**People's Democratic Republic of Algeria**
**Ministry of Higher Education and Scientific Research**
**University Echahid Cheikh Larbi Tebessi- Tebessa**

**Faculty of Exact Sciences and Natural and Life Sciences**
**Department of computer science/mathematics**
**Domiciliation laboratory: laboratory of mathematics, informatics and systems (LAMIS)**

# Thesis

## Presented and Publicly Defended for Obtaining the Diploma of Doctorate in Third Cycle

**By:**
**Merouane Boudraa**

**Domain:** mathematics and computer science **Stream:** computer science
**Specialty**: Information Systems

## *Title*

**Contribution to historical documents classification**

Defended on : 15/05/2025 Before the jury composed of:

| Full name | Rank | University | |
|---|---|---|---|
| **Mr Hakim BENDJENNA** | PROF | Univ. of Larbi Tebessi | President |
| **Mr Akram BENNOUR** | PROF | Univ. of Larbi Tebessi | Supervisor |
| **Mr Tahar MEKHAZNIA** | PROF | Univ. of Larbi Tebessi | Co- supervisor |
| **Mr Makhlouf DERDOUR** | PROF | Univ. of Oum el Bouaghi | Examiner |
| **Mr Mohamed GASMI** | MCA | Univ. of Larbi Tebessi | Examiner |
| **Mr Soufiane KHEDAIRIA** | MCA | Univ. of Souk ahras | Examiner |

**Academic year**: 2024/2025

# Abstract

Historical documents play a crucial role in unraveling the mysteries of history. However, studying these documents is a complex, time-consuming, and resource-intensive task, requiring significant expertise. With the rise of digital humanities and the availability of digital versions of documents, automating tasks related to these documents has become a possibility. One of the key challenges in this domain is "*the classification of historical documents*". Early studies relied on traditional machine learning techniques that required manual feature extraction, but these methods often fell short in terms of accuracy. The advent of deep learning has opened new avenues for improving classification results, with various deep learning architectures being explored in this field. Despite the potential of deep learning, further exploration and the development of more efficient techniques are needed.

This realization led us to contribute to the literature with our study on historical document classification. Our research provides a comprehensive exploration of the topic from both historical and computational perspectives. We began by understanding the complexities of historical document classification and analyzing existing studies. This analysis allowed us to identify key challenges and difficulties, guiding the direction of our contribution.

Our approach begins with precise preprocessing techniques, including denoising via the Non-Local Means algorithm and binarization using the Canny edge detector. After these initial steps, we move into feature detection by applying the Harris corner detector to identify keypoints within the manuscript. These keypoints are then clustered using the k-means algorithm, allowing us to extract meaningful patches of predefined dimensions. This process not only isolates significant visual features but also supports systematic data augmentation to enhance the depth and diversity of our dataset. In the final stage, we harness the capabilities of vision transformers—a powerful deep learning architecture—finetuned to our specific classification tasks. By integrating automatically extracted handcrafted features, the model evolves into a robust and highly effective classification system for historical document analysis. As an added layer of refinement, we deploy a majority vote mechanism on image patches, meticulously engineered to heighten system accuracy. These findings have been validated and published in leading journals and conference proceedings, providing a strong foundation for future research in historical document analysis.

**Keywords**: Historical Documents, Machine Learning, Deep Learning, Classification, Dating, Writer Identification, Script Classification, Convolutional Neural Networks, Vision Transformers, Features Extraction, Computer Vision, Pattern Recognition.

# Resumé

Les documents historiques jouent un rôle crucial dans le décryptage des mystères de l'histoire. Cependant, l'étude de ces documents est une tâche complexe, chronophage et exigeante en ressources, nécessitant une expertise considérable. Avec l'émergence des humanités numériques et la disponibilité de versions numériques des documents, l'automatisation des tâches liées à ces documents est devenue possible. L'un des principaux défis dans ce domaine est « la classification des documents historiques ». Les premières études se sont appuyées sur des techniques traditionnelles d'apprentissage automatique nécessitant une extraction manuelle des caractéristiques, mais ces méthodes manquaient souvent de précision. L'avènement de l'apprentissage profond a ouvert de nouvelles perspectives pour améliorer les résultats de classification, avec diverses architectures d'apprentissage profond explorées dans ce domaine. Malgré le potentiel de l'apprentissage profond, des recherches supplémentaires et le développement de techniques plus efficaces sont encore nécessaires.

Cette prise de conscience nous a conduits à contribuer à la littérature avec notre étude sur la classification des documents historiques. Notre recherche propose une exploration complète du sujet, à la fois d'un point de vue historique et informatique. Nous avons commencé par comprendre les complexités de la classification des documents historiques et analyser les études existantes. Cette analyse nous a permis d'identifier les défis et difficultés clés, orientant la direction de notre contribution.

Notre approche commence par des techniques de prétraitement précises, incluant le débruitage via l'algorithme Non-Local Means et la binarisation à l'aide du détecteur de contours de Canny. Une fois ces étapes initiales accomplies, nous procédons à la détection de caractéristiques en appliquant le détecteur de coins de Harris pour identifier les points clés dans le manuscrit. Ces points clés sont ensuite regroupés par l'algorithme k-means, ce qui nous permet d'extraire des fragments pertinents de dimensions prédéfinies. Ce processus permet non seulement d'isoler des caractéristiques visuelles significatives, mais aussi de soutenir une augmentation systématique des données afin d'enrichir la profondeur et la diversité de notre ensemble de données. Dans la phase finale, nous exploitons les capacités des vision transformers — une architecture puissante d'apprentissage profond — ajustée avec finesse à nos tâches spécifiques de classification. En intégrant des caractéristiques faites main extraites automatiquement, le modèle devient un système de classification robuste et hautement efficace pour l'analyse de documents historiques. En guise de raffinement supplémentaire, nous déployons un mécanisme de vote majoritaire sur les fragments d'image, soigneusement conçu pour améliorer la précision du système. Ces résultats ont été validés et publiés dans des revues et conférences de renom, constituant une base solide pour les recherches futures en analyse de documents historiques.

**Mots-clés :** Documents historiques, apprentissage automatique, apprentissage profond, classification, datation, identification d'écrivains, classification de scripts, réseaux neuronaux convolutifs, transformateurs de vision, extraction de caractéristiques, la vision par ordinateur, la reconnaissance de formes.

# ملخص

تلعب المخطوطات التاريخية دورًا حيويًا في كشف أسرار التاريخ. ومع ذلك، فإن دراسة هذه المخطوطات تُعد مهمة معقدة وتستهلك الكثير من الوقت والموارد، وتتطلب خبرة كبيرة. مع ظهور العلوم الإنسانية الرقمية وتوفر النسخ الرقمية للمخطوطات، أصبح من الممكن أتمتة المهام المتعلقة بهذه الوثائق. يُعد أحد التحديات الرئيسية في هذا المجال هو "تصنيف المخطوطات التاريخية". اعتمدت الدراسات الأولى على تقنيات التعلم الآلي التقليدية التي تتطلب استخراجًا يدويًا للميزات، لكن هذه الأساليب غالبًا ما كانت تعاني من نقص في الدقة. فتح ظهور التعلم العميق آفاقًا جديدة لتحسين نتائج التصنيف، حيث تم استكشاف العديد من الهياكل المعمارية للتعلم العميق في هذا المجال. على الرغم من إمكانات التعلم العميق، لا تزال هناك حاجة لمزيد من الاستكشاف وتطوير تقنيات أكثر كفاءة.

قادنا هذا الإدراك إلى المساهمة في الأدبيات من خلال دراستنا حول تصنيف المخطوطات التاريخية. تقدم أبحاثنا استكشافًا شاملاً للموضوع من الناحيتين التاريخية والحاسوبية. بدأنا بفهم تعقيدات تصنيف المخطوطات التاريخية وتحليل الدراسات الموجودة. سمح لنا هذا التحليل بتحديد التحديات والصعوبات الرئيسية، مما وجه مسار مساهمتنا.

تبدأ منهجيتنا بتقنيات معالجة أولية دقيقة، تشمل إزالة الضوضاء باستخدام خوارزمية Non-Local Means، والتثنية (تحويل الصورة إلى أبيض وأسود) باستخدام كاشف الحواف .Canny بعد الانتهاء من هذه الخطوات الأولية، ننتقل إلى مرحلة اكتشاف السمات من خلال تطبيق كاشف الزوايا Harris لاستخراج النقاط المفتاحية داخل المخطوطة. يتم بعد ذلك تجميع هذه النقاط باستخدام خوارزمية k-means، مما يتيح لنا استخراج أجزاء ذات أبعاد محددة مسبقًا. لا تساعد هذه العملية فقط في عزل السمات البصرية المهمة، بل تدعم أيضًا زيادة ممنهجة للبيانات بهدف تعزيز عمق وتنوع مجموعة البيانات. في المرحلة الأخيرة، نستفيد من قدرات بنية التعلم العميق القوية المعروفة بـ Vision Transformers، حيث يتم تعديلها بدقة لتناسب مهام التصنيف الخاصة بنا. ومن خلال دمج السمات المُعدّة يدويًا والمستخرجة تلقائيًا، يتحول النموذج إلى نظام تصنيف قوي وفعّال للغاية لتحليل الوثائق التاريخية. وكطبقة إضافية من التحسين، نعتمد آلية التصويت بالأغلبية على أجزاء الصور، والتي صُممت بعناية لرفع دقة النظام. وقد تم التحقق من هذه النتائج ونشرها في مجلات ومؤتمرات مرموقة، مما يوفر أساسًا قويًا لأبحاث مستقبلية في مجال تحليل الوثائق التاريخية.

**الكلمات المفتاحية:** المخطوطات التاريخية، التعلم الآلي، التعلم العميق، التصنيف، التأريخ، التعرف على الكاتب، تصنيف النصوص، الشبكات العصبية الالتفافية، محولات الرؤية، استخراج الميزات، رؤية الكمبيوتر، التعرف على الأنماط.



# Acknowledgment

First and foremost, Alhamdulillah, the Lord of all creation, for His countless blessings and guidance in helping me reach this milestone.

I extend my deepest gratitude to my supervisor, Dr. Bennour Akram, whose support, encouragement, and belief in my abilities went beyond that of a supervisor and truly embodied the role of a mentor and coach. His invaluable guidance and optimism made this journey not only manageable but rewarding.

My sincere appreciation also goes to the LAMIS Laboratory, particularly to Prof. Benjenna Hakim, for creating a supportive research environment and for their continuous efforts in aiding my work.

I am deeply grateful to my parents for their endless patience, support, and sacrifices. Their unwavering belief in me has been my strength throughout this journey. My heartfelt thanks also go to my brothers and sisters, who have always stood by me with encouragement and understanding.

Finally, to my friends and colleagues who have supported me in various ways, thank you for making this journey meaningful and memorable.





# *Dedication*

*To my beloved parents, for their endless love, sacrifices, and unwavering faith in me.*

*To my brothers and sisters, for their constant encouragement and support.*

*And to all those who dream, work hard, and never give up.*



# Table of Contents


1. **GENERAL INTRODUCTION ..... I**
2. **Understanding 'Historical documents' ..... 5**
   - 1.1 Introduction ..... 5
   - 1.2 Defining Historical documents ..... 5
   - 1.3 Material of Historical documents ..... 6
   - 1.4 Types of historical documents ..... 8
   - 1.5 Role of Historical documents ..... 9
   - 1.6 Characteristics of Historical documents ..... 10
     - 1.6.1 Moral Characteristics ..... 10
     - 1.6.2 Physical Characteristics ..... 11
   - 1.7 Paleography ..... 12
     - 1.7.1 Paleographic Features ..... 12
     - 1.7.2 Role of Paleography ..... 13
     - 1.7.3 Digital Paleography ..... 14
   - 1.8 Conclusion ..... 15
3. **Historical documents classification ..... 17**
   - 2.1 Introduction ..... 17
   - 2.2 Understanding *"Historical documents Classification"* ..... 17
   - 2.3 Documents classification tasks ..... 18
     - 2.3.1 Dating ..... 18
     - 2.3.2 Writer identification ..... 19
     - 2.3.3 Script Type Classification ..... 20
   - 2.4 Historical documents classification methods ..... 21
     - 2.4.1 Physical Methods ..... 22
     - 2.4.2 Paleographic methods ..... 28
     - 2.4.3 Computer-based methods ..... 32
     - 2.4.4 Conclusion ..... 32
4. **Computer based methods ..... 33**
   - 3.1 Introduction ..... 33
   - 3.2 What are computer-based methods? ..... 33

3.3 Documents Databases ..... 34
3.4 Artificial Intelligence ..... 35
3.5 Preprocessing Techniques ..... 37
3.5.1 Binarization ..... 37
3.5.2 Noise Reduction ..... 39
3.5.3 Contrast Enhancement ..... 42
3.5.4 Resizing and Cropping ..... 43
3.6 Feature Extraction ..... 43
3.6.1 Handcrafted Features ..... 44
3.6.2 Deep Learning Features ..... 47
3.7 Classification ..... 50
3.8 Conclusion ..... 52
5. **Related Works ..... 53**
4.1 Introduction ..... 53
4.2 Datasets ..... 53
4.2.1 Medieval Paleographical Scale (MPS) ..... 53
4.2.2 Svenskt Diplomatariums Huvudkartotek (SDHK) ..... 54
4.2.3 KERTAS dataset ..... 54
4.2.4 DEEDS (Document of Early England Data Set) ..... 55
4.2.5 ICFHR 2016 Competition on the Classification of Medieval Handwritings in Latin Script ..... 56
4.2.6 ICDAR 2017 Competition on the Classification of Medieval Handwritings in Latin Script ..... 57
4.2.7 ICDAR 2021 Competition on Historical Document Classification ..... 58
4.2.8 ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI) ..... 59
4.2.9 ICDAR 2019 Competition on Image Retrieval for Historical Handwritten Documents ..... 59
4.2.10 ICFHR 2020 Competition on Image Retrieval for Historical Handwritten Fragments ..... 60
4.2.11 GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification ..... 61
4.3 Writer identification related works ..... 62
4.4 Dating documents related works ..... 63

4.5 Script type recognition related works ..... 65
4.6 Analysis of existing methodologies ..... 67
4.7 Conclusion ..... 68
6. **Contributions ..... 69**
5.1 Introduction ..... 69
5.2 System Overview ..... 69
5.3 Preprocessing ..... 71
5.3.1 Documents denoising ..... 72
5.3.2 Documents binarization ..... 76
5.4 Feature extraction ..... 80
5.4.1 FAST (Features from Accelerated Segment Test) ..... 80
5.4.2 SIFT (Scale-Invariant Feature Transform) ..... 81
5.4.3 Harris Corner Detection ..... 83
5.4.4 Clustering ..... 84
5.4.5 Patch extraction ..... 85
5.5 Classification ..... 87
5.5.1 Convolutional Neural Network (CNN) ..... 87
5.5.2 Vision Transformers ..... 89
5.5.3 Transfer Learning ..... 91
5.5.4 Majority Vote ..... 93
5.6 Contribution to writer identification task ..... 94
5.7 Contribution to Script recognition task ..... 96
5.8 Contribution to Dating task ..... 98
5.9 Conclusion ..... 99
7. **Experiments & Results ..... 100**
6.1 Introduction ..... 100
6.2 Datasets & Evaluation metrics ..... 100
6.3 Denoising techniques experiment ..... 103
6.3.1. Visual Quality and Denoising Efficiency ..... 105
6.3.2 Impact on Binarization ..... 105
6.3.3 Implications for Specific Tasks ..... 106
6.4 Binarization techniques experiments ..... 106

6.4.1 Comparison of Binarization Techniques ..... 107
6.4.2 Impact on Feature Detection and Key-Point Distribution ..... 108
6.4.3 Implications for Classification Tasks ..... 108
6.5 keypoints detection techniques experiment ..... 109
6.5.1 Comparison of Keypoint Detection Methods ..... 109
6.5.2 Implications for Classification Tasks ..... 110
6.6 Hyper-parameters experiment ..... 110
6.7 Impact of preprocessing ..... 112
6.8 Impact of patch size ..... 113
6.9 Impact of number of clusters ..... 115
6.10 Model Selection ..... 117
6.11 Final system results ..... 120
6.12 Performance comparison ..... 124
6.13 Conclusion ..... 127
8. **Conclusion & Perspectives ..... 128**
9. **References ..... 130**

# List of Figures

**FIGURE 1.1**. HISTORICAL DOCUMENTS (HANDWRITTEN). ..... 6
**FIGURE 1.2**. HISTORICAL DOCUMENTS(PRINTED). ..... 6
**FIGURE 1.3**. DIFFERENT WRITING MATERIALS OF HISTORICAL DOCUMENTS. ..7
**FIGURE 1.4** PALAEOGRAPHICAL FEATURES. **[32]** ..... 12
**FIGURE 2.1** HISTORICAL DOCUMENTS TASKS ..... 18
**FIGURE 2.2** TYPES OF HISTORICAL DOCUMENTS CLASSIFICATION METHODS 21
**FIGURE 2.3** TYPES OF HISTORICAL DOCUMENTS CLASSIFICATION METHODS 22
**FIGURE 2.4** TYPES OF HISTORICAL DOCUMENTS CLASSIFICATION METHODS **[35]** ..... 23
**FIGURE 2.5** BRIGHT-FIELD TEM IMAGE OF A FRAGMENT **[54]** ..... 24
**FIGURE 2.6** RGB (LEFT), MULTISPECTRAL (CENTER)., HYPERSPECTRAL IMAGING (RIGHT) **[55]** ..... 27
**FIGURE 2.7** MULTISPECTRAL IMAGE OF THE HMML PALIMPSEST, WITH AREAS OF SYRIAC UNDERWRITING CIRCLED (IMAGE: MB TOTH) **[56]** ..... 28
**FIGURE 3.1** COMPUTER BASED METHODS CATEGORIES FOR CLASSIFICATION OF HISTORICAL DOCUMENTS ..... 34
**FIGURE 3.2** EXAMPLE OF HISTORICAL DOCUMENTS DATABASE **[57]** ..... 34
**FIGURE 3.3** ARTIFICAIL INTELLIGENVE SUBFIELDS RELATED TO HISTORICAL DOCUMENTS CLASSIFICATION USING IMAGE BASED METHODS ..... 36
**FIGURE 3.4** BINARIZATION OF ANCIENT DOCUMENT USING SAUVOLA'S THRESHOLDING **[58]** ..... 37
**FIGURE 3.5** BINARIZATION TYPES OF TECHNIQUES ..... 38
**FIGURE 3.6** NOISE FORMS IN HISTORICAL DOCUMENTS **[66]** ..... 40
**FIGURE 3.7** DENOISING TYPES OF TECHNIQUES ..... 40
**FIGURE 3.8** DEMONSTRATION OF RESIZING AND CROPPING TECHNIQUES ..... 43
**FIGURE 3.9** FEATURE EXTRACTION TECHNIQUE CATEGORIES ..... 44
**FIGURE 3.10** STRUCTURE OF CHARACTERS IN HISTORICAL DOCUMENTS OVER TIME **[75]** ..... 45
**FIGURE 3.11** EXAMPLE OF STRUCTURAL FEATURES FROM OLD DOCUMENT **[75]** ..... 46
**FIGURE 3.12** ARCHITECTURE OF A CNN **[81]** ..... 48
**FIGURE 3.13** RNN ARCHITECTURE **[86]** ..... 49

