère est le plus utilisé. Il est basé sur la mesure de l'erreur quadratique moyenne (MSE) calculée entre les pixels originaux et dégradés [25]:

$$\mathrm{MSE} = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(I(m,n) - \hat{I}(m,n)\right)^2$$

Où (M × N) est la taille de l'image, et I_P et $\hat{I}_P$ sont respectivement les amplitudes des pixels sur les images originale et dégradée. Il est vraisemblable que l'œil tienne beaucoup plus compte des erreurs à grandes amplitudes, ce qui favorise la mesure quadratique.

1.5.2 Rapport crête signal sur bruit (PSNR)

Au lieu de mesurer la distorsion, cette valeur (Peak Signal to Noise Ratio, PSNR) mesure la fidélité, puisqu'elle est proportionnelle à la qualité. Tout de même, elle est une fonction de MSE; sa définition et son utilisation proviennent du domaine du traitement de signal [26] :

$$\mathrm{PNSR} = 10\log_{10}\left(\frac{{I_{max}}^2}{\mathrm{MSE}}\right)$$

Pour une image à niveau de gris, Imax désigne la luminance maximale possible. Une valeur de PSNR infini correspond à une image non dégradée. Et cette valeur décroît en fonction de la

dégradation. Le PSNR relie donc le MSE à l'énergie maximale de l'image.

1.5.3 Rapport signal sur bruit (SNR)

On utilise parfois une autre variante du rapport signal sur bruit (Signal to Noise Ratio : SNR), qui relient le MSE à l'énergie moyenne de l'image [27].

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\frac{1}{N}\sum I^2}{\mathrm{MSE}}\right)$$

1.6 Conclusion

La restauration et la reconnaissance des documents anciens dégradés reposent sur un ensemble de techniques complémentaires permettant d'optimiser l'exploitation de contenus historiques difficilement lisibles. Ce chapitre a mis en lumière l'importance de la numérisation adaptée aux contraintes des supports anciens, ainsi que le rôle déterminant de l'IA dans la restauration visuelle et la lecture automatique du texte via l'OCR.

Chapitre 02:

Extraction d'Informations et Archivage Numérique Structuré

2.1 Introduction

L'extraction d'informations pertinentes à partir de documents anciens représente un enjeu majeur dans les domaines de la recherche historique, de la gestion des archives et de la valorisation du patrimoine écrit. Grâce aux avancées du Traitement Automatique du Langage Naturel (TALN) et de l'apprentissage automatique, il est désormais possible d'analyser et de structurer automatiquement des contenus textuels issus d'images ou de manuscrits dégradés.

Ce chapitre explore les différentes méthodes d'extraction d'informations, notamment la reconnaissance d'entités nommées historiques et les approches hybrides combinant TALN et apprentissage automatique. Par ailleurs, la pérennisation des documents numériques nécessite une réflexion sur les formats d'archivage adaptés aux besoins de longévité, d'accessibilité et de lisibilité. Enfin, une évaluation rigoureuse des performances de ces méthodes est essentielle pour garantir la fiabilité des systèmes d'extraction, en tenant compte à la fois de mesures automatiques et d'analyses humaines.

2.2 Extraction d'informations à partir de textes anciens

2.2.1 Reconnaissance d'entités nommées historiques

La restauration et le problème de la reconnaissance des entités nommées (NER) dans les textes historiques a été étudié par plusieurs chercheurs jusqu'à présent.

En (2008) ont développé un système NER basé sur des règles, capable de reconnaître les noms de lieux et de personnes dans des archives numérisées des débats parlementaires britanniques provenant de deux périodes différentes : la fin du XVIIe siècle et le début du XIXe siècle. Ils se sont concentrés sur les difficultés liées à la nature des textes historiques (comme la grande variabilité dans l'utilisation des majuscules en début de mot) et à l'utilisation de la technologie de reconnaissance optique des caractères (OCR). Ils ont constaté que la reconnaissance des noms de personnes obtenait de meilleurs résultats que celle des noms de lieux. Autrement dit, l'identification de schémas pour la reconnaissance des noms de personnes était plus simple et donc plus résistante aux erreurs d'OCR.

D'un autre côté, ils ont également identifié d'autres problèmes causés par les erreurs d'OCR : mauvaise interprétation de la mise en page, division incorrecte des toques et segmentation erronée des paragraphes (au milieu d'un ton). Ils ont atteint un score F1 de 71,81 % pour la période 1814-1817 et de 70,35 % pour la période 1685-1691.

En (2010) ont expérimenté la reconnaissance des noms de personnes à l'aide de données OCR bruyantes. Ils ont testé trois approches différentes et évalué leurs résultats par rapport à des données tests annotés manuellement. Ils ont montré que les erreurs au niveau des caractères dans les données

OCR a un impact moindre sur la NER par rapport aux erreurs dans l'ordre des mots.

En (2012) ont évalué quatre outils différents pour la NER dans les textes historiques :

a) OpenGL

b) Stanford NER

c) Finkl

d) Open Calais

Ils ont utilisé les documents du projet EHRI comme entrée et défini trois types d'entités nommées : personne, lieu et organisation. Ils ont montré que Stanford NER obtenait la meilleure performance globale (en particulier pour les entités de type personne et lieu), tandis que Finkl fonctionnait le mieux pour les entités de type organisation dans les textes corrigés manuellement. OpenGL a affiché la plus faible précision.

En (2015) ont appliqué le système Stanford NER aux 155 millions d'articles Ocrisés issus de journaux historiques australiens afin de reconnaître les entités nommées de type personne, lieu et organisation. Ils ont montré comment ces données pouvaient être exploitées à l'aide d'une méthode de clustering.

En (2016) a créé un corpus ouvert pour la NER dans les journaux historiques Ocrisés en néerlandais, français et allemand, dans le cadre du projet Europe ana Europe ana. Ils ont utilisé Stanford NER pour le prétraitement des données en allemand, tandis que les autres langues ont été annotées manuellement. Ils se sont concentrés uniquement sur trois types d'entités nommées : personne, lieu et organisation.

En (2018) ont évalué plusieurs outils de NER (NER-Tagger, Stanford NER, Edinburgh Genopasse, Spacey, Polyglotte et leurs combinaisons) en fonction de leur capacité à identifier automatiquement les noms de lieux dans des correspondances historiques. Ils ont démontré quels systèmes de NER sont dépendants du corpus, et que le prétraitement et la traduction en anglais moderne n'affectaient pas significativement les résultats de la NER.

L'étude de la NER dans les textes historiques tchèques est encore peu explorée. En (2011) ont développé un analyseur morphologique pour le tchèque ancien . Ils ont utilisé une transformation du tchèque moderne vers le tchèque ancien et vice-versa. Ils ont ainsi « vieilli » un analyseur pour le tchèque moderne afin de l'utiliser pour annoter un corpus modernisé du tchèque ancien. Ensuite, ils ont remplacé les formes modernisées des mots par leurs formes originales, ce qui a permis d'obtenir le corpus final utilisé pour entraîner un tagger pour le tchèque ancien. L'avantage de cette approche est qu'elle minimise le travail spécifique à la langue, mais elle a donné de moins bons résultats (seulement 74 % de précision) par rapport aux méthodes plus traditionnelles, telles que le tagger More, qui a atteint 94 % de précision [28].

2.2.2 Définition du Traitement Automatique du Langage Naturel (TALN)

L'étude Le traitement automatique des langues (TAL ou TALN), en anglais Natural langage procession ou NLP, est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement du langage naturel pour diverses applications. Il ne doit pas être confondu avec la linguistique informatique, qui vise à comprendre les langues au moyen d'outils informatiques.

Le TALN est sorti des laboratoires de recherche pour être progressivement mis en œuvre dans des applications informatiques nécessitant l'intégration du langage humain à la machine [29].

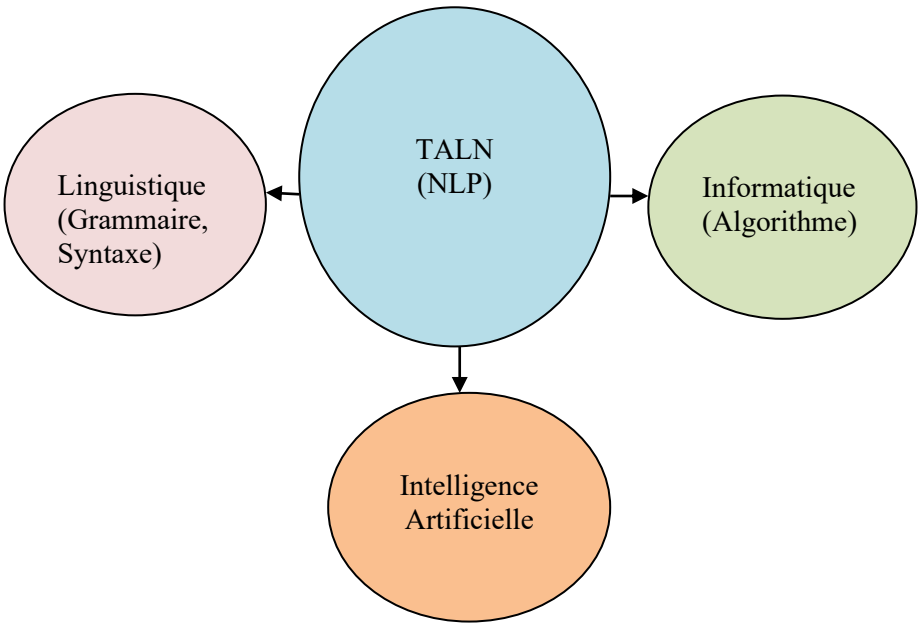


Figure 2. 1:TALN avec ses trois domaines fondamentaux

2.2.3 Définition de l'apprentissage automatique

Est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. On parle d'apprentissage statistique car l'apprentissage consiste à créer un modèle dont l'erreur statistique moyenne est la plus faible possible.

L'apprentissage automatique comporte généralement deux phases. La première consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. L'estimation du modèle consiste à résoudre une tâche pratique, telle que traduire un discours, estimer une densité de probabilité, reconnaître la présence d'un chat dans une photographie ou participer à la conduite d'un véhicule autonome. Cette

phase dite « d'apprentissage » ou « d'entraînement » est généralement préalable à l'utilisation pratique du modèle. La seconde phase est la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée [30].

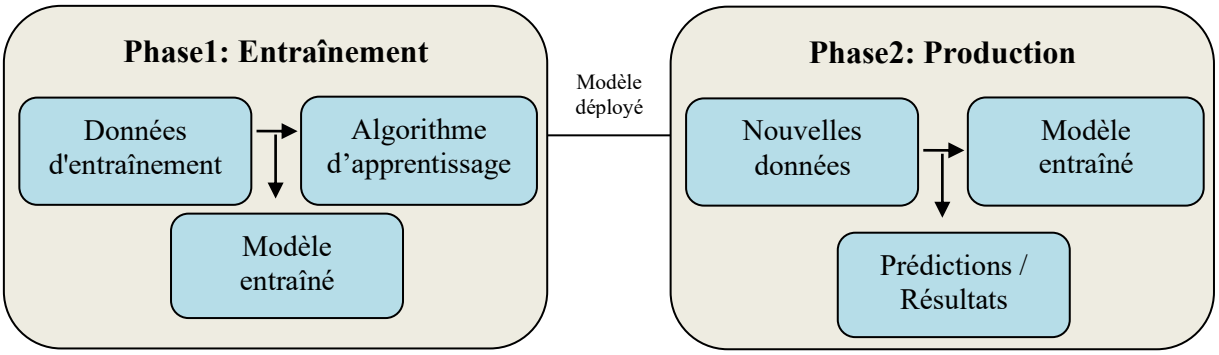


Figure 2. 2: Les deux phases essentielles d'apprentissage automatique

2.2.4 Approches hybrides combinant TALN et apprentissage automatique

L'approche hybride, qui combine le meilleur des approches basées sur des règles et de l'apprentissage automatique, intègre le NLP, le ML et l'intervention humaine. L'expérience humaine permet une analyse précise, tandis que l'apprentissage automatique facilite l'évolutivité de cette analyse. Les organisations peuvent utiliser à la fois le ML et les approches basées sur des règles de manière hybride, leur permettant ainsi de tirer parti des avantages des deux méthodes. L'approche hybride en NLP offre plusieurs options pour déterminer le meilleur chemin pour modéliser un texte et prendre des décisions commerciales plus rapidement. L'utilisation d'une stratégie hybride pour analyser des données textuelles non structurées est généralement une bonne pratique, en fonction de l'objectif recherché.

En utilisant l'apprentissage semi-supervisé pour automatiser l'étiquetage des données en fonction des entrées humaines dans les données d'entraînement, l'apprentissage automatique peut encore réduire le travail manuel de construction du modèle. Cette méthode ne nécessite même pas une grande quantité de données d'entraînement. Elle améliore la transparence du système pour garantir une expérience client positive et permet un suivi des indicateurs de performance clés (KPIs). De plus, cette approche offre une plus grande flexibilité, intégrabilité et rapidité, ce qui réduit la charge sur les ressources. Elle est adaptée aux projets conversationnels et orientés tâches.

Le langage humain est un problème complexe pour lequel les entreprises recherchent depuis longtemps une solution optimale. Cela peut enfin être résolu grâce à une approche NLP mixte. Les techniques hybrides peuvent améliorer les résultats de nos applications NLP dans divers cas. Par exemple, si nous créons un correcteur grammatical, nous aurons besoin d'un module capable de

reconnaître des expressions idiomatiques comme "passer l'arme à gauche", ainsi que d'un module basé sur des règles qui détecte les erreurs et propose des corrections appropriées. C'est un exemple d'une approche hybride [31].

2.3 Préservation Numérique et archivage

2.3.1 Formats modernes d'archivage numérique

À l'ère numérique, où les documents papier sont progressivement remplacés par des fichiers électroniques, l'archivage électronique est devenu essentiel, voire obligatoire, pour de nombreuses entreprises et organisations. Pour garantir la légalité, l'intégrité et la durabilité des documents numériques, il est primordial d'utiliser des formats d'archivage appropriés. Dans cet article, nous explorerons les différents formats d'archivage électronique légal les plus couramment utilisés, en mettant l'accent sur leurs avantages et leurs applications.

PNG : est un format image, mono page et sans intérêt particulier pour la numérisation.

JPEG : défini par et pour la photographie, ce format présente des inconvénients pour la numérisation de documents. Les techniques de compression utilisées par JPEG ne sont pas optimales pour le texte et ce format n'est pas multipage [32].

PDF (Portable Document Format) : est l'un des formats d'archivage électronique les plus répandus et les plus populaires. Développé par Adobe Systems, le PDF est conçu pour préserver la mise en page, les polices, les images et les liens hypertextes, offrant ainsi une représentation fidèle des documents originaux. Il est largement utilisé pour archiver des contrats, des rapports financiers, des factures et d'autres documents juridiques. Un aspect essentiel du PDF est sa non-modifiabilité, ce qui garantit que les informations archivées restent inchangées et authentiques.

PDF\A (Archival) : est une variante du format PDF spécialement conçue pour l'archivage à long terme. Conformément aux normes ISO 19005, le PDF/A élimine les fonctionnalités potentiellement problématiques du PDF standard, telles que les scripts et les liens vers des ressources externes. En conséquence, il garantit une conservation plus sûre et plus durable des documents électroniques. Les organisations gouvernementales, les bibliothèques et les archives préfèrent souvent le PDF/A pour garantir l'intégrité des documents sur de nombreuses décennies.

PDF\X (Exchange) : est une norme spécifique du PDF utilisée principalement dans l'industrie de l'impression. Bien que son objectif principal soit de garantir que les fichiers sont prêts pour la production d'impression, il peut également être utilisé pour l'archivage électronique de documents graphiques tels que des illustrations, des maquettes de conception et des projets créatifs.

PDF\UA (Universal Accessibility):est un format d'archivage électronique destiné à garantir l'accessibilité des documents aux personnes en situation de handicap. Conformément aux directives

de l'ISO 14289, le PDF/UA encourage la création de documents électroniques qui peuvent être utilisés avec des technologies d'assistance, telles que les lecteurs d'écran, pour les utilisateurs malvoyants ou malentendants.

TIFF (Tagged Image File Format) : est couramment utilisé pour l'archivage électronique d'images et de documents visuels. Il prend en charge une large gamme de résolutions, de profondeurs de couleurs et de compressions, ce qui en fait un format polyvalent pour la conservation d'images haute qualité. Le TIFF est couramment utilisé dans les industries créatives, les services d'archivage de photographies et d'autres domaines où la qualité visuelle est cruciale.

XML (Extensible MarkupLanguage) : est un format de données structuré qui facilite le stockage, la transmission et l'interprétation des informations. En tant que format d'archivage électronique, XML est particulièrement adapté aux documents qui nécessitent des métadonnées et une structuration détaillée. Les documents archivés au format XML sont faciles à traiter automatiquement, ce qui les rend idéaux pour les grandes organisations qui gèrent d'importants volumes de données.

EML (Email Message Format): L'archivage électronique de courriers électroniques est devenu essentiel pour de nombreuses entreprises et organisations. Le format EML (Email Message Format) est largement utilisé pour archiver des courriers électroniques, y compris les pièces jointes et les métadonnées. En préservant la structure et le contenu des messages électroniques, le format EML assure l'intégrité des communications importantes. Cela est particulièrement utile pour les entreprises qui doivent se conformer à des réglementations spécifiques concernant la conservation des communications électroniques.

CSV (Comma-Separated Values) : est couramment utilisé pour l'archivage électronique de données tabulaires. Enregistrant les données sous forme de texte simple, où chaque ligne correspond à un enregistrement et les valeurs sont séparées par des virgules, le format CSV offre une solution simple et légère pour stocker des ensembles de données structurées. Il est souvent utilisé pour archiver des feuilles de calcul, des données de base de données et des rapports d'activités.

DOCX (Microsoft World Open XML): est le format de fichier utilisé par Microsoft Word depuis la version 2007. Bien que souvent utilisé pour créer et éditer des documents, le format DOCX peut également être un bon choix pour l'archivage électronique. En stockant les documents sous forme de fichiers XML, le DOCX facilite la récupération des données structurées et permet une meilleure compatibilité avec d'autres applications de traitement de texte [33].

2.4 Évaluation des performances d'extraction d'information

2.4.1 Métriques classiques d'évaluation (précision, rappel, F1-score)

Les métriques classiques telles que la précision, le rappel et le F1-score sont fondamentales pour évaluer la performance des systèmes d'extraction d'information. Elles permettent de mesurer respectivement la proportion de résultats pertinents parmi les résultats proposés, la proportion de résultats pertinents correctement identifiés, et la moyenne harmonique de la précision et du rappel [34].

Tableau 2. 1: Tableau de contingence

	Classe prédite : +	**Classe prédite : +**
Classe réelle : +	VP (Vrai Positif)	FN (Faux Négatif)
Classe réelle : –	FP (Faux Positif)	VN (Vrai Négatif)

La précision et le rappel sont mesurés par la qualité d'un classificateur :

- Précision pour les positifs $=\frac{v_p}{v_p+f_p}$

- Précision pour les négatifs $=\frac{v_n}{v_n+f_n}$

- Rappel pour les positifs $=\frac{v_p}{v_p+f_n}$

- Rappel pour les négatifs $=\frac{v_n}{v_n+f_p}$

• La précision mesure la proportion d'exemples vraiment positifs (resp. négatifs) parmi ceux qui sont classés comme positifs (resp. négatifs).

• Le rappel mesure la proportion d'exemples vraiment positifs(resp. négatifs) parmi tous les exemples de classes positive (resp. négative).

• Il est toujours plus pratique de manipuler un seul nombre qui synthétise les autres. Ainsi la mesure -F est définie par :

$$F = \frac{2 \times \text{Rappel} \times \text{Précision}}{\text{Rappel} + \text{Précision}} = \frac{2 \times v_p}{2 \times v_p + f_p + f_n}$$

2.4.2 Évaluation humaine vs automatique

L'évaluation humaine reste précieuse pour valider des aspects subtils ou contextuels de l'information extraite, tandis que l'évaluation automatique permet une mesure rapide, reproductible

et objective des performances des modèles. Le tableau ci-dessous résume les différences entre ces deux approches [35].

Tableau 2.2: comparatif entre Évaluation humaine et Évaluation automatique

Critère	**Évaluation humaine**	**Évaluation automatique**
Fiabilité	Dépend des compétences et biais humains.	Objective, basée sur des métriques mesurables.
Temps nécessaire	Long, fastidieux.	Rapide, instantané après le traitement.
Cohérence	Variable selon l'annotateur.	Constante sur tout l'ensemble de test.
Coût	Élevé (nécessite des experts).	Faible (automatisé une fois implémenté).
Flexibilité	Peut juger des aspects complexes (style, lisibilité, etc.).	Limité aux critères définis par l'algorithme.
Adaptabilité multilingue	Moyenne à bonne (si annotateurs compétents).	Bonne si les outils prennent en charge plusieurs langues.
Exemples d'utilisation	Vérification manuelle de résultats OCR ou NER.	Calcul automatique du F1-score, précision, rappel, etc.

2.5 Conclusion

Ce chapitre a mis en lumière les enjeux techniques et méthodologiques liés à l'extraction d'informations à partir de documents anciens et à leur archivage structuré. L'efficacité de ces systèmes repose sur une combinaison d'outils linguistiques, d'algorithmes intelligents et de choix judicieux en matière de formats numériques. L'évaluation continue des performances permet d'orienter les améliorations futures, en particulier dans les contextes multilingues ou faiblement dotés en ressources.

Chapitre 03:

Approche Proposée : Conception et mise en œuvre

3.1 Introduction

La reconnaissance optique de caractères (OCR) multilingue représente un défi majeur dans le contexte de la digitalisation documentaire actuelle. Les solutions traditionnelles présentent des limitations importantes : traitement monolingue, qualité d'extraction variable et formats d'export limités.

Ce chapitre présente une approche innovante pour développer un système OCR multilingue hybride, capable de traiter simultanément des documents en français, anglais et arabe. Notre proposition s'articule autour d'une architecture modulaire intégrant des technologies de pointe en traitement d'image et reconnaissance de caractères.

3.2 Notre Approche Proposée

3.2.1 Vision Globale

Notre approche consiste à développer un système OCR multilingue hybride combinant plusieurs technologies de reconnaissance optique de caractères pour optimiser l'extraction de texte à partir de documents numériques. Le système proposé intègre une architecture modulaire permettant de traiter efficacement des documents en français, anglais et arabe, tout en offrant des fonctionnalités avancées de prétraitement d'images et d'évaluation de qualité.

3.2.2 Objectifs de l'Approche

L'approche proposée vise à surmonter les défis majeurs rencontrés dans le domaine de la reconnaissance optique de caractères (OCR) multilingue, en mettant en place un système avancé et performant. Parmi les principaux objectifs figurent le multilinguisme avancé**,** qui permet un support natif pour trois langues l'arabe, le français et l'anglais en utilisant des algorithmes spécialisés adaptés aux spécificités des écritures latines et arabe.

Le système intègre également une fonction d'optimisation automatique grâce à des techniques de traitement d'image adaptatives, visant à améliorer la qualité des documents numérisés avant leur analyse OCR. En outre, il assure une flexibilité de formats, en traitant uniformément divers types de fichiers tels que les images (PNG, JPG, JPEG, GIF) et les PDF multipages. Une attention particulière est portée à l'évaluation quantitative, avec des métriques permettant de mesurer objectivement la qualité du texte extrait ainsi que l'efficacité du prétraitement. Enfin, une exportation polyvalente vers plusieurs formats standards (PDF, DOCX, TXT, JSON, XML) garantit une préservation fidèle de la mise en forme originale, facilitant l'utilisation ultérieure des résultats.

3.2.3 Innovation de l'Approche

Cette approche se distingue par plusieurs innovations technologiques et méthodologiques. La première réside dans l'architecture hybride adoptée, combinant intelligemment deux moteurs OCR

puissants : EasyOCR pour le traitement du texte arabe et PaddleOCR pour les langues latines (français et anglais), afin d'exploiter au mieux leurs performances respectives. Cette combinaison offre une solution robuste et précise pour les documents multilingues. De plus, un prétraitement adaptatif basé sur des techniques avancées telles que CLAHE (Contrast Limited Adaptive Histogram Equalization) et le débruitage bilatéral a été mis en œuvre, optimisé selon les caractéristiques spécifiques de chaque document. Cela améliore considérablement la qualité visuelle et facilite une extraction OCR efficace. Par ailleurs, une évaluation métrique innovante a été développée, intégrant des algorithmes propriétaires capables de quantifier objectivement la clarté et la performance globale du processus d'extraction. Enfin, la gestion contextuelle du système permet un traitement différencié des documents en fonction de leur type et de leur langue détectée automatiquement, ce qui renforce l'automatisation et l'adaptabilité du système à différents cas d'utilisation.

3.3 Architecture du Modèle Proposé

L'architecture ci-dessous représente la conception détaillée de notre système :

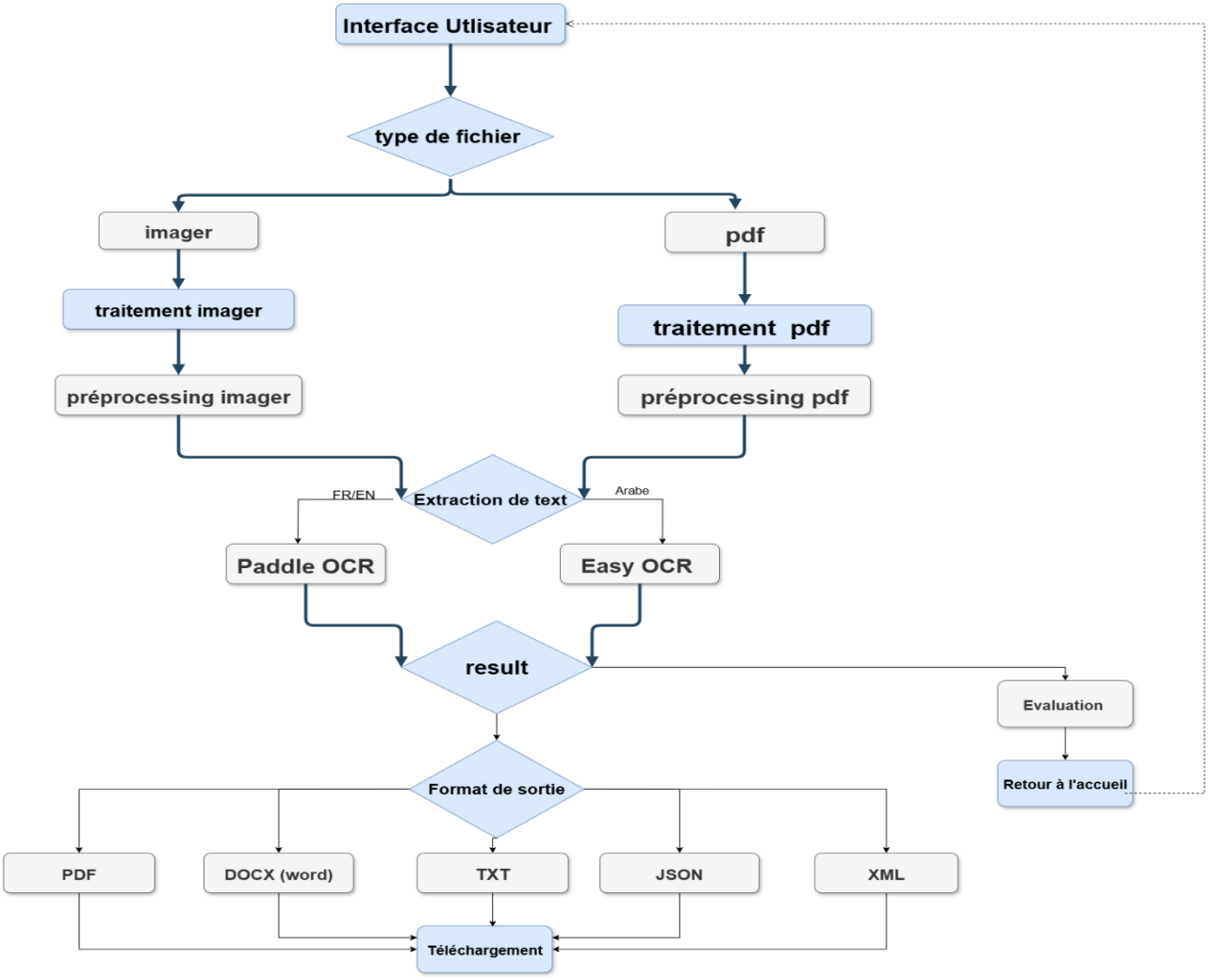


Figure 3. 1: Architecture générale du système proposé

3.4 Étapes de l'approche proposée

3.4.1 Étape 1 : Réception et validation des fichiers

Lorsqu'un utilisateur sélectionne un fichier via l'interface web Flask, le système vérifie d'abord que le fichier existe. Ensuite, il valide son extension parmi une liste blanche comprenant PNG, JPG, JPEG, GIF et PDF. Le nom du fichier est sécurisé avec la fonction secure_filename() pour éviter toute injection malveillante. Puis, le fichier est sauvegardé dans un répertoire sécurisé avec un identifiant unique généré via UUID. Enfin, le type du document (image ou PDF) est détecté automatiquement pour orienter le traitement. La figure suivante représente une explication du processus :

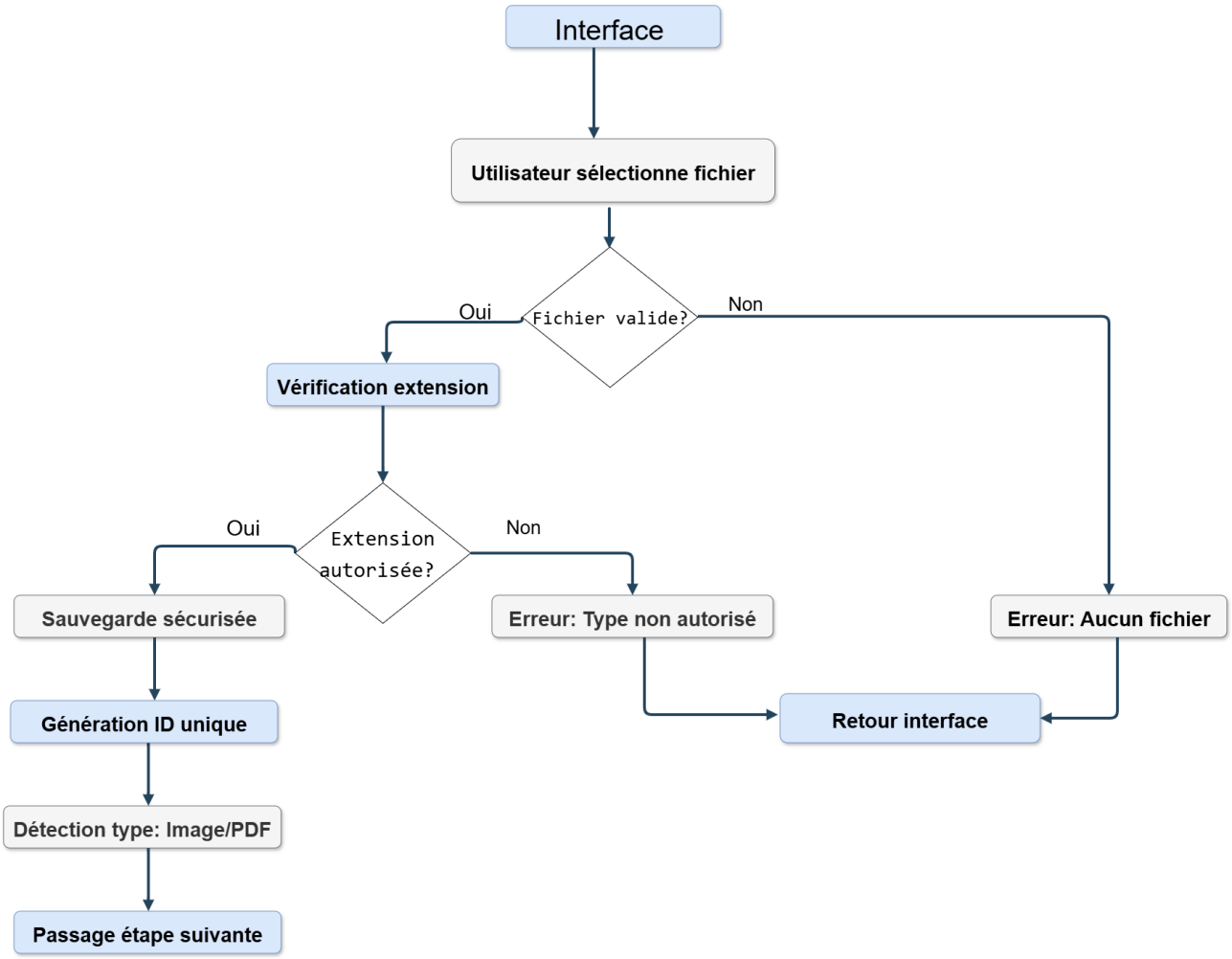


Figure 3.2: Architecture de Réception et Validation des Fichiers

3.4.2 Étape 2 : Prétraitement et Amélioration d'Image

3.4.2.1 CLAHE (Contrast Limited Adaptive Histogram Equalization)

Est une technique de traitement d'image utilisée pour améliorer le contraste d'une image. CLAHE fonctionne sur de petites régions de l'image, appelées tuiles, plutôt que sur l'image entière. L'application de l'histogramme est distincte de l'histogramme global, car la méthode adaptative consiste à calculer une fonction du contraste local pour modifier chaque tuile. Le contraste de chaque tuile est ensuite augmenté à l'aide de l'histogramme, selon une approximation de la région basée sur une valeur redistribuée. Les tuiles voisines sont ensuite fusionnées pour éviter l'apparition de bords artificiels entre les régions. Le contraste, en particulier dans les zones homogènes, est limité afin de ne pas amplifier le bruit qui pourrait être présent dans l'image [36].

3.4.2.2 Filtre bilatéral

Est une technique simple de combinaison non linéaire, il est défini comme une moyenne pondérée des pixels proches, d'une manière très semblable à la convolution gaussienne. La différence c'est que le filtre bilatéral prend en compte la différence de valeur avec les voisins pour conserver les contours (les bords) tout en lissant l'image. L'idée clé de ce filtre est que pour qu'un pixel influence sur un autre pixel, il devrait non seulement occuper un lieu voisin mais aussi avoir une valeur similaire [37]. Donc, les coefficients de ce filtre dépendent non seulement de la distance euclidienne au point à lisser, mais aussi des similitudes photométriques (la ressemblance des intensités entre le pixel central et le pixel du masque) [38].

L'expression de ce filtre est définie comme suit :

$$\hat{u} = \frac{1}{c(x)} \sum_{\beta\rho} \exp - \frac{|x-y|^2}{2\rho^2} \exp - \frac{|y(x)-y(y)|^2}{2h^2} u(y)$$

Avec :

$$c(x) = \sum_{\beta\rho} \exp - \frac{|x-y|^2}{2\rho^2} \exp - \frac{|y(x)-y(y)|^2}{2h^2}$$

Où :

$\hat{u}$(x): la valeur estimée du pixel à la position x.

C(x): Le coefficient de normalisation qui normalise la somme des poids des pixels (égale à 1).

βp: Représente la fenêtre de convolution.

x: la position 2D d'un pixel dans le plan de l'image.

y: est un ensemble de coordonnées 2D dans le masque.

(x): représente l'intensité du pixel à la position x dans l'image bruitée.

(y): représente l'intensité du pixel à la position y dans l'image bruitée.

ρet h : sont respectivement la déviation standard (écart type) de la distribution gaussienne des poids géométriques (voisinage) et la déviation standard de la distribution gaussienne des poids d'intensités.

Nous avons utilisé dans deuxième étape le prétraitement améliore la qualité de l'image pour optimiser la reconnaissance OCR. L'image est convertie du format BGR vers l'espace LAB, permettant de travailler sur la luminosité indépendamment des couleurs. Une égalisation adaptative d'histogramme (CLAHE) est appliquée au canal de luminosité pour améliorer le contraste local. Ensuite, un filtre bilatéral réduit le bruit tout en préservant les contours nets. Cette étape assure une meilleure précision lors de l'extraction de texte. Comme le montre dans la figure ci-dessous :

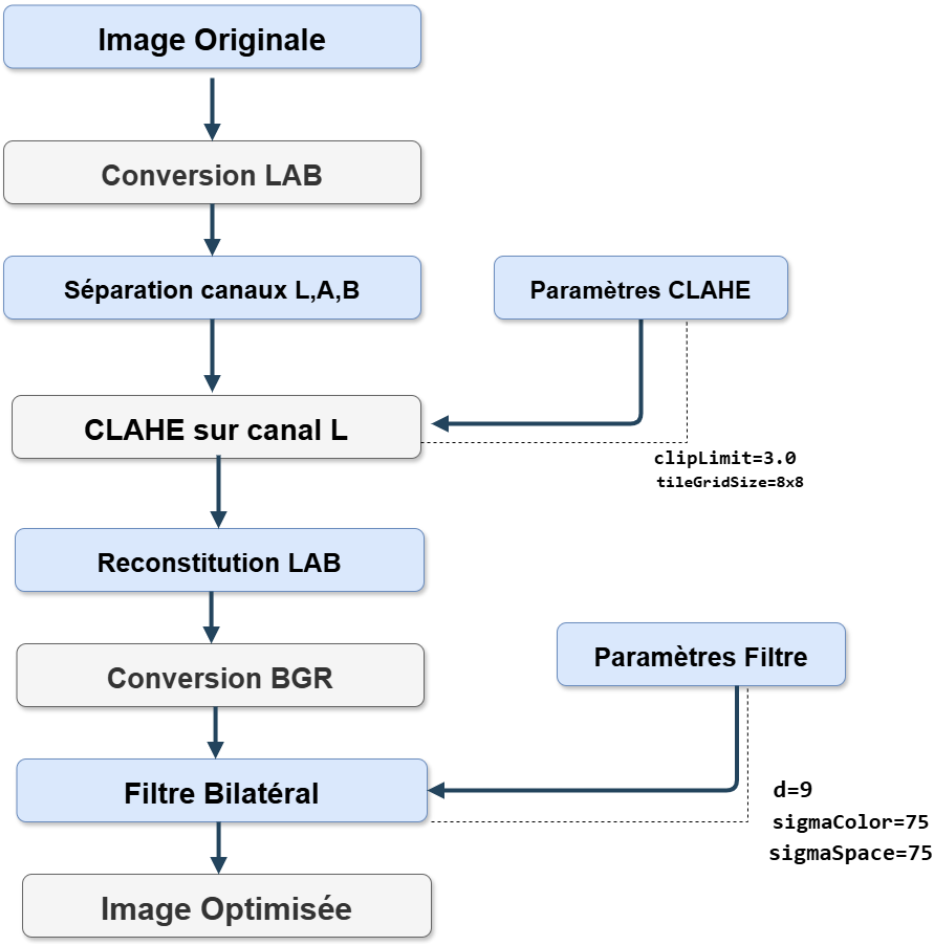


Figure 3.3: Architecture de prétraitement et amélioration d'image

3.4.3 Étape 3 : Traitement Différencié par Type de Document

3.4.3.1 EasyOCR

EasyOCR est un outil d'Optical Character Recognition (OCR) open-source développé par Jaided AI, une startup basée en Thaïlande. C'est une bibliothèque basée sur Python qui permet aux machines de lire et d'interpréter du texte à partir d'images, de fichiers PDF et d'autres documents.

EasyOCR a connu une grande popularité grâce à sa simplicité, sa facilité d'utilisation et sa prise en charge de plusieurs langues[39].

3.4.3.2 PaddleOCR

Paddle OCR est un kit d'outils open-source de reconnaissance optique de caractères (OCR) développé par PaddlePaddle, une plateforme d'intelligence artificielle et d'apprentissage profond. Le kit Paddle OCR est conçu pour être une solution d'OCR rapide et précise, permettant de reconnaître et d'extraire du texte à partir d'images et de documents [40].

Nous avons divisé la troisième étape en deux sections :

Pour les images, EasyOCR est utilisé pour les textes arabes grâce à ses modèles spécialisés, tandis que PaddleOCR est choisi pour les langues latines comme le français et l'anglais. Chaque moteur est adapté à sa langue pour garantir des résultats précis. Comme le montre dans la figure :

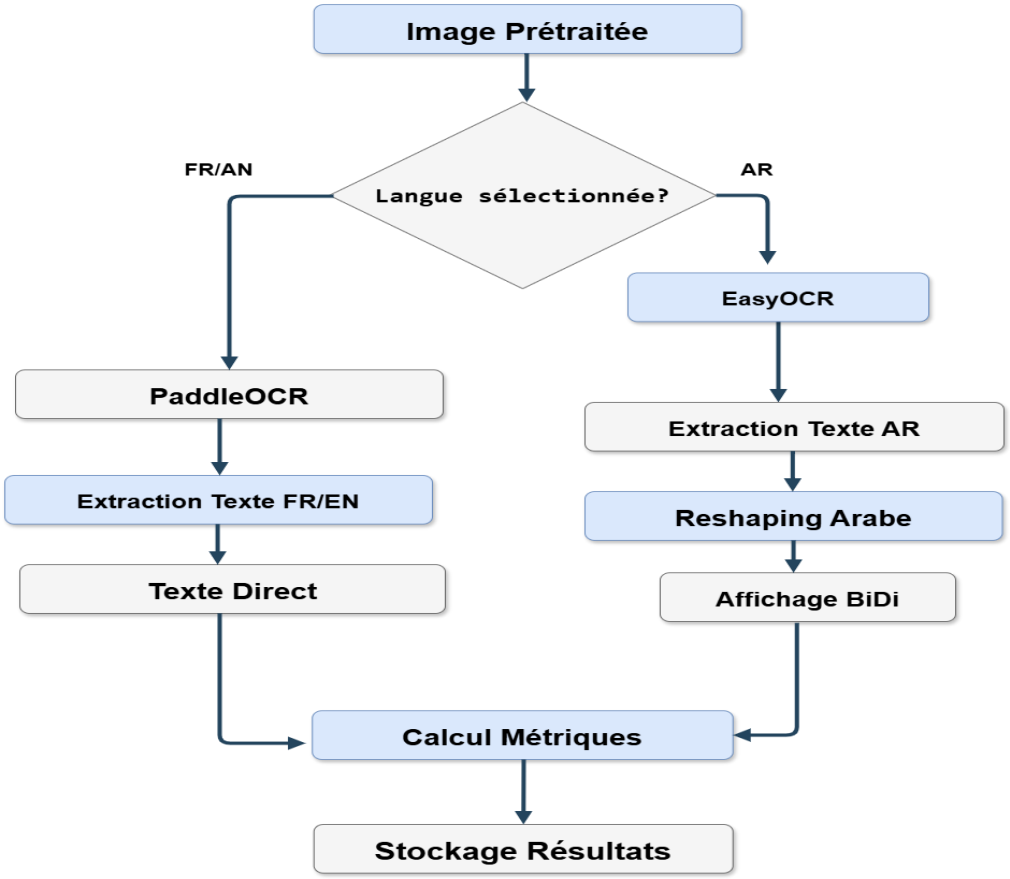


Figure 3.4: Architecture de traitement différencié par type d'image

Pour les PDF, une extraction directe du texte est tentée. Si elle échoue ou si le texte est insuffisant, chaque page est convertie en image haute résolution avant un traitement OCR similaire aux images. Les résultats sont agrégés pour former le document final. Comme le montre la figure :

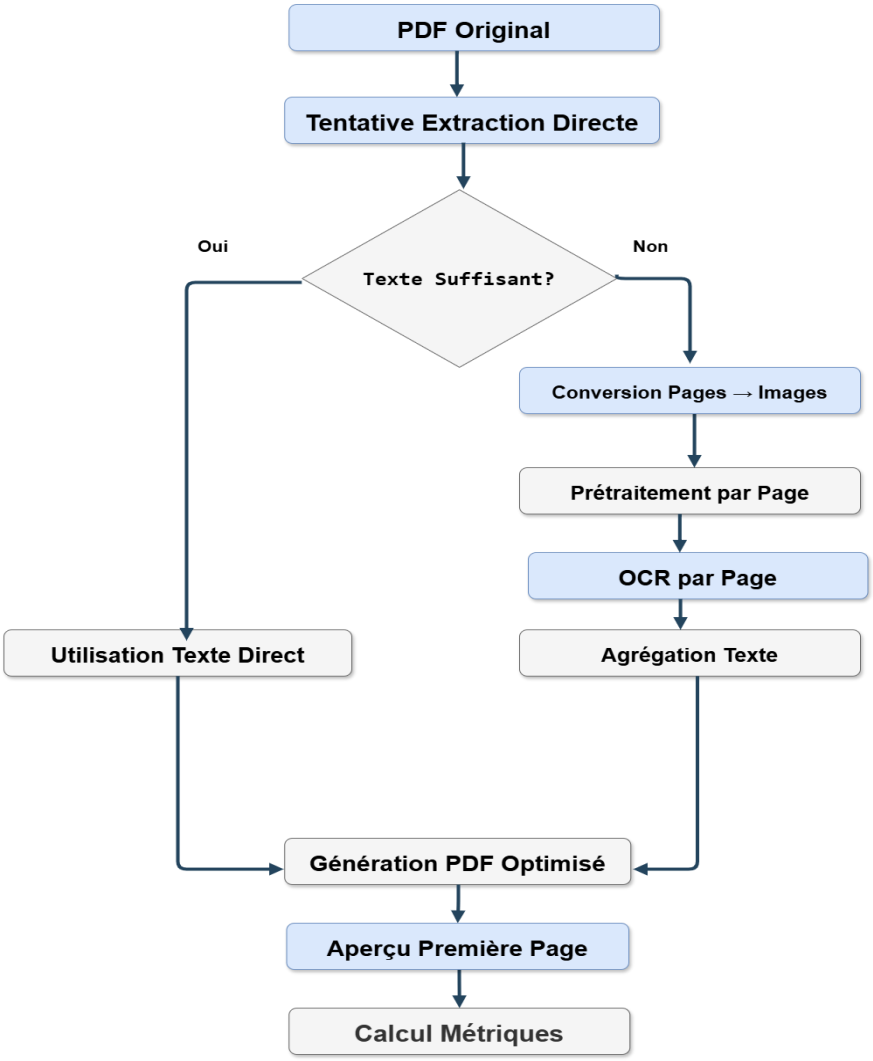


Figure 3. 5: Architecture de traitement différencié par type de PDF

3.4.4 Étape 4 : Évaluation de la qualité

La qualité de l'image est mesurée via la variance du Laplacien, qui indique la netteté des contours. Cette mesure est normalisée sur une échelle de 0 à 100 pour faciliter l'interprétation. La qualité du texte extrait est évaluée en analysant sa longueur, la cohérence des structures et la présence de ponctuation. Ces indicateurs aident à juger la performance globale du processus OCR (Figure 3.6).

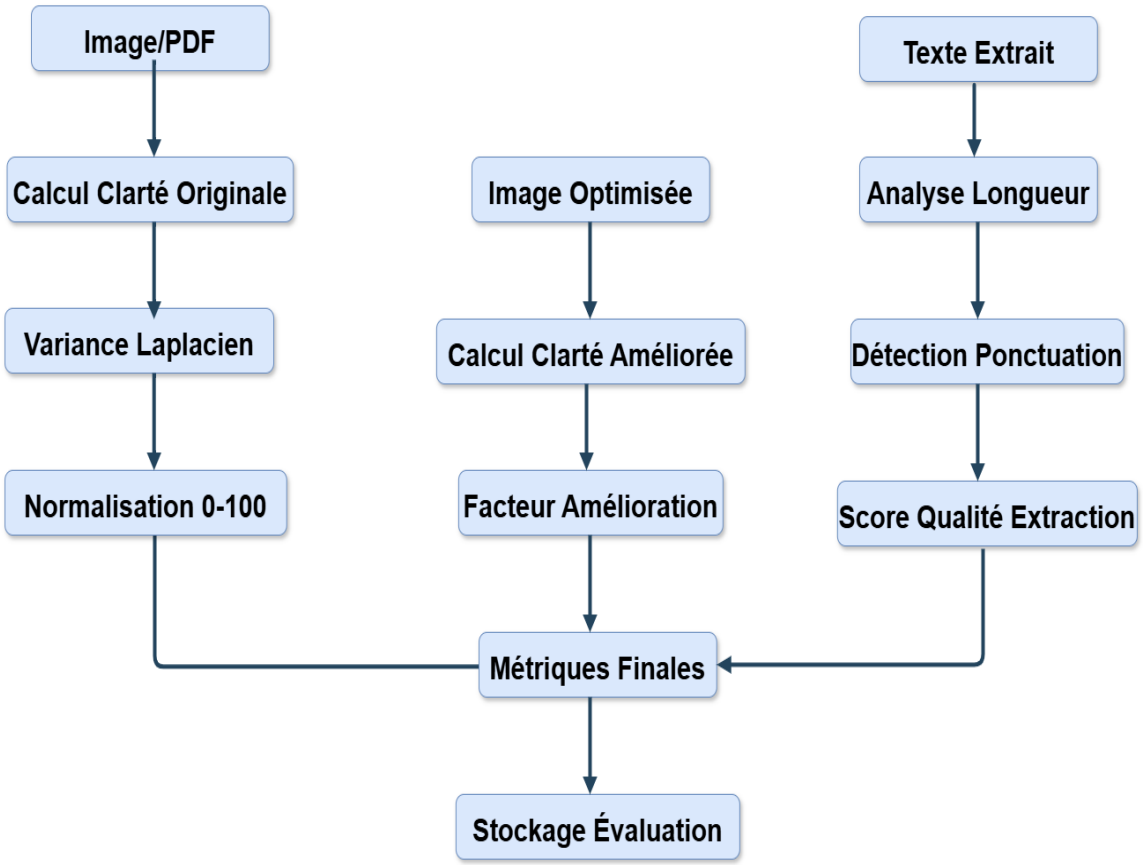


Figure 3. 6: Architecture d'évaluation de la qualité

3.4.5 Étape 5 : Post-traitement et formatage

Après l'extraction, le texte subit un post-traitement adapté à la langue. Pour l'arabe, ArabicReshaper corrige les formes contextuelles des lettres, et l'algorithme BiDi gère l'affichage bidirectionnel (de droite à gauche). Pour les langues latines, le texte est nettoyé des erreurs courantes d'espacement et de caractères. Le contenu est ensuite structuré selon le format demandé, avec encodage UTF-8 et organisation logique. Comme le montre dans la figure :

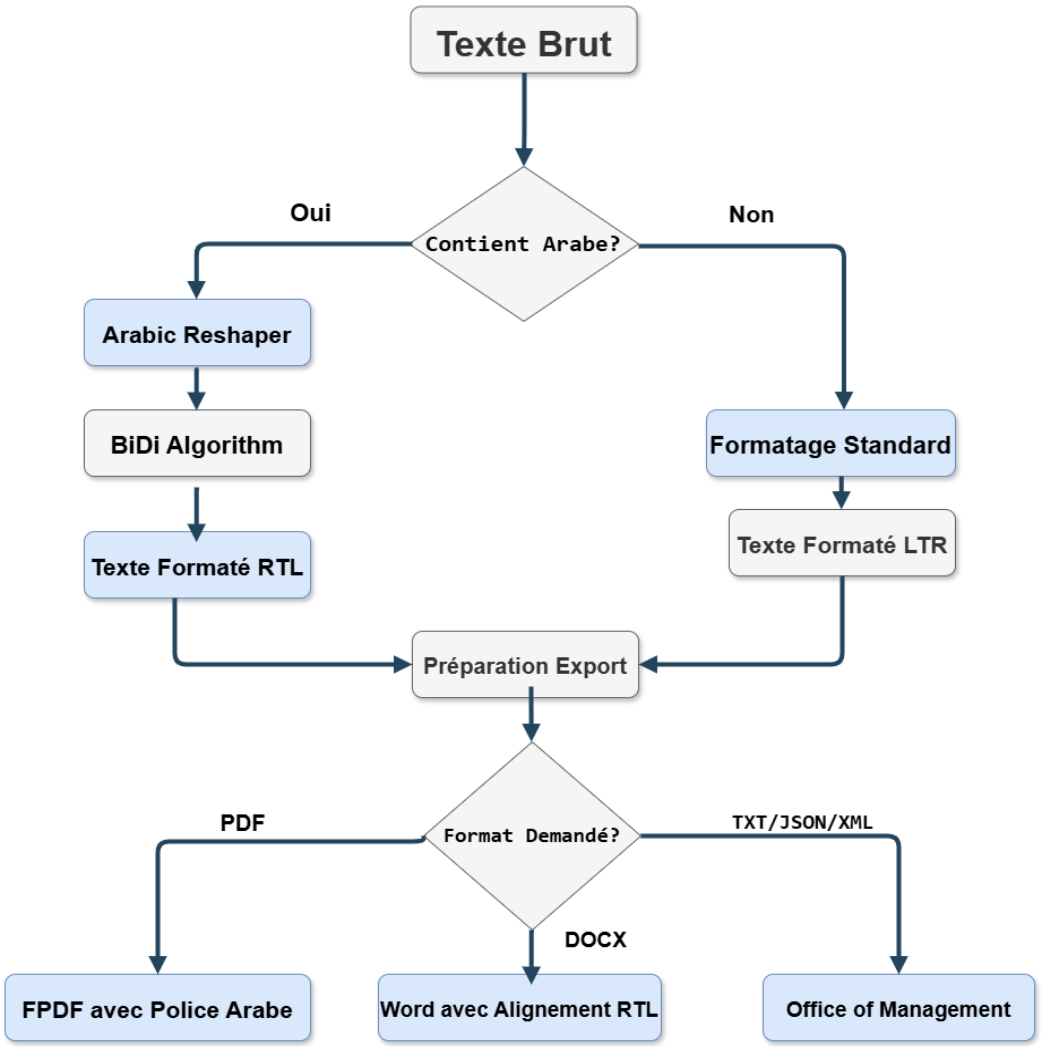


Figure 3. 7: Architecture de Post-traitement et Formatage

3.4.6 Étape 6 : Génération et export des résultats

Les résultats sont exportés dans les formats souhaités (PDF, DOCX, TXT, JSON, XML). Pour PDF, FPDF gère l'alignement et les polices arabes ou latines. DOCX utilise python-docx pour un formatage correct dans Word. JSON/XML inclut des métadonnées enrichies pour faciliter l'intégration. TXT préserve l'encodage et la structure. Chaque format contient également des informations sur le traitement effectué. Comme le montre dans la figure :

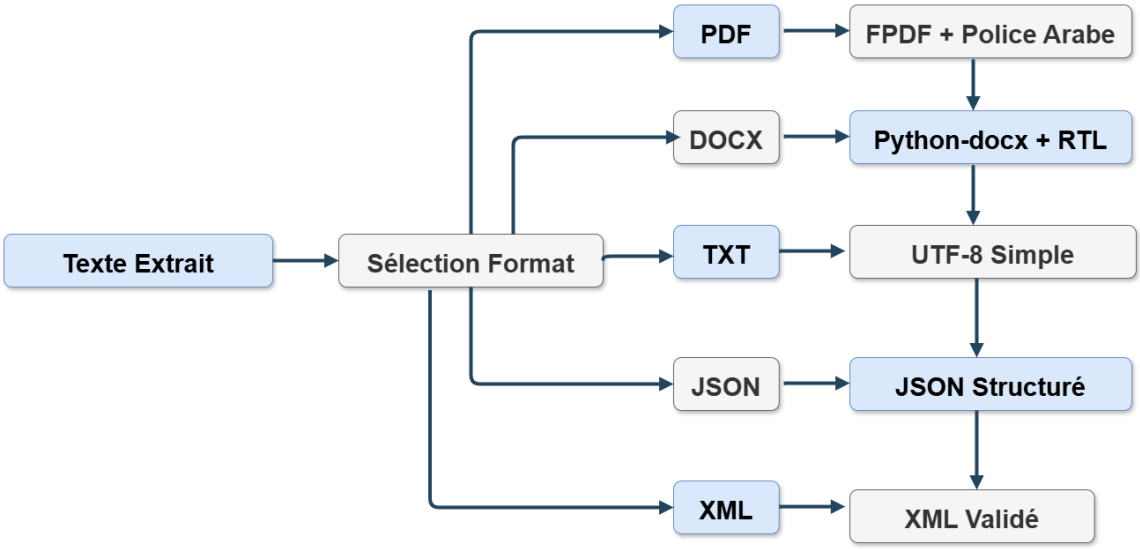


Figure 3. 8: Architecture de Génération et Export des Résultats

❖ **Gestion des erreurs et optimisations**

Le système gère les erreurs de manière robuste, en distinguant leur type et leur gravité. Des solutions automatiques ou des alertes utilisateurs sont proposées. Les moteurs OCR sont mis en cache pour éviter de les redémarrer à chaque traitement. Enfin, le système libère les ressources après chaque opération, assurant une utilisation efficace de la mémoire.

3.5 Conclusion

L'approche proposée constitue une avancée significative dans le domaine de l'OCR multilingue. En combinant EasyOCR et PaddleOCR dans une architecture hybride, notre système offre une solution robuste pour le traitement de documents multilingues. Les innovations introduites - prétraitement adaptatif CLAHE, filtrage bilatéral, évaluation quantitative de qualité et gestion contextuelle - permettent d'atteindre des niveaux de précision supérieurs aux solutions existantes. L'architecture modulaire en six étapes garantit une approche optimisée depuis la validation des fichiers jusqu'à l'export multiformat. Cette contribution ouvre la voie à de nouveaux développements dans le domaine de l'OCR multilingue, établissant les bases pour des systèmes plus performants et adaptés aux besoins professionnels actuels.

Chapitre 04 :

Réalisation et Évaluation des performances

4.1 Introduction

Ce chapitre présent l'implémentation pratique de notre application de reconnaissance optique de caractères (OCR) multilingue. Nous détaillons l'environnement de développement utilisé, comprenant les outils matériels et logiciels qui ont permis la réalisation de ce projet. L'application développée vise à traiter des documents en trois langues principales : l'arabe, le français et l'anglais, en proposant une interface utilisateur intuitive et des fonctionnalités de prétraitement d'images pour améliorer la qualité de reconnaissance.

Notre approche se base sur l'utilisation de technologies modernes telles que Google Colab pour le développement, Python comme langage de programmation principal, et diverses bibliothèques spécialisées pour le traitement d'images et la reconnaissance de texte.

4.2 Environnement de travail

4.2.1 Matériel utilisé

Notre travail s'est effectué sur un micro-ordinateur ayant les caractéristiques suivantes :

- Processeur : Intel(R) Core (TM) i3-4030U CPU @ 1,90 GHz.
- RAM : 4 Go.
- Système d'exploitation : Windows 10 Professionnel.

4.2.2 Outils de développement

4.2.2.1 Google Colab

Colab (abréviation de Colaboratory) est un produit proposé par Google Research qui permet chercheurs en apprentissage automatique pour travailler sur des projets dans le navigateur. Similaire à Google Docs, Il nous permet de partager des projets entre plusieurs personnes, et le meilleur de tous, qu'il offre un accès gratuit aux GPU afin que nous puissions former des modèles rapidement sans aucun enregistrement.

Comme définition générale, Colab est un environnement de notebook gratuit qui s'exécute entièrement dans le cloud. Mieux encore, il ne nécessite aucune configuration et les blocsnotes que nous créons peuvent être modifiés simultanément par les membres de notre équipe, tout comme nous modifions des documents dans Google Docs. Colab prend en charge de nombreuses bibliothèques de machine learning populaires qui peut être facilement chargées sur notre ordinateur [41].La figure ci-dessous présente l'interface de googlecolab.

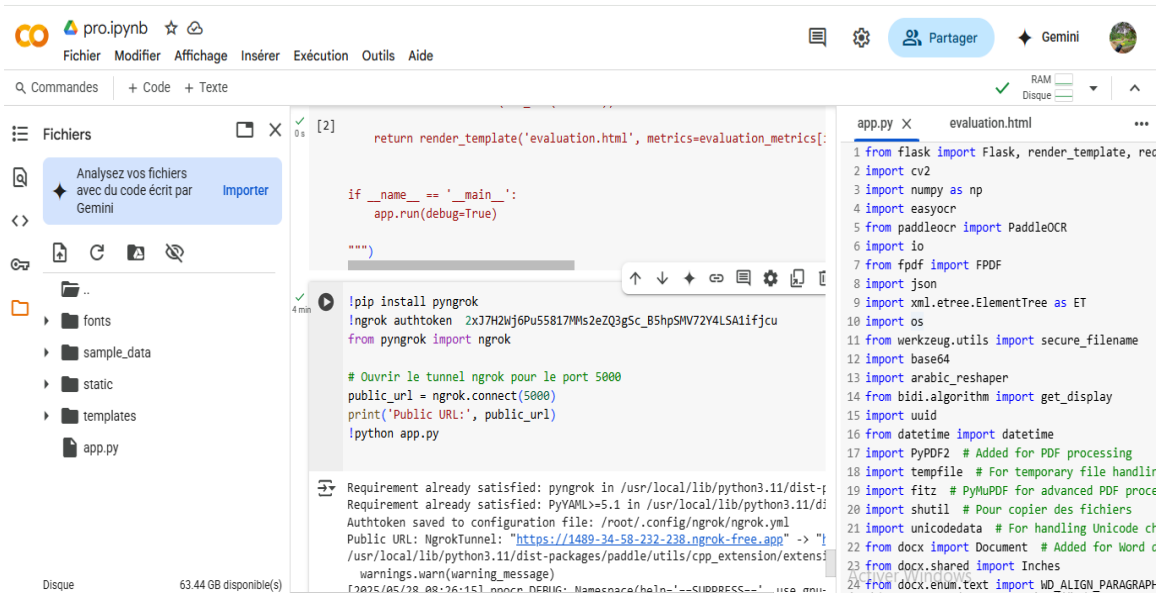


Figure 4. 1: Google Colab capture d'écran

4.2.2.2 NGROK:

Ngrok est la porte d'entrée de notre application. Il est un reverse proxy (serveur mandataire inversé) distribué à l'échelle mondiale qui sécurise, protège et accélère vos applications et services réseau, peu importe où vous les exécutez. ngrok prend en charge la diffusion d'applications basées sur HTTP, TLS ou TCP.

Vous pouvez utiliser ngrok en phase de développementpour tester des webhooks, ou en environnement de productioncomme une passerelle API (API Gateway), un contrôleur d'entrée Kubernetes (KubernetesIngress), ou un proxy sensible à l'identité (Identity-Aware Proxy). Vous pouvez également exécuter ngrok pour créer facilement une connexion sécurisée vers des API situées sur les réseaux de vos clients ou sur des appareils déployés sur le terrain [42]. La figure 4.2. présente l'interface de ngrok.

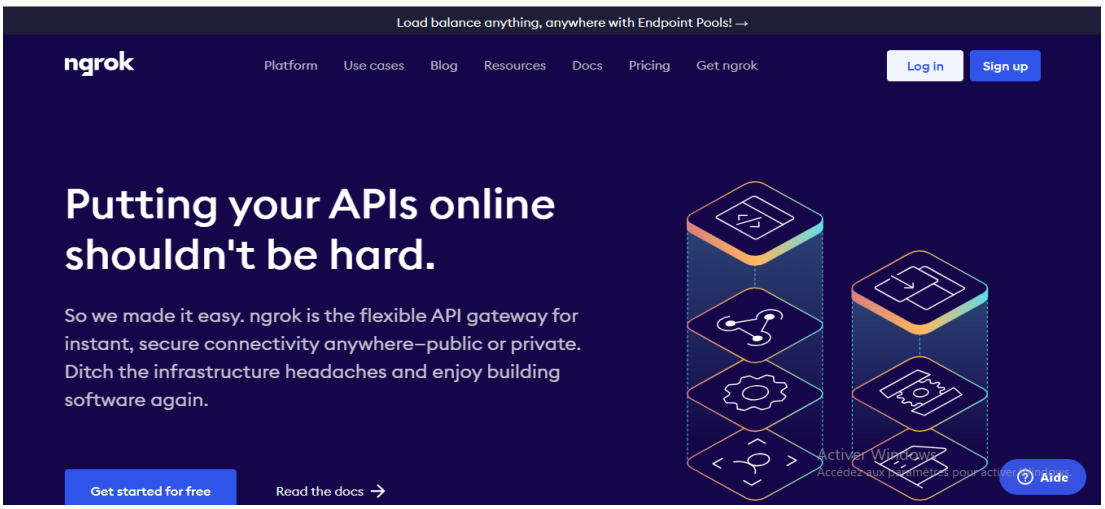


Figure 4. 2: Capture d'écran Ngrok

4.2.2.3 Langage de programmation

Python est le langage de programmation open source le plus fréquemment employé par les professionnels de l'informatique. Ce langage se distingue dans les domaines tels que la gestion des infrastructures, l'analyse de données ou encore le développement logiciel. Effectivement, grâce à ses atouts, Python offre aux développeurs la possibilité de se focaliser sur leur travail plutôt que sur la méthode employée. Il libère les développeurs des restrictions formelles des anciens langages qui affligeaient leur époque. Ainsi, coder en Python est plus rapide comparé à d'autres langages de programmation.

- Bibliothèques en Python :

Normalement, une bibliothèque est une collection de livres ou une pièce ou un lieu où de nombreux livres sont stockés pour être utilisés plus tard. De même, dans le monde de la programmation, une bibliothèque est une collection de codes précompilés qui peuvent être utilisés ultérieurement dans un programme pour certaines opérations spécifiques bien définies. Outre les codes précompilés, une bibliothèque peut contenir de la documentation, des données de configuration, des modèles de message, des classes et des valeurs, etc [43].

- Bibliothèques utilisé:

Nous en avons utilisé beaucoup, notamment les suivants :

Flask : est un framework Web python. Cela signifie que flask vous fournit des outils, des bibliothèques et des technologies qui vous permettent de créer une application Web. Cette application Web peut être des pages Web, un blog, un wiki ou devenir aussi grande qu'une application de calendrier Web ou un site Web commercial [44].

openCV : est une bibliothèque open source utilisée pour la vision par ordinateur et le traitement d'images. Elle supporte des langages comme Python, C++ et Java, et permet de détecter des objets, des visages ou de la reconnaissance manuscrite. Associée à Numpy, elle offre des opérations puissantes pour manipuler efficacement les images et les données [45].

4.3 Présentation de l'application

Au démarrage de l'application, la fenêtre principale de notre application s'affiche à l'écran comme le montre la figure ci-dessous.

Figure 4. 3: Interface principale de notre application

Cette interface nous permet d'accéder aux différentes fonctionnalités de l'application. Ci-après un utilisateur interagit avec cette interface.

Sélection de la langue

- L'utilisateur clique sur le bouton "Langue" pour ouvrir le menu déroulant.
- Il choisit une langue (Arabe, Français, Anglais).

Chargement du Fichier :

- L'utilisateur clique sur le bouton "Choisir un fichier".
- Il sélectionne un fichier image ou PDF à traiter.

Démarrage du Processus :

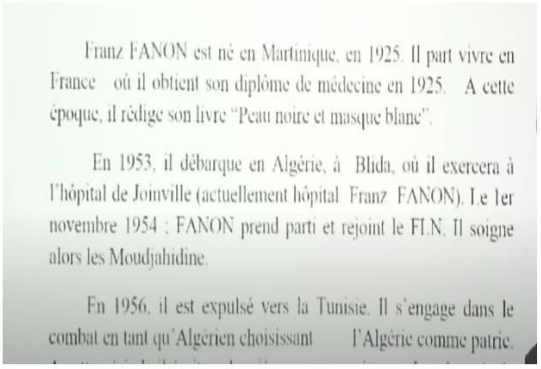
- Après avoir sélectionné la langue et le fichier, l'utilisateur clique sur le bouton "Démarrer le processus".

Apres le processus les résultats sont renvoyés à l'utilisateur sous la forme d'une nouvelle page.

Des Exemples réelles sur les résultats:

❖ **Image/Français**

Image d'origine

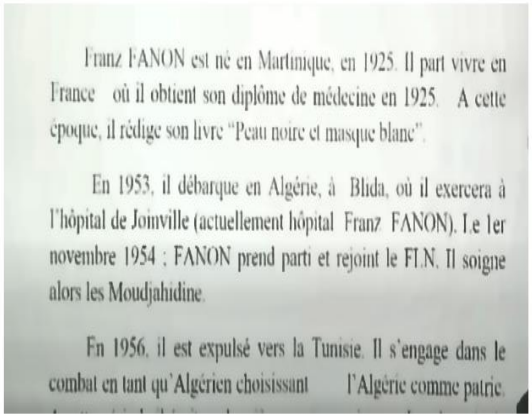


Franz FANON est né en Martinique, en 1925. Il part vivre en France où il obtient son diplôme de médecine en 1925. A cette époque, il rédige son livre "Peau noire et masque blanc".

En 1953, il débarque en Algérie, à Blida, où il exercera à l'hôpital de Joinville (actuellement hôpital Franz FANON). I.e 1er novembre 1954 ; FANON prend parti et rejoint le FI.N. Il soigne alors les Moudjahidine.

Fn 1956, il est expulsé vers la Tunisie. Il s'engage dans le combat en tant qu'Algérien choisissant l'Algérie comme patrie.

Étape 1 : Amélioration de texte



Franz FANON est né en Martinique, en 1925. Il part vivre en France où il obtient son diplôme de médecine en 1925. A cette époque, il rédige son livre "Peau noire et masque blanc".

En 1953, il débarque en Algérie, à Blida, où il exercera à l'hôpital de Joinville (actuellement hôpital Franz FANON). I.e 1er novembre 1954 ; FANON prend parti et rejoint le FI.N. Il soigne alors les Moudjahidine.

Fn 1956, il est expulsé vers la Tunisie. Il s'engage dans le combat en tant qu'Algérien choisissant l'Algérie comme patrie.

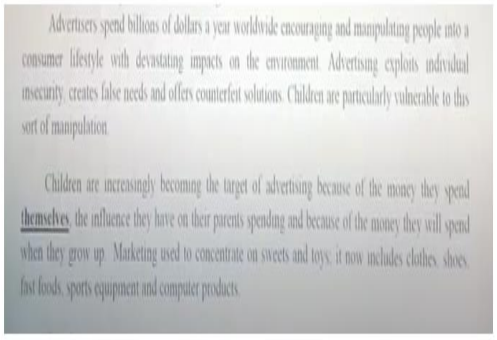
Étape 2 : Extraction de texte

```
Franz FANON est ne cn Martinique, en 1925. Il part vivre en
Franceou il obtient son diplome de medecine en 1925.A cette
epoque, il redige son livre "Peau noire ct masque blane".
En 1953, il debarque en Algerie, a Blida, ou il exercera a
l'hopital de Joinville (actuellement hopitalFranz FANON).I.e 1er
novembre 1954 : FANON prend parti et rejoint le FI.N. Il soigne
alors les Moudjahidine
Fn 1956. il est expulse vers la Tunisie.I1 s'engage dans le
combat cn tantqu'Algericn choisissant
l'Algeric comme patric
```

Figure 4. 4: Exemple de résultat Image/Français

❖ **Image/Anglais**

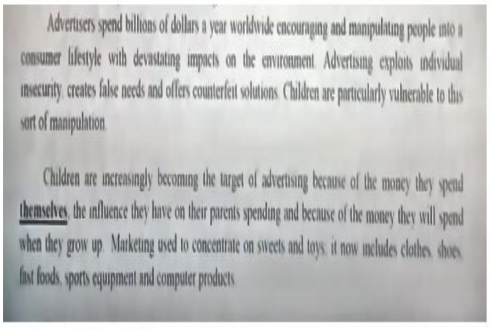
Image d'origine



Advertisers spend billions of dollars a year worldwide encouraging and manipulating people into a consumer lifestyle with devastating impacts on the environment. Advertising exploits individual insecurity, creates false needs and offers counterfeit solutions. Children are particularly vulnerable to this sort of manipulation.

Children are increasingly becoming the target of advertising because of the money they spend themselves, the influence they have on their parents spending and because of the money they will spend when they grow up. Marketing used to concentrate on sweets and toys; it now includes clothes, shoes, fast foods, sports equipment and computer products.

Étape 1 : Amélioration de texte



Advertisers spend billions of dollars a year worldwide encouraging and manipulating people into a consumer lifestyle with devastating impacts on the environment. Advertising exploits individual insecurity, creates false needs and offers counterfeit solutions. Children are particularly vulnerable to this sort of manipulation.

Children are increasingly becoming the target of advertising because of the money they spend themselves, the influence they have on their parents spending and because of the money they will spend when they grow up. Marketing used to concentrate on sweets and toys; it now includes clothes, shoes, fast foods, sports equipment and computer products.

Étape 2 : Extraction de texte

```
Adverusers spend billions of dollars a year worldwide cncouraging and manipulating people into a
consumer lifestyle with devastating impacts on the environment. Advertisng cxploits individual
insecurity. creates false needs and offers counterfeit solutions. Children are partcularly vulnerable to this
sort of manipulation
Children are increasingly becoming the target of advertsing because of the money they spend
themselves, the influence they have on their parents spending and because of the money they will spend
when they grow up. Marketing used to concentrate on sweets and toys: it now includes clothes. shoes
fast foods.sports equipment and computer products
```

Figure 4. 5: Exemple de résultat image/Anglais

❖ **PDF/Arabe**

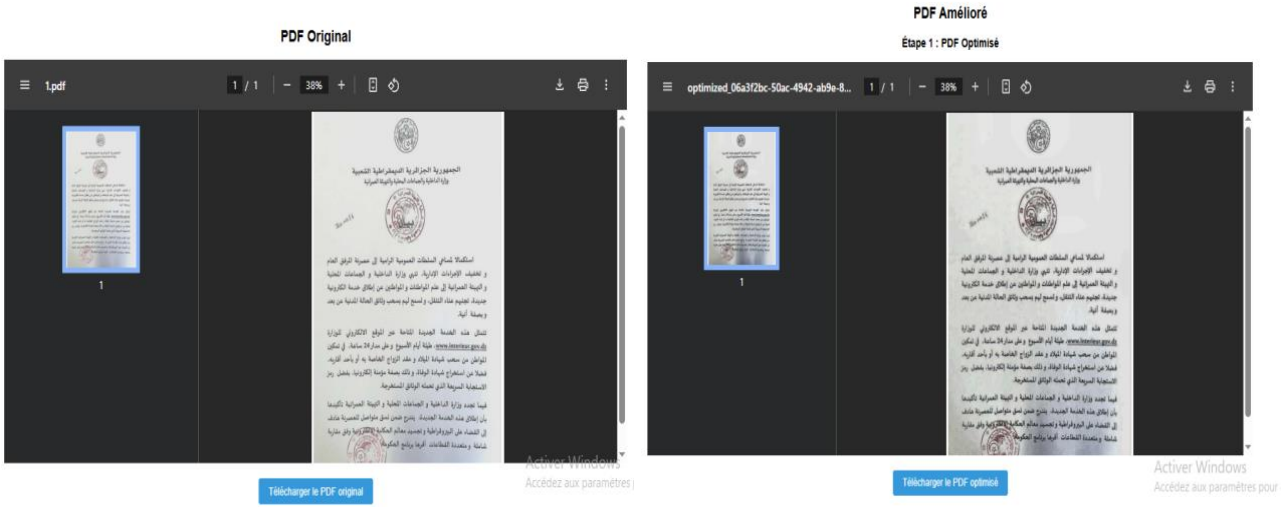


Figure 4. 6: Exemple de résultat PDF/Arabe

➢ Grâce à l'efficacité de notre approche, nous avons également réussi à détecter expérimentalement des textes manuscrits sur plusieurs images. La figure suivante illustre un exemple de ce type de détection :

Figure 4. 7: Exemple sur les textes manuscrits

4.4 Evaluation de l'Efficacité de Notre Approche Proposée

D'après les résultats obtenus, nous avons effectué une comparaison entre plusieurs documents afin d'évaluer l'efficacité de notre approche en termes de clarté des images et de qualité d'extraction du texte, pour différentes langues : arabe, anglais et français.

Les tableaux ci-dessous résument les performances moyennes obtenues sur cinq documents différents par langue, avec les mesures suivantes :

Image originale : qualité visuelle initiale sans traitement.

Image améliorée : qualité visuelle après application des techniques de prétraitement.

Qualité d'extraction : taux de réussite du processus d'extraction du texte via OCR (Reconnaissance Optique de Caractères).

Pour évaluer l'efficacité nous avons utilisé, plusieurs techniques de calcul ont été utilisées, principalement autour de trois axes :

- **Erreur quadratique moyenne (MSE) :**

$$\mathrm{MSE} = \sum_{m=1}^{M} \sum_{n=1}^{N} \left(I(m,n) - \hat{I}(m,n) \right)^2$$

- **Rapport crête signal sur bruit (PSNR) :**

$$\text{PNSR} = 10\log_{10}\left(\frac{{I_{max}}^2}{\text{MSE}}\right)$$

- **Rapport signal sur bruit (SNR) :**

$$\text{SNR} = 10\log_{10}\left(\frac{\frac{1}{N}\sum I^2}{\text{MSE}}\right)$$

Tableau 4. 1: Statistiques de clarté sur plusieurs documents arabes

	1	**2**	**3**	**4**	**5**	**Moyenne**
Image originale	55.5%	5%	16.4%	46.4%	60.4%	36.74%
Image améliorée	62.2%	24%	32.4%	60.7%	74.3%	50.72%
Qualité d'extraction	79.0%	76.4%	76.3%	74.0%	72.7%	75.68%

Analyse :

Nous observons une nette amélioration de la qualité visuelle des images après le prétraitement, avec une augmentation moyenne de +14 % (passage de 36.74% à 50.72%). Cela montre que nos techniques de nettoyage et d'amélioration d'image sont efficaces. En ce qui concerne l'extraction du texte, le taux moyen est élevé (75.68%),ce qui indique EasyOcr une bonne reconnaissance du texte malgré la complexité de la langue arabe (écriture cursive, caractères liés...).

Tableau 4.2: Statistiques de clarté sur plusieurs documents anglais

	1	**2**	**3**	**4**	**5**	**Moyenne**
Image originale	57.3%	56.3%	47.4%	16.9%	56.5%	46.88%
Image améliorée	68.1%	81.3%	76.0%	41.9%	81.4%	69.74%
Qualité d'extraction	76.0%	79.0%	76.3%	79.0%	79.0%	77.86%

Analyse :

Les résultats obtenus pour les documents anglais sont très satisfaisants. L'amélioration visuelle est significative avec un gain moyen de +22.86%**.** La qualité d'extraction atteint une moyenne de 77.86%**,** ce qui traduit une grande précision du système PaddleOcr sur cette langue. Les variations observées (notamment document n°4) peuvent être attribuées à des problèmes de netteté ou de format de l'image source.

Tableau 4. 3: Les statistiques de clarté sur plusieurs documents français

	1	2	3	4	5	Moyenne
Image originale	33.7%	10.4%	57.9%	32.7%	31.5%	33.24%
Image améliorée	51.4%	24.7%	62.2%	69.2%	64.9%	54.48%
Qualité d'extraction	72.8%	71.0%	79.0%	77.3%	79.0%	75.82%

Analyse :

Concernant les documents en français, on constate également une amélioration notable de la qualité visuelle (moyenne passant de 33.24% à 54.48%). Cette amélioration a eu un impact positif sur la qualité d'extraction, qui s'élève à 75.82%. Toutefois, certains documents restent plus difficiles à traiter, notamment ceux dont la qualité initiale était très faible (ex. document n°2).

La figure suivante représente une vue graphique des performances moyennes pour chaque langue :

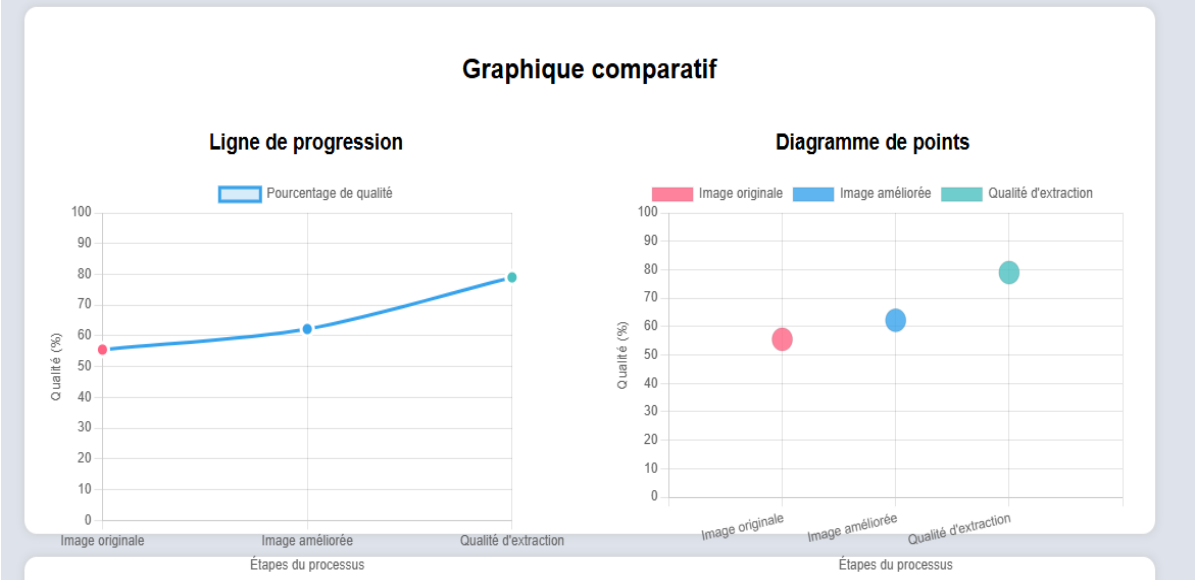


Figure 4. 8: Capture d'un graphique comparatif

Ce graphique permet de visualiser clairement les gains apportés par le prétraitement d'image ainsi que la qualité finale d'extraction du texte selon la langue traitée. On note globalement une amélioration systématique de la qualité visuelle et une extraction performante, confirmant l'efficacité de notre solution.

4.5 Evaluation basé sur les métriques de performance

Les tableaux ci-dessous présentent les performances en termes de précision, rappel et F1-score, qui sont couramment employées dans les évaluations en TALN (Traitement Automatique des langages Naturel), pour illustrer la qualité d'extraction du texte à partir des cinq documents

anciens arabes , anglais et français . Ces métriques sont définies comme suit :

- **Précision(P)**: Proportion de caractères/mots correctement extraits par rapport au total extrait.

$$P = \frac{\text{Caractères/mots correctement extraits}}{\text{Caractères/mots correctement extraits} + \text{Caractères/mots incorrectement extraits}}$$

- **Rappel (R)**: Proportion de caractères/mots correctement extraits par rapport au total attendu (référence manuelle).

$$R = \frac{\text{Caractères/mots correctement extraits}}{\text{Caractères/mots correctement extraits} + \text{Caractères/mots absents de l'extraction}}$$

- **F1-Score** : Moyenne harmonique entre la précision et le rappel, donnant une vue globale de la performance globale du système d'extraction.

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Tableau 4. 4: Evaluation basé sur les métriques de performance des documents arabes

	1	**2**	**3**	**4**	**5**
Précision	0.76	0.75	0.77	0.73	0.79
Rappel	0.73	0.74	0.75	0.71	0.77
F1-Score	0.74	0.74	0.76	0.72	0.78

Analyse :

- La précision varie entre 0.73 et 0.79, ce qui indique que le système extrait correctement environ 73 à 79 %du texte arabe.
- Le rappel est compris entre 0.71 et 0.77, montrant que le système réussit à retrouver 71 à 77 %du texte attendu.
- Le F1-score, qui est une moyenne harmonique entre précision et rappel, oscille entre 0.72 et 0.78.
- Globalement, les performances sont bonnes, malgré la complexité de la langue arabe.

Tableau 4. 5: Evaluation basé sur les métriques de performance des documents anglais

	1	**2**	**3**	**4**	**5**
Précision	0.77	0.78	0.76	0.79	0.77
Rappel	0.78	0.79	0.75	0.80	0.78
F1-Score	0.77	0.78	0.76	0.79	0.78

Analyse :

- La précision varie entre 0.76 et 0.79, ce qui montre une grande fiabilité du système pour extraire uniquement des textes pertinents.
- Le rappel est compris entre 0.75 et 0.80, ce qui signifie que le système retrouve jusqu'à 80% du texte attendu.
- Le F1-score se situe entre 0.76 et 0.79, avec un maximum atteint par le document n°4 (F1-score = 0.79).

Tableau 4. 6: Evaluation basé sur les métriques de performance des documents français

	1	**2**	**3**	**4**	**5**
Précision	0.74	0.72	0.78	0.76	0.79
Rappel	0.73	0.71	0.77	0.76	0.78
F1-Score	0.73	0.72	0.77	0.76	0.78

Analyse :

- La précision varie entre 0.72 et 0.79, ce qui est assez bon mais légèrement inférieur à celui de l'anglais.
- Le rappel est compris entre 0.71 et 0.78, montrant une bonne couverture des données attendues.
- Le F1-score varie entre 0.72 et 0.78, avec le meilleur résultat pour le document n°5 (F1-score = 0.78).

4.6 Conclusion

L'évaluation exhaustive de notre système démontre la validité de notre approche d'amélioration d'images pour l'OCR multilingue. Les résultats quantitatifs révèlent une amélioration systématique des performances d'extraction textuelle, avec des gains significatifs observés pour toutes les langues testées. L'analyse comparative entre images originales et images traitées confirme l'efficacité de nos algorithmes d'amélioration, particulièrement pour les documents de faible qualité initiale. Ces résultats positionnent notre solution comme un outil viable pour le traitement automatisé de documents multilingues, ouvrant des perspectives d'application dans divers domaines nécessitant une numérisation précise de contenus textuels.

Conclusion générale

À l'issue de cette étude approfondie portant sur l'amélioration et l'extraction d'informations à partir de documents anciens, il est clair que les techniques modernes de traitement d'image, de reconnaissance optique des caractères (OCR) et de traitement automatique du langage naturel (TALN) jouent un rôle central dans la préservation et la valorisation du patrimoine documentaire. Face aux défis posés par la dégradation physique des documents, les méthodes numériques offrent aujourd'hui des solutions efficaces pour restaurer, analyser et structurer le contenu textuel de ces archives précieuses.

Notre recherche s'est concentrée sur la conception et la mise en œuvre d'un système hybride multilingue (arabe, français, anglais) intégrant des technologies avancées telles que CLAHE pour l'amélioration du contraste, les filtres bilatéraux pour la réduction du bruit, et l'utilisation combinée de moteurs OCR performants (EasyOCR pour l'arabe et PaddleOCR pour les langues latines). Ce système propose également un post-traitement intelligent, une évaluation quantitative de la qualité visuelle et textuelle, ainsi qu'une exportation des résultats dans divers formats standards (PDF, DOCX, TXT, JSON, XML), facilitant leur réutilisation dans des contextes professionnels ou académiques.

Les résultats expérimentaux obtenus montrent une amélioration significative de la lisibilité des images traitées, ainsi qu'une extraction textuelle précise malgré les défauts courants des documents anciens (tâches, déchirures, faibles contrastes). L'évaluation comparative entre les images originales et les images améliorées a confirmé l'efficacité des techniques de prétraitement, avec des gains moyens de qualité visuelle allant jusqu'à 20 % selon la langue traitée. En outre, la qualité d'extraction OCR atteint des taux moyens supérieurs à 75 %, ce qui témoigne de la robustesse de notre solution face à la complexité linguistique et typographique des documents historiques.

Cette étude souligne également l'importance d'une architecture modulaire et flexible, capable de s'adapter à différents types de documents (images fixes ou PDF multipages) et de gérer automatiquement les erreurs et les variations de qualité. L'intégration d'une interface utilisateur intuitive permet aux utilisateurs non spécialistes de bénéficier facilement des fonctionnalités avancées du système, tout en conservant la possibilité d'une analyse plus approfondie grâce aux formats structurés générés.

En perspective, plusieurs axes d'amélioration peuvent être envisagés : l'extension du système à d'autres langues, notamment celles disposant de moins de ressources numériques ; l'intégration de modèles OCR entraînables spécifiquement sur des corpus historiques pour augmenter la précision ;

L’automatisation accrue de l’archivage numérique structuré via l’analyse sémantique des textes extraits ; ainsi que l’optimisation des performances en environnement embarqué ou mobile pour faciliter l’accès aux archives dans des contextes variés.

En somme, cette contribution constitue une réponse innovante et concrète aux besoins croissants de numérisation, de restauration et d’exploitation des documents anciens. Elle offre des perspectives prometteuses pour la recherche historique, la gestion des archives publiques et privées, ainsi que pour la sauvegarde du patrimoine culturel écrit dans un monde de plus en plus tourné vers le numérique.

Bibliographies

[1] Dira, F. *Contribution à la restauration des images de documents anciens*. Thèse de Doctorat, Institut National des Sciences Appliquées de Lyon, 2007.

[2] Dira, L. "Apport du traitement des images à la numérisation des documents manuscrits anciens." *Document numérique*, vol. 7, no. 3, pp. 13-26, 2003.

[3] Poole, Frazer G. "Integrability: Aspects of the Conservation Problem in Archives." *The American Archivist*, vol. 40, no. 2, pp. 163-172, 1977.

[4] *Digital Archives*. Digital Preservation, Terms & Definitions. Archivemati.ca, http://archivemati.ca/2005/11/08/digital-archives/, consulté le 20 mai 2025.

[5] Luquet, Jean. "La numérisation dans les services d'archives publics en France, une révolution pragmatique." *La Gazette des archives*, no. 204, pp. 233-253, 2006.

[6] Brown, Heather; Baker, John; Cybulski, Walter; Fenton, Andy; Glover, John; Negus, Paul; Palm, Jonas. "The Role of Microfilm in Digital Preservation." *Microform & Digitization Review*, vol. 41, no. 2, pp. 54-68, 2012.

[7] Berisha, F. "Archives Management and Permanent Storage of Documents." *Atlanti*, vol. 25, no. 1, pp. 153-161, 2015.

[8] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, SherjilOzair, Aaron Courville, et Yoshua Bengio. "Generative Adversarial Nets." *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 2672-2680, 2014.

[9] Wang, Zhihao; Chen, Jian; Hoi, Steven C. H. "Deep Learning for Image Super-resolution: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365-3387, 2021.

[10] Brownlee, Jason. *A Gentle Introduction to Generative Adversarial Networks (GANs)*. Machine Learning Mastery, 2019.

[11] Szeliski, Richard. *Computer Vision: Algorithms and Applications*. 2e édition, Springer Nature, 2022.

[12] Park, Sung Cheol, Min Kyu Park, et Moon Gi Kang. "Super-resolution image reconstruction: a technical overview." *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21-36, 2003.

[13] Mishra, V., Agarwal, S., etPuri, N. "Comprehensive and comparative analysis of neural network." *International Journal of Computer Applications*, vol. 175, no. 2, pp. 1-6, 2018.

[14] Zianou, A. S. *Évaluation de la qualité d'image*. Thèse de Doctorat, Université de Mentouri–Constantine, 2012

[15] IBM. *Reconnaissance optique de caractères*. https://www.ibm.com/fr-fr/think/topics/optical-character-recognition, consulté le 1 mai 2025.

[16] El Gajoui, K. *Reconnaissance optique des caractères : Cas de la langue amazighe*. Thèse de Doctorat, Université Mohammed V, Rabat, Laboratoire de Recherche en Informatique et Télécommunications, 2021.

[17] Coquenet, D., Chatelain, C., etPaquet, T. "End-to-end handwritten paragraph text recognition using a vertical attention network." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 508-524, 2022.

[18] Zhou, Xinyu, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, et Jiajun Liang. "East: an efficient and accurate scene text detector." *Proc. CVPR*, pp. 2642–2651, 2017.

[19] Soykan, Gurkan; Yuret, Deniz; Sezgin, Tevfik Metin. "A Comprehensive Gold Standard and Benchmark for Comics Text Detection and Recognition." *arXiv preprint*, Computation and Language (cs.CL); Artificial Intelligence (cs.AI), 2022

[20] Vedhaviyassh, D.R.; Sudhan, R.; Saranya, G.; Safa, M.; Arun, D. "Comparative Analysis of EasyOCR and TesseractOCR for Automatic License Plate Recognition using Deep Learning Algorithm." *2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECATECH)*, Coimbatore, 1-3 décembre 2022.

[21] Feng, Lei; Ke, Zongwu; Wu, Na. "ModelsKG: A Design and Research on Knowledge Graph of Multimodal Curriculum Based on PaddleOCR and DeepKE." *2022 14th International Conference on Advanced Computational Intelligence (ICACI)*, Guilin, 15-17 juillet 2022.

[22] Berisha, F. "Archives Management and Permanent Storage of Documents." *Atlanti*, vol. 25, no. 1, pp. 153-161, 2015.

[23] BerSingh, Jaskirat; Bhushan, Bharat. "Real-Time Indian License Plate Detection using Deep Neural Networks and Optical Character Recognition using LSTM Tesseract." *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, 18-19 octobre 2019.

[24] Anitha, R., Rajeev, R. R., Nazeem, M., etNavaneeth, S. *Open-Source OCR Libraries: A Comprehensive Study for Low Resource Language*. International Centre for Free and Open Source Solutions (ICFOSS), Government of Kerala, Thiruvananthapuram, Kerala, India.

[25] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[26] Horé, Alain; Ziou, Djemel. "Image Quality Metrics: PSNR vs. SSIM." *20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turquie, 23-26 août 2010, IEEE, pp. 2366-2369, 2010.

[27] Pitas, Ioannis. *Digital Image Processing Algorithms and Applications*. Wiley-Interscience, New York, 2000.

[28] Kova, H. *Hub Kova recognition in Cecchi historia textes usine à CNN-BiLSTM neural*

network model. Thèse de Doctorat, Université Charles de Prague, 2019.

[29] Cherniak, Eugene. *Introduction to artificial intelligence*. Addison-Wesley, 1984.

[30] *Vocabulaire de l'intelligence artificielle (liste de termes, expressions et définitions adoptés)*. Journal officiel de la République française no. 0285, 9 décembre 2018.

[31] Algoscale. *What is Hybrid Approach in NLP?*https://algoscale.com/blog/what-is-hybrid-approach-in-nlp/, consulté le 4 avril 2025.

[32] Archimag. *Numérisation : quel format pour conserver vos documents ?*https://www.archimag.com/demat-cloud/2021/03/29/numerisation-format-conserver-documents/, consulté le 4 avril 2025.

[33] Stratow. *Format d'archivage*. https://www.stratow.com/post/format-archivage, consulté le 4 avril 2025.

[34] Allili, M. S. *Techniques d'apprentissage automatique*. Cours de 2e cycle, Université du Québec en Outaouais (UQO), Québec, Canada, Hiver 2015.

[35] Van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, Sander Wubben, et EmielKrahmer. "Best practices for the human evaluation of automatically generated text." *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 355-368, 2019.

[36] Ramadhani, A. L., P. Musa, et E. P. Wibowo. "Human face recognition application using PCA and eigenface approach." *2017 Second International Conference on Informatics and Computing (ICIC)*, Jayapura, Indonésie, pp. 1–5, novembre 2017.

[37] Paris, Sylvain, Pierre Kornprobst, Jack Tumblin, et Frédo Durand. "Bilateral filtering: Theory and applications." *Foundations and Trends in Computer Graphics and Vision*, vol. 4, no. 1, pp. 1–73, 2008.

[38] Mukherjee, Sumita, HritikTyagi, PurushautamTyagi, Nikita Singh, etShraddha [nom completmanquant]. "OCR Using Python and Its Application." *Journal of Advanced Zoology*, vol. 44, no. S-3, pp. 1083-1092, 2023.

[39] Hébert, Patrick. *Traitement de base d'une image (partie 1)*. Dernière révision octobre 2008.

[40] IronSoftware. *EasyOCR vs Tesseract: A ComprehensiveComparison*. https://ironsoftware.com/csharp/ocr/blog/ocr-tools/easyocr-vs-tesseract/, consulté le 11 mai 2025.

[41] Plugger.ai. *What is PaddleOCR? A Complete Guide*. https://www.plugger.ai/blog/what-is-paddle-ocr, consulté le 12 mai 2025.

[42] Nedjar, H. *Traitement d'images en utilisant la plateforme 'Google Colab'*. Mémoire de fin d'études, Sciences et Technologies – Automatique et informatique industrielle, 26 juin 2022.

[43] Ngrok. *Documentation*. https://ngrok.com/docs, consulté le 22 mai 2025

[44] GeeksforGeeks. *Libraries in Python*. https://www.geeksforgeeks.org/libraries-in-python/, consulté le 25 mai 2025.

[45] Das, Kushal. "Introduction to Flask." *Python for you and me 0.4.beta1 documentation*.