République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Larbi Tébessi –Tébessa
Faculté des Sciences Exactes et des Sciences
de la Nature et de la Vie
Département de Mathématiques et d'Informatique

Mémoire de master

Domaine : Mathématiques et Informatique
Filière : Informatique
Option : Systèmes et Multimédias

Thème :

Contribution à l'analyse et la classification de documents

Présenté par :

Hamel Aymene Abdelkouddous

Devant le jury :

Mr. Mekhaznia Tahar	*Maitre de conférences A*	*Université de Tébessa*	*Président*
Mr. Laimech Lakhdar	*Maitre de conférences B*	*Université de Tébessa*	*Examinateur*
Mr. Djeddi Chawki	*Maitre de conférences A*	*Université de Tébessa*	*Rapporteur*

Année Universitaire : 2021-2022

Résumé

Dans ce manuscrit, un système automatique pour classification et la datation de manuscrits anciens est introduit. Le système proposé comprend deux étapes principales : l'extraction de caractéristiques et la classification (datation du manuscrit). Dans la première étape, des caractéristiques basées sur les polygones et les chaînes de Freeman sont extraites à partir d'un manuscrit ancien. Dans la deuxième étape, nous avons utilisé les séparateurs à vastes marges (SVMs), les forêts aléatoires, un algorithme de boosting adaptatif (Adaboost) ainsi qu'un classificateur à renforcement de gradient (GBT) pour la classification. Les expérimentations sont menées sur un ensemble de données qui comprend plus de 2000 manuscrits arabes anciens et les résultats enregistrés sont encourageants.

Keywords : Datation de manuscrits, documents arabes anciens, chaînes de Freeman, polygones, SVM, forêts aléatoires, AdaBoost, GBT.

Abstract

In this manuscript, an automatic system for classification and dating of ancient manuscripts is introduced. The proposed system includes two main steps: feature extraction and classification (manuscript dating). In the first step, features based on polygons and Freeman's chains are extracted from an ancient manuscript. In the second step, we used Support vector machines (SVMs), random forests, an adaptive boosting algorithm (Adaboost) and a gradient-boosting tree classifier (GBT) for classification. The experiments are being carried out on a dataset that includes more than 2000 ancient Arabic manuscripts and the recorded results are encouraging.

Keywords: Manuscript dating, ancient Arabic documents, freeman chain code, polygons, SVM, random forests, AdaBoost, GBT.

الملـخص

في هذا العمل، تم تقديم نظام آلي لتصنيف و تأريخ المخطوطات القديمة. يتضمن النظام المقترح خطوتين رئيسيتين: استخلاص المميزات و تصنيفها (تأريخ المخطوطة). في الخطوة الأولى، يتم استخلاص السمات القائمة على سلاسل فريمان و المضلعات من مخطوطة قديمة. في الخطوة الثانية، استخدمنا مصنفات آلات المتجهات الداعمة، الغابات العشوائية، خوارزمية التعزيز التكيفي و مصنف شجرة تعزيز التدرج من أجل التصنيف. أجريت التجارب على مجموعة بيانات تضم أكثر من 2000 مخطوطة عربية قديمة والنتائج المسجلة مشجعة.

الكلمات المفتاحية : تأريخ المخطوطات، مخطوطات عربية قديمة، سلاسل فريمان، المضلعات، آلات المتجهات الداعمة، الغابات العشوائية، خوارزمية التعزيز التكيفي، مصنف شجرة تعزيز التدرج.

REMERCIEMENTS

Ce travail n'a pu être mené bien qu'avec le soutien de plusieurs personnes que je voudrais, à travers ces quelques lignes, remercier du fond du cœur.

Premièrement, nous remercions Dieu source de toute connaissance.

Je voudrais adresser toute ma gratitude au encadreur Dr. DJEDDI.C et aussi Dr. GAHMOUSSE.A, des enseignants au département de mathématiques et d'informatique, faculté des sciences exactes et sciences naturelles et de la vie de l'université de Tébessa, pour la patience, la disponibilité et surtout les bons conseils, qui ont contribué à alimenter ma réflexion.

Je remercie Dr.MEKHAZNIA.T et Dr.LAIMECHE.L, enseignants au département de mathématiques et d'informatique, faculté des sciences exactes et sciences naturelles et de la vie de l'université de Tébessa, d'avoir accepté d'être les jurys de cette thèse.

Par ailleurs, je tiens à remercier tous les membres du corps professoral de la Faculté de mathématiques et d'informatique de l'Université Larbi Tebessi pour avoir fourni des informations précieuses pour le succès de la recherche à l'université.

Enfin, je remercie mes amis mes et collègues universitaires pour leurs supports et leurs conseils.

Merci pour tout.

Table des matières

Liste des abréviations

DSS	Dead sea souls
MPS	Micro physiology systems
SVM	Support Vector Machines
RF	Random Forest
Adaboost	Adaptive boosting
GBT	Gradient boosting tree
JC	Jesus Christ
SDHK	Svenskt Diplomatariums huvudkartotek
CNN	Convolutional neural network

Liste des tableaux

Liste des figures

INTRODUCTION GENERALE

1.1. Contexte du travail

Les manuscrits arabes constituent une partie importante du patrimoine arabe et musulman dans le monde. Les bibliothèques nationales hébergent des centaines de milliers d'images numériques ; cependant, de nombreux documents n'indiquent pas expressément quand ils ont été rédigés. La datation des documents historiques aidera à les relier à un événement important et à déterminer leur importance historique.

Les styles d'écriture en arabe ont évolué avec le temps. Chaque siècle islamique possède son propre ensemble de scripts d'écriture, donnant aux différents styles d'écriture des caractéristiques distinctes. Certains styles d'écriture ont évolué au cours des siècles, en conservant leurs caractéristiques générales tout en incorporant un nouvel ensemble de personnalités. L'état dégradé des documents historiques, ainsi que la similitude des styles d'écriture, rendent difficile la datation du document historique.

1.2. Objectifs du travail

Dans ce mémoire, nous montrerons comment les méthodes d'apprentissage automatique peuvent être utilisées pour l'analyse historique dans les humanités numériques. Partout dans le monde, des manuscrits pré-modernes sont photographiés et numérisés à grande échelle à des fins d'héritage culturel. La numérisation supprime efficacement la nécessité d'un accès physique pour effectuer des recherches sur les écrits historiques. Cependant, pour rendre possible la recherche sur l'ensemble des documents conservés, des méthodes de calcul doivent être utilisées.

Nous présentons notre tentative réussie d'utiliser des techniques d'apprentissage automatique ainsi que quelques méthodes de traitement d'images développées à l'origine pour l'identification de scripteurs, pour soutenir ce nouveau type passionnant de recherche dans les sciences

humaines. Nous avons utilisé les séparateurs à vastes marges (SVMs), les forêts aléatoires, un algorithme de boosting adaptatif (Adaboost) ainsi qu'un classificateur à renforcement de gradient (GBT) pour l'estimation des dates de production des manuscrits arabes anciens.

L'évaluation du système proposé est effectuée sur un ensemble de manuscrits arabes historiques appelé KERTAS. Il contient plus de 2000 images acquises du 1er au 14ème siècle. Nous montrons que l'apprentissage automatique est applicable à la tâche de datation de manuscrits anciens et que les performances sont en moyenne comparables à celles d'un expert humain.

1.3. Organisation du mémoire

Ce mémoire est structuré en deux parties. La première est consacrée à la présentation des principaux concepts, outils et travaux relatifs à l'étude entreprise. Dans la deuxième partie de ce mémoire, nous abordons de manière détaillée nos choix conceptuels, la mise en œuvre ainsi que les résultats obtenus par le système proposé pour la datation automatique de manuscrits arabes anciens.

Chapitre 2. Analyse et classification de documents : concepts, outils et état de l'art

Ce chapitre est consacré à la présentation des concepts liés directement avec le problème étudié. Il présente une tâche pertinente dans l'estimation de la date de fabrication des manuscrits anciens. Nous concluons ce chapitre en présentant les applications possibles du domaine de recherche étudié.

Chapitre 3. Caractéristiques texturales pour l'estimation de la date de production de manuscrits arabes anciens

Ce chapitre se détache des aspects théoriques abordés dans la première partie et s'oriente vers la présentation de notre contribution qui consiste en un système automatique pour l'estimation de la date de production d'un manuscrit arabe ancien en se basant sur la caractérisation des différentes images de manuscrits considérés par des caractéristiques basées sur les polygones et les chaînes de Freeman. Nous décrivons la base de données utilisée avant de nous focaliser sur la présentation des méthodes d'extraction de caractéristiques adoptées, les séparateurs à vastes marges (SVMs), les forêts aléatoires, un algorithme de boosting adaptatif (Adaboost) ainsi qu'un classificateur à renforcement de gradient (GBT) sont employés pour la classification. Les

expérimentations effectuées seront aussi présentées. A la fin de ce chapitre, les résultats sont exposés et discutés.

A la fin de ce mémoire, nous émettons nos conclusions sur le travail que nous avons entrepris dans le domaine de la datation de manuscrits anciens. Nous présentons aussi les perspectives d'extensions futures du travail que nous avons présenté dans ce document.

PARTIE 1

2. Datation de manuscrits historiques : Concepts, outils et état de l'art

Ce chapitre est consacré à la présentation des quelques concepts liés directement avec le problème étudié. Il présente la définition d'un manuscrit historique, le processus de datation de manuscrits historiques ainsi que les différents types de supports d'écritures utilisés pour la réalisation de manuscrits historiques. Nous décrivons également les différentes techniques de datation de manuscrits historiques à savoir les techniques physiques, paléographiques et modernes (informatiques) ainsi que certaines bases de données qui sont utilisées dans le domaine de recherche considéré. Nous terminons ce chapitre la présentation d'un état de l'art.

2.1. Introduction

Il existe un grand nombre de manuscrits historiques non datés numérisés et non numérisés de l'Antiquité. Les dates de production de ces manuscrits historiques intéressent les historiens, conservateurs, paléographes, universitaires et autres parties prenantes. Ce chapitre traite de la tendance des approches et des techniques de datation des manuscrits historiques, des méthodes traditionnelles aux méthodes modernes (informatiques). De nombreuses méthodes de datation des manuscrits historiques ont été développées. Elles se divisent en 3 grandes catégories:

- Méthodes paléographiques,
- Méthodes physiques,
- Méthodes modernes (informatiques).

Les méthodes de datation des manuscrits historiques traditionnelles nécessitent des échantillons de manuscrits réels, prennent beaucoup de temps et elles sont coûteuses. Ces inconvénients ont encouragé le passage aux méthodes modernes (informatiques). Ces dernières utilisent des images numérisées de

manuscrits, les préservant ainsi, et elles sont relativement moins coûteuses, plus rapides, pratiques et applicables à grande échelle. Il est à noter que peu de recherches ont été effectuées sur la datation des manuscrits historiques, en particulier sur les méthodes modernes (informatiques) applicables à grande échelle.

2.2. Manuscrits historiques

Les manuscrits historiques sont des documents manuscrits réalisés au cours des époques historiques de l'histoire humaine. Ce sont d'excellentes sources d'information, de culture, de patrimoine et d'origine de personnes, de sociétés, des organisations et des gouvernements par rapport aux périodes historiques [1]. Un manuscrit est composé de 2 parties: un support d'écriture et un contenu textuel [2]. Le support d'écriture est un objet dont la surface est écrite avec un texte ou dessinée avec des graphiques. Le contenu textuel comprend des caractères et des graphiques écrits ou dessinés sur un support d'écriture.

L'étude et l'analyse des manuscrits historiques donnent de grandes informations sur l'état des choses et la vie avant les époques modernes [3]. En raison de leur origine lointaine, les manuscrits historiques souffrent de dégradations physiques notamment sur le support d'écriture et l'encre. Par conséquent, lors du traitement et de l'étude des manuscrits historiques, toute méthode utilisée impliquant leur manipulation physique doit être évitée dans la mesure du possible car elle accélère la dégradation et les dommages supplémentaires.

Par conséquent, les manuscrits historiques sont numérisés pour une meilleure gestion, traitement, partage, accès, préservation et étude [4]. Dans l'étude des manuscrits historiques, divers acteurs s'intéressent à différents aspects : les historiens cherchent à déchiffrer la culture des peuples anciens, les linguistes veulent savoir comment la langue a évolué au fil du temps en étudiant la sémantique du contenu textuel, les philologues visent à comprendre la structure de la langue [5], les paléographes cherchent à connaître la date/l'âge, l'origine et l'évolution des styles d'écriture au fil du temps [6], et la communauté scientifique cherche principalement à authentifier, connaître la composition élémentaire des manuscrits et quand ils ont été créés [7].

Les principaux aspects de l'étude des manuscrits historiques qui recoupent toutes les parties prenantes sont la date de création, la nature/l'origine du contenu et les auteurs des manuscrits. De tous ceux-ci, la date de production

des manuscrits historiques a attiré le plus d'attention à l'heure actuelle, dans ce qu'on appelle la datation des manuscrits historiques.

2.3. Datation de manuscrits historiques

La datation des manuscrits historiques est le processus de détermination de l'âge ou du temps/année de production des manuscrits historiques. Au cours du processus de datation, en fonction de l'approche utilisée, le résultat peut être l'âge du manuscrit ou le temps de production. En plaçant les manuscrits historiques dans leurs contextes temporels, des périodes de temps (ou des époques) sont souvent utilisées. Il y a 3 époques principales : les périodes pré-médiévale, médiévale et moderne.

L'ère pré-médiévale est la période allant du temps d'origine (du monde) au 5ème siècle après JC. L'ère médiévale est la période du 5ème au 15ème siècle après JC tandis que l'ère moderne est la période du 15ème siècle à nos jours. Compte tenu de la valeur des manuscrits historiques, connaître leur date de production est essentiel [8] pour donner aux manuscrits historiques l'importance et la valeur monétaire [2], aider à la recherche pour les utilisateurs à l'aide d'informations temporelles comme l'âge ou le temps de production, faciliter l'archivage, la classification et le catalogage des manuscrits [3], aider à l'étude et à la compréhension d'autres manuscrits connexes, et aider à comprendre le contexte des manuscrits en relation avec la culture, l'histoire et le lieu d'origine ou l'auteur [3]. La datation des manuscrits historiques a été et continue d'être une tâche de recherche difficile. Les attributs du contenu textuel et des supports d'écriture des manuscrits historiques sont utilisés pour estimer leur âge ou leur temps de production.

2.4. Supports d'écritures

Il existe 5 principaux types de supports d'écriture utilisés pour la réalisation de manuscrits historiques:

1- Le papyrus est un support d'écriture obtenu grâce à la transformation des tiges d'une plante africaine, appelée également papyrus, en une surface souple, lisse, de plusieurs mètres de long, facile à replier en rouleau qui a été le support principal de l'écriture et de la peinture dans le bassin méditerranéen durant l'Antiquité et le Haut Moyen Âge [9].

2- Parchemin: Il a été utilisé pour la première fois vers le 2ème siècle avant JC. Il est fabriqué à partir de peau d'animaux comme les moutons, les

antilopes, les chèvres et d'autres animaux [10]. Il est solide, stable et flexible par rapport au papyrus. Le parchemin était un matériau d'écriture courant aux IIe-XVe siècles.

3- Vélin : C'est la même chose que le parchemin sauf qu'il est fabriqué à partir de peaux de jeunes animaux comme les veaux, les chevreaux et les agneaux. Il est de meilleure qualité et plus cher que le parchemin [10].

4- Tissu : Pièce de vêtement utilisée comme support d'écriture.

5- Papier : il a été inventé vers le IIe siècle.

2.5. Techniques de datation de manuscrits historiques

La datation des manuscrits historiques peut être divisée en 3 grandes classes: les techniques physiques, les techniques paléographiques et les techniques modernes (informatiques). Les techniques physiques et paléographiques constituent des méthodes traditionnelles de datation des manuscrits historiques, tandis que les techniques informatiques sont des méthodes modernes de datation des manuscrits historiques.

2.5.1. Techniques physiques

Ces techniques de datation des manuscrits historiques utilisent des mesures d'échantillons de manuscrits historiques réels. Elles sont très objectives, plus précises et plus fiables que d'autres méthodes. La datation de manuscrits historiques à l'aide de techniques physiques comprend les 3 étapes suivantes [2]:

1- Tout d'abord, un échantillon est extrait du manuscrit historique faisant l'objet de l'enquête. L'échantillon peut être un support d'écriture (matériel sur lequel l'écriture est faite) ou une partie manuscrite (du contenu écrit) du manuscrit.

2- Ensuite, les composants (ou éléments) constitutifs de l'échantillon sont mesurés et identifiés à l'aide d'un instrument spécialisé.

3- Enfin, les mesures sont analysées pour obtenir/déduire l'âge/le temps de production du manuscrit étudié.

2.5.2. Techniques paléographiques

Dans cette technique, les experts paléographes utilisent leur expertise et leur intuition pour déterminer les dates des manuscrits historiques [11]. Les paléographes sont des personnes impliquées dans l'étude des documents historiques. La technique paléographique de datation des manuscrits historiques comprend 2 étapes principales: l'extraction des attributs paléographiques et l'inférence. Les informations paléographiques peuvent être obtenues par examen visuel ou par métadonnées. Les informations paléographiques incluent l'auteur (s), la classe de contenu, la langue du contenu, la mise en page, entre autres. Les paléographes estiment le temps de production en fonction des attributs visibles et des informations paléographiques des manuscrits historiques.

L'âge ou le temps de production des manuscrits historiques est déduit de l'étude des tendances des styles d'écriture, de la structure de la langue et de la structure des caractères au fil du temps. Un paléographe doit avoir une connaissance approfondie de ces caractéristiques linguistiques en évolution et des époques (périodes) auxquelles elles ont existé [11]. Les caractéristiques et les attributs du langage qui évoluent avec le temps comprennent la structure des caractères, la grammaire, les styles d'écriture, les abréviations manuscrites de divers mots, les vocabulaires/mots, les ligatures et les ponctuations. La nature et le type de matériel d'écriture utilisé peuvent également être utilisés pour déduire la date de production des manuscrits. Le processus de datation paléographique est subjectif et conduit souvent à des dates contradictoires entre différents paléographes [12]. La technique paléographique de datation des manuscrits historiques se divise en 2 approches principales : les approches basées sur l'examen visuel et les approches basées sur les métadonnées.

2.5.2.1. Approches basées sur l'examen visuel

Dans cette approche, un paléographe utilise les attributs visibles des manuscrits pour donner leurs dates de production probables. Les attributs visibles utilisés comprennent le style d'écriture, la mise en page des manuscrits, la nature et le type de matériel d'écriture (support), la nature de l'encre utilisée, l'instrument d'écriture probable (ou stylo) utilisé, le contenu des manuscrits et la nature de la reliure, le cas échéant [11]. Un paléographe utilise les attributs visibles en conjonction avec son expertise pour obtenir un indice de l'année de production. Le succès est plutôt déterminé par le niveau d'expertise du paléographe [11].

2.5.2.2. Approches basées sur les métadonnées

Dans cette approche, les informations/attributs paléographiques (métadonnées) associés à des manuscrits historiques individuels sont récupérés à partir des archives et utilisés pour estimer la date de production. Les métadonnées utilisées incluent le nom du ou des auteurs, le lieu d'origine, le contenu, la classe de contenu, la taille, la langue utilisée et la mise en page du manuscrit.

Pour les approches basées sur l'examen visuel et sur les métadonnées de la datation des manuscrits historiques, la date de production est déterminée de manière déductive, c'est-à-dire sur la base des caractéristiques et des attributs tirés du manuscrit.

L'inférence sur la date de production est effectuée de deux manières :

1- déduction directe de la date de production à partir des caractéristiques et des attributs obtenus,

2- référence textuelle chronologique, c'est là qu'un texte du manuscrit qui fait référence à un événement connu ou daté ou datable (comme une cérémonie), des personnes notables (comme des rois ou des célébrités), des événements (comme des catastrophes, des guerres, des pandémies ou des épidémies) est utilisé pour estimer la date de production d'un manuscrit [11].

3- comparaison paléographique ou parallélisme où un manuscrit non daté est comparé à d'autres manuscrits similaires (manuscrits de référence) datés ou datables [12].

La comparaison est effectuée à l'aide de caractéristiques visuelles telles que les styles d'écriture manuscrite. La date de production estimée attribuée au manuscrit non daté provient de la même époque/période que celle d'un manuscrit similaire et daté (datable). Les méthodes de datation basées sur la paléographie sont manuelles, non computationnelles, prennent du temps et ne sont réalisables que pour un petit nombre de manuscrits. Ils ne sont pas réalisables à grande échelle.

2.5.3. Techniques modernes (informatiques)

Les techniques modernes (informatiques) de datation des manuscrits historiques deviennent de plus en plus des alternatives aux techniques physiques et paléographiques avec des avantages et des commodités

attrayants. La datation automatique de manuscrits historiques est une approche intéressante où des méthodes automatiques sont utilisées pour extraire les caractéristiques des manuscrits, qui à leur tour sont utilisées pour estimer/déduire la date de production par des techniques statistiques ou par ‘entrainement de modèles de datation. Les techniques de datation modernes sont automatiques, objectives, moins sujettes aux erreurs, pratiques, plus rapides et moins coûteuses comparés aux méthodes traditionnelles (méthodes physiques et paléographiques) [3]. Elles sont aussi plus faciles à partager et applicables à grande échelle.

Les ensembles de données annotés d'images numérisées de manuscrits historiques sont des outils nécessaires pour créer des modèles de datation de manuscrits historiques. Pour développer un modèle de datation automatique, les caractéristiques sont d'abord extraites à partir de manuscrits datés [9]. Ces caractéristiques permettent de capturer les principaux attributs des manuscrits, tels que l'utilisation de mots/phrases ou le style d'écriture. Un classificateur approprié est ensuite entrainé avec les caractéristiques extraites pour obtenir un modèle permettant de prédire l'année de production de manuscrits non datés. Lors de la prédiction de la date de production des manuscrits historiques non datés, leurs caractéristiques sont extraites puis introduites dans un modèle entraîné de datation [9]. La date de production prévue est générée par le modèle.

En utilisant une approche informatisée, la datation des manuscrits historiques peut être considérée comme une tâche de classification ou de régression. Lorsque la datation historique des manuscrits est abordée comme une tâche de classification, la période du corpus est divisée en unités temporelles, qui sont des partitions temporelles de taille uniforme sans chevauchement couvrant la période complète du corpus considéré. La période du corpus est la plage de temps allant du premier au dernier moment de production des manuscrits dans un corpus. La taille de l'unité temporelle peut varier entre 1 et 100 ans en fonction de la période du corpus. Pour un corpus donné, chaque manuscrit (document) ne correspond qu'à une unité temporelle. Dans le cadre de la datation historique des manuscrits, l'unité temporelle correspond à la durée de production des manuscrits. Un classificateur approprié est entrainé pour attribuer un manuscrit non daté à l'une des unités ou classes temporelles. Dans le modèle de classification, les unités ou classes temporelles correspondent au temps/période de publication, Lorsque la datation d'un manuscrit historique est entreprise en tant que tâche de régression, l'année de production d'un manuscrit est directement déterminée via un modèle/une technique de régression.

2.6. Bases de données disponibles

Cette section présente diverses bases de données contenant des manuscrits historiques avec des dates de production connues. La plupart d'entre elles ont été utilisées pour développer des systèmes automatiques de datation de manuscrits anciens.

2.6.1. Base de données MPS

Elle se compose de 2858 images numérisées de chartes collectées dans 4 villes [13]. MPS couvre la période médiévale de 1300 à 1550 après JC. Les manuscrits historiques dans la base MPS décrivent l'évolution progressive des styles d'écriture au fil du temps. Les manuscrits originaux ont été rédigés par des professionnels.

2.6.2. Base de données SDHK

Elle contient plus de 40000 lettres suédoises médiévales datées de 817 à 1540 après JC. Elle se compose à la fois de formes originales et de transcriptions de lettres. Cette base de données est conservée et mis à jour par les archives nationales suédoises [9].

2.6.3. Base de données DEEDS

Est une collection de numérisations de chartes et d'actes de propriété foncière associés à des documents d'échange de propriété légale dans le comté d'Essex en Angleterre à l'époque médiévale. Il s'agit d'un projet de recherche établi en 1975 au Département d'études historiques et culturelles de l'Université de Toronto. Les dates de production des manuscrits de la base de données DEEDS vont du XIe au XVe siècle après JC. Elle consiste en 3300 manuscrits datés et 5000 non datés [14].

2.6.4. La base de données BH2M

Elle est composée d'images numérisées de documents de mariage couvrant 5 siècles [15]. Les documents ont été obtenus à partir d'un registre de mariage conservé à la cathédrale de Barcelone. Le registre se compose de 244 livres

avec des actes de mariages tenus entre 1451 et 1905 après JC. Les livres sont écrits par des auteurs différents.

2.6.5. La base de données CLaMM

Elle est constituée d'images d'écritures médiévales annotées et issues de catalogues français [16]. Elle a été utilisée en 2017 lors d'une compétition scientifique sur la datation de manuscrits anciens en marge de la Conférence internationale sur l'analyse et la reconnaissance de documents (ICDAR). La base de données CLaMM est composée de 9500 images et 15 classes pour la tâche de datation de manuscrits. Ces 15 classes couvrent la période 500-1600 après JC.

2.6.6. La base de données KERTAS

Elle est composée de plus de 2000 images de manuscrits arabes manuscrits historiques couvrant plus de 14 siècles islamiques [17]. Les manuscrits proviennent principalement de la Bibliothèque nationale du Qatar. Parmi les autres contributeurs figurent les universités de Tubingue, de Berlin et de Tokyo. Les manuscrits de la base de données KERTAS couvrent des sujets comme l'histoire, les mathématiques, la métaphysique, la physique et l'histoire islamique.

2.7. Etat de l'art

Les recherches actuelles dans le domaine de la datation de manuscrits historiques utilisent des descripteurs visuels extraits à partir des images de manuscrits. Certaines méthodes proposées dans la littérature reposent uniquement sur le contenu du manuscrit, mais il existe également des suggestions pour l'utilisation de méthodes indépendantes du contenu. En général, ces méthodes se répartissent en deux catégories: les approches basées sur la classification automatique et celles basées sur l'apprentissage profond [18].

Plusieurs études ont proposé diverses méthodes d'estimation de date de production d'un manuscrit ancien à l'aide de bases de données MPS telle que [16, 27, 28].

Dans [19], les auteurs ont estimé la date de production de documents historiques en utilisant une méthode de régression qui emploie des

caractéristiques de niveau local et globale. La méthode utilise des caractéristiques de charnière(en anglais : Hinge) et de fragments (en anglais : fragelets).

He et al. dans [20] ont proposés une méthode basée sur les codebooks entraînée en combinant les fragments des contours locaux (kCF) et des fragments de lignes (kSF) pour estimer l'âge des documents historiques.

Dans [21] un algorithme de regroupement permettant de relier les descripteurs visuels de bas niveau du document historique à leurs étiquettes dans la base de données MPS a été proposé par les auteurs. La méthode a montré des corrélations entre les descripteurs d'images et les étiquettes.

Sur la base de statistiques physiques, Wahlberg et al. dans [22] ont présenté des techniques de datation automatique pour les images en niveaux de gris non. Les techniques proposées ont été testées sur la base de données "Svenskt diplomatariums huvudkartotek", constituée d'images scannées de chartes médiévales conservées aux Archives nationales de Suède.

Dans [23], les auteurs ont utilisé un réseau neuronal convolutionnels (CNN) pour prédire les dates des documents imprimés à partir de la base de données de livres de Google.[24].

Hamid et al. dans [25] suggèrent que la combinaison de certaines caractéristiques donnera de meilleures performances que l'utilisation de caractéristiques individuelles. Les auteurs ont utilisé une combinaison de filtres de Gabor avec des motifs locaux binaires (LBP).

Dans [26] , les auteurs ont présenté une approche basée sur l'apprentissage profond utilisant l'apprentissage par transfert avec des réseaux neuronaux convolutionnels (CNN) prétraités.

Studer et al. dans [27] ont présenté une technique pour dater des documents historiques utilisant l'apprentissage par transfert d'un réseau neuronal pré-entraîné sur la base de données ImageNet [35, 38].

Rahiche et al. dans [32] ont réalisé l'un des travaux les plus récents sur la datation des documents historiques, qui ont introduit une technique indépendante du contenu basée sur les propriétés optiques des documents historiques, telles que la décoloration et les changements dans le matériel écrit. La méthode proposée recueille des informations temporelles sur l'encre métallique à l'aide d'une technique d'imagerie à multiples facettes combinée à une approche de classificateur de régression (KDLOR).

Dans un autre travail récent [33], les auteurs ont proposé d'utiliser une méthode basée sur les graphèmes avec une carte temporelle auto-organisée (SOTM) comme codebook.

Dans [18], un ensemble de caractéristiques traditionnelles et celles extraites par un réseau résiduel (ResNet) sont fusionnées dans une approche hiérarchique utilisant une représentation éparse commune. Pour résoudre le problème de bruit lors de la fusion, les auteurs ont proposé une nouvelle approche basée sur plusieurs sous-ensembles de caractéristiques est envisagée. De nombreuses caractéristiques sont prises en compte. Ensuite, des classificateurs supervisés et non supervisés sont utilisés pour la classification.

2.8. Conclusion

Dans ce chapitre, nous avons présenté dans un premier temps, quelques concepts qui sont en liaison directe avec le problème étudié, ces concepts incluent : la définition d'un manuscrit historique, le processus de datation de manuscrits historiques ainsi que les différents types de supports d'écritures utilisés pour la réalisation de manuscrits historiques, nous nous sommes concentré ensuite sur les différentes techniques de datation de manuscrits historiques à savoir les techniques physiques, paléographiques et modernes (informatiques), puis, nous avons décrit certaines bases de données qui sont utilisées dans le domaine de recherche considéré. Nous avons terminé ce chapitre par une conclusion précédée par les travaux connexes dans le domaine de la datation de manuscrits historiques.

PARTIE 2

3. Caractéristiques texturales pour l'estimation de la date de production de manuscrits arabes anciens

Dans ce chapitre, nous présentons notre contribution qui consiste en un système automatique pour la datation de manuscrits Arabes anciens en se basant sur la caractérisation des différentes images de manuscrits considérés par des caractéristiques basées sur les polygones et les chaînes de Freeman. Nous décrivons la base de données utilisée avant de nous focaliser sur la présentation des méthodes d'extraction de caractéristiques adoptées, les séparateurs à vastes marges (SVMs), les forêts aléatoires, un algorithme de boosting adaptatif (Adaboost) ainsi qu'un classificateur à renforcement de gradient (GBT) sont employés pour la classification. Les expérimentations effectuées seront aussi présentées. A la fin de ce chapitre, les résultats sont exposés et discutés.

3.1. Introduction

La date d'écriture d'un manuscrit historique est toujours importante pour les historiens et paléographes. Cela peut les aider à comprendre le contexte textuel sous-jacent du document et aider à comprendre les références culturelles et historiques présentées dans le texte. Savoir quand le manuscrit a été écrit peut également aider les chercheurs à catégoriser les manuscrits historiques avec plus de précision et d'efficacité. En combinant l'efficacité des paléographes avec la rapidité des techniques de l'intelligence artificielle, le domaine de datation de manuscrits historiques pourra être automatisé et plus précis.

Le présent chapitre vise à mettre au point un système automatique d'estimation de la date de production de manuscrits historiques et qui pourra être très utile aux historiens et paléographes. Le système proposé est conçu pour fonctionner sur des manuscrits Arabes historiques. Ce chapitre est organisé comme suit. Dans la section suivante, nous décrivons la base de

données utilisée. Nous présentons ensuite la description des techniques d'extraction de caractéristiques proposés dans la section 3.3, suivies par la description des classifieurs utilisés dans la section 3.4. Les résultats expérimentaux et leur analyse sont présentés dans la section 3.5, tandis que la dernière section conclut ce chapitre.

3.2. Base de données utilisé

La base de données KERTAS [17], [18] est une base de données de manuscrits arabes historiques écrits entre le 1er et le 14ème siècles islamiques, elle se compose de plus de 2000 images numérisées avec une haute qualité et une haute résolution. Les manuscrits de la base KERTAS sont divisés en 14 classes, et chaque classe elle contient des manuscrits du même siècle. Le tableau 1 donne un aperçu de la répartition des manuscrits de la base KERTAS sur les ensembles d'apprentissage et de test. La figure 3.1 montre deux échantillons extraits à partir de la base considérée dans le cadre du présent chapitre et le tableau 3.1 présente la répartition des échantillons sue les différentes classes (siècles).

Siècle	**Nombre de documents**	**Documents d'apprentissage**	**Documents de test**
1	58	39	19
2	46	32	14
3	144	98	46
4	592	406	186
5	164	113	51
6	126	87	39
7	184	125	59
8	110	75	35
9	152	104	48
10	73	50	23
11	169	116	53
12	147	100	47
13	119	81	38
14	17	12	5
Total	2101	1438	663

Tableau 3.1 – Distribution des documents de la base de données KERTAS

La figure 1 montre également deux exemples de bases de données :

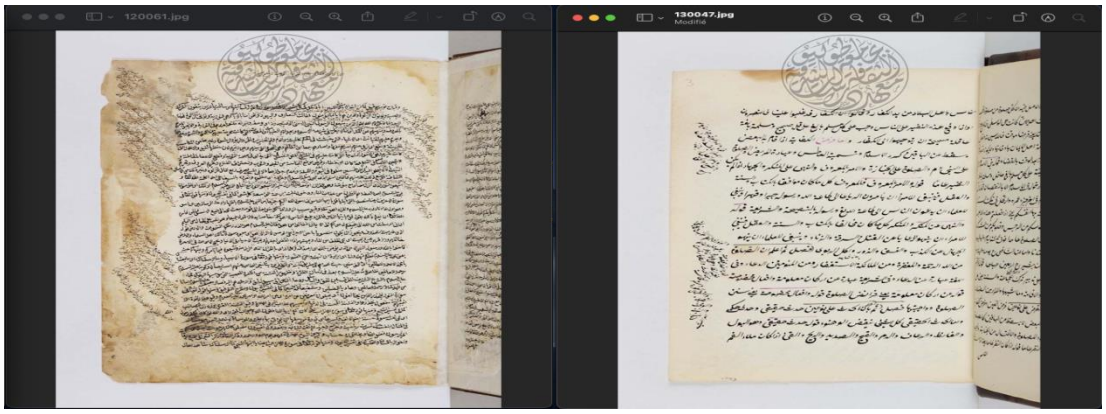

Figure 3.1 – Exemples de base de données KERTAS[18]

3.3. Extraction des caractéristiques

Le système de datation de manuscrits proposé est basé sur un ensemble des mesures texturales basées sur les polygones et les chaînes de Freeman extraites à partir des images de manuscrits Arabes anciens. Ces caractéristiques sont décrites dans les sous sections suivantes.

3.3.1. Caractéristiques polygonales

Cette méthode est utilisée pour extraire un vecteur de caractéristiques polygonales. Ces caractéristiques sont destinées à préserver uniquement les caractéristiques de la police en supprimant les détails. Effectuez une estimation du contour à travers une série de segments de ligne en utilisant l'algorithme de polygonisation séquentielle [34].

L'algorithme nécessite un paramètre T qui contrôle la précision de l'approximation. Plus la valeur de T est élevée, plus le segment est long, et inversement. Nous avons utilisé une valeur de T égale à 2 choisie empiriquement.

La figure 3 montre l'estimation polygonale du contour d'un mot manuscrit pour différentes valeurs de T :

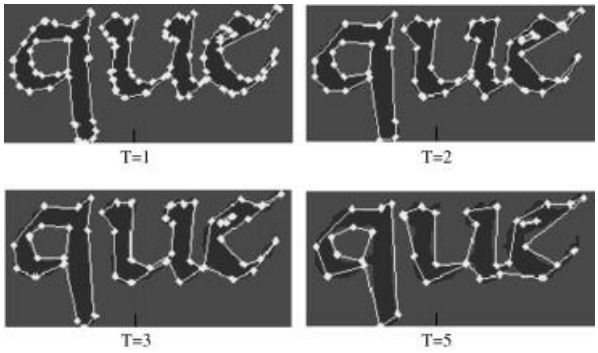

Figure 3.2 – Polygonisation à différentes valeurs de T [35]

Tout d'abord, nous calculons la pente de chaque segment de droite et nous utilisons sa distribution pour la caractérisation. Chaque ligne est identifiée comme appartenant à l'une des classes représentées sur la Figure 4. Ces classes sont sélectionnées de manière à ce que les segments qui sont approximativement dans la même direction que la direction principale (verticale, horizontale, etc.) soient classés dans leurs classes respectives. Par exemple, tous les segments compris entre 12° et 12° sont classés presque horizontalement.[36]

Nous calculons également la distribution pondérée par la longueur du gradient. Ici, pour chaque segment du gradient i, la position i augmente de la longueur du segment. Enfin, la distribution est normalisée par la longueur totale du segment dans l'image [36].

Nous calculons ensuite les angles entre les deux segments connectés et utilisons la distribution de ces angles pour estimer la courbure.

3.3.2. Chaînes de Freeman Globales

Afin de capturer l'information d'orientation [36], nous partons de l'histogramme de distribution des codes de Freeman (fonction de densité de la pente des contours où les éléments de l'histogramme représentent les contributions en pourcentage des huit directions principales. En outre, nous trouvons également les histogrammes des distributions des différentiels du premier (et du second) ordre qui sont calculés en soustrayant chaque élément du code de Freeman du précédent et en prenant le résultat modulo la connectivité (8 dans notre cas). Ces histogrammes représentent la répartition des angles entre pixels de texte successifs et les variations de ces angles au fur et à mesure de la progression du trait. Les distributions des codes de Freeman et leurs différentiels donnent une idée globale sur les formes de l'écriture, mais ils peuvent ne pas être très efficaces pour capturer les détails fins de l'écriture; nous proposons donc de prendre en considération non seulement les occurrences des directions individuelles des codes de Freeman

mais également les distributions de paires de codes de Freeman, dans un histogramme, illustré pour deux écritures dans la figure 3.3.

Les éléments (i,j) de l'histogramme (8x8) représentent la contribution en pourcentage de la paire (i,j) dans la séquence de code de Freeman des contours. En se basant sur le même principe, nous calculons également l'histogramme (8x8x8) des triplets de code de Freeman f5. Il est important de préciser que toutes les 64 paires possibles et 512 triplets possibles ne peuvent pas exister en traçant les contours et nous pouvons avoir un total de 44 paires et 236 triplets.

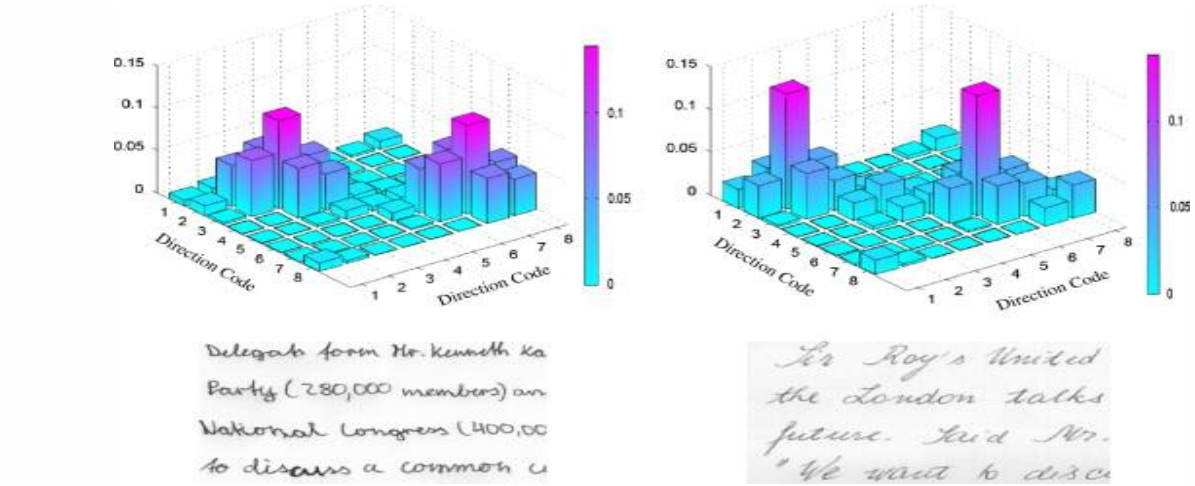

Figure 3.3 – Deux écritures et leurs distributions respectives de paires de codes de chaîne [36]

3.4 Classification

Une fois que les images de manuscrits à comparer sont représentées par leurs caractéristiques, nous procédons à l'utilisation de ces vecteurs de caractéristiques pour l'entrainement ou le test. L'entrainement et la classification sont effectués en utilisant l'un les classifieurs décrits dans les sous sections suivantes

3.4.1 Les machines à vecteur de support

Le classifieur SVM [37] est basé sur l'hypothèse que, pour un espace donné, il existe un classificateur linéaire (appelé hyperplan) pour distinguer deux classes d'espace (+ /). Le but de cette méthode est d'apprendre, à partir d'un ensemble d'exemples d'apprentissage (apprentissage supervisé), une fonction qui prédit des classes pour de nouveaux objets. Plus précisément, il s'agit de

trouver l'hyperplan optimal, qui sépare les données et maximise la distance entre les deux couches. L'hyperplan optimal est celui parmi tous les hyperplans valides, réalisant l'amplitude maximale entre les points des deux classes. C'est pourquoi on l'appelle un délimiteur de marge large. Les points les plus proches de la frontière entre deux couches et utilisés pour déterminer l'hyperplan optimal sont appelés vecteurs de support. L'hyperplan optimal est celui qui classifiera le mieux les nouveaux exemples. Une nouvelle classification d'un exemple inconnu est donnée selon sa position par rapport à l'hyperplan optimal. Face au cas de la décomposition non linéaire (c'est-à-dire la plupart des problèmes pratiques), les méthodes SVMs utilisent une fonction de multiplication pour effectuer une transformation non linéaire des données. Le résultat de cette transformation, appelé espace de ré-description, est un espace de dimension supérieure. Dans notre cas on a utilisé un classificateur SVM avec un noyau 'rbf', comme montré sur la figure 3.4:

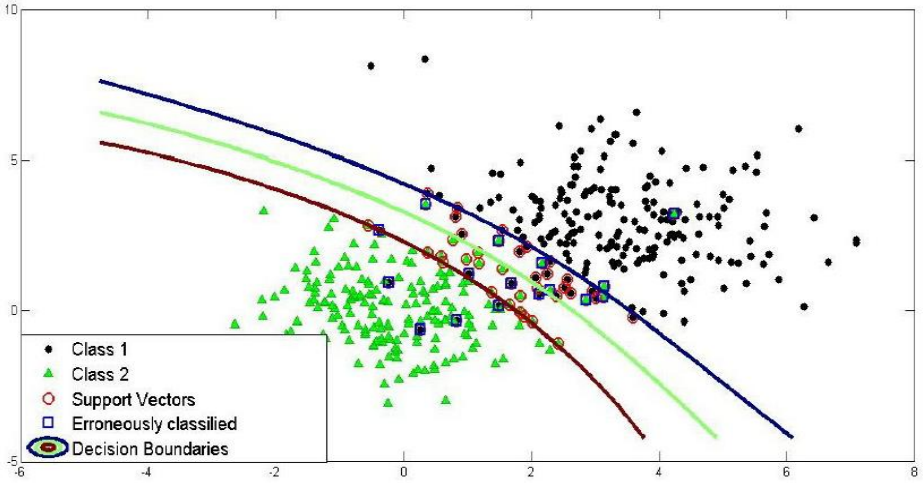

Figure 3.4 – Un exemple de classification SVM utilisant le noyau RBF. [38]

3.4.2. Les forêts aléatoires

Le classificateur de forêt aléatoire [39] est composé d'un certain nombre d'arbres, chacun planté en utilisant une certaine forme de caractère aléatoire. Les nœuds feuilles de chaque arbre sont marqués avec des estimations de la distribution postérieure sur les couches d'image du manuscrit ancien. Chaque nœud interne contient un test pour la meilleure division de l'espace de données classifié.

L'image d'un manuscrit ancien est classée en l'envoyant à chaque arbre et en résumant les distributions des feuilles. Le hasard peut être injecté à deux moments du processus d'apprentissage; sous-échantillonner les données d'apprentissage de sorte que chaque arbre soit développé en utilisant un sous-

ensemble différent; et dans la sélection des boutons de test. La figure 3.5 est un exemple qui montre le principe de classificateur RF :

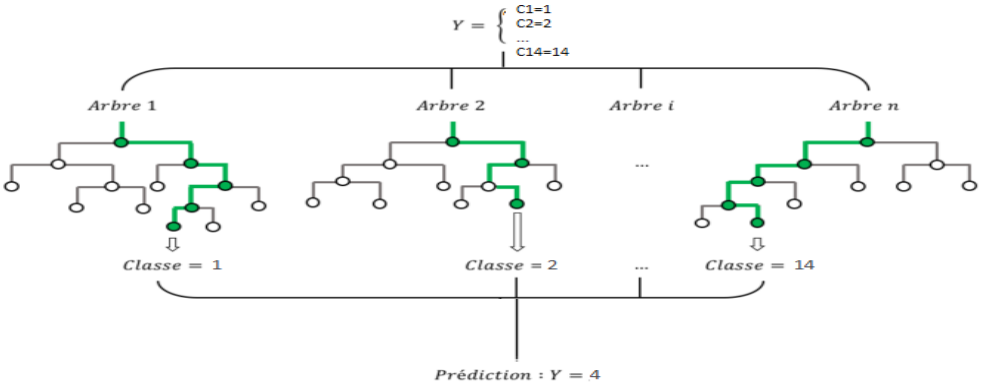

Figure 3.5 – Principe du classifieur RF[40]

3.4.3 Boosting adaptatif (Adaboost)

Le boosting adaptatif [41] est un algorithme d'apprentissage d'ensembles utilisé pour résoudre des problèmes de classification complexes en intégrant des classificateurs faibles et forts simples. Le modèle d'apprentissage d'ensemble est conçu autour de deux approches : Boosting et Bagging. Adaboost est l'un des algorithmes BOOSTRING dans lequel les poids des instances individuelles sont déterminés à plusieurs reprises en fonction de la précision du résultat final de la classification. La figure 3.6 est un exemple montrant le principe du classifieur Adaboost:

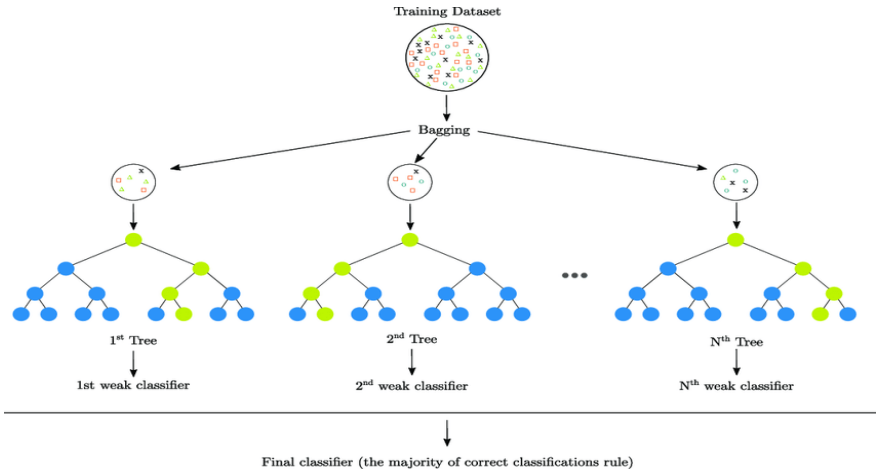

Figure 3.6 – Principe du classificateur Adaboost [40]

3.4.4. Arbre de renforcement du gradient (Gradient boosting tree)

L'arbre de renforcement du gradient est similaire à AdaBoost dans le sens sens qu'ils utilisent tous les deux un ensemble d'arbres de décision pour prédire les étiquettes cibles. Cependant, contrairement à AdaBoost, les arbres de renforcement du gradient ont une profondeur supérieure à 1. En pratique, on les utilise avec un nombre maximum de feuilles entre 8 et 32.

L'algorithme GBT [42] est un modèle de prédiction basé sur le principe de la combinaison de plusieurs arbres de régression. En particulier, les arbres de régression sont des modèles caractérisés soit par un biais élevé et des erreurs de variance faibles si l'arbre est peu profond, soit par un biais faible et des erreurs de variance élevées si l'arbre est profond. Pour résoudre ce problème, il existe deux familles d'algorithmes qui combinent plusieurs arbres de régression pour réduire les erreurs élevées.

Les arbres de renforcement du gradient sont basés sur le principe du boosting [43], c'est-à-dire la combinaison de modèles avec un biais élevé et une erreur de variance faible afin de réduire le biais tout en conservant une variance faible. En détail, au lieu d'utiliser des arbres profonds et différents ensembles de données d'entraînement, les GBT emploient des arbres peu profonds qui sont entraînés dans le même ensemble de données, où chaque arbre est spécialisé dans une caractéristique spécifique de la relation entrée-sortie. En particulier, des arbres superficiels successifs sont formés en série, l'arbre étant ensuite formé dans le but de réduire les erreurs de prédiction de l'arbre n-1 précédent. La figure 3.7 est un exemple de modèle de boosting de gradient :

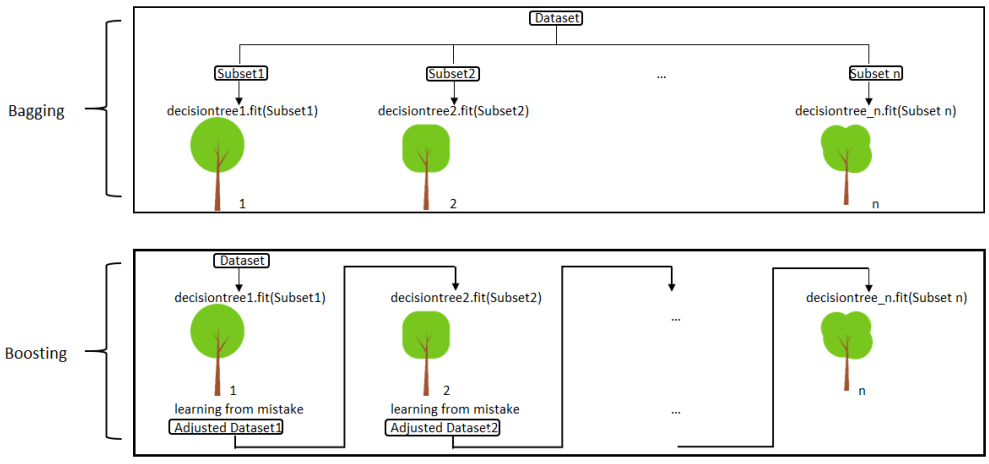

Figure 3.7 – Principe de fonctionnement de l'arbre de boosting de gradient pour une tâche de détection simplifiée.[40]

3.4.5. La recherche par grille

La méthode traditionnelle d'optimisation des hyperparamètres est une recherche par grille, qui effectue simplement une recherche complète sur un sous-ensemble donné de l'espace des hyperparamètres de l'algorithme d'apprentissage (figure 3.8). Comme l'espace des paramètres de l'algorithme d'apprentissage automatique peut inclure des espaces avec des valeurs réelles ou illimitées pour certains paramètres, il est possible que nous devions spécifier une limite pour appliquer une recherche par grille. La recherche par grille souffre des espaces de grande dimension, mais peut souvent être facilement parallélisée, puisque les valeurs des hyperparamètres avec lesquels l'algorithme travaille sont généralement indépendantes les unes des autres.[44]

Nous définissons manuellement une série de paramètres possibles et l'algorithme effectue une recherche complète sur ces paramètres :

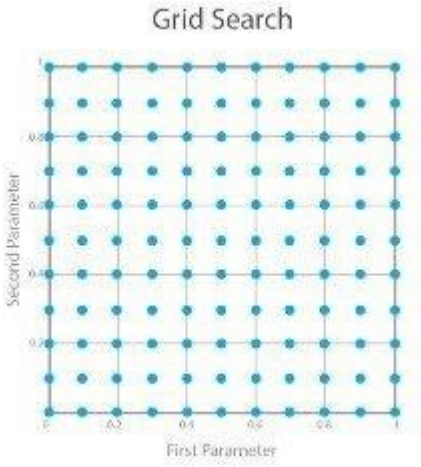

Figure 3.8 – Une illustration d'un espace de recherche en grille. [44]

3.5. Résultats et discussion

Dans cette section, nous présentons et analysons les performances des caractéristiques proposées dans l'estimation de la date de production de manuscrits anciens. Nous avons envisagé un seul scénario d'évaluation où nous avons choisi 67% des images de manuscrits de chaque classe (siècle) pour l'apprentissage et 33% des images de manuscrits restants de chaque classe (siècle) pour le test. Le tableau 3.2 présente les taux globaux d'identification enregistrés avec les caractéristiques basées sur les polygones et les chaînes de Freeman en utilisant les quatre classifieurs choisis.

Classifieurs	Caractéristiques	
	Polygones	Chaînes de Freeman
SVM	86%	87%
RF	85%	89%
Adboost	85%	89%
GBT	83%	90%

Tableau 3.2 – Taux de précision en % enregistrés avec les classifieurs : SVM, Random Forest, Adaboost et Gradient Boosting Tree

3.5.1 Matrices de confusion

La matrice de confusion est une technique permettant de résumer les performances d'un algorithme de classification. Le calcul de la matrice de confusion nous donne une meilleure idée du fonctionnement du modèle de classification et du type d'erreur qui se produit.

Les figures suivantes présentent les matrices de confusion obtenue pour chaque classifieur avec les méthodes d'extractions proposées :

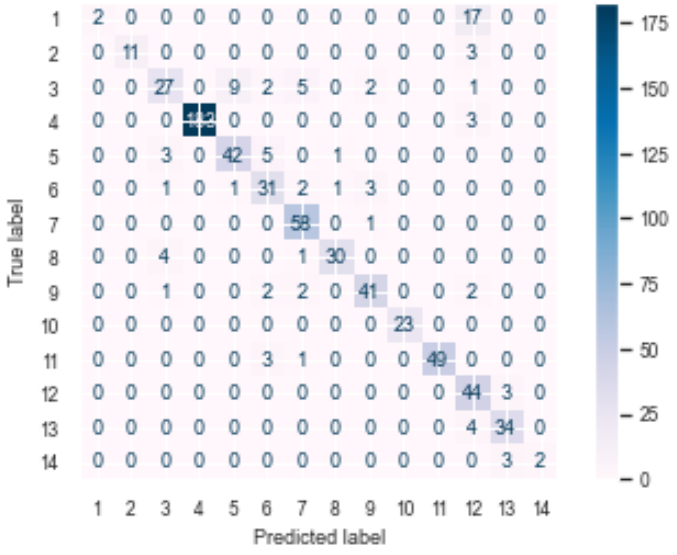

Figure 3.9 – Matrice de confusion en utilisant les chaines de Freeman avec le classifieur SVM

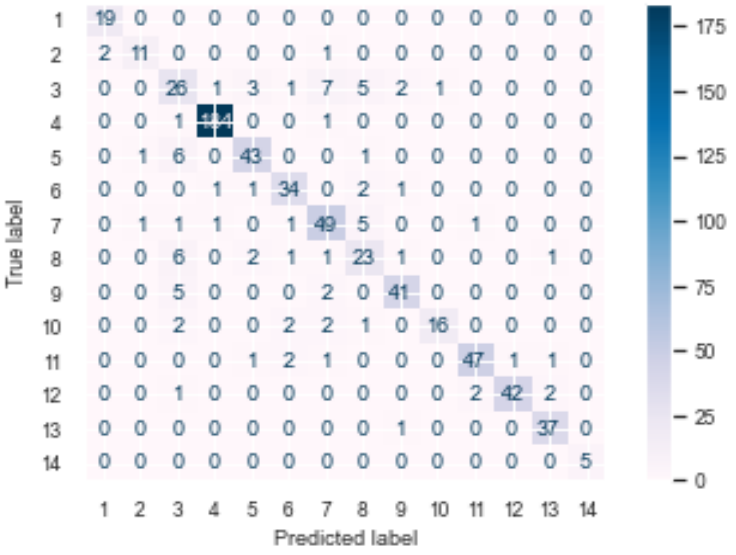

Figure 3.10 – Matrice de confusion en utilisant les caractéristiques polygonales avec le classifieur SVM

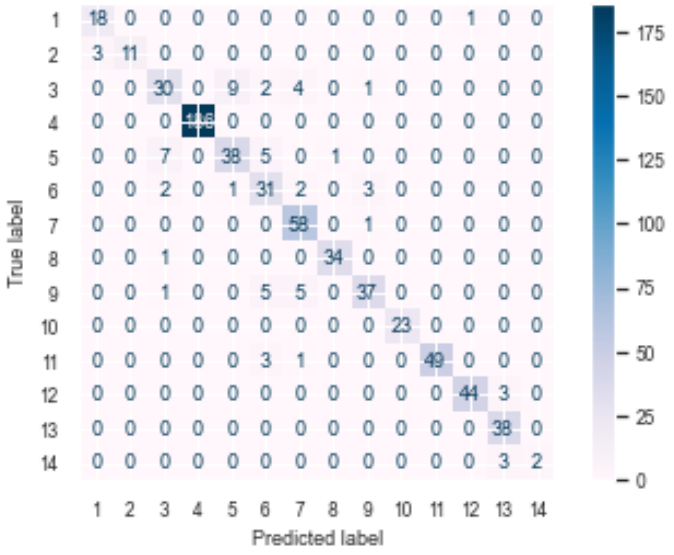

Figure 3.11 – Matrice de confusion en utilisant les chaines de Freeman avec le classifieur forêts aléatoires.

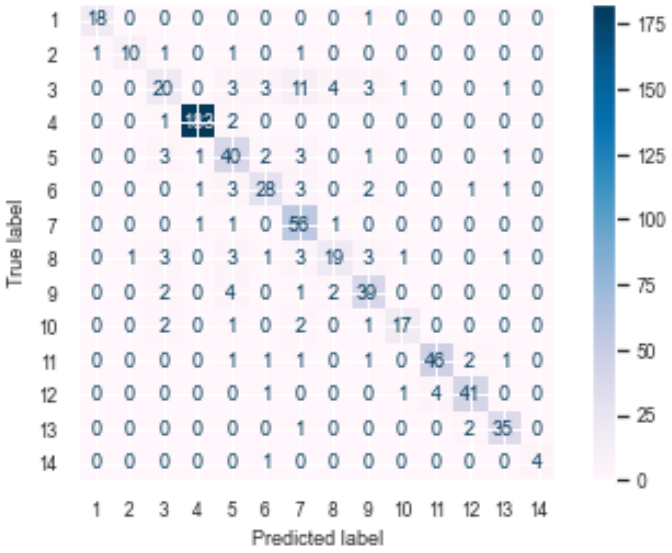

Figure 3.12 – Matrice de confusion en utilisant les caractéristiques polygonales avec le classifieur forêts aléatoires.

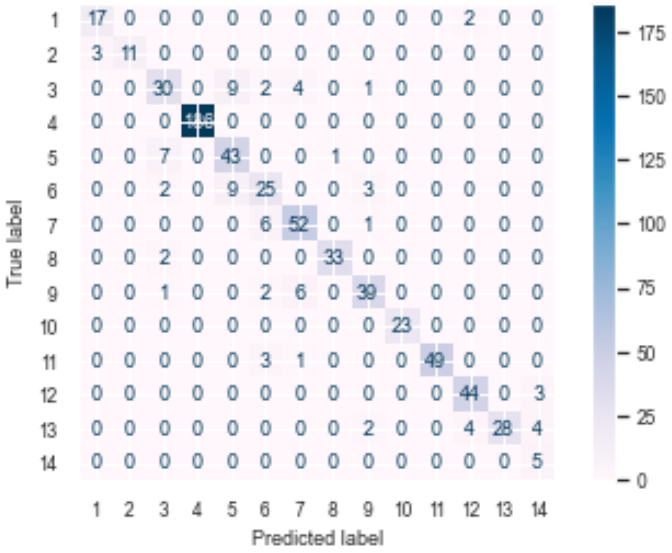

Figure 3.13 – Matrice de confusion en utilisant les caractéristiques chaines de Freeman avec le classifieur Adaboost.

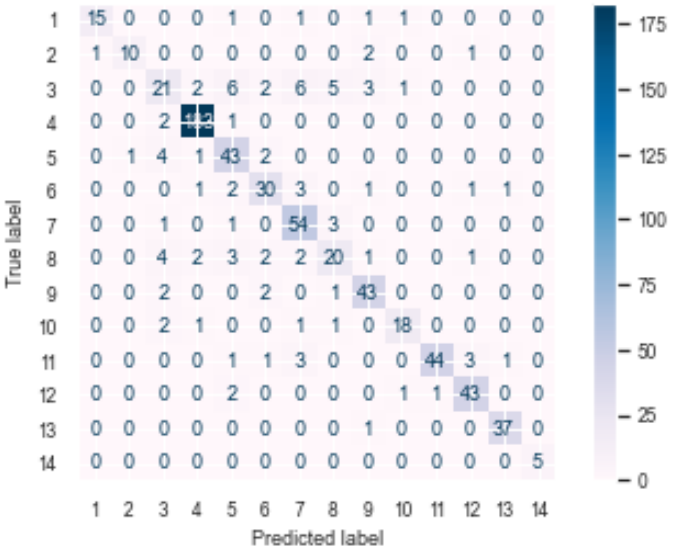

Figure 3.14 – Matrice de confusion en utilisant les caractéristiques polygonales avec le classifieur Adaboost

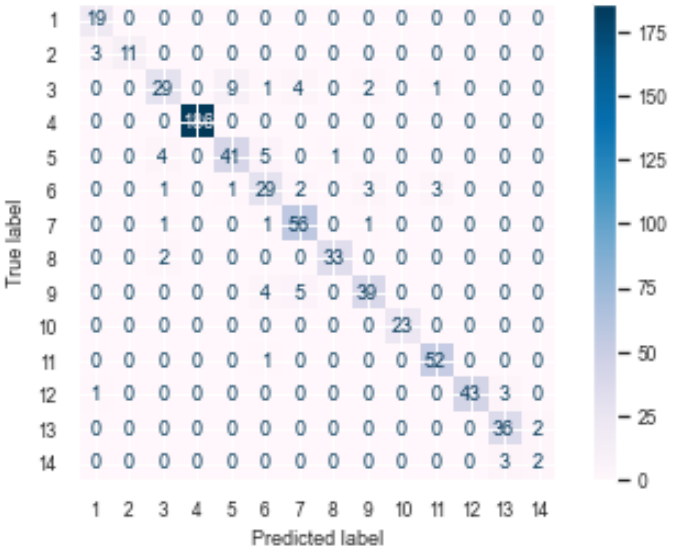

Figure 3.15 – Matrice de confusion en utilisant les chaines de Freeman avec le classifieur GBT.

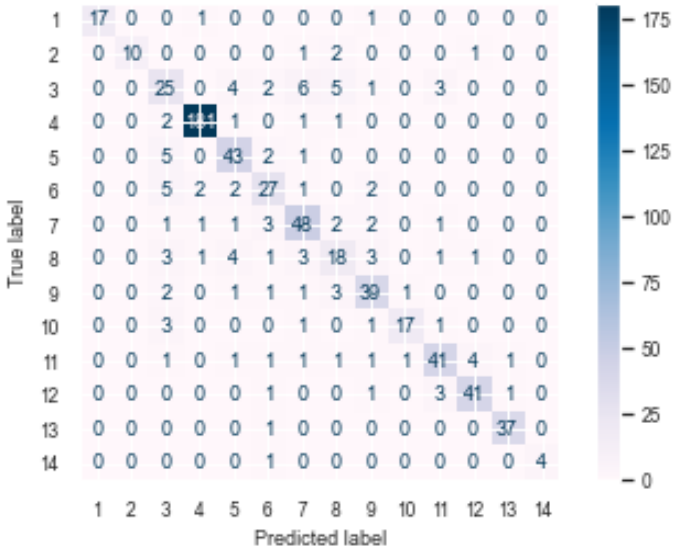

Figure 3.16 – Matrice de confusion en utilisant les caractéristiques polygonales avec le classifieur GBT.

Enfin, nous pouvons dire que les résultats obtenus pour l'estimation de la date de production de manuscrits historiques démontrent clairement le potentiel des caractéristiques proposées.

3.6. Conclusion

Ce travail avait pour objectif de présenter une méthode pour la datation de manuscrits anciens. Nous avons utilisé des caractéristiques texturales qui ont montré des résultats prometteurs. Les évaluations ont été effectuées sur une base de données contenant des échantillons de manuscrits Arabes anciens contenant plus de 2000 images différentes. Les résultats obtenus sont encourageants et reflètent l'efficacité des caractéristiques texturales sur des manuscrits anciens. La contribution que nous avons proposée dans le cadre de ce mémoire nous ont permis d'aboutir à des résultats prometteurs, mais nous ont aussi ouvert plusieurs voies pouvant être exploitées dans le futur.

CONCLUSION GÉNÉRALE

Ce travail a abordé le problème de la datation de manuscrits historiques en utilisant des caractéristiques basées sur les polygones et les chaînes de Freeman. Les séparateurs à vastes marges (SVMs), les forêts aléatoires, un algorithme de boosting adaptatif (Adaboost) ainsi qu'un classificateur à renforcement de gradient (GBT) ont été employés pour la classification. La méthode proposée a été évaluée en utilisant une base de données contenant 2000 manuscrits anciens où des performances très encourageantes ont été réalisées en particulier lors de l'utilisation d'un classificateur à renforcement de gradient (GBT).

Les études ultérieures sur ce sujet seront destinées à introduire des caractéristiques supplémentaires et ensuite appliquer un mécanisme de sélection de caractéristiques pour savoir quelles sont les caractéristiques les plus discriminantes pour ce problème et pour des problèmes similaires. Il est nécessaire de rappeler que la performance du système proposé ne dépend pas seulement de la technique de classification utilisée, mais aussi des caractéristiques choisies.

Dans ce cadre, il serait très intéressant d'exploiter la combinaison de caractéristiques proposées dans ce mémoire afin d'améliorer les performances de la méthode proposé. Pour les techniques de classification utilisées, nous pensons qu'il serait intéressant d'envisager l'utilisation d'autres techniques de classification que celles que nous avons adoptées dans le présent mémoire. Il serait aussi très intéressant aussi d'envisager et d'expérimenter des possibilités de combinaison de techniques de classification et de considérer des bases de données plus volumineuses que celle utilisée dans le cadre du présent travail.

Bibliographie

[1] Adam K, Al-Maadeed S, Bouridane A (2017) Letter-based classification of Arabic scripts style in ancient Arabic manuscripts. IEEE International Workshop on Arabic Script Analysis and Recognition (ASAR), pp 95–98.

[2] Nesmerak K, Nemcova I (2012), Dating of historical manuscripts using spectrometric methods: a MiniReview. Anal Lett 45(4):330–344.

[3] Hamid A, Bibi M, Moetesum M, Siddiqi I (2019) Deep learning based approach for historical manuscript dating. International Conference on Document Analysis and Recognition (ICDAR), pp 967–972.

[4] Bourgeois FL, Trinh E, Allier B, Eglin B, Emptoz H (2004) Document images analysis solutions for digital libraries. In: Proceedings of First International Workshop on Document Image Analysis for Libraries. IEEE, pp 2–24.

[5] Moalla I, LeBourgeois F, Emptoz H, Alimi AM (2006) Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography. In: Bunke H, Spitz AL (eds) Document Analysis Systems VII. DAS 2006. Lecture Notes in Computer Science, vol 3872. Springer, Berlin.

[6] Rehbein M, Sahle P, Schaßan T (2009) Codicology and Palaeography in the Digital Age. Books on Demand GmbH, Norderstedt, pp 219–235.

[7] Duran A, Perez-Rodriguez JL, Espejo T, Franquelo ML, Castaing J, Walter P (2009) Characterization of illuminated manuscripts by laboratory-made portable XRD and micro-XRD systems. Anal Bioanal Chem 395:1997–2004.

[8] He S, Samara P, Burgers J, Schomaker L (2016a) Image-based historical manuscript dating using contour and stroke fragments. Pattern Recogn 58:159–171.

[9] Omayio, E.O., Indu, S. & Panda, J. Historical manuscript dating: traditional and current trends. Multimed Tools Appl (2022).

[10] Metzger BM (1981) Manuscripts of the greek bible: an introduction to greek paleography. Oxford University Press, Inc., pp 14.

[11] Garain U, Parui SK, Paquet T, Heutte L (2007) Machine dating of handwritten manuscripts. 9th International Conference on Document Analysis and Recognition (ICDAR).

[12] Head PM (1995) The date of the Magdalen papyrus of Mathew (p.MAGD.GR. 17¯ P64): A response to C.P. Thiende, Tyndale Bullet 46:251–285.
[13] He S, Samara P, Burgers J, Schomaker L (2014) Towards style-based dating of historical documents. In: International Conference of Frontiers in Handwriting Recognition (ICFHR), pp 265–270.
[14] Feuerverger A, Hall P, Tilahun G, Gervers M (2008) Using statistical smoothing to date medieval manuscripts. Inst Math Stat Collect 1:321–331.
[15] Fernandez-Mota D, Almazan J, Cirera N, Forn´es A, Llad´os J (2014) BH2M: the Barcelona Historical Handwritten Marriages database. 22nd International Conference on Pattern Recognition, pp 256–261.
[16] Stutzmann D (2016) Clustering of medieval scripts through computer image analysis: towards an evaluation protocol. Digit Medievalist J 10.
[17] Kalthoum Adam, Asim Baig, Somaya Al-Maadeed, Ahmed Bouridane, Sherine El-Menshawy, "KERTAS: dataset for automatic dating of ancient Arabic manuscripts," 8 September 2018, Sep. 2018.
[18] Kalthoum Adam, Somaya Al-Maadeed, Younes Akbari, "Hierarchical Fusion Using Subsets of Multi-Features for Historical Arabic Manuscript Dating," Mar. 2022.
[19] He, S.; Sammara, P.; Burgers, J.; Schomaker, L., Towards style-based dating of historical documents. In Proceedings of the 2014 14th International Conference on Frontiers in Handwriting Recognition. Hersonissos, Greece, 2014.
[20] He, S.; Samara, P.; Burgers, J.; Schomaker, L, "Image-based historical manuscript dating using contour and stroke fragments. Pattern Recognit.," pp. 58, 159-171., 2016.
[21] He, S.; Samara, P.; Burgers, J.; Schomaker, L., "A multiple-label guided clustering algorithm for historical document dating and localization. IEEE Trans. Image Process.," p. 25,5252-5265, 2016.
[22] Wahlberg, F.; Mårtensson, L.; Brun, A, "Large scale style based dating of medieval manuscripts. In Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing," Nancy, France, pp. 107–114, Aug. 22, 2015.
[23] Li, Y.; Genzel, D.; Fujii, Y.; Popat, A.C., "Publication date estimation for printed historical documents using convolutional neural networks. In Proceedings of the 3rd International Workshop

on Historical Document Imaging and Processing.," Nancy, France, pp. 99–106, Aug. 22, 2015.
[24] Vincent, L, Google book search: Document understanding on a massive scale. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2. Curitiba, Brazil, 2007.
[25] Hamid, A.; Bibi, M.; Siddiqi, I.; Moetesum, M., "Historical manuscript dating using textural measures. In Proceedings of the 2018 International Conference on Frontiers of Information Technology (FIT)," Islamabad, Pakistan, pp. 235–240, Dec. 17, 2018.
[26] Hamid, A.; Bibi, M.; Moetesum, M.; Siddiqi, I, "Deep Learning Based Approach for Historical Manuscript Dating. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)," Sydney, NSW, Australia, pp. 967–972, Sep. 20, 2019.
[27] Studer, L.; Alberti, M.; Pondenkandath, V.; Goktepe, P.; Kolonko, T.; Fischer, A.; Liwicki, M.; Ingold, R, "A comprehensive study of ImageNet pre-training for historical document image analysis. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR).," Sydney, NSW, Australia, pp. 720–725, Sep. 20, 2019.
[28] Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; Ha, D., "Deep learning for classical Japanese literature. arXiv 2018, arXiv:1812.01718.," 2018.
[29] Cloppet, F.; Eglin, V.; Helias-Baron, M.; Kieu, C.; Vincent, N.; Stutzmann, D., "Icdar2017 competition on the classification of medieval handwritings in latin script. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Volume 1.," Kyoto, Japan, pp. 1371–1376, Nov. 09, 2017.
[30] Fiel, S.; Kleber, F.; Diem, M.; Christlein, V.; Louloudis, G.; Nikos, S.; Gatos, B, "Icdar2017 competition on historical document writer identification (historical-wi). In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR); Volume 1," Kyoto, Japan, pp. 377–1382, Nov. 09, 2017.
[31] Simistira, F.; Seuret, M.; Eichenberger, N.; Garz, A.; Liwicki, M.; Ingold, R. Diva-hisdb, "A precisely annotated large dataset of challenging medieval manuscripts. In Proceedings of the 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).," Shenzhen, China, pp. 471–476, Oct. 23, 2016.

[32] Rahiche, A.; Hedjam, R.; Al-maadeed, S.; Cheriet, M, "Historical documents dating using multispectral imaging and ordinal classification.," J. Cult. Herit, pp. 45, 71–80, 2020.
[33] Dhali, M.A.; Jansen, C.N.; de Wit, J.W.; Schomaker, L., "Feature-extraction methods for historical manuscript dating based on writing style development. Pattern Recognit. Lett.," pp. 131, 413–420, 2020.
[34] Wall, K., Danielsson P., "A fast sequential method for polygonal approximation of digitized curves. Computer Vision, Graphics and Image Processing, vol.28," pp. 220–227, 1984.
[35] I. Siddiqi and N. Vincent, "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features," Pattern Recognition, vol. 43, no. 11, pp. 3853–3865, Nov. 2010, doi: 10.1016/j.patcog.2010.05.019.
[36] Imran Siddiqi, Nicole Vincent, "Combining Contour Based Orientation and Curvature Features for Writer Recognition," Laboratoire CRIP5 –SIP, Paris Descartes University, 45 Rue des Saint Pères, 75006 France, 2009.
[37] SURYANNARAYANA CHANDAKA, AMITAVA CHATTERJEE, "Cross-correlation aided support vector machine classifier for classification of EEG signals, ELSEVIER .," 2009.
[38] Periklis Gogas,Theophilos Papadimitriou, "Emerging Methodologies in Economics and Finance," Department of Economics, Democritus University of Thrace, Greece.
[39] Pal, M, "RANDOM forest classifier for remote sensing classification, International Journal of Remote Sensing .," 2003.
[40] Nikolaos Sapountzoglou, Jesus Lago, and Bertrand Raison, "Fault diagnosis in low voltage smart distribution grids using gradient boosting trees," pp. 3–4, May 2020.
[41] SCHAPIRE, Yoav Freund Robert E, "Experiments with a New Boosting Algorithm, AT&T Research," 1996.
[42] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and DataMining – KDD'16, ACM Press, San Francisco, California, USA, 2016, pp. 785–794.
[43] T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, SpringerSeries in Statistics, Springer New York, New York, NY, 2001.
[44] Petro Liashchynskyi and Pavlo Liashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS," 2019, p. 3,6.