**FIGURE 3.14** TRANSFORMER ARCHITECTURE ..... 49
**FIGURE 3.15** DIFFERENT CLASSIFICATION ALGORITHMS ..... 51
**FIGURE** 4.**1** DOCUMENT SAMPLE FROM MPS DATASET ..... 54
**FIGURE** 4.**2** DOCUMENT SAMPLE FROM SDHK DATASET ..... 54
**FIGURE** 4.**3** DOCUMENT SAMPLE FROM KERTAS DATASET ..... 55
**FIGURE** 4.**4** DOCUMENT SAMPLE FROM DEEDS DATASET ..... 56
**FIGURE** 4.**5** DOCUMENT SAMPLE FROM CLAMM 2016 DATASET ..... 57
**FIGURE** 4.**6** DOCUMENT SAMPLE FROM CLAMM 2017 DATASET ..... 58
**FIGURE** 4.**7** DOCUMENT SAMPLE FROM HDC 2021 DATASET ..... 58
**FIGURE** 4.**8** DOCUMENT SAMPLE FROM HISTORICAL-WI 2017 DATASET ..... 59
**FIGURE** 4.**9** DOCUMENT SAMPLE FROM ICDAR 2019 DATASET ..... 60
**FIGURE** 4.**10** DOCUMENT SAMPLE FROM ICFHR 2020 DATASET ..... 61
**FIGURE** 4.**11** DOCUMENT SAMPLE FROM GRK-PAPYRI DATASET ..... 61
**FIGURE 5.1** HISTORICAL DOCUMENTS CLASSIFICATION SYSTEM DESIGN ..... 70
**FIGURE 5.2** PROBLEMS WITH CLAMM SAMPLES: (A) WORN PAPER, (B) INK TRACES, (C) INDISTINCT TEXT. ..... 72
**FIGURE 5.3:** INFLUENCE OF DENOISING ON THE BINARIZATION PROCESS. ..... 73
**FIGURE 5.4** EXAMPLE OF PATCHES EXTRACTED FROM DIFFERENT SAMPLES 87
**FIGURE 5.5** CONVOLUTION NEURAL NETWORKS ARCHITECTURE ..... 89
**FIGURE 5.6** VISION TRANSFORMER ARCHITECTURE (VIT). ..... 91
**FIGURE 5.7** TRADITIONAL CNN, RESNET, DENSENET ARCHITECTURES. ..... 92
**FIGURE 5.8** SHIFTED WINDOWS TRANSFORMER ARCHITECTURE (SWIN). ..... 92
**FIGURE 5.9** DATA EFFICIENT IMAGE TRANSFORMER ARCHITECTURE (DEIT) ..93
**FIGURE 5.10** MAJORITY VOTE TECHNIQUE. ..... 94
**FIGURE 5.11** KEY STEPS IN THE PROPOSED SYSTEM. ..... 95
**FIGURE 5.12** SCRIPT CLASSIFICATION SYSTEM DESIGN ..... 97
**FIGURE 5.13** DATING SYSTEM DESIGN. ..... 99
**FIGURE 6.1** ORIGINAL SAMPLES FROM HISTORICAL-WI DATASET [95]. ..... 101
**FIGURE 6.2** SCRIPT SAMPLES FROM CLAMM DATASET. ..... 102
**FIGURE 6.3** SCRIPT SAMPLES FROM MPS DATASET. ..... 103
**FIGURE 6.4** CHART OF PATCH SIZE IMPACT WITH METRICS BASED ON EACH TASK (WRITER, SCRIPT, DATE) ..... 114
**FIGURE 6.5** CHART OF CLUSTERS NUMBER IMPACT WITH METRICS BASED ON EACH TASK (WRITER, SCRIPT, DATE) ..... 116

**FIGURE 6.6** CHART OF MODELS PERFORMANCE ON WRITER TASK.....118
**FIGURE 6.7** CHART OF MODELS PERFORMANCE ON SCRIPT TASK .....119
**FIGURE 6.8** CHART OF MODELS PERFORMANCE ON DATING TASK.....120
**FIGURE 6.9** FINAL SYSTEM LOSS AND ACCURACY FOR WRITER TASK. .....121
**FIGURE 6.10** CONFUSION MATRIX OF 10 CLASSES FOR WRITER TASK. .....121
**FIGURE 6.11** CONFUSION MATRIX FOR SCRIPT TASK, (A) VIT, (B) CNN. .....122
**FIGURE 6.12** CONFUSION MATRIX FIR DATING TASKS AT BOTH PATCH AND DOCUMENT LEVELS).....123
**FIGURE 6.13** GRAPHS ILLUSTRATING THE FINAL SYSTEM'S LOSS AND ACCURACY AT THE PATCH-LEVEL FOR DATING TASK.....123

# List of Tables

**TABLE 6.1** SCRIPT NUMBER OF SAMPLES PER CLASS..... 102
**TABLE 6.2** NUMBER OF SAMPLES IN EACH PERIOD OF THE MPS DATASET..... 103
**TABLE 6.3** OUTCOMES OF DIFFERENT DENOISING TECHNIQUES .....104
**TABLE 6.4** OUTCOMES OF BINARIZATION METHODS ON DIFFERENT SAMPLES..... 107
**TABLE 6.5** COMPARISON BETWEEN DIFFERENT FEATURE DETECTION METHODS.... 109
**TABLE 6.6** WRITER TASK HYPERPARAMETERS EXPERIMENT RESULTS USING CNN. 111
**TABLE 6.7** SCRIPT TASK HYPERPARAMETERS EXPERIMENT RESULTS FOR BOTH CNN & VIT..... 111
**TABLE 6.8** DATING TASK HYPERPARAMETERS EXPERIMENT RESULTS USING VIT...111
**TABLE 6.9** IDENTIFICATION PERFORMANCE WITH AND WITHOUT PREPROCESSING STEPS..... 112
**TABLE 6.10** PATCH SIZE IMPACT ON MODEL PERFORMANCE ACROSS DIFFERENT TASKS..... 113
**TABLE 6.11** PERFORMANCE ON DIFFERENT PRE-TRAINED MODELS FIR WRITER TASK..... 118
**TABLE 6.12** MODEL'S EXPERIMENT RESULTS FOR SCRIPT TASK..... 119
**TABLE 6.13** PERFORMANCE ON DIFFERENT PRE-TRAINED MODELS FOR DATING TASK..... 120
**TABLE 6.14** FINAL SYSTEM RESULTS FOR WRITER TASK (PATCH-LEVEL VS. DOCUMENT-LEVEL PERFORMANCE)..... 121
**TABLE 6.15** FINAL SYSTEM RESULTS FOR SCRIPT TASK (DOCUMENT-LEVEL PERFORMANCE)..... 122
**TABLE 6.16** FINAL SYSTEM RESULTS FOR DATING TASK (PATCH-LEVEL VS. DOCUMENT-LEVEL PERFORMANCE)..... 123
**TABLE 6.17** PERFORMANCE COMPARISON OUR SYSTEM FOR WRITER TASK WITH OTHER TECHNIQUES..... 124
**TABLE 6.18** PERFORMANCE COMPARISON OUR SYSTEM FOR SCRIPT TASK WITH OTHER TECHNIQUES..... 125
**TABLE 6.19** PERFORMANCE COMPARISON OUR SYSTEM FOR DATING TASK WITH OTHER TECHNIQUES..... 126

# GENERAL INTRODUCTION

Technological advancement has profoundly impacted every aspect of our lives, driven by the unprecedented speed of innovation witnessed in this century. Today, technology, particularly artificial intelligence (AI), is deeply embedded in everything we use. AI involves teaching machines to perform tasks autonomously, often surpassing human capabilities in specific areas due to advanced and innovative algorithms. One of the most significant breakthroughs in AI is machine learning (ML), which entails training machines to learn tasks based on explicit features. This development has made machines capable of learning and adapting, paving the way for further advancements. Deep learning, a subset of ML, has revolutionized the field of artificial intelligence. Unlike traditional ML, which relies on manually selected features, deep learning enables machines to learn from vast amounts of data and identify patterns that are not immediately apparent to humans. This ability has allowed machines to compete with and even surpass human performance in numerous tasks. The implications of these advancements are far-reaching, transforming how we approach problem-solving and opening doors to innovative ideas. The impact of AI spans various domains, including education **[1]**, medicine **[2]**, sports **[3]**, and beyond, highlighting its potential to drive global progress and improve humanity.

One of the complex domains that can be approached by artificial intelligence is history. For a long time, it seemed that this domain was completely unrelated to technology. However, with the rise of AI, historians have increasingly embraced AI tools to solve their intricate problems **[4]**. One of the most sophisticated areas of historical study is the analysis of historical documents **[5]**. Historians often need to examine thousands of documents individually, a task that requires considerable time, expertise, and resources, and does not always yield sufficient results. These documents, often handwritten on materials such as parchment, paper, papyrus, and more, hold significant and valuable information that can unlock many historical mysteries. They have the potential to answer critical questions and make discoveries across various domains **[6]**. Each document is a repository of valuable data, and their study encompasses tasks such as fraud verification, dating, writer identification, script classification, and layout segmentation. Therefore, studying these documents is of utmost importance. Traditionally, historians have relied on physical and chemical methods **[7,8]**, as well as the expertise of linguists familiar with ancient languages, to analyze these documents. These traditional methods, however, demand substantial time, money, and specialized knowledge, often limiting the ability to study more than one document thoroughly. In this context, artificial intelligence offers a promising alternative, capable of accelerating and enhancing the study of historical documents.

The digitization of documents over the last century has resulted in a significant number of document databases **[9,10,11]**, sparking increased interest in applying technology to this domain. The primary benefit of digitization has been the preservation of these documents

through the creation of digital versions, which allows simultaneous access for many scientists and researchers, thereby accelerating the study process. This digitization has also facilitated the development of algorithms and testing systems across various problems such as ***"historical documents classification".*** This involves categorizing documents based on various criteria such as writer, date, script...etc. The classification task is essential for better understanding and organizing historical data, which can lead to new insights and discoveries in the field of history.

Several approaches have been proposed in the literature to address the classification of historical documents, which can be broadly divided into two main categories: text-based approaches and image-based approaches. The primary distinction between these two lies in the features used for classification. Text-based approaches focus on the content of the documents, involving semantic analysis using natural language processing (NLP) algorithms **[12,13]**. This method, however, presents several challenges. Firstly, the quality of the documents is often poor, with many documents containing unclear or damaged text. Secondly, understanding the semantics of these historical documents requires a high level of expertise in both language and history, making it difficult for computers to accurately interpret the content without extensive domain knowledge. Due to these difficulties, most studies in the field have leaned towards image-based approaches for the classification of historical documents **[14]**. Image-based methods analyze the visual features of the documents, which can be categorized into textural-based features and grapheme-based features. These features are primarily extracted from the handwritten style of the document, which, according to experts, is a significant characteristic that can reveal a wealth of information about the document. This is where paleography, the study of ancient writing, plays a crucial role **[15]** Focusing on the visual attributes of handwriting, such as stroke patterns, letter shapes, and overall writing style, image-based approaches can provide a more reliable means of classification, bypassing the need for clear text and deep semantic understanding. Paleography is the study of ancient handwriting, focusing on the analysis and interpretation of historical scripts. Using image-based approaches to automate the task of paleography has demonstrated significant success in the classification of historical documents. These approaches extract paleographical features, including handwritten styles, curves, strokes, and other unique characteristics of the writing. Numerous studies in the literature have highlighted the effectiveness of these methods, which can be summarized in the context of the ICDAR competitions involving historical documents **[16,17,18]**.

Early studies in this field utilized traditional machine learning algorithms, which involved manually extracting features through statistical and mathematical techniques **[19]**. These methods proved to be effective compared to the physical and manual efforts of historians, offering a significant aid in the study of documents. However, with the advent of deep learning, recent studies have shifted towards solutions that include automatic feature extraction **[20]**. Deep learning approaches have significantly increased the accuracy of document classification, achieving impressive results. Despite these advancements, there is still a need for further development to fully harness the potential of deep learning in paleography and historical document classification. Analyzing the existing solutions in the

literature has allowed us to identify the overall strengths, weaknesses, and challenges inherent in these approaches. Despite notable advancements, a significant lack of accuracy remains a persistent issue. We can summarize the main difficulties into three primary points: lack of sufficient samples, poor documents quality, and complex feature-extraction algorithms. These challenges have led us to conceive a system that addresses these critical problems, aiming for optimal accuracy across various tasks.

Our image-based solution encompasses three key stages: preprocessing, feature extraction, and classification. To address issues stemming from poor document quality that hinders feature extraction, our preprocessing phase involves two primary sub-steps: denoising and binarization. These steps are crucial for enhancing document clarity and preparing them for subsequent processing. Next, we employ feature detection to identify regions of interest and clustering techniques to group keypoints, enabling the extraction of small patches from each document. These patches contain a wealth of features, significantly improving both data quality and quantity, thereby bolstering the effectiveness of our classification algorithm in the subsequent phase. For classification, we leverage deep learning techniques, specifically employing transfer learning tailored to the specific task of document classification under study. This approach ensures our system learns from existing knowledge and adapts it to achieve optimal classification performance. Finally, a majority vote approach is employed on the small patches to determine the classification of the final document. This method aggregates predictions from individual patches to arrive at a final decision, leveraging the collective insights gained from different parts of the document. This contribution led us to produce several publications encompassing different tasks, including writer identification **[21],** dating **[22,23]**, and script classification **[24,25]**. This dissertation is designed to comprehensively study the classification of historical documents, providing the necessary knowledge to the reader in both technical and historical aspects. This will enable a thorough understanding of our contributions, which will be explained in detail throughout the dissertation.

The structure of the dissertation is as follows:

**Chapter One** introduces the primary material of study: historical documents. This chapter covers the definition, characteristics, and importance of historical documents, establishing a foundation for the reader.

**Chapter Two** delves into the concept of "classification" in the context of historical documents. It explores various tasks associated with classification and the existing methods used, including physical methods and paleographical methods.

**Chapter Three** provides an in-depth explanation of computer-based methods for document classification. It discusses the two broad categories: text-based and image-based approaches, with detailed explanation of image-based approaches.

**Chapter Four** reviews the most relevant works in the literature, highlighting the advancements and limitations of existing solutions.

**Chapter Five** outlines our contributions and designed solution, detailing each component of our system and the rationale behind our approach.

**Chapter Six** presents the experiments, results, and discussion of our system's performance. This chapter focuses on fine-tuning our system and comparing it to state-of-the-art solutions.

**In Conclusion**, we end this document by discussing the work accomplished and discussing our future directions.

# Chapter 1

# Understanding ‘Historical documents’

## 1.1 Introduction

Before delving into the technical analysis of historical documents from a computer-based perspective, it is essential for us, as computer scientists, to establish a foundational understanding of our study's material: ***“historical documents”*** from a historical perspective. This entails clarifying the origin and precise meaning of the term to achieve clarity about the material under examination. Through a comprehensive definition of the term, our goal is to encapsulate our comprehension of the topic. Our inquiry will extend to understanding the methodologies of document creation, including the materials utilized and the origin of the tools employed in the writing process. This scrutiny aims to provide valuable insights into the intrinsic features of these documents, fostering a nuanced comprehension of the subject matter. Furthermore, we will categorize and define the types of historical documents, acknowledging their significance and importance in contributing to our understanding of history and culture. As we conclude this chapter, we will undertake the crucial task of defining the characteristics inherent in historical documents. Finally, we will explore the concepts of paleography and digital paleography, elucidating their relevance to our study and their significant contributions to the study of historical documents.

## 1.2 Defining Historical documents

We will begin our journey by dissecting the term ***"historical documents"*** Understanding its etymology and exact meaning is pivotal for our computational approach. Delving into the historical context of this term will not only solidify our grasp on the subject matter but will also guide our subsequent technical analyses. To gain a comprehensive understanding of historical documents, it is instructive to deconstruct this compound term into its constituent elements: *"historical"* and *"document"* By parsing these components, we can unearth the profound significance that historical documents hold:

- ***“Historical”*****: Chronicles of the Past:** The first element "*historical*", relates to, or is concerned with, the past, especially events, developments, or phenomena that occurred in the past and are significant or noteworthy. It encompasses anything related to history, which is the study and interpretation of past events, societies, and civilizations **[26]**.
- ***“Document”*****: The Craft of Handwriting:** The second component, "*document*," delves into the method of inscription, the term originates from the Latin words "manu" (meaning "by hand") and "scriptum" (meaning "written" or "to write"). These are not typewritten or digitally produced documents but the handwritten manifestations of human thought and expression. In every stroke of the pen or quill, we encounter the

artistry of the scribe. Documents are the tangible connection between the author and the reader, bearing the distinct character of the writer's hand **[27]**.

- ***"Historical documents"*:** Historical documents, those venerable custodians of the past, are far more than mere written records. They are the tangible echoes of bygone eras, the fragile vessels of human knowledge, and the intimate handwritten narratives of time itself **[28]**. Before we journey deeper into the realm of these extraordinary artifacts, we must first seek to define what distinguishes them from ordinary documents. At its core, a historical document is a handwritten document, inscribed by human hands in times long past (**Figure 1.1**). This manual transcription, by the very act of recording, imprints the identity of the scribe upon the page. Unlike printed books or modern typed texts (**Figure 1.2)**, historical documents bear the imprint of the author's hand – the swirls, strokes, and idiosyncrasies of their unique penmanship.



**Figure 1.1**. Historical documents (handwritten).

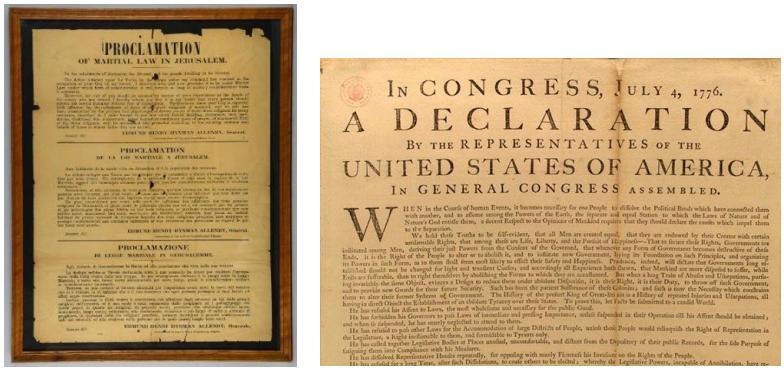



**Figure 1.2**. Historical documents(printed).

## 1.3 Material of Historical documents

In this section, we will unravel the intricate process of how historical documents are written. Examining the materials used, such as parchment, vellum, paper, and the tools employed, including quills and ink, will provide us with profound insights into the physical creation of these documents. The material of historical documents can vary widely depending on the time period, geographical location, and purpose of the document (**Figure 1.3)**. The choice of materials often reflects the technological and cultural characteristics of the time in which the document was produced:



**Figure 1.3**. Different writing materials of historical documents.

- **Parchment:** Parchment is one of the earliest writing materials and was made from the treated skins of animals, usually sheep or goats. It was durable and provided a smooth surface for writing.

- **Vellum:** Similar to parchment, vellum is made from animal skins, often of young calves. It is finer and more costly than parchment, and it was frequently used for high-quality documents, including illuminated documents.

- **Paper:** With the invention of paper, primarily in China, this material gradually replaced parchment and vellum in many regions. Paper was more affordable and easier to produce, making it a popular choice for documents.

- **Papyrus:** In ancient times, especially in Egypt, papyrus was commonly used as a writing material. It is made from the pith of the papyrus plant, and sheets were woven together to create scrolls or sheets.

- **Ink:** Historically, ink for writing was derived from various sources, including plant extracts, minerals, and animal products. Iron gall ink, made from gall nuts and iron sulfate, was widely used in medieval documents.

- **Illumination Materials:** Illuminated documents often used precious materials for decoration, including gold leaf, silver, and vibrant pigments made from minerals, plants, and insects.

- **Leather and Wood (Book Covers):** The covers of bound documents were often made from materials like leather, wood, or a combination of both. These covers provided protection for the fragile pages inside.

- **Metal Clasps and Fittings:** Some medieval documents featured metal clasps and fittings to keep the book securely closed. These were both functional and decorative elements.

- **Textile Materials (Bindings):** In addition to leather, textiles such as silk or velvet were sometimes used in the binding s of documents, adding a decorative touch.

## 1.4 Types of historical documents

A thorough exploration of the various types of historical documents is crucial for comprehending the breadth and depth of our study. We will categorize documents based on their content, purpose, and historical significance:

- **Religious Documents:** Includes sacred texts, scriptures, and religious commentaries. Examples include medieval illuminated documents of the Bible, Quranic documents, and Buddhist sutras.

- **Literary Documents:** Comprises handwritten copies of literary works, including novels, poems, plays, and other creative writings. Examples include medieval documents of works by authors like Chaucer or Shakespeare.

- **History Documents:** Includes chronicles, historical narratives, and official records. Examples are historical annals, royal charters, and chronicles detailing events and figures from the past.

- **Scientific Documents:** Encompasses handwritten documents related to scientific research, observations, and theories. Examples include documents by early scientists like Galileo or Newton.

- **Legal Documents:** Consists of legal documents such as contracts, wills, and deeds. These documents provide insights into legal practices and societal structures.

- **Philosophical Documents:** Contains handwritten works on philosophy, ethics, and metaphysics. Examples include philosophical treatises and dialogues from various periods.

- **Medical Documents:** Involves documents related to medical knowledge, treatments, and observations. Examples include medieval medical texts and early anatomical drawings.

- **Artistic Documents:** Encompasses documents created for artistic purposes, including illuminated documents, illustrated poetry collections, and calligraphy samples.

- **Travel Documents:** Includes handwritten accounts of travels, explorations, and geographical observations. Examples include medieval travelogues and expedition journals.
- **Musical Documents:** Involves handwritten scores, compositions, and musical notations. Examples include documents of classical music compositions.

- **Personal Correspondence:** Comprises letters, diaries, and personal writings. These documents offer insights into the lives and thoughts of individuals from various periods.

- **Educational Documents:** Includes documents related to education and learning, such as textbooks, instructional guides, and scholarly commentaries.

- **Scientific Illustrations and Diagrams:** Documents featuring detailed drawings and diagrams accompanying scientific or technical explanations.

- **Ephemeral Documents:** Temporary or short-lived documents, such as notes, drafts, or sketches, that were not necessarily intended for long-term preservation.

These categories are not mutually exclusive, and many documents may fall into multiple classifications. The diversity of document types reflects the rich tapestry of human knowledge, creativity, and communication across different historical and cultural contexts.

## 1.5 Role of Historical documents

Recognizing the importance of these documents lays the groundwork for justifying our computational exploration:

- **Primary Sources of Information:** Historical documents serve as primary sources of information, offering firsthand accounts of events, people, and cultures. They provide a direct connection to the past and offer a unique perspective that may not be available through secondary sources.

- **Cultural and Social Insights:** Documents offer insights into the cultural and social aspects of a particular time period. They often include details about daily life, customs, traditions, and societal norms, contributing to a richer understanding of different historical periods.

- **Historical Record Keeping:** Documents are vital records for historical record-keeping. They include official documents, legal agreements, and administrative records that document the governance, legal systems, and societal structures of bygone eras.

- **Literary and Artistic Heritage:** Many historical documents are literary or artistic works that contribute to the cultural heritage of a society. These may include documents of literary classics, religious texts, illuminated documents, and other forms of creative expression.
- **Preservation of Language and Writing Styles:** Historical documents preserve languages, writing styles, and linguistic nuances that may have evolved or become obsolete over time. Studying these documents helps linguists trace language evolution and understand the development of writing systems.
- **Source for Historical Research:** Scholars and researchers rely on historical documents for academic study and historical research. They use these documents to reconstruct

historical events, analyze societal changes, and gain insights into the motivations and perspectives of individuals from the past.

- **Documenting Change Over Time:** Documents provide a timeline of historical changes, capturing the evolution of societies, technologies, and ideologies. This documentation is valuable for understanding how and why certain changes occurred.

- **Material Culture and Palaeography:** Historical documents contribute to the study of material culture, providing information about the materials used in writing, bookbinding techniques, and the physical aspects of documents. Palaeography, the study of ancient writing systems and scripts, is another area enriched by historical documents.

- **Legal and Genealogical Research:** Legal documents found in historical documents, such as wills, contracts, and property records, are valuable for legal and genealogical research. They provide evidence of familial relationships, property ownership, and legal transactions.

## 1.6 Characteristics of Historical documents

Each document possesses a distinct set of characteristics that not only defines its physical nature but also offers profound insights into its historical context, cultural origins, and the intentions of its creators:

### 1.6.1 Moral Characteristics

- **Content ("What"):** Historical documents often transcend mere information, embodying moral purpose. Whether conveying ethical teachings, religious doctrines, or philosophical principles, the documents provide a moral compass that guided the societies of their time.

- **Authorship ("Who"):** The identity of the writer or scribe is a cornerstone of a document's character. It not only offers insights into the individual behind the pen but also introduces potential biases, perspectives, and motivations that shaped the document's content.

- **Temporal Context ("When"):** The temporal context of a document, encompassing both the date of creation and any subsequent modifications, provides a chronological framework. This temporal dimension enables scholars to trace the evolution of ideas and perspectives over time.

- **Geographical Origin ("Where"):** Rooted in specific locations, documents bear the imprints of regional influences, cultural contexts, and historical events. Understanding the geographical origin is essential for unraveling the diverse tapestry of human civilization.
- **Script and Language:** The script and language employed in a document are vital characteristics, revealing not only the linguistic choices of its creators but also offering

clues about regional or cultural affiliations. Paleographers scrutinize scripts to decipher historical nuances.

## 1.6.2 Physical Characteristics

- **Material Composition:** The physicality of a document, manifested through the materials used, including paper, ink, and additional elements, speaks volumes about the technological capabilities and resource availability of the historical period.

- **Dimensions and Format:** The physical size and format of a document influence its practicality, purpose, and visual appeal. Folios, quartos, or octavos not only convey information about the intended use but also reflect cultural and historical conventions.

- **Condition and Conservation Needs:** The physical condition of a document, its wear and tear, holds valuable information. Conservation needs unveil the ongoing dialogue between the document and the passing of time, prompting scholars to safeguard these fragile narratives.

- **Binding and Cover Materials:** Documents wear the craftsmanship of their bindings. Whether adorned in leather, wood, or textiles, the binding materials provide a tactile connection to the document's past, revealing the aesthetic preferences of its creators.

- **Watermarks and Paper Origin:** Watermarks imprinted on paper, if present, serve as historical fingerprints. These subtle marks unveil the origin of the paper, aiding in dating and unraveling the provenance of the document.

- **Colors, Illumination, and Artistic Features:** Documents often transcend mere words, incorporating vibrant colors, illuminations, and artistic elements. These aesthetic features not only enhance the visual appeal but also offer cultural and artistic insights.

- **Provenance Marks and Ownership History:** Tracing the provenance marks and ownership history of a document unveils its journey through time. Marks, inscriptions, or bookplates serve as signposts in the document's odyssey, connecting it to various collectors and custodians.

When historians and archaeologists possess the original document, examining the physical characteristics becomes a relatively straightforward task, especially when scrutinizing the physical copy directly. This direct examination allows for the determination of the type of paper, ink, and all materials employed in creating the document. However, a significant challenge emerges when historians are faced with unknown or undocumented documents. Deciphering the moral characteristics, such as determining the author, date, script, and location of a document, demands a high level of expertise and specialization. In this domain, the experts known as paleographers play a pivotal role.

## 1.7 Paleography

Paleography, as a discipline within historical document studies, holds a paramount position in unraveling the intricacies of handwritten texts **[29]**. In the context of our study, paleography emerges as a key science bridging the historical and computational domains **(Figure 1.4)**. This section explores the significance of paleography and its pivotal role in understanding and classifying historical documents.

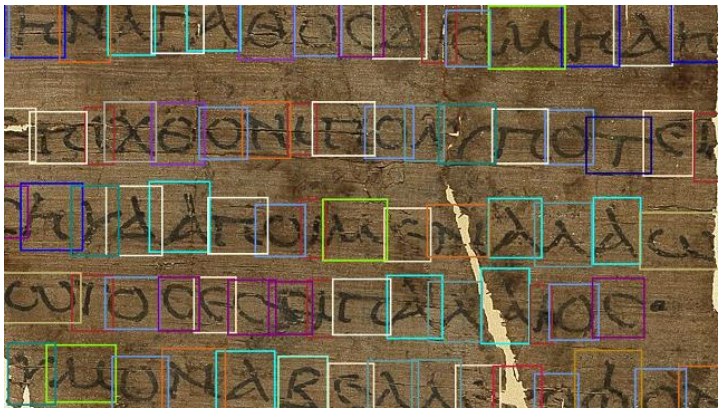


**Figure 1.4** Paleographical features. **[32]**

### 1.7.1 Paleographic Features

Paleographers utilize a variety of features within ancient handwriting to analyze and interpret documents. These features provide clues about the script's style, era, geographic origin, and the individual characteristics of the scribe. Some of the key paleographic features include:

- **Letterforms**: Paleographers examine the shapes and forms of individual letters. This includes variations in the size, curvature, and angles of letters, as well as the presence of ascenders (strokes that rise above the main body of a letter) and descenders (strokes that extend below the baseline).

- **Script Styles**: Different historical periods and regions had distinct script styles. Paleographers identify and categorize these styles, such as uncial, half-uncial, Carolingian minuscule, Gothic script, etc., based on the overall appearance and characteristics of the writing.

- **Abbreviations and Ligatures**: Many documents employed abbreviations and ligatures to save space and time. Paleographers analyze these shorthand forms and ligatures to understand the conventions used by scribes and to decipher the abbreviated text.

- **Diacritics and Punctuation**: Diacritical marks (such as accents, dots, and strokes) and punctuation symbols offer insights into pronunciation, grammar, and syntax. Paleographers study the placement and usage of these marks to reconstruct the intended meaning of the text.

- **Ink and Writing Instruments**: The type of ink used and the writing instrument (e.g., quill, reed pen, brush) leave distinctive marks on the writing surface. Paleographers analyze the color, consistency, and texture of the ink, as well as the width and pressure of the strokes, to infer details about the writing process and materials.

- **Layout and Formatting**: The layout and formatting of a document, including margins, spacing between lines and words, and arrangement of text on the page, provide valuable clues about scribal practices, document production techniques, and the intended audience of the text.

- **Palaeographic Dating Criteria**: Paleographers consider specific features that are characteristic of certain time periods or regions to help determine the approximate date of a document. These criteria may include the presence of certain letterforms, script styles, or decorative elements that are typical of a particular era.

- **Individual Handwriting Characteristics**: Each scribe had their own unique handwriting style, influenced by factors such as training, personal preference, and regional conventions. Paleographers examine the individual quirks and idiosyncrasies of a scribe's handwriting to attribute authorship and identify scribal hands within a document.

### 1.7.2 Role of Paleography

Paleography serves as a crucial means to delve into the historical narratives contained within documents including different languages such as, Latin, Hebrew, Arabic...Etc. **[30]**. Paleographers primarily aim to ascertain the date, script type, authorship, and origin of a document. Given the variation in historical writing styles across script type, region, date, and author, deciphering the writing characteristics from documents yields invaluable insights, including:

- **Dating the Document**: Paleographers can determine the approximate date of a document by comparing its script style with known examples from different time periods. Changes in handwriting styles over time can provide clues to when a document was written, aiding historians in establishing timelines and understanding historical contexts.

- **Identifying Script Types**: Different cultures and time periods had distinct script types. Paleographers are skilled at recognizing these variations and can identify whether a document is written in Roman cursive, Carolingian minuscule, Gothic script, or other script types. This identification helps place the document within its cultural and historical context.

- **Attributing Authorship**: Handwriting can be as distinctive as a fingerprint. Paleographers analyze the individual quirks and characteristics of handwriting to attribute authorship to anonymous or disputed texts. By comparing the handwriting in a

document with known examples from the purported author, paleographers can provide insights into the authorship of a document.

- **Determining Geographic Origin**: Regional variations in handwriting styles provide clues about the geographic origin of a document. Paleographers compare the script of a document with examples from different regions to pinpoint where it may have been produced. This geographical information contributes to our understanding of trade routes, cultural exchanges, and the dissemination of knowledge across different regions.

- **Revealing Cultural and Social Contexts**: Handwriting reflects not only individual style but also broader cultural and social conventions. Paleographers consider factors such as writing implements, writing surfaces, abbreviations, and punctuation conventions to interpret the cultural and social context in which a document was produced. These insights shed light on literacy rates, educational practices, and social hierarchies of the time.

- **Detecting Forgeries and Alterations**: Paleographers play a crucial role in authenticating documents and detecting forgeries or alterations. By closely examining the consistency and coherence of handwriting throughout a document, paleographers can identify inconsistencies or anomalies that may indicate tampering or forgery attempts.

- **Understanding Scribal Practices**: Paleographers study the techniques and practices of scribes to understand how documents were produced. This includes examining aspects such as layout, spacing, corrections, and marginalia. Insights into scribal practices provide valuable information about the process of document production, dissemination, and reception in historical contexts.

- **Preserving Cultural Heritage**: Paleographers contribute to the preservation of cultural heritage by documenting and studying ancient handwriting traditions. Their work helps ensure that valuable historical documents are accurately transcribed, interpreted, and preserved for future generations, thereby safeguarding our collective cultural heritage.

### 1.7.3 Digital Paleography

Digital paleography is an interdisciplinary field that merges traditional paleographic methods with digital technologies to study historical handwriting and documents **[31]**. It encompasses the use of computational tools, image processing techniques, and machine learning algorithms to analyze and interpret handwritten texts in digital format. Digital paleography leverages advancements in digitization, data analysis, and artificial intelligence to enhance the study of paleography and make historical documents more accessible for research and scholarship.

- **Digitization of Documents:** Digital paleography begins with the digitization of historical documents, transforming physical documents into digital images or text files. Documents may be digitized using high-resolution scanners, cameras, or specialized imaging equipment to capture detailed images of handwritten pages. Digital images are

stored in digital repositories or databases, allowing scholars to access and study documents remotely without the need for physical proximity to the original documents.

- **Image Processing and Enhancement:** Digital paleographers utilize image processing techniques to enhance the quality and legibility of digitized document images. This may involve removing noise, adjusting contrast and brightness levels, correcting distortions, and sharpening edges to improve the clarity of handwritten text. Image processing algorithms help enhance the visibility of faint or damaged ink, making it easier for paleographers to analyze and transcribe handwritten characters.

- **Automated Script Recognition:** Digital paleography employs machine learning algorithms and pattern recognition techniques to automate the recognition and classification of script styles within digitized documents. These algorithms analyze the visual characteristics of handwritten text, such as letter shapes, stroke patterns, and spacing, to identify distinct script types and scribal hands. Automated script recognition accelerates the process of script analysis and facilitates large-scale studies of document collections.

- **Script Transcription and Annotation:** Digital paleography facilitates the transcription and annotation of handwritten texts using digital tools and software applications. Paleographers use digital transcriptions to convert handwritten characters into machine-readable text, enabling textual analysis, searchability, and indexing of document content. Digital annotations allow scholars to add metadata, notes, and commentary to digitized documents, enriching their scholarly value and facilitating collaborative research efforts.

- **Data Analysis and Visualization:** Digital paleography employs data analysis techniques and visualization tools to extract insights from large-scale document collections. Scholars use statistical analysis, text mining, and computational linguistics to identify textual patterns, linguistic features, and scribal characteristics within digitized documents. Data visualization techniques, such as heatmaps, dendrograms, and network graphs, help visualize relationships and trends within document data, facilitating interpretation and discovery.

- **Crowdsourcing and Collaborative Projects:** Digital paleography fosters collaboration and crowdsourcing initiatives to engage scholars, students, and the public in the study of historical documents. Crowdsourcing platforms allow volunteers to transcribe, annotate, and analyze digitized documents, contributing to large-scale transcription projects and scholarly databases. Collaborative projects bring together researchers from diverse disciplines to share expertise, resources, and data, advancing our understanding of paleography and document culture.

## 1.8 Conclusion

In conclusion, this chapter has provided us with a robust comprehension of the subject matter under study. As computer scientists, this foundational understanding serves as a vital bridge to the historical context essential for navigating the technical complexities ahead. The

insights gleaned from this chapter will play a pivotal role, steering our transition from historical analysis to the conceptualization and implementation of a system.

In the next chapiters, we will delve deeper into the technical specifications and implementation details of our system, elucidating how it addresses the unique challenges faced by historians and paleographers. Moreover, we will explore the motivational aspects driving our research, highlighting the broader significance of our endeavors and the potential impact on scholarly practices within the field.

# Chapter 2

# Historical documents classification

## 2.1 Introduction

Building upon the foundational knowledge acquired in Chapter 1 regarding historical documents, this chapter delves deeper into our research problem: the classification of historical documents. We will embark on a journey to understand the significance of classifying historical documents from both historical and computer science perspectives. Elucidating the tasks associated with this classification process, we aim to provide a comprehensive framework for approaching this complex endeavor. Furthermore, we will explain in general manner, existing approaches employed in tackling these classification tasks, beginning with physical methods and paleographical methods before transitioning to the realm of computer-based methods. Within this domain, we will differentiate between semantic and syntactic methods, laying the groundwork for our exploration of classification techniques in subsequent sections.

## 2.2 Understanding *"Historical documents Classification"*

The concept of documents classification serves as the cornerstone of our study, bridging the gap between historical scholarship and computational methods. It is essential to clarify the relationship between "documents" and "classification" to delineate the goals and evaluation criteria of our research. The interpretation of these terms can vary between historians and computer scientists, necessitating a unified understanding for effective collaboration.

For computer scientists, the term ***"classification"*** holds technical significance, referring to a type of problem encountered in training machine learning models. It involves categorizing data into predefined classes or categories based on features or attributes. Conversely, for historians, classification embodies its literal meaning: organizing documents according to specific criteria such as authorship, date, script type, and geographical origin.

The genesis of our research problem stems from the challenges historians face in manually classifying documents. Traditional methods rely on the expertise of domain specialists to discern and categorize documents based on various characteristics. However, this manual approach is labor-intensive, time-consuming, and prone to subjective biases. As a result, historians encounter difficulties in accurately and efficiently classifying documents, hindering their research endeavors.

Recognizing the limitations of manual classification methods, our study seeks to automate this task leveraging the capabilities of computers and machine learning techniques. Harnessing the computational power of algorithms, we aim to streamline the classification process, making it more efficient, accurate, and scalable. In essence, our research problem revolves around automating the manual classification of historical documents, thereby empowering historians with computational tools to enhance their research methodologies and advance our understanding of the past.

## 2.3 Documents classification tasks

Classifying historical documents involves delineating distinct tasks based on various criteria, each representing a unique challenge in the classification process. These tasks serve as focal points for developing methodologies and approaches tailored to address specific classification objectives. Some of the most significant tasks in the classification of historical documents include: Authorship Attribution, Dating Documents, Script Identification, Geographical Origin Determination, Content Classification…etc**.** Each of these classification tasks indeed presents its own set of challenges, necessitating specialized methodologies, datasets, and expertise for effective resolution. Comprehensively understanding and delineating these tasks, researchers can develop targeted approaches to document classification, thereby contributing to a deeper understanding of historical texts and their cultural significance. In our research, we have chosen to focus on three main tasks: Dating, Writer Identification, and Script Type Classification. Figure 2.1 demonstrate a number of tasks related to historical documents **[33]**.

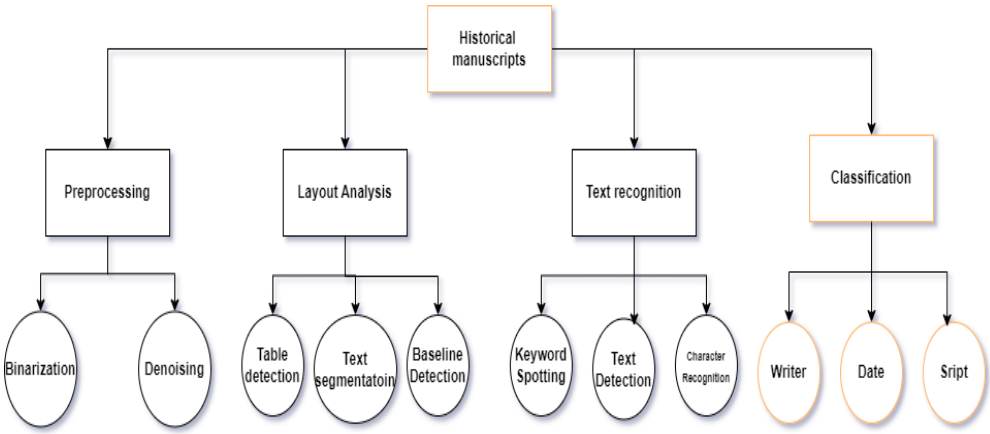



**Figure 2.1** Historical documents tasks

### 2.3.1 Dating

Dating historical documents stands as a paramount task in the realm of historical research, offering insights into the chronological contexts and cultural landscapes of ancient texts. It encompasses a multifaceted approach, drawing upon a diverse array of methodologies to discern the age and temporal origins of ancient texts. Dating documents poses several challenges, stemming from the complexities inherent in analyzing ancient texts and materials:

- **Script Variation:** Variations in script styles within a single document or across different regions and time periods complicate dating efforts. Deciphering evolving script forms and discerning regional script traditions require keen paleographic expertise.

- **Material Deterioration:** Deterioration of document materials over time, including ink fading, parchment degradation, and physical damage, can obscure paleographic features crucial for dating. Non-invasive imaging and conservation techniques are employed to mitigate material deterioration and enhance legibility.

- **Interdisciplinary Collaboration:** Effective dating of documents often necessitates interdisciplinary collaboration between paleographers, historians, archaeologists, and scientists. Integrating diverse expertise and methodologies is essential for comprehensive dating analyses.

This task holds profound significance for historical research and scholarship, yielding valuable insights into various aspects of the past:

- **Chronological Context:** Accurate dating of documents establishes chronological contexts for historical events, cultural movements, and intellectual developments, enriching our understanding of the past.

- **Textual Transmission:** Dating documents elucidates the transmission and dissemination of literary, religious, and scholarly texts across time and space, tracing the trajectories of ideas and cultural exchange.

- **Cultural Heritage Preservation:** Dating documents contributes to the preservation and conservation of cultural heritage, enabling informed decisions regarding document digitization, conservation, and exhibition strategies.

Dating documents serves as a fundamental endeavor in historical research, unlocking the temporal dimensions of ancient texts and enriching our comprehension of human history. Employing rigorous methodologies and interdisciplinary collaboration, researchers continue to illuminate the timelines of the past, ensuring that the voices of bygone eras resonate with clarity and resonance in the present day.

### 2.3.2 Writer identification

Writer identification stands as a pivotal task in the domain of document studies, offering insights into the individuals responsible for shaping historical texts. Writer identification encompasses a diverse array of methodologies, drawing upon paleographic analysis, linguistic features, and historical context to discern the unique traits of individual scribes. This task presents several challenges, stemming from the complexities inherent in analyzing handwriting and discerning individual stylistic traits:

- **Variability in Handwriting:** Handwriting styles can exhibit variability within a single document or across different works attributed to the same author. Distinguishing between genuine variations and stylistic inconsistencies poses a challenge for paleographers.

- **Scribe Imitation:** Scribes often emulated the handwriting styles of renowned authors or copied existing documents, complicating efforts to attribute works to specific writers. Discerning between authentic authorial hands and imitative scripts requires careful analysis and expertise.

- **Limited Biographical Information:** In many cases, limited biographical information is available about historical scribes, making it challenging to establish definitive attributions based solely on handwriting analysis. Integration of additional evidence from historical records and textual studies is essential for robust writer identification.

Writer identification serves as a cornerstone of document scholarship, enabling researchers to uncover the hands behind the quill and illuminate the voices of bygone authors. Through rigorous analysis and interdisciplinary collaboration, scholars continue to unravel the mysteries of document production, enriching our understanding of literary history and cultural heritage.

### 2.3.3 Script Type Classification

Script type classification is a fundamental task in the study of historical documents, offering insights into the evolution and diversity of writing styles throughout history. Script type classification encompasses a range of methodologies, drawing upon paleographic analysis, typological features, and historical context to discern the distinctive characteristics of script styles. Script type classification presents several challenges, stemming from the complexities inherent in analyzing handwriting and discerning regional and temporal variations:

- **Regional Variation:** Script styles exhibit variability across different regions and cultural traditions, making it challenging to establish standardized classification criteria. Distinguishing between regional script traditions and localized variations requires expertise in paleography and historical linguistics.

- **Temporal Evolution:** Script styles evolve over time in response to cultural, technological, and aesthetic influences, resulting in chronological variations within script families. Deciphering the chronological development of script types and identifying transitional forms pose challenges for paleographers.

- **Interdisciplinary Collaboration:** Effective script type classification often necessitates interdisciplinary collaboration between paleographers, historians, linguists, and art historians. Integrating diverse expertise and methodologies is essential for comprehensive script analysis and classification.

Script type classification holds profound significance for historical research, linguistic studies, and cultural heritage preservation, offering insights into various aspects of document production and textual transmission:

- **Cultural Context**: Script type classification provides insights into the cultural, religious, and educational contexts in which documents were produced. Analysis of script styles illuminates the dissemination of writing systems, literacy rates, and cultural exchange networks.

- **Historical Linguistics:** Comparative analysis of script styles contributes to the study of historical linguistics, tracing the evolution of writing systems, language contact phenomena, and dialectal variations over time.

- **Cultural Heritage Preservation:** Script type classification contributes to the preservation and authentication of cultural heritage, enabling the identification and cataloging of document collections according to script traditions and styles.

## 2.4 Historical documents classification methods

Having established the importance of documents classification and identified the key classification tasks relevant to historical documents, we now turn our attention to the existing approaches employed by researchers to address these tasks. These methods can be categorized based on the type of approach utilized, namely physical methods, paleographical methods, and computer-based methods (Figure 2.2). Exploring these methods in detail, researchers can gain insights into the diverse approaches available for classifying historical documents and identify the most appropriate techniques for addressing specific classification tasks. Integrating multiple methods and interdisciplinary approaches can enhance the accuracy and reliability of document classification, advancing our understanding of historical texts and their cultural significance.

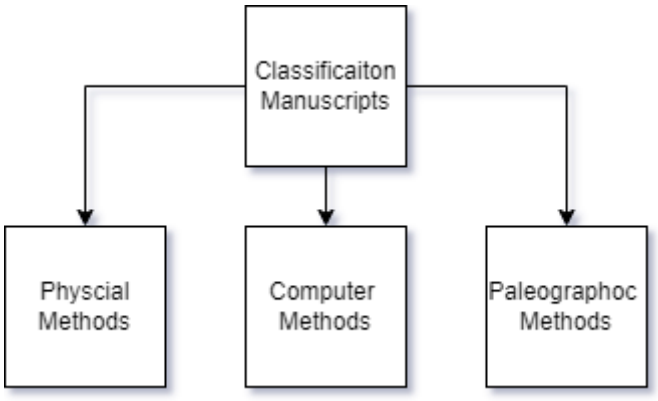



**Figure 2.2** Types of historical documents classification methods

## 2.4.1 Physical Methods

Physical methods encompass a wide array of techniques that involve the direct examination and analysis of document materials (Figure 2.3), offering invaluable insights into their composition, production techniques, and historical context. Through meticulous examination and scientific analysis, researchers utilize various tools and methodologies to unravel the mysteries concealed within the physical attributes of historical documents.

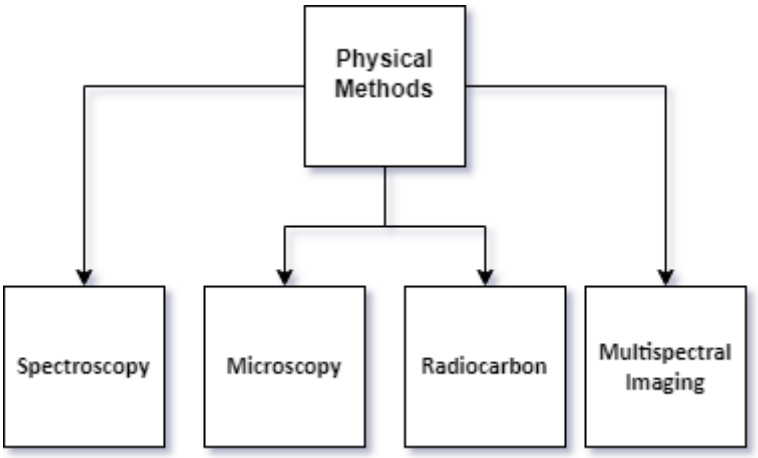



**Figure 2.3** Types of historical documents classification methods

### 2.4.1.1 Spectroscopy

Spectroscopy is a scientific technique used to study the interaction of matter with electromagnetic radiation. It involves analyzing the light emitted, absorbed, or scattered by a material across different wavelengths or frequencies. This interaction provides valuable information about the composition, structure, and properties of the material being studied **[34]**.

In spectroscopy, a sample is exposed to electromagnetic radiation, typically in the form of light, which can range from ultraviolet (UV) to visible to infrared (IR) wavelengths, and sometimes even beyond these ranges. The sample may absorb certain wavelengths of light, emit light at specific frequencies, or scatter light in various directions, depending on its chemical composition and molecular structure. The resulting spectrum, which is a plot of intensity versus wavelength or frequency, contains characteristic features that can be analyzed to identify the substances present in the sample and to elucidate their chemical bonds, molecular arrangements, and other properties (Figure 2.4).

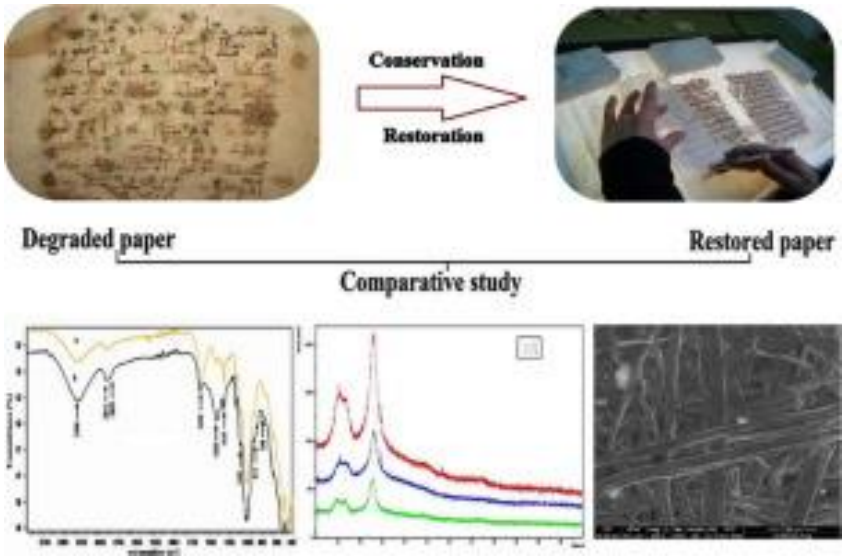



**Figure 2.4** Types of historical documents classification methods **[35]**

Spectroscopy is utilized in the study of historical documents primarily for material analysis and characterization. Examining the interaction of electromagnetic radiation with the materials present in the document **[36,37,38]**, spectroscopic techniques provide valuable information about their composition, structure, and state of preservation:

- **Identification of Ink and Pigments:** Raman spectroscopy and infrared spectroscopy are commonly employed to analyze the chemical composition of inks and pigments used in historical documents. These techniques can identify the specific compounds present in the ink or pigment, helping to determine their origin, age, authenticity and writer. For example, by analyzing the spectral signatures of pigment samples extracted from a document, researchers can identify the type of ink used and potentially uncover clues about the document's production process.

- **Characterization of Parchment and Paper:** Infrared spectroscopy is often used to analyze the composition of parchment and paper materials. Studying the infrared absorption spectra of these materials, researchers can identify the presence of organic compounds such as cellulose, protein, and lignin, which are characteristic of parchment and paper fibers. This information can provide insights into the manufacturing techniques used to produce the document and its overall condition.

- **Detection of Degradation and Alterations**: Spectroscopic techniques can also be employed to assess the degradation and alterations present in historical documents. For example, infrared spectroscopy can identify chemical changes in the document caused by exposure to light, humidity, or pollutants. Additionally, fluorescence spectroscopy can detect the presence of modern materials or alterations made to the document over time.

- **Dating and Authentication:** While not as common as other dating methods like radiocarbon dating, spectroscopic techniques can provide indirect evidence for dating historical documents. Analyzing the materials and pigments used in the document,

researchers can infer its age or authenticity based on the presence of specific compounds or manufacturing techniques characteristic of certain time periods.

### 2.4.1.2 Microscopy

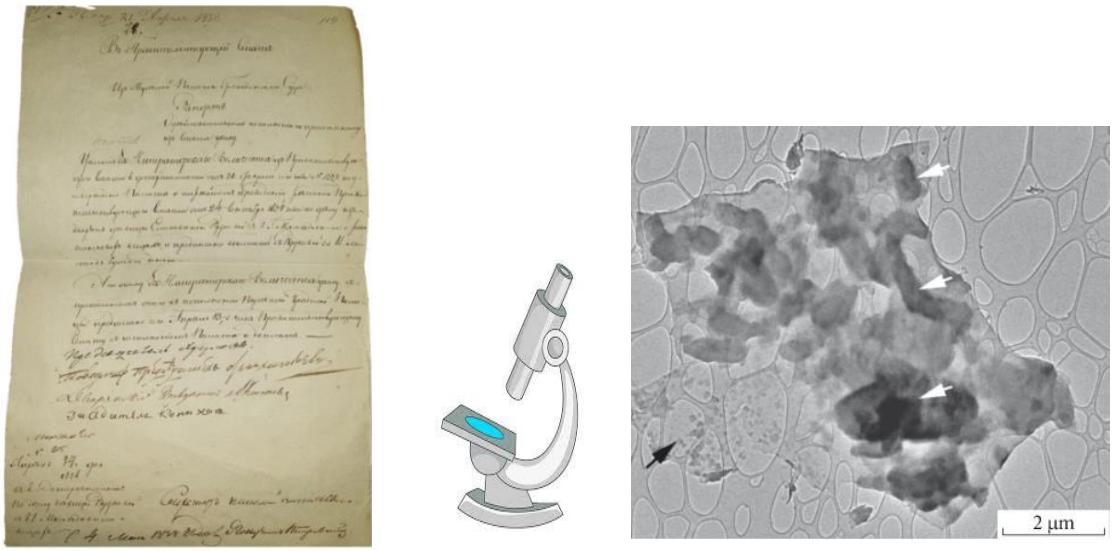



**Figure 2.5** Bright-field TEM image of a fragment **[54]**

Microscopy is a scientific technique used to observe and study objects or materials at a microscopic level (Figure 2.5), providing detailed images and insights into their structure, morphology, and composition **[39]**. In the context of historical documents, microscopy is an invaluable tool for examining the physical characteristics of the materials used in their creation, as well as for investigating any features or alterations present on the document surface **[40,41]**. This is how microscopy is utilized in the study of historical documents:

- **Observation of Material Composition:** Microscopic analysis allows researchers to examine the composition of document materials such as parchment, paper, ink, and pigments. Optical microscopy, which uses visible light, can provide high-resolution images of the surface and cross-sections of document samples. This enables researchers to identify the fibers, particles, and other constituents present in the materials, aiding in their characterization and classification.

- **Detection of Writing and Illustrations:** Microscopy is used to observe and analyze the writing and illustrations present on the surface of historical documents. Magnifying the document surface, researchers can examine the ink or pigment particles, as well as the strokes and patterns of handwriting or artistic techniques. This allows for the identification of different writing styles, script types, and artistic elements, contributing to the study of document production and cultural context.

- **Examination of Degradation and Damage:** Microscopic analysis is crucial for assessing the condition of historical documents and detecting any signs of degradation or damage. Inspecting the surface and structure of the document under high magnification, researchers can identify issues such as ink fading, pigment

degradation, mold growth, insect damage, or physical wear. This information helps conservators and curators develop appropriate preservation and restoration strategies to safeguard the documents for future generations.

- **Investigation of Alterations and Forgeries:** Microscopy can be used to detect alterations, additions, or forgeries present on historical documents. Closely examining the surface of the document, researchers can identify discrepancies in handwriting, ink composition, or artistic style that may indicate tampering or manipulation. This forensic analysis helps authenticate documents and distinguish original content from later modifications.

- **Study of Document Structure and Manufacture:** Microscopy allows researchers to study the structural features and manufacturing techniques used in historical documents. Examining cross-sections or thin slices of document samples, researchers can observe the layering, fiber orientation, binding materials, and other structural elements. This provides insights into the production methods, materials sourcing, and technological advancements of document production throughout history.

#### 2.4.1.3 Radiocarbon

Radiocarbon, also known as carbon or radiocarbon analysis, is a scientific method used to determine the age of organic materials based on the decay of radioactive carbon isotopes. Radiocarbon dating relies on the principle that living organisms absorb carbon-14 (14C) from the atmosphere through the process of photosynthesis. Once an organism dies, it stops absorbing carbon, and the carbon-14 within it begins to decay at a known rate. Carbon-14 is a radioactive isotope of carbon with a half-life of approximately 5,730 years. This means that over time, half of the carbon-14 atoms in a sample will decay into nitrogen-14 (14N) **[42]**.

Radiocarbon is utilized in the study of historical documents to establish the chronological framework of the materials used in their creation **[43,44,45]**, aiding researchers in understanding their historical context and provenance:

- **Dating Parchment or Paper:** Radiocarbon dating can be used to determine the age of parchment or paper documents by analyzing the carbon content within the material. Researchers extract small samples from the document for analysis, typically from areas with minimal impact on the document's integrity. These samples undergo chemical processing to isolate the carbon-containing compounds, which are then analyzed to measure the ratio of carbon-14 to carbon-12 isotopes.

- **Estimating Document Age:** Comparing the measured ratio of carbon isotopes in the document samples to the known atmospheric levels of carbon-14 over time, scientists can calculate an estimated age range for the document. This age estimate provides valuable chronological information about when the parchment or paper was created or when the document was written.

- **Contextualizing Historical Events:** Radiocarbon dating of historical documents can help contextualize significant historical events or periods by establishing the

timeframe in which the documents were produced. For example, radiocarbon dating of documents from a particular archaeological site can provide insights into the timeline of human occupation or cultural activities in that area.

- **Verifying Document Authenticity:** Radiocarbon dating can be used to verify the authenticity of historical documents and detect potential forgeries. Comparing the radiocarbon dates obtained from document samples to the expected chronological context based on historical records or other archaeological evidence, researchers can assess the consistency of the document's age with its purported origin or historical context.

- **Identifying Document Reuse or Parchment Origins:** Radiocarbon dating can also help identify instances of document reuse, where older parchment or paper materials were recycled for later documents. Analyzing multiple samples from different sections of a document or comparing radiocarbon dates from related documents, researchers can detect patterns of parchment reuse or trace the origins of parchment materials to specific geographical regions or time periods.

While radiocarbon dating provides valuable age estimates, it is subject to certain limitations and sources of error. Factors such as fluctuations in atmospheric carbon-14 levels over time, contamination from modern carbon sources, and sample size can affect the accuracy of the results. To address these issues, radiocarbon dates are often calibrated using known age samples or by applying statistical corrections to account for these factors.

#### 2.4.1.4 Multispectral Imaging

Multispectral imaging is a non-invasive imaging technique that captures images of an object or scene at various wavelengths across the electromagnetic spectrum. It provides valuable information beyond what is visible to the human eye by detecting wavelengths of light that are invisible or difficult to perceive **[46]**. Multispectral imaging works on the principle that different materials interact with light in unique ways, absorbing, reflecting, or emitting specific wavelengths. Capturing images at multiple wavelengths, multispectral imaging reveals details and features that may not be visible in standard photographs or under normal lighting conditions.

Multispectral imaging systems are equipped with sensors capable of capturing images at specific wavelengths or wavelength ranges across the electromagnetic spectrum (Figure 2.6). Researchers can select wavelengths based on the properties of the materials they wish to study and the information they seek to extract from the images. It involves capturing a series of images of an object or scene at different wavelengths. Each image represents the object's appearance when illuminated with light of a specific wavelength. These images are then processed and combined to create a multispectral dataset containing information about the object's spectral response at each wavelength.

Once the multispectral images are acquired, researchers analyze the data to extract valuable information about the object's composition, condition, and characteristics. This may involve techniques such as image enhancement, spectral analysis, and pattern recognition to identify features of interest or detect subtle variations in the material properties.

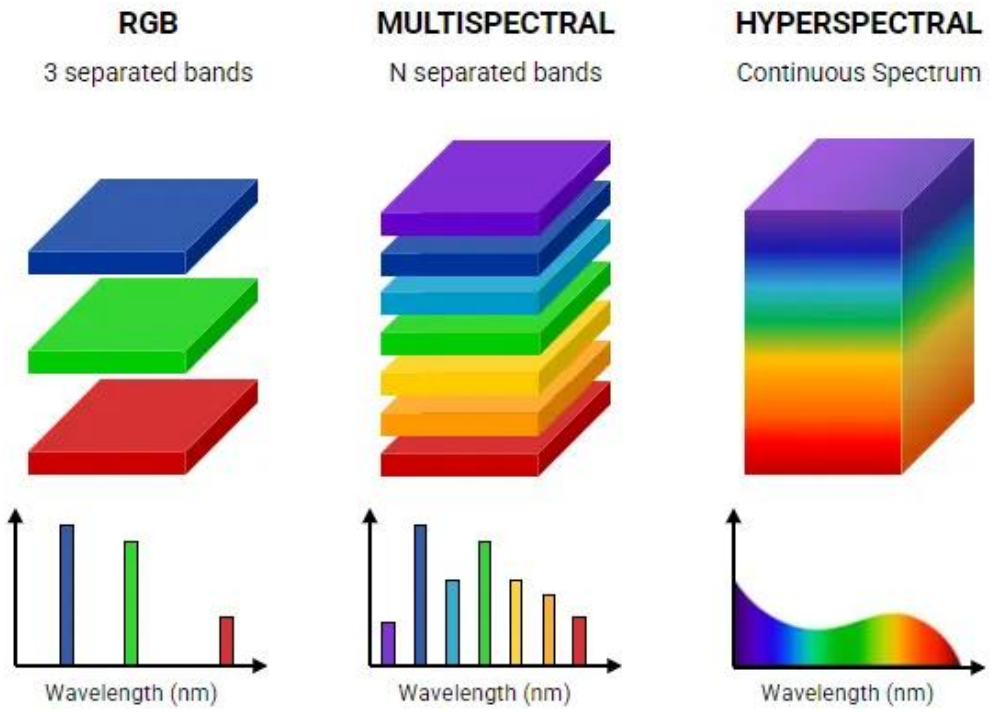



**Figure 2.6** RGB (left), multispectral (center)., hyperspectral imaging (right) **[55]**

Multispectral imaging is extensively used in the study and preservation of historical documents due to its ability to reveal hidden or obscured features that are not visible to the naked eye **[47,48,49]**.

- **Detection of Faded or Erased Text:** Over time, the ink used in historical documents can fade or be intentionally erased, making the text difficult or impossible to read using conventional methods. Multispectral imaging can capture images of documents under different wavelengths of light, including ultraviolet and infrared, which can penetrate the surface layers and reveal traces of faded or erased text (Figure 2.7).

- **Identification of Different Ink Types:** Historical documents often contain text or illustrations written with various types of ink, such as iron gall ink, carbon-based ink, or organic dyes. Multispectral imaging allows researchers to distinguish between different ink types based on their spectral characteristics.

- **Visualization of Underdrawings or Annotations:** Some historical documents may contain hidden drawings, sketches, or annotations beneath the surface layers of the parchment or paper. Multispectral imaging can penetrate these layers and capture images of the underlying features, revealing hidden details that are invisible to the naked eye.

- **Detection of Alterations or Forgeries:** Multispectral imaging is instrumental in detecting alterations, additions, or forgeries in historical documents. Capturing images of the document under different lighting conditions, multispectral imaging can highlight inconsistencies in handwriting, ink composition, or surface texture that may indicate tampering or manipulation. Researchers can use multispectral images to compare different sections of the document and identify discrepancies that may suggest the presence of alterations or forgeries.

- **Preservation and Conservation:** Multispectral imaging is used in the preservation and conservation of historical documents by providing valuable information about the condition and stability of the materials.

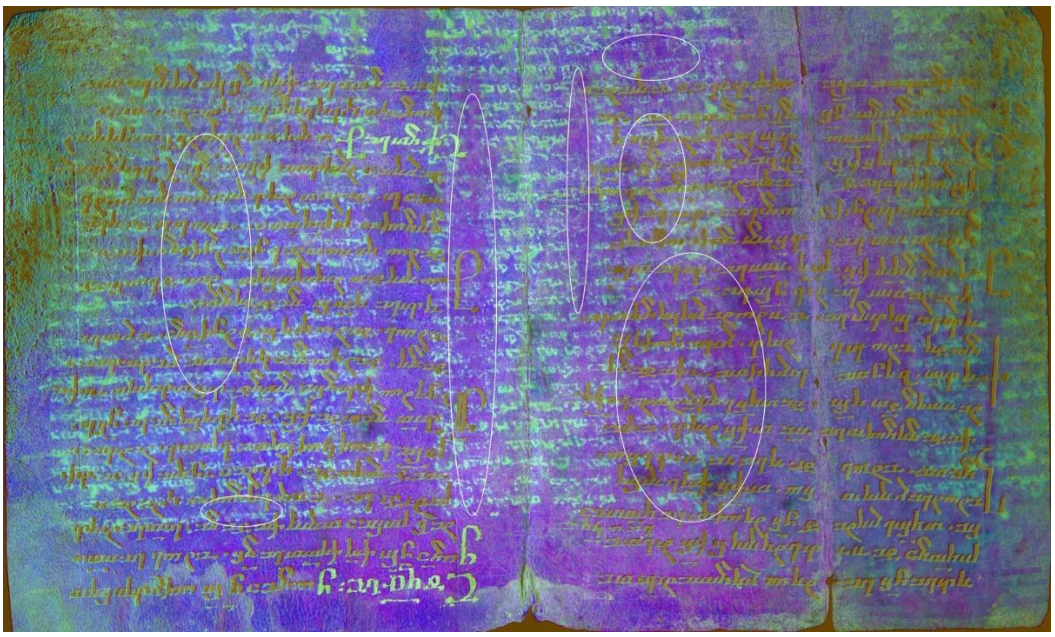
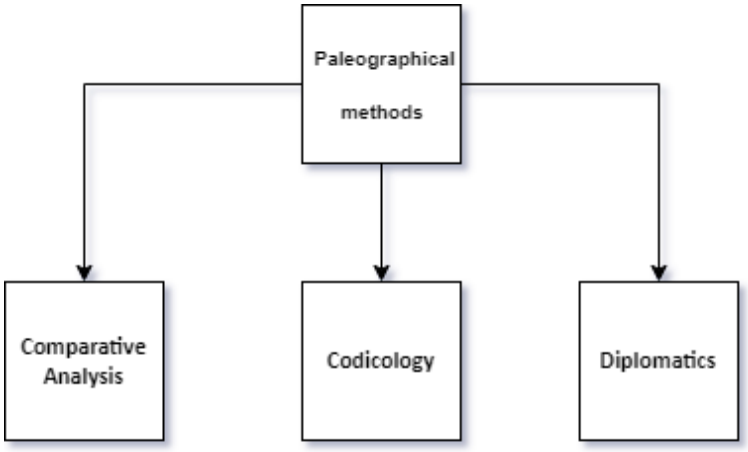


**Figure 2.7** Multispectral image of the HMML Palimpsest, with areas of Syriac underwriting circled (image: MB Toth) **[56]**

## 2.4.2 Paleographic methods

Paleographical methods represent a cornerstone in the intricate realm of document analysis, delving deep into the nuances of handwriting and script styles to unravel the historical narratives embedded within each document. At its core, paleography is the art of deciphering the subtle intricacies of script, transforming strokes and curves into windows that peer into the past **[50]**. Central to paleographical analysis is the meticulous examination of individual letters and script styles, akin to a linguistic archaeologist scrutinizing ancient artifacts. Paleographers meticulously study the shapes, proportions, and strokes of each letter, discerning the unique characteristics that distinguish one script from another. Through this painstaking process, they navigate the labyrinth of regional variations and chronological developments, tracing the evolution of scripts across time and space.




#### 2.4.2.1 Comparative Analysis

Comparative analysis in paleography involves the systematic examination and comparison of handwriting and script styles across multiple documents **[51]**. This approach allows paleographers to discern patterns, variations, and similarities in script characteristics, facilitating the classification and identification of documents based on shared features.

- **Collecting Document Samples:** Comparative analysis begins with the collection of document samples that represent a specific period, region, writers or script type of interest. These samples may be sourced from archival collections, libraries, museums, or digital repositories. Paleographers select documents with diverse provenance and content to ensure a comprehensive representation of the script styles under study.

- **Detailed Examination of Handwriting:** Paleographers meticulously examine the handwriting and script styles present in each document sample. They analyze the shapes, strokes, proportions, and spacing of individual letters, as well as the overall layout and presentation of the text, closely scrutinizing these features, paleographers identify distinctive traits and characteristics that define each script type.

- **Identification of Common Features:** Through comparative analysis, paleographers identify common features and shared patterns among document samples. This may include similarities in letter forms, ligatures, abbreviations, punctuation marks, or decorative elements. Paleographers also pay attention to regional variations and chronological developments within script styles, noting differences that may indicate changes over time or geographical influences.

- **Classification and Typology:** Based on the observed similarities and variations, paleographers classify documents into distinct script types or categories. They develop typologies or classification systems that group documents according to shared script features, regional characteristics, or historical periods. This classification process helps paleographers organize and categorize document collections, facilitating further study and analysis.

- **Comparative Studies:** Comparative analysis often extends beyond individual documents to include broader comparative studies of script development and evolution. Paleographers may compare document samples from different regions, periods, or cultural contexts to trace the dissemination of script styles and the exchange of writing traditions. These comparative studies provide insights into the interconnectedness of script cultures and the transmission of writing practices across geographical and historical boundaries.

Comparative analysis is a fundamental tool in paleography, enabling scholars to decipher the complexities of handwriting and script styles within historical documents, systematically comparing document samples and identifying shared features, paleographers contribute to the classification, identification, and understanding of document traditions and textual cultures across time and space.

#### 2.4.2.2 Codicology

Codicology is a multidisciplinary field of study that focuses on the physical aspects of documents, including their materials, structure, and production techniques **[52]**. Codicologists examine documents as physical artifacts, seeking to understand their material composition, construction, and historical context.

- **Materials and Construction:** Codicologists study the materials used in the production of documents, including parchment, paper, ink, pigments, and bindings. They analyze the quality, origin, and preparation of parchment or paper, as well as the types of ink and pigments employed in writing and illustration. Additionally, codicologists examine the construction of documents, including the arrangement of folios (individual leaves), the sewing and binding techniques used to assemble the document, and the presence of covers or protective elements.

- **Physical Features:** Codicologists investigate the physical characteristics of documents, such as their size, shape, pagination, and layout. They examine the dimensions of the document, the ruling patterns used to guide the placement of text, and the organization of content into sections, chapters, or folios. Codicologists also analyze decorative elements, such as illuminated initials, marginalia, and illustrations, which provide insights into the aesthetic and artistic preferences of the document's creators.

- **Conservation and Preservation:** Codicologists assess the condition of documents, identify signs of damage, degradation, or deterioration, and develop strategies for their care, repair, and restoration.

#### 2.4.2.3 Diplomatics

Diplomatics is a scholarly discipline that focuses on the study of historical documents, particularly their forms, structures, and functions within their historical contexts. Diplomatics seeks to understand the creation, transmission, and authentication of documents, as well as their roles in legal, administrative, and cultural contexts **[53]**.

- **Forms and Structures of Documents**: Diplomatics examines the external and internal characteristics of historical documents, including their physical appearance, layout, and textual components. Diplomatic analysis involves identifying the typical forms and structures of documents, such as charters, deeds, contracts, letters, and diplomas. Scholars study the conventions and protocols governing document creation, including the use of specific formulae, legal language, and authentication marks.

- **Authentication and Authority:** Diplomatics investigates the mechanisms of authentication and authority inherent in historical documents. Scholars examine the seals, signatures, notarial marks, and other authentication devices used to validate the authenticity and legality of documents. Diplomatic analysis also considers the hierarchical relationships between issuers and recipients, as well as the social and institutional contexts in which documents were produced and exchanged.

- **Documentary Practices and Procedures:** Diplomatics explores the bureaucratic and administrative procedures involved in document creation, circulation, and preservation. Scholars examine the roles and responsibilities of scribes, chancellors, notaries, and other officials involved in document production. Diplomatic analysis considers the conventions governing document format, language, and transmission, as well as the methods of storage, cataloging, and archival management employed by custodial institutions.

- **Cultural and Social Contexts:** Diplomatics situates historical documents within their broader cultural, social, and institutional contexts. Scholars examine the cultural norms, legal customs, and administrative practices that shaped document production and usage in different historical periods and regions. Diplomatic analysis considers the roles of religious institutions, governmental bodies, legal systems, and private individuals in document creation, dissemination, and preservation.

In today's digital age, the intersection of paleography and technology represents a transformative opportunity for historical research. By harnessing the capabilities of computational tools and deep learning algorithms, researchers can delve deeper into the intricate nuances of ancient handwriting, unlocking hidden insights and shedding new light on the past. Digital paleography serves as a linguistic bridge between bygone eras and the computational present, facilitating the analysis, interpretation, and dissemination of paleographic data with unprecedented precision and efficiency. Through the digitization of documents and the application of advanced image processing techniques, researchers can access high-resolution images of handwritten texts, enabling detailed examination of script styles, letterforms, and other paleographic features. However, the true promise of digital paleography lies in its integration with machine learning and deep learning algorithms. By training machine learning models on vast repositories of annotated documents, researchers can develop algorithms capable of automatically identifying and classifying script styles, deciphering abbreviations, and even attributing authorship with remarkable accuracy. This marriage of paleography and deep learning opens up new avenues for research which is computer-based methods.

### 2.4.3 Computer-based methods

Computer-based methods have emerged as a predominant trend in the study of historical documents, driven by the proliferation of digitized document databases. These databases provide researchers across various fields with access to vast collections of digitized historical documents, enabling collaborative research and interdisciplinary studies. In the realm of computer science, researchers have devised different methodologies to analyze and interpret these documents, with two main approaches gaining prominence: text-based methods and image-based methods.

Text-based methods involve analyzing the textual content of documents to classify documents based on their linguistic features. However, this approach often encounters challenges and limitations, such as variations in language, handwriting styles, and textual complexities, which can hinder its performance and accuracy.

On the other hand, image-based methods treat documents as visual artifacts and focus on analyzing the visual characteristics, such as texture, layout, and writing style, of digitized document images. This approach bypasses the complexities of textual analysis and directly analyzes the visual representations of documents. Various methods have been developed under this approach, leveraging techniques from statistics, machine learning, deep learning, pattern recognition, and computer vision.

In the upcoming chapter, we will delve deeper into these methods, exploring their principles, algorithms, and applications in the classification and analysis of historical documents. Understanding these methodologies in detail will provide insights into existing research efforts and lay the groundwork for our own contributions in this domain.

### 2.4.4 Conclusion

In conclusion, this chapter has provided a comprehensive overview of historical document classification, delineating the concept, tasks, and various methods involved in the analysis of these invaluable artifacts. We have elucidated the significance of classification as a fundamental step in understanding and organizing historical documents, emphasizing the diverse tasks encompassed within this process. Additionally, we have discussed the multitude of methods utilized in the analysis of historical documents, ranging from physical techniques to paleographical methods and computer-based approaches.

Moving forward, the subsequent chapter will delve deeper into the realm of computer-based methods, which have emerged as a prominent avenue for the study and classification of historical documents, exploring these methods in detail, we aim to gain a comprehensive understanding of their underlying principles, techniques, and applications. Through this exploration, we endeavor to elucidate the role of computational approaches in advancing our understanding of historical documents and contributing to the broader field of document studies.

# Chapter 3

# Computer based methods

## 3.1 Introduction

Building upon the foundational knowledge established in the previous chapter regarding historical document classification approaches, this chapter delves deeper into the computer-based approaches. Given that our study is predominantly focused on this category, this chapter serves as a crucial exploration of the methodologies and techniques employed in the automated classification of historical documents. In this chapter, we will explore the two primary methods within the computer-based approach: text-based and image-based focusing more on the second one due to its popularity in the community. Each of these methods share a range of concepts, including artificial intelligence, machine learning, and deep learning, which form the backbone of computational analysis in document studies. Moreover, we will elucidate the key steps involved in each method, from dataset acquisition and preprocessing to feature detection and classification.

## 3.2 What are computer-based methods?

Computer-based methods refer to the utilization of computers, algorithms, and computational techniques to automate tasks that were traditionally performed manually. In the historical document classification, computer-based methods involve leveraging computational power to analyze digitized document data, extract relevant features, and classify documents based on features. These methods encompass a wide range of algorithms and techniques related to artificial intelligence, machine learning, deep learning, natural language processing, pattern recognition, and computer vision.

Computer-based techniques can be broadly classified into two main categories: text-based and image-based techniques (Figure 3.1). Most studies in the literature are based on image-based methods due to the nature of documents, which can be challenging to process using text-based methods. Text-based approaches require clear text, high-quality documents, and expertise in ancient languages, all of which can be difficult to achieve with historical documents. Therefore, we will focus on the image-based techniques within our study.

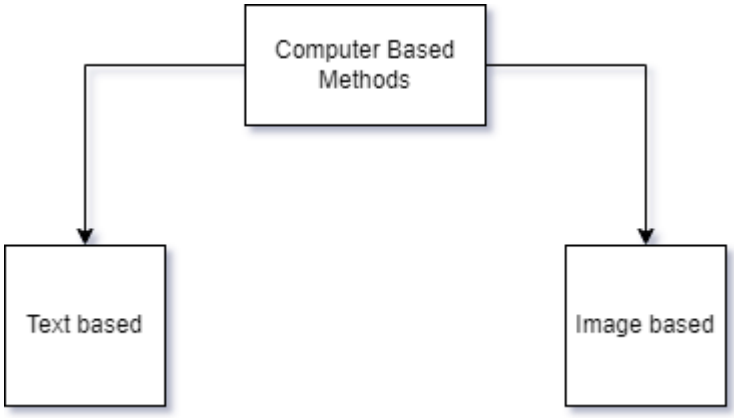



**Figure 3.1** Computer based methods categories for classification of historical documents

## 3.3 Documents Databases

The utilization of computer-based methods began with the digitalization of historical documents, enabling historians to study these invaluable artifacts using digital screens. These digitalized documents are stored in databases, which are referred to as datasets [**9,10,11]**. These databases contain a diverse array of old documents presented as digital information in different formats, providing researchers with the opportunity to analyze them using advanced algorithms and computational techniques (Figure 3.2).

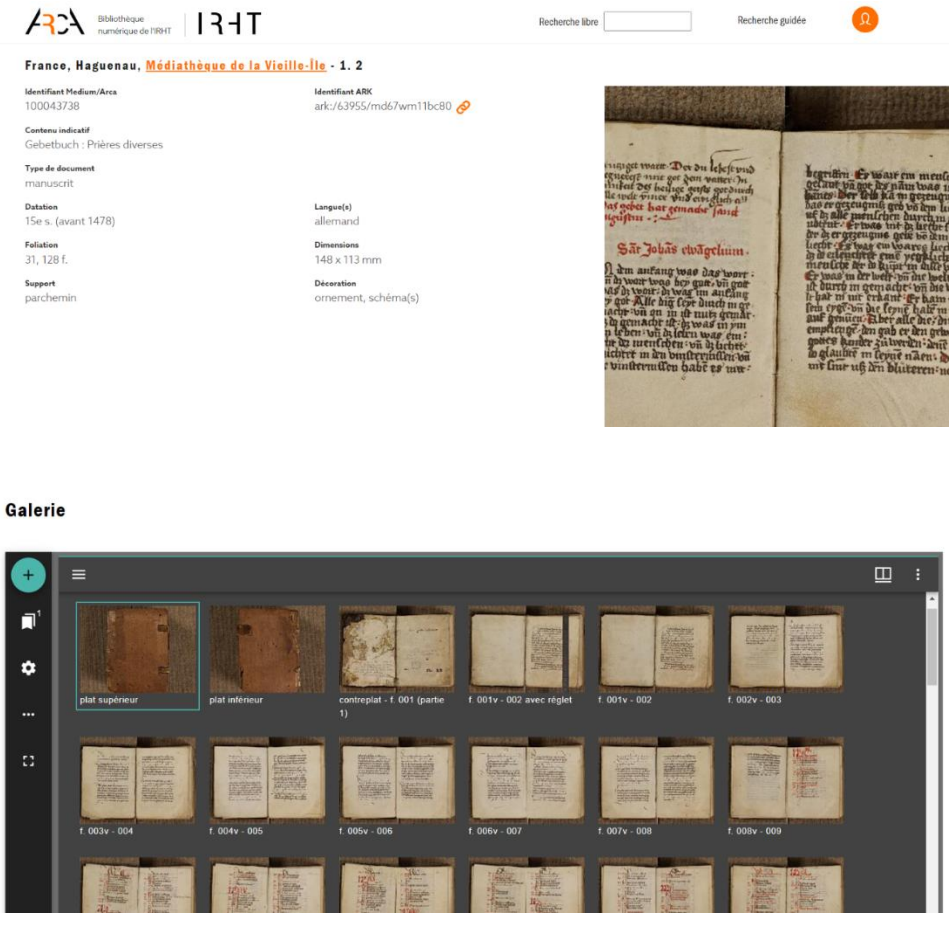



**Figure 3.2** Example of historical documents database **[57]**

The availability of these digital documents has paved the way for further digitalization efforts, leading to the creation of hundreds of datasets containing historical document digital information that are readily accessible for research purposes online. These databases provide researchers, scholars, and students with access to digitized images of historical documents, along with tools for analysis, search, and visualization through:

- **Digitized Documents:** contain digitized images of handwritten documents, often captured through high-resolution scanning or digital photography. These images preserve the visual characteristics of original documents, including ink color, stroke patterns, and paper texture. Digitized documents are organized and indexed within the database, allowing users to browse, view, and study individual pages or entire documents remotely via digital platforms.

- **Metadata and Cataloging:** incorporate metadata and cataloging information to describe the characteristics and provenance of digitized documents. Metadata may include details such as document title, author, date, language, script type, genre, and physical dimensions. Cataloging facilitates the organization, retrieval, and management of document collections within the database, enabling users to navigate and explore documents based on specific criteria or parameters.

- **Search and Retrieval Tools:** feature search and retrieval tools that allow users to locate specific documents, texts, or textual features within the database. Search functionality may include keyword search, advanced search filters, faceted browsing, and full-text search capabilities. Retrieval tools enable users to access and retrieve digitized images, transcriptions, or metadata associated with documents matching their search criteria.

- **Analysis and Visualization:** provide tools for analysis and visualization of document data, facilitating the exploration and interpretation of handwritten texts. Users can analyze textual patterns, linguistic features, and script characteristics using computational tools, statistical methods, and data visualization techniques. Visualization tools may include heatmaps, dendrograms, network graphs, and interactive visualizations that reveal patterns, trends, and relationships within document collections.

- **Collaborative and Crowdsourced Projects:** support collaborative and crowdsourced initiatives that engage researchers, scholars, and volunteers in the study of handwritten texts. Crowdsourcing platforms allow users to contribute to transcription, annotation, and analysis tasks, accelerating the digitization and study of document collections. Collaborative projects foster interdisciplinary collaboration, knowledge sharing, and scholarly exchange within the paleographic community.

## 3.4 Artificial Intelligence

**Artificial intelligence (AI)** refers to the development of computer systems that can perform tasks typically requiring human intelligence. Machine learning is a key component

of artificial intelligence (AI). It involves the development of algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed. In document classification, machine learning algorithms analyze historical document datasets to identify patterns, correlations, and trends within the data. Machine learning algorithms used in document classification include supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning algorithms are trained on labeled datasets, where each document is associated with a known category or class. These algorithms learn to map input features to output labels and can then classify new documents based on their learned knowledge. Unsupervised learning algorithms, on the other hand, work with unlabeled data and aim to discover hidden patterns or structures within the document dataset. Semi-supervised learning combines aspects of both supervised and unsupervised learning, leveraging a combination of labeled and unlabeled data for training.

Deep learning is a subset of machine learning that involves the use of artificial neural networks with multiple layers to learn features automatically from large amounts of data. In the context of historical document classification, deep learning algorithms excel in analyzing complex image data. For example, CNNs are commonly used for image-based document classification tasks, where they can automatically learn hierarchical features from document images, such as textures, shapes, and patterns.

These methods typically progress through three primary stages: preprocessing, feature extraction, and classification (Figure 3.3). In the forthcoming sections, we will delve into each stage, uncovering the methodologies and techniques employed within them.

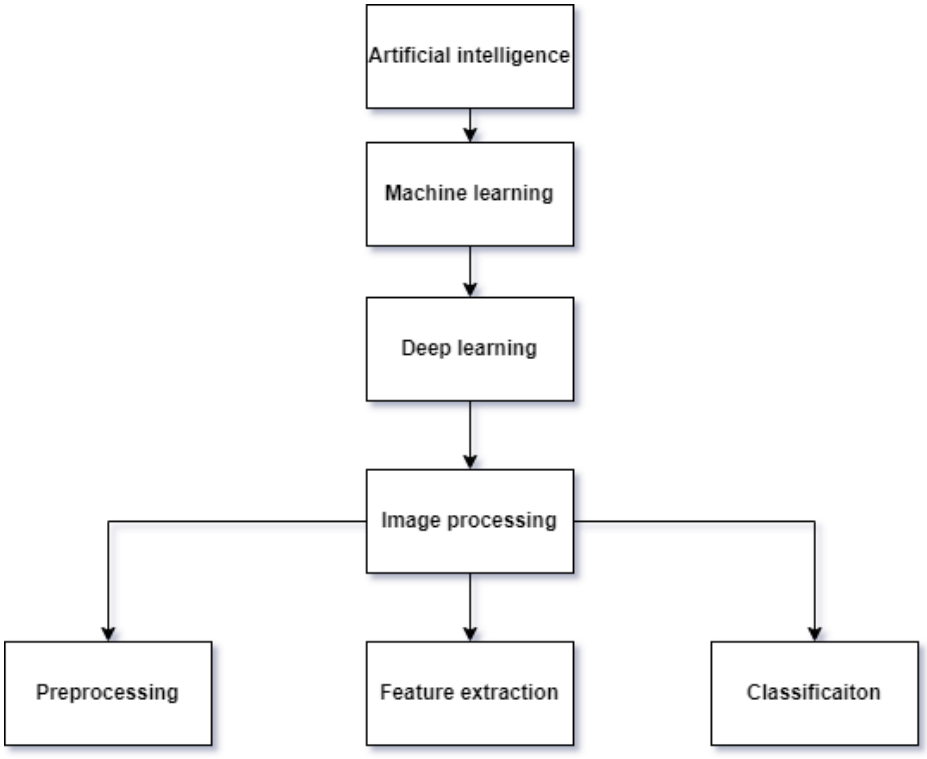



**Figure 3.3** Artificial intelligence subfields related to historical documents classification using image based methods

## 3.5 Preprocessing Techniques

Preprocessing is a critical component within machine learning systems, tasked with applying specialized techniques to improve the quality of data being analyzed. In the realm of historical documents, these preprocessing techniques hold particular importance, given their potential to profoundly influence the performance of classification systems. This section delves into an array of preprocessing techniques commonly used in historical documents.

Image-based preprocessing techniques are instrumental in addressing the challenges posed by digitized document images in historical document classification. These techniques are designed to enhance the quality and clarity of images, ultimately improving the accuracy of subsequent image analysis and classification tasks. As we concentrate on image-based methods for historical document classification in our study, we will explore in detail the preprocessing techniques utilized within this method.

### 3.5.1 Binarization

Binarization is a crucial preprocessing step in image-based approaches for historical document classification. It involves converting a grayscale image into a binary image, where the pixels are classified as either foreground (text) or background. This step enhances the document's readability and simplifies the subsequent feature extraction and classification tasks (Figure 3.4). Binarization techniques can be broadly categorized into several types as Figure 3.5 shows. Each method has its strengths and weaknesses, particularly when applied to the unique challenges presented by historical documents **[59]**.

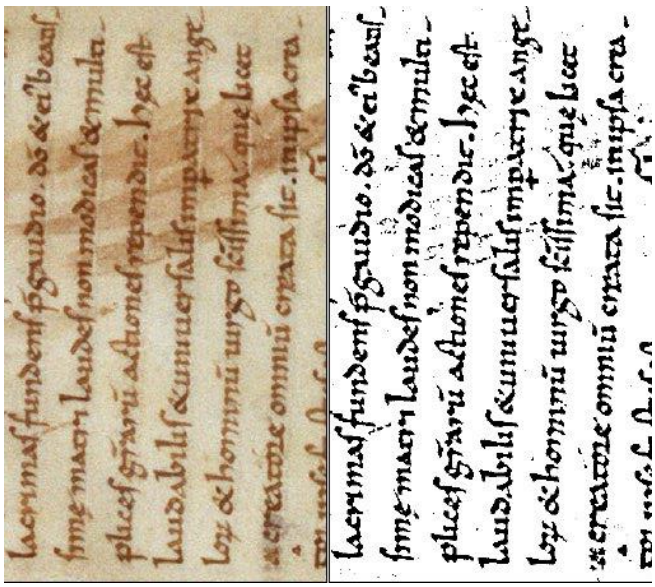


**Figure 3.4** Binarization of ancient document using Sauvola's thresholding **[58]**

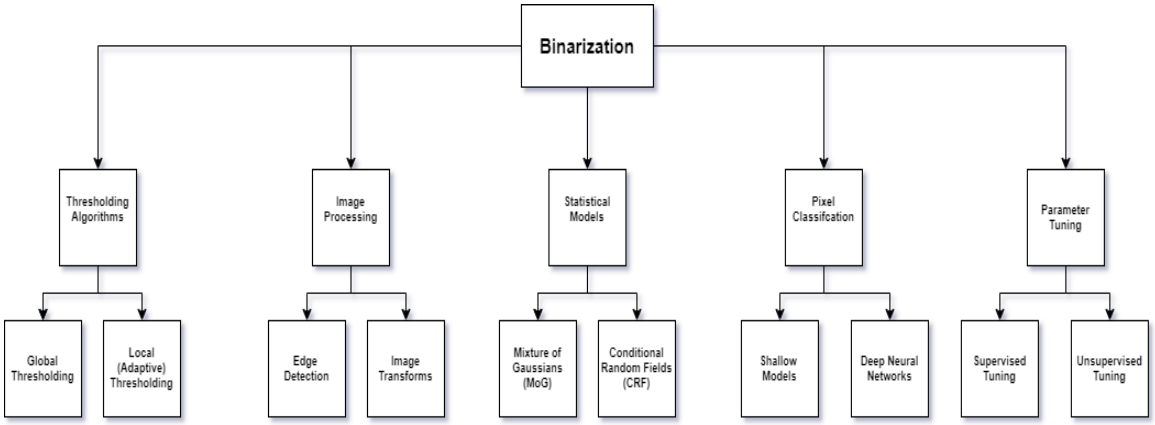



**Figure 3.5** Binarization types of techniques

#### 3.5.1.1 Classic Thresholding Algorithms

- **Global Thresholding:** Global thresholding such as Otsu's Method **[60]**, involves selecting a single threshold value for the entire image. Pixels with intensity values above the threshold are classified as foreground (white), while those below are classified as background (black) as Equation (1) shows.

$$I_b(x,y) = \begin{Bmatrix} 1 \; if \; I(x,y) \geq T \\ 0 \; if \; I(x,y) < T \end{Bmatrix} \qquad \textbf{(1)}$$

**Where:**

- $I_b(x,y)$ is the binarized image
- $I(x,y)$ is the grayscale image
- T is the global threshold

- **Local (Adaptive) Thresholding:** Local thresholding such as Sauvola's Method **[61]** calculates a threshold for each pixel based on its local neighborhood (Equation (2)).

$$I_b(x,y) = \begin{Bmatrix} 1 \; if \; I(x,y) \geq T(x,y) \\ 0 \; if \; I(x,y) < T(x,y) \end{Bmatrix} \qquad \textbf{(2)}$$

**Where** $T(x,y)$ is the local threshold computed from the local neighborhood of the pixel (x,y)

#### 3.5.1.2 Image Processing Techniques

- **Edge Detection:** Edge detection techniques, such as the Canny edge detector **[62]**, identify edges within the image, which can then guide the binarization process.
- **Image Transforms:** Image transform techniques, such as the Fourier transform or wavelet transform **[63]**, enhance features in the frequency domain before binarization. After enhancing specific frequency components, the inverse transform is applied, followed by binarization.

#### 3.5.1.3 Statistical Models

- **Mixture of Gaussians (MoG):** MoG models the distribution of pixel intensities as a mixture of several Gaussian distributions based on Equation (3) **[64]**.

$$P(I(x,y)) = \sum_{k=1}^{K} \pi_k N(I(x,y) \mid u_k, \sigma_k^2) \qquad \textbf{(3)}$$

**Where:**

- $\pi_k$ are the mixture weights
- $u_k$ and $\sigma_k^2$ are the mean and variance of the k-th Gaussian component.

- **Conditional Random Fields (CRF):** CRFs model the binarization problem as a labeling problem, considering the contextual relationships between neighboring pixels as demonstrated in Equation (4) **[65]**.

$$P(Y \mid I) = \frac{1}{Z(I)} \exp\left(\sum_{c \in C} \phi_c(Y_c, I)\right) \qquad \textbf{(4)}$$

**Where** $Z(I)$ is the partition function, and $\phi_c$ are the potential functions defined over cliques C.

#### 3.5.1.4 Pixel Classification

- **Shallow Models:** Shallow models, such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), classify pixels based on handcrafted features.
- **Deep Neural Networks:** Deep learning approaches, particularly Convolutional Neural Networks (CNNs), learn features directly from the image data to perform binarization.

#### 3.5.1.5 Parameter Tuning

- **Supervised Tuning:** Supervised tuning involves training models using labeled datasets to optimize the binarization process.
- **Unsupervised Tuning:** Unsupervised tuning optimizes the binarization process without labeled data, often using clustering techniques like k-means.

### 3.5.2 Noise Reduction

Noise Reduction, also known as denoising, is a crucial preprocessing step in the classification of historical documents. It involves the removal or reduction of unwanted data or artifacts from the dataset, which can significantly impact the performance of the classification system. In the context of historical documents, denoising is particularly important due to the presence of various types of noise and distortions in digitized images (Figure 3.6).

One of the key challenges in historical document classification is dealing with noise introduced during the digitization process. This noise can manifest as random pixel variations, scanning artifacts, ink smudges, paper texture imperfections, and other

irregularities that affect the quality and clarity of the document image. Denoising techniques aim to enhance the quality of digitized document images by filtering out unwanted noise while preserving important features relevant to classification tasks. These techniques can be categorized into several types as shown in Figure 3.7.

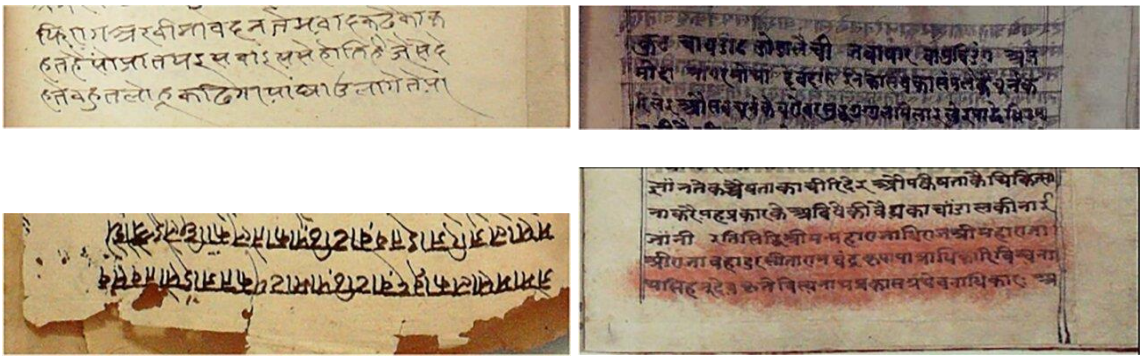


**Figure 3.6** Noise forms in historical documents **[66]**

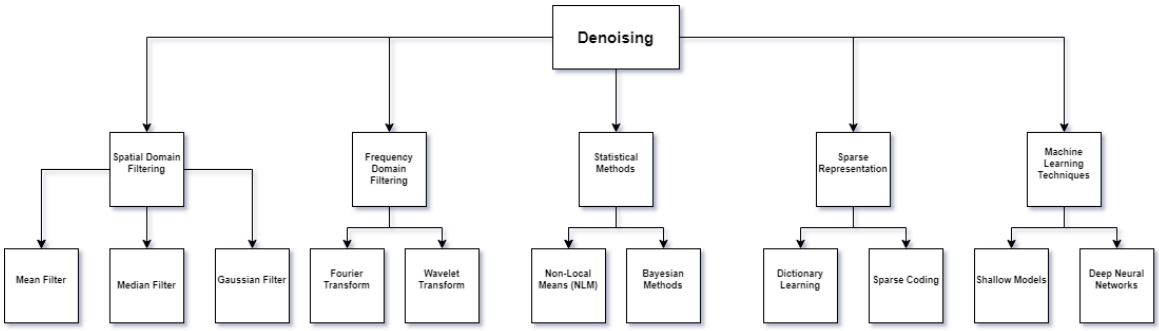



**Figure 3.7** Denoising types of techniques

### 3.5.2.1 Spatial Domain Filtering

Spatial domain filtering methods such as **[67,68],** operates directly on the image pixels and involves convolution operations with a kernel or mask that defines the filtering operation. These techniques are used to smooth or enhance an image by averaging or modifying the pixel values within a local neighborhood.

**Example:**

- **Mean Filtering:** Replaces each pixel value with the average value of its neighboring pixels, helping to reduce random noise (Equation (5)).

$$g(x,y) = \frac{1}{|W|}\sum_{(i,j)\in W} f(x+i, y+j) \qquad \textbf{(5)}$$

**Where** $f(x,y)$ is the original image, $g(x,y)$ is the filtered image, and $W$ is the set of neighboring pixels.

Spatial domain filtering helps in removing noise and enhancing the readability of degraded documents, making it easier for subsequent image-based analysis tasks like classification and feature extraction.

#### 3.5.2.2 Frequency Domain Filtering

Frequency domain filtering methods such as **[69,70],** involves transforming the image into the frequency domain using a transformation like the Fourier Transform, performing operations, and then transforming it back to the spatial domain.

**Example:**

- **Low-Pass Filtering:** Attenuates high-frequency components, which correspond to noise, while preserving low-frequency components (important image structures) as shown in Equation (6).

$$G(u,v) = H(u,v) \cdot F(u,v) \qquad \textbf{(6)}$$

**Where** $F(u,v)$ is the Fourier transform of the image, $H(u,v)$ is the filter function, and $G(u,v)$ is the result.

Frequency domain filtering can be used to enhance the edges and fine details in historical documents, aiding in tasks like script identification and writer recognition.

#### 3.5.2.3 Statistical Methods

Statistical methods such as **[71]** use statistical properties of the image to perform denoising, often considering noise as a probabilistic model and estimating the true signal.

**Example:**

- **Wiener Filtering:** Uses the image's local mean and variance to estimate the true image, minimizing the mean square error using Equation (7).

$$g(x,y) = \mu + \frac{\sigma^2 - \eta^2}{\sigma^2}\,(f(x,y) - \mu) \qquad \textbf{(7)}$$

**Where** μ is the local mean**,** $\sigma^2$ is the local variance, and $\eta^2$ is the noise variance.

- **Bayesian Denoising:** Applies Bayesian inference to estimate the true image based on the observed noisy image and a prior model of the image (Equation (9)).

$$p(I \mid Y) = \frac{p(Y|I)p(I)}{P(Y)} \qquad \textbf{(8)}$$

**Where** $I$ is the true image, $Y$ is the observed noisy image, $p(Y \mid I)$ is the likelihood, $p(I)$ is the prior, and $p(Y)$ is the evidence.

Statistical methods are particularly useful for denoising historical documents where the noise characteristics can be modeled probabilistically, leading to better preservation of fine details and textures.

#### 3.5.2.4 Sparse Representation

Sparse representation such as **[72]** involves representing the image as a sparse combination of basic functions or atoms, which can effectively separate noise from the true signal.

**Examples:**

- **K-SVD:** An iterative method for designing overcomplete dictionaries for sparse representation, which can be used for image denoising.
- **Wavelet Transform:** Represents the image in terms of wavelets, allowing for effective denoising by thresholding the wavelet coefficients.

Sparse representation methods are effective for denoising historical documents as they can preserve significant features and structures while removing noise, crucial for tasks like script and writer identification.

#### 3.5.2.5 Machine Learning Techniques

Machine learning techniques such as **[73]**, involve training models to learn the noise characteristics and remove it from the images. These can range from shallow models to deep neural networks.

**Examples:**

- **Autoencoders:** Neural networks trained to encode the input image into a lower-dimensional representation and then decode it back to the original image, learning to remove noise in the process.
- **Convolutional Neural Networks (CNNs):** Used to learn a mapping from noisy images to clean images through supervised learning.

Machine learning techniques, especially deep learning models, have shown great promise in denoising historical documents by learning complex noise patterns and effectively removing them, improving the quality and readability of the documents for further analysis.

### 3.5.3 Contrast Enhancement

Contrast enhancement techniques adjust the contrast levels of images, making the textual content more visible and improving overall image clarity. This is particularly beneficial for enhancing faint or degraded textual content that may be difficult to discern. By adjusting the contrast, these techniques can reveal hidden details and make the text stand out against the background, facilitating better recognition and analysis.

**Example:**

- **Histogram Equalization [74]:** This technique redistributes the intensity values of an image to enhance contrast using Equation (9).

$$p(r_k) = \frac{n_k}{n} \; and \; s_k = \sum_{j=0}^{k} p(r_j) \qquad \textbf{(9)}$$

Where $p(r_k)$ is the probability of intensity level $r_k$, $n_k$ is the number of pixels with intensity $r_k$ and n is the total number of pixels.

## 3.5.4 Resizing and Cropping

Resizing and cropping images (Figure 3.8) to standard dimensions help in aligning and optimizing image data for analysis. Standardizing image sizes ensures consistency across the dataset, which is crucial for efficient processing and analysis. This step is particularly important in machine learning applications where uniform input dimensions are required for training and inference.

- **Resizing**: Adjusting the dimensions of an image to a specified size using interpolation methods.
- **Cropping**: Removing unwanted outer areas from an image to focus on the region of interest.




**Figure 3.8** demonstration of resizing and cropping techniques

## 3.6 Feature Extraction

Feature extraction is a critical stage in machine learning models, encompassing the detection and extraction of key information that significantly contributes to the classification processes, it involves identifying and isolating relevant information from raw data, which is crucial for accurate and effective classification due to the complexity and diversity of textual and visual content found in these documents. It allows:

- **Capturing Unique Attributes:** Historical documents exhibit a wide range of unique attributes, including writing styles, script variations, language nuances, historical context, and content themes. Feature extraction helps in capturing these distinct

characteristics by identifying key patterns, shapes, textures, and structures present in the textual or image data. For example, features related to handwriting strokes, ink color variations, parchment texture, and script layouts can be extracted to differentiate between documents from different time periods or geographical regions.

- **Enhancing Classification Accuracy:** The extraction of relevant features, machine learning algorithms can build robust models that accurately classify historical documents into predefined categories or classes.

- **Facilitating Pattern Recognition:** Feature extraction facilitates pattern recognition by transforming raw data into meaningful representations.

- **Enabling Interpretability and Insights:** Extracted features provide interpretability to machine learning models, allowing researchers and historians to understand the underlying factors influencing classification decisions.

- **Adapting to Data Variability:** Feature extraction techniques enable algorithms to adapt to this variability by capturing relevant information while disregarding noise or irrelevant data.

In image-based approaches for historical document classification, feature extraction techniques are essential for identifying visual patterns and elements that contribute to the differentiation of documents based on their visual attributes. These techniques focus on analyzing the structural and textural characteristics of digitized document images to extract meaningful features. Importantly, there are two main categories of feature extraction methods: handcrafted features and deep learning-based features (Figure 3.9).

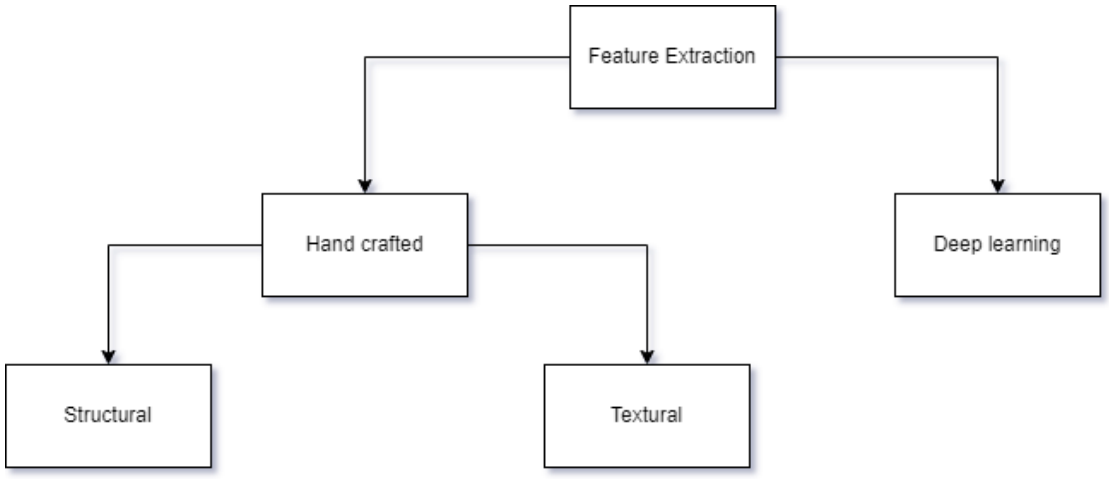



**Figure 3.9** Feature extraction technique categories

## 3.6.1 Handcrafted Features

Handcrafted feature extraction involves manually designing algorithms to identify specific patterns and characteristics within the images. This approach relies on human

expertise to determine which features are relevant for the task. Common handcrafted features used in historical document classification include:

#### 3.6.1.1 Structural Features

Structural features capture the geometric properties of a document's text, focusing on the shapes, strokes, and curvatures that define handwriting styles. These features are essential in distinguishing different handwriting styles, which is crucial for tasks like writer identification, script classification, and dating historical documents (Figure 3.10).

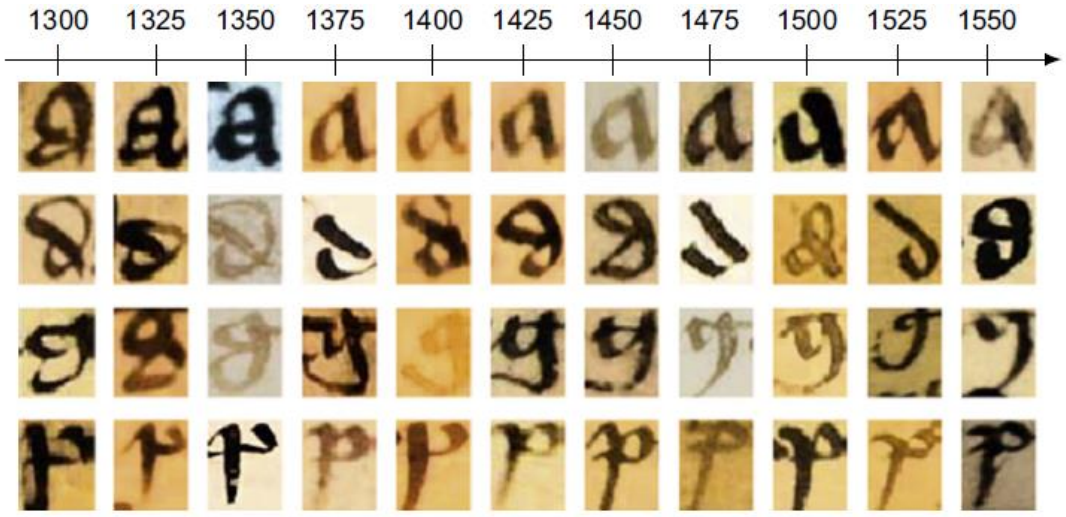



**Figure 3.10** Structure of characters in historical documents over time **[75]**

In historical document analysis, the structure of the handwriting provides significant clues about the writer, the time period, and the cultural context. The geometric properties of the text can reveal:

- **Writer Identification:** Each individual's handwriting has unique structural characteristics. By analyzing the shapes and strokes of letters, it is possible to identify the author of a document.
- **Script Classification:** Different scripts have distinct geometric patterns. Structural features help in classifying documents into categories such as Gothic, Carolingian, or Byzantine scripts.
- **Dating and Provenance:** The evolution of handwriting styles over time can assist in dating documents and determining their geographical origin.

Several structural features are commonly used in the analysis of handwriting in historical documents (Figure 3.11):

- **Character Shapes:** The overall form and design of individual characters, including their height, width, and slant.
- **Strokes**: The direction, curvature, and pressure of pen strokes used to form characters.

- **Curvature**: The smoothness and sharpness of curves within characters and between connected letters.
- **Ascenders and Descenders:** The parts of letters that extend above or below the typical character height, such as the top of 'd' or the bottom of 'g'.



**Figure 3.11** Example of structural features from old document **[75]**

To effectively utilize structural features, various computational methods can be employed. Here is an example:

- **Histogram of Oriented Gradients (HOG) [76]:** Captures the distribution of gradient directions in localized portions of an image, effectively describing the structure of text using Equation (10).

$$H(i) = \sum_{(x,y)} \frac{I(x,y)}{G(x,y)} \cdot \delta(\theta(x,y) - \theta_i) \qquad \textbf{(10)}$$

Where $H(i)$ is the histogram of gradients, $I(x,y)$ is the image intensity, $G(x,y)$ s the gradient magnitude, and $\theta(x,y)$ is the gradient orientation.

While structural features are powerful, they come with challenges such as:

- **Degraded Documents:** Poor quality and damaged documents can make it difficult to accurately extract structural features.

- **Variability:** Handwriting styles can vary widely even within the same document, requiring robust methods to handle this diversity.

- **Scalability:** Efficiently processing large collections of documents requires scalable feature extraction techniques.

#### 3.6.1.2 Textural features

Textural features describe the texture of a document's surface, capturing details such as ink density, paper grain, and other surface irregularities. These features are essential in

analyzing historical documents, as they provide critical information that can be used for classification.

The texture of a document encompasses various physical properties that can reveal a wealth of information about its origin, condition, and authenticity:

- **Ink Density:** Variations in ink density can indicate different writing instruments, ink compositions, and even the pressure applied by the scribe.

- **Paper Grain:** The texture of the paper or parchment can provide clues about its manufacturing process, age, and geographical origin.

- **Surface Irregularities:** Features such as wrinkles, stains, and other imperfections can offer insights into the document's history and the environmental conditions it has endured.

Several textural features are commonly used in the analysis of documents:

- **Gray-Level Co-occurrence Matrix (GLCM) [77]**: Calculates how often pairs of pixels with specific values and in a specified spatial relationship occur in an image aiming to capture the spatial relationship between pixels in an image to describe texture.

- **Local Binary Patterns (LBP) [78]:** Describes the local texture around each pixel comparing each pixel to its neighboring pixels and encodes the result as a binary number (Equation (11)).

$$LBP(x,y) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad \textbf{(11)}$$

  Where $g_c$ is the gray value of the center pixel, $g_p$ is the gray value of neighboring pixels, and s is a binary function.

- **Gabor Filters [79]:** Captures texture by analyzing the frequency and orientation of the local image structure convolving the image with a set of Gabor filters with various scales and orientations.

## 3.6.2 Deep Learning Features

Deep learning-based feature extraction leverages neural networks to automatically learn and extract features from data, offering significant advantages over traditional handcrafted methods. This approach has gained popularity due to its ability to discover complex patterns and representations without manual intervention, making it particularly effective for tasks involving large and diverse datasets, such as historical document classification. Advantages of Deep Learning-Based Features:

- **Automatic Feature Learning:** Unlike traditional methods that require manual feature engineering, deep learning models can automatically learn relevant features from raw data, reducing the need for domain-specific knowledge.

- **Complex Pattern Recognition:** Deep learning models can capture intricate and non-linear relationships within the data, making them capable of identifying complex patterns that may be missed by traditional methods.

- **Scalability:** Deep learning approaches can scale effectively with the availability of more data, often improving performance as more training data is provided.

Several deep learning architectures are commonly used for feature extraction in image-based tasks, including:

- **Convolutional Neural Networks (CNNs) [80]:** CNNs are specialized neural networks designed for processing grid-like data, such as images. They use convolutional layers to automatically extract spatial hierarchies of features.

  **Key layers:** Convolutional layers, pooling layers, and fully connected layers (Figure 3.12).

  **Example**: VGGNet **[82]**, ResNet **[83]**, and Inception **[84]** are popular CNN architectures used for image classification and feature extraction.

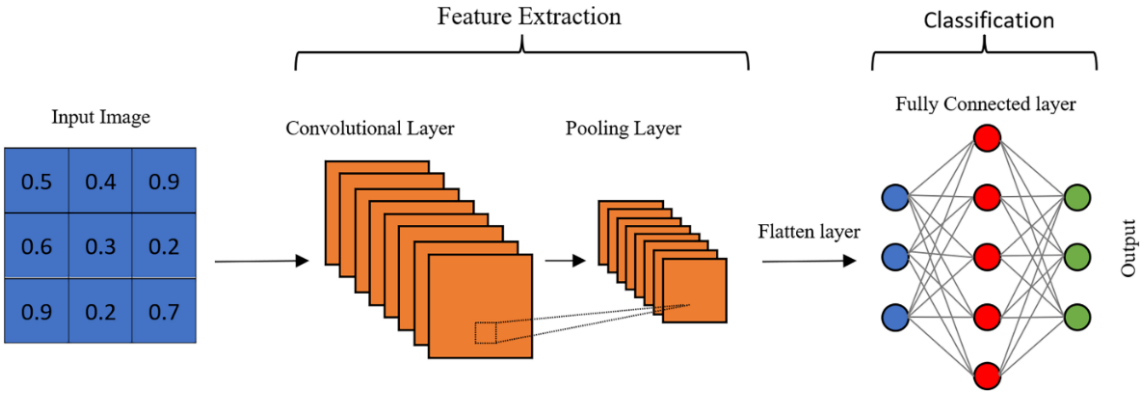



**Figure 3.12** Architecture of a CNN **[81]**

- **Recurrent Neural Networks (RNNs) [85]:** RNNs are designed for sequential data and can capture temporal dependencies. While not typically used for raw image data, RNNs can be applied to sequences of image features.

  **Key Layers:** Recurrent layers, such as LSTM or GRU cells **(Figure 13)**.

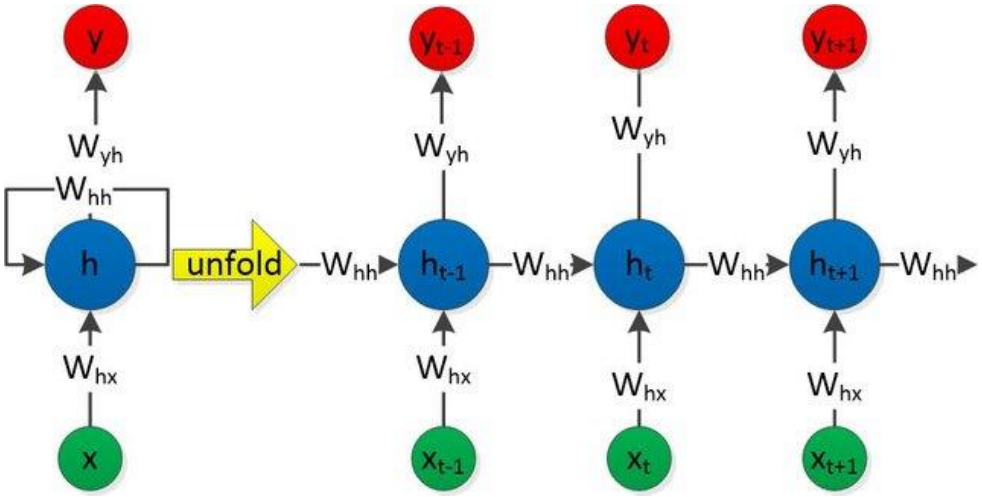



**Figure 3.13** RNN architecture **[86]**

- **Transformers:** Transformers **[87]** are neural networks that rely on self-attention mechanisms to process sequential data. Recently, Vision Transformers (ViTs) have been adapted for image processing tasks.
  **Key Layers:** Self-attention layers, multi-head attention, and feed-forward neural networks (Figure 3.14).

  **Example:** Vision Transformer (ViT) [88] for image classification and feature extraction.

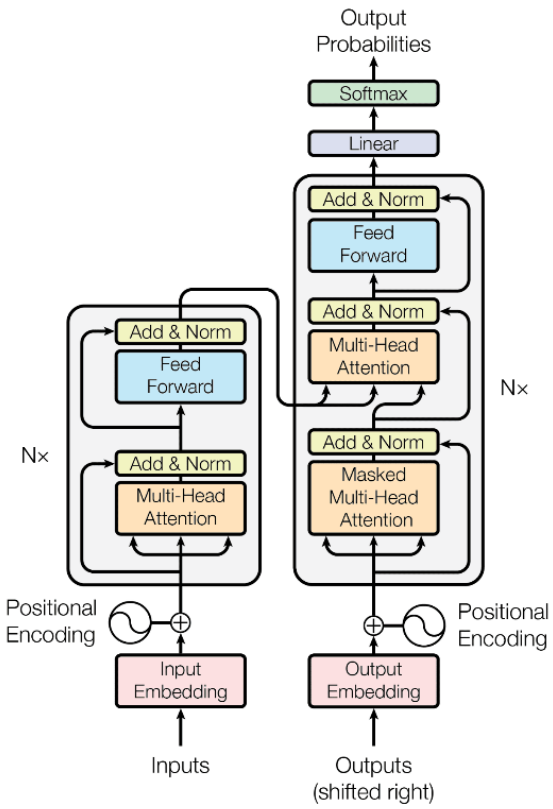



**Figure 3.14** Transformer architecture

Despite their advantages, deep learning-based feature extraction methods face several challenges:

- **Data Requirements:** Deep learning models require large amounts of labeled data for training, which can be a limitation in the context of historical documents where labeled datasets are scarce.

- **Computational Resources:** Training deep learning models, especially large architectures like transformers, demands significant computational power and memory.

- **Interpretability:** Deep learning models are often considered "black boxes," making it challenging to interpret the learned features and understand the decision-making process.

In summary, feature extraction is a foundational step in the classification of historical documents, enabling the identification of relevant visual patterns and elements. By employing a combination of handcrafted and deep learning-based techniques, we can effectively analyze the structural and textural characteristics of documents, paving the way for accurate and efficient classification.

## 3.7 Classification

After the preprocessing and feature extraction stages, the classification stage plays a pivotal role in the overall process of historical document classification. This stage involves the application of machine learning algorithms to assign a label or category to each document based on its extracted features. The classification process aims to automate the identification and categorization of documents, making it easier for historians, researchers, and archivists to manage and analyze large volumes of historical documents. Several classification algorithms can be employed in historical document classification, depending on the nature of the extracted features and the complexity of the classification task (Figure 3.15). Some common classification algorithms include:

- **Support Vector Machines (SVM):** SVM is a powerful supervised learning algorithm used for binary and multiclass classification tasks. It works by finding the optimal hyperplane that separates different classes in the feature space while maximizing the margin between classes. SVMs are effective for handling high-dimensional data and nonlinear decision boundaries.

- **Decision Trees:** Decision trees are hierarchical tree-like structures that make decisions based on feature values at each node. They recursively split the data into subsets based on the most discriminative features, leading to a tree structure where each leaf node represents a class label. Decision trees are interpretable and can handle both numerical and categorical data.

- **Random Forest:** Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. It creates a forest of decision trees by training each tree on a random subset of features and aggregating their predictions through voting or averaging. Random Forests are robust, scalable, and suitable for large datasets.

- **K-Nearest Neighbors (KNN):** KNN is a simple yet effective lazy learning algorithm used for classification tasks. It classifies new data points by comparing them to the k closest training examples in the feature space and assigning the majority class label among its neighbors. KNN is non-parametric and adaptable to varying data distributions.

- **Naive Bayes:** Naive Bayes is a probabilistic classifier based on Bayes' theorem and the assumption of feature independence. It calculates the posterior probability of each class given the input features and selects the class with the highest probability. Naive Bayes is efficient, particularly for text classification tasks.

- **Neural Networks:** Neural networks, including networks (CNNs), can also be used for classification stage.

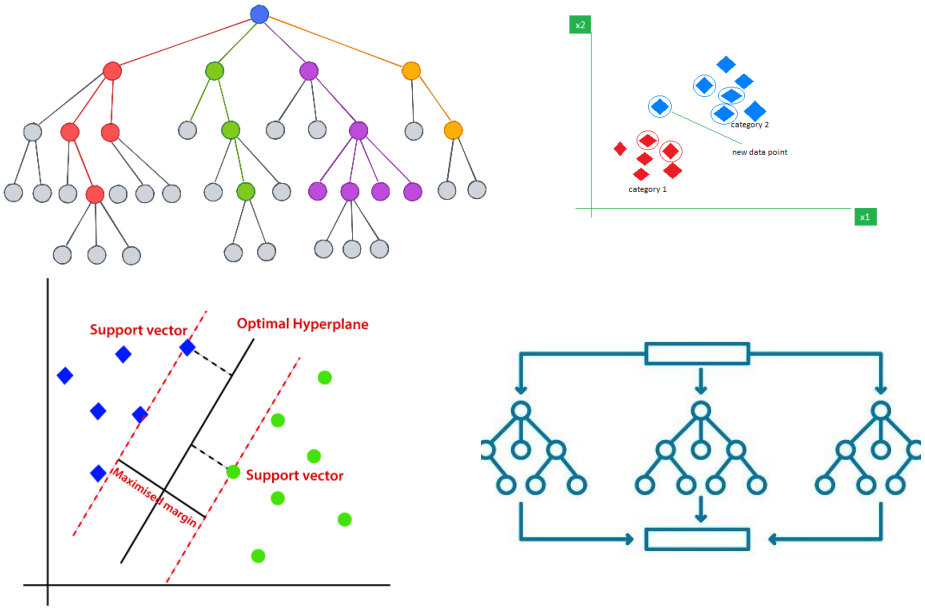



**Figure 3.15** Different classification algorithms

The classification stage is a crucial step in historical document classification, where machine learning algorithms leverage extracted features to automate the categorization of documents. By employing diverse classification algorithms and evaluating model performance using appropriate metrics, researchers and historians can enhance the accuracy, efficiency, and interpretability of historical document classification systems.

## 3.8 Conclusion

In summary, this chapter provides a comprehensive exploration of the principles relevant to historical document classification within the domain of image-based methodologies. We have elucidated the fundamental techniques employed in this approach, including machine learning and deep learning, while also delineating the overarching phases of these techniques as applied to our historical document classification problem. Specifically, we have detailed the preprocessing methodologies, the process of feature extraction, and the various techniques for classification. This chapter serves as a foundational stepping stone for comprehending subsequent chapters, particularly in the realm of related works, where we delve into prior research utilizing these aforementioned methodologies. Furthermore, it sets the stage for our forthcoming contribution, which will be unveiled in the final chapter.

# Chapter 4

# Related Works

## 4.1 Introduction

Having delved into the core concepts surrounding historical documents classification and the primary methodologies utilized for addressing classification challenges, we have established a foundational understanding of our research problem. The preceding chapters have equipped us with insights into the prevalent approaches and techniques applied in this field. In this chapter, we shift our focus towards an examination of existing literature, particularly emphasizing state-of-the-art datasets and techniques that are closely aligned with our research goals, from traditional to deep learning methods. Our primary interests lie in areas such as writer identification, dating methods, and script classification. Exploring related works in these domains, we aim to lay the groundwork for our forthcoming contribution, which will be elaborated upon in the subsequent chapter.

## 4.2 Datasets

Datasets are the lifeblood of machine learning and deep learning systems. In the realm of historical document classification, a plethora of datasets is available to the research community. These datasets have been meticulously curated and compiled by various entities such as institutions, laboratories, governmental bodies, and individual contributors. They serve as repositories of invaluable historical document data, facilitating experimentation, model training, and algorithm development within the field. In this section, we will explore some notable datasets that have been utilized by researchers in the literature to address various historical document classification problems.

### 4.2.1 Medieval Paleographical Scale (MPS)

The MPS dataset **[89]**, consisting of 3267 images of dated charters from four cities spanning the period between 1300 and 1550, plays a crucial role in historical document classification. These charters, being formal declarations of legal or financial transactions, offer a rich source of diverse writing hands without major chronological gaps (Figure 4.1). The dataset's focus on charters in a consistent script type, primarily 'cursive' script or "53ursive," allows for the establishment of a Paleographic scale. This scale serves as a foundation for comparing and extrapolating script development in charters to other script types, such as "textualis" used in books. Despite inherent imbalances in the dataset due to historical record-keeping, its extensive collection of dated documents provides valuable data for analyzing and classifying historical document scripts, aiding in the dating task.

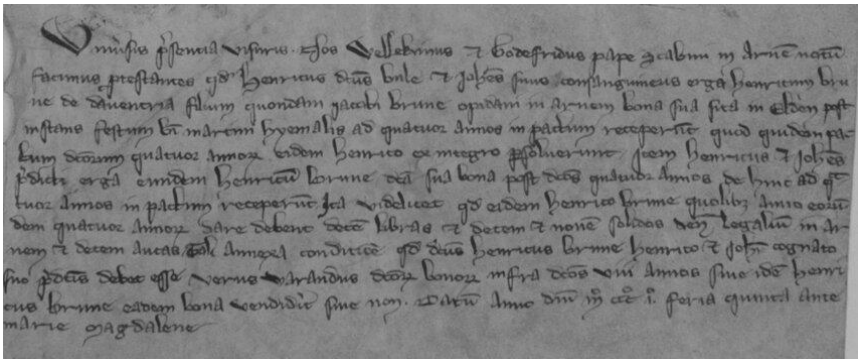


**Figure** 4.1 Document sample from MPS dataset

## 4.2.2 Svenskt Diplomatariums Huvudkartotek (SDHK)

The Swedish Diplomatarium's main directory (SDHK) **[90]** has extensive collection of Swedish medieval letters spanning a significant time period. These letters provide valuable textual and contextual information that can be leveraged for training machine learning models in historical document classification tasks (Figure 4.2). The database's search capabilities based on personal names, place names, dates, and important facts allow researchers to extract specific datasets relevant to their classification studies. Furthermore, the inclusion of detailed information about each letter's tradition and preservation history enhances the reliability and accuracy of the dataset. The availability of digital images linked to original letters also offers opportunities for image-based classification tasks. Therefore, SDHK serves as a rich and comprehensive resource for historical document classification research, contributing to the development of effective classification models and methodologies.



**Figure** 4.2 Document sample from SDHK dataset

## 4.2.3 KERTAS dataset

The KERTAS dataset **[91]** is highly relevant to our historical document classification problem, specifically focusing on writer identification and age detection. Its organization

by Islamic centuries aligns with historical document structures, ensuring a historical context for analysis (Figure 4.3). The dataset's meticulous verification of writing dates and handling of complexities like copied documents provide a reliable foundation for algorithm training and testing. Moreover, the diverse content, ranging from mathematics to Islamic history, offers a comprehensive set of documents for algorithm evaluation. This diversity includes figures, tables, marginal comments, and various writing styles, mirroring the intricacies found in historical documents. Therefore, the KERTAS dataset serves as a valuable resource for developing and testing robust algorithms in writer identification, dating, document segmentation, and text analysis within the historical document domain.



**Figure** 4.3 Document sample from KERTAS dataset

## 4.2.4 DEEDS (Document of Early England Data Set)

DEEDS **[92]** is a dataset featuring a diverse range of early English documents, including charters, documents, and legal texts (Figure 4.4). Researchers leverage this dataset to study English paleography, textual content, and historical context dating back to the medieval period.



**Figure** 4.4 Document sample from DEEDS dataset

### 4.2.5 ICFHR 2016 Competition on the Classification of Medieval Handwritings in Latin Script

This dataset was created specifically for the ICFHR 2016 competition focusing on the classification of medieval handwritings in Latin script **[93]**. It includes a variety of handwritten samples for training and evaluation of classification algorithms (Figure 4.5). It consists of grayscale images in TIFF format at 300 dpi, capturing a 100 x 150 mm section of document pages. It is divided into three sets: a training set containing 2000 images, a task 1 set with 1000 images, and a task 2 set with 2000 images. These images were primarily sourced from French catalogues of dated and datable documents, supplemented with additional documentation from BVMM and Gallica to ensure balanced classes. The training set images are classified into 12 labels based on morphological distinctions observed in Latin scripts, such as Uncial, Half-uncial, Caroline, Humanistic, Humanistic Cursive, Praegothica, Southern Textualis, Semitextualis, Textualis, Hybrida, Semihybrida, and Cursiva. The distribution of script classes varies between task 1 and task 2, as detailed in Tables I and II respectively, reflecting the varying cardinality of script types in each task-specific dataset.

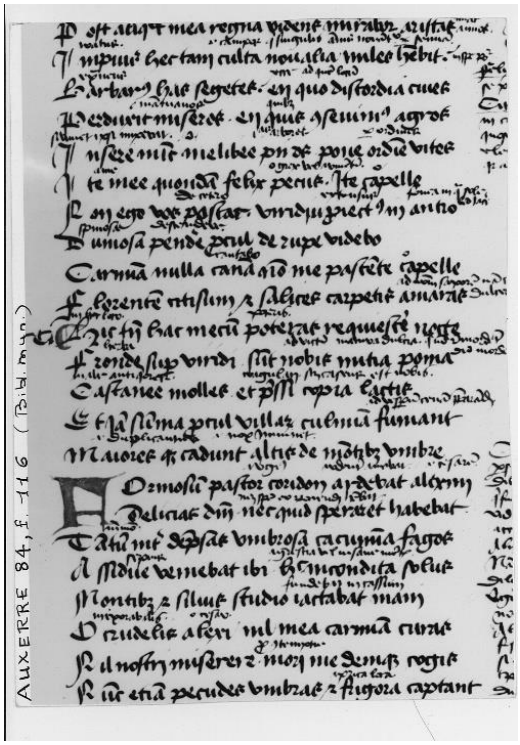



**Figure** 4.5 Document sample from ClaMM 2016 dataset

### 4.2.6 ICDAR 2017 Competition on the Classification of Medieval Handwritings in Latin Script

The ICDAR 2017 dataset **[94]** is akin to the ICFHR16 CLaMM dataset, serving as a benchmark for evaluating models focused on classifying medieval Latin script handwritings. This dataset includes both training and test datasets for tasks 1 and 3, consisting of grayscale images in TIFF format at 300 dpi, capturing a 100 x 150 mm portion of document pages (Figure 4.6). Conversely, the test dataset for tasks 2 and 4 comprises heterogeneously encoded data, featuring varying formats (JPEG/TIFF), resolutions (300/400 dpi), and color representations (grayscale and color images). The training set contains 3540 images, while tasks 1 and 3 have 2000 images each, and tasks 2 and 4 have 1000 images each. These images are sourced from French catalogues of dated and datable documents, supplemented by documentation from BVMM and Gallica. Similar to previous datasets, the training set images are categorized into 12 labels based on morphological differences observed in Latin scripts.

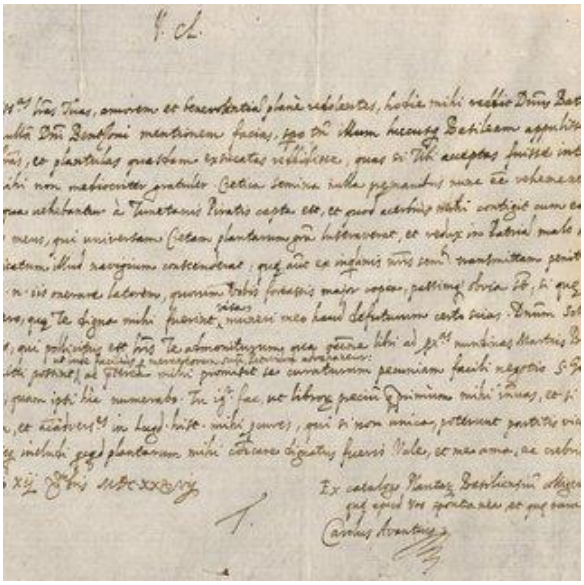


**Figure** 4.6 Document sample from ClaMM 2017 dataset

### 4.2.7 ICDAR 2021 Competition on Historical Document Classification

The ICDAR 2021 competition data **[95]** includes: font group classification, script type classification along with a dating and location classification tasks. For font group classification, the training data consists of over 35,000 document images categorized into ten font group classes, with two noise classes, while the test data comprises document images obtained from the Bayerische Staatsbibliothek München (Figure 4.7). The test data has been balanced to include rarer fonts by discarding pages with extremely frequent font groups. Additionally, the dataset has been augmented for robustness testing, resulting in a total of 5506 images for testing. The script type classification task utilizes the CLaMM 2017 dataset for training, with test data obtained from the same documents, augmented to avoid biases. The dating task uses scripts from e-codices, and the dataset contains 1698 documents with unambiguous dating intervals, split into training and test subsets based on their centuries of origin. Overall, the dataset includes 11,810 pages for training and testing across all tasks

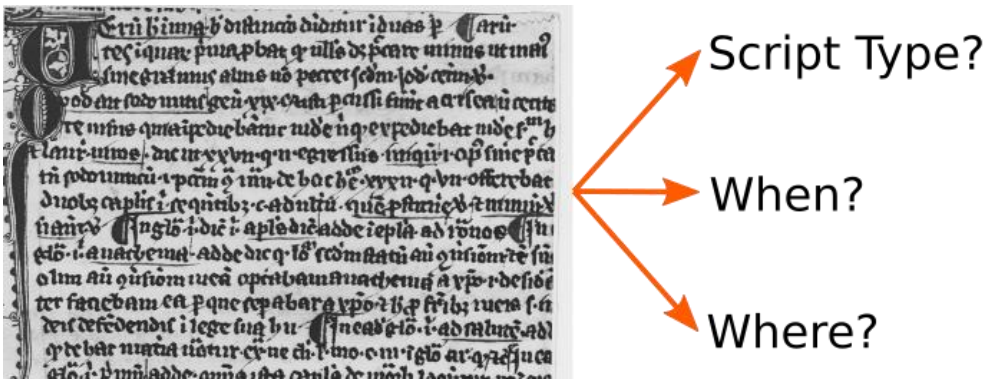



**Figure** 4.7 Document sample from HDC 2021 dataset

### 4.2.8 ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)

The ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI) **[96]** utilized a dataset sourced from the electronic library of the Universitätsbibliothek Basel, comprising 140,000 images under the Public Domain Mark (Figure 4.8). These images encompass various document types, including correspondences, notes, books, drawings, music scores, photographs, envelopes, and technical drawings. The documents are in different languages, predominantly German and French, with instances of Arabic handwriting as well. Filtering of pages was automated, initially checking METS files for author names and dates of birth/death to ensure consistency. Text occurrence was estimated using SIFT features distribution for text region approximation. Following image selection, a test set with five document images per writer and a training set with three document images per writer were defined, resulting in 720 writers contributing to the test set (3600 pages) and 394 writers remaining for the training set (1182 pages). All images were of 300 dpi quality, stored in JPEG or PNG format for binarized images.



**Figure** 4.**8** Document sample from Historical-WI 2017 dataset

### 4.2.9 ICDAR 2019 Competition on Image Retrieval for Historical Handwritten Documents

The ICDAR 2019 Competition on Image Retrieval for Historical Handwritten Documents **[97]** presents a diverse dataset categorized into documents, letters, and charters, drawn from various repositories and collections. The documents section focuses on writers from the European Middle Ages (9th to 15th century CE), with a majority being anonymous (Figure

4.9). The corpus includes consecutive pages within a book to represent a single writer, alongside script samples from known individuals. The letters subset includes digitized letters from Christoph Jacob Trew, obtained from the University Library of Friedrich-Alexander University Erlangen-Nurnberg, and correspondences from the University Library Basel. Charters are sourced from Monasterium.net, emphasizing distinct date information to avoid duplication from the same scribe. Dataset splits include a training/validation set utilizing datasets like "ICDAR17 Historical-WI" and additional resources such as the ICFHR2016 dataset. The testing set comprises 20,000 images, featuring isolated documents and contributions from writers providing one, three, or five pages each.



**Figure** 4.**9** Document sample from ICDAR 2019 dataset

## 4.2.10 ICFHR 2020 Competition on Image Retrieval for Historical Handwritten Fragments

The ICFHR 2020 Competition on Image Retrieval for Historical Handwritten Fragments **[98]** utilized a dataset originating from the publicly available Historical-IR19 test set for the training set, encompassing documents, charters, and letters from various sources (Figure 4.10). The test set consisted of new data, including documents and letters not used in prior competitions. The documents in the dataset were predominantly from the European Middle Ages (9th to 15th century CE) and were often anonymous, with sequences of consecutive pages selected assuming they were written by a single person, validated by a paleographer. Letters were sourced from the University Library Basel, with metadata guiding the selection and discarding of samples based on text quantity and quality. The training set underwent fragmentation of the central image part, ensuring visible text in the fragments, while the test set employed bounding boxes to label the main text area for fragmenting, guaranteeing text presence. The dataset split comprised 101,706 training fragments from 8,717 writers and 20,019 test fragments from 1,152 writers, with no overlap between the training and test sets to maintain distinct writer representation.



**Figure** 4.**10** Document sample from ICFHR 2020 dataset

## 4.2.11 GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification

The GRK-Papyri Dataset **[99]** comprises 50 handwriting samples in Greek on papyrus, originating from the 6th century A.D. These samples are sourced from The Bank of Papyrus Images of Byzantine Aphrodite (BIPAb), curated by Prof. Jean-Luc Fournet and supported by the Association Internationale de Papyrologues and Strasbourg University (Figure 4.11). Each image is accompanied by proper photographic credit and copyright information. The dataset consists of contracts penned by ten different notaries, with each scribe contributing between 4 to 7 samples. Experts have validated the ground-truth of these samples using palaeographic evidence and scribe signatures on the contracts. This dataset is freely accessible for non-commercial research related to writer identification tasks. The selection of images is meticulously done based on their relevance to scribe identification research in Papyrology. Only images with a substantial amount of handwritten text are considered, and they have been cropped to include only the relevant handwritten portions. The dataset is available in two versions: one for the leave-one-out scenario with all 50 images and another for the training-test scenario with 20 images in the training set (two images per scribe for balanced training) and 30 images in the test set with varying sample counts per scribe. Future plans include expanding the dataset's size and incorporating samples from different geographical regions and historical periods.

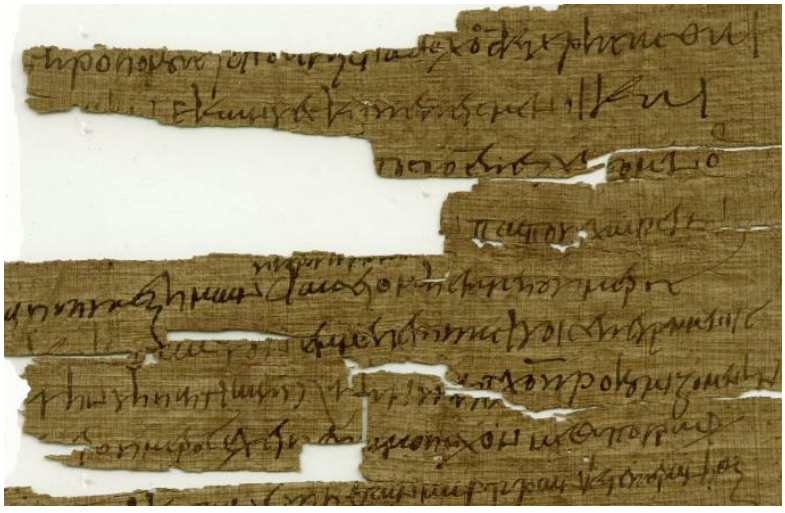


**Figure** 4.**11** Document sample from GRK-Papyri dataset

These datasets collectively provide an extensive pool of historical document data, enabling researchers to delve into and advance the field of historical document classification using

machine learning, deep learning, and other computational methods. In the upcoming section, we will delve into the related methodologies pertinent to each task (writer, date, script) of our historical document classification problem.

## 4.3 Writer identification related works

Over the years, numerous studies have delved into the writer identification problem, utilizing a range of approaches including classical machine learning as well as more recent deep learning techniques.

From the viewpoint of hand-crafted features-based methods, textural measures have been a popular and effective choice for many researchers in capturing writing style. Among them, local binary patterns (LBP) and its various variants have been commonly employed **[100–103]**.

Authors in **[104]** explore the used of the Implicate shape codebook combined with SVM to recognize the document writer. Likewise, a combination of shape and texture-based features is explored in **[105]**.

In another similar work, **[106]**, authors explore the combination of multiple features including histogram of oriented gradients and local binary patterns to improve the identification performance.

In **[107]** a digital paleographic approach was proposed for writer identification from the digitized versions of the Dead Sea Scrolls when the authors investigate a number of standard textural features in the study and compare their effectiveness.

Among other studies, **[108]** employed pathlet and SIFT features for writer identification in historical documents.

In **[109]**, the authors utilized a technique inspired by the functionality of the immune system for recognizing complex patterns. Specifically, they applied the Clonal Selection Classification Algorithm for Arabic writer identification. These features were specifically adapted to capture the shape and structural information in handwriting, including slant, curvature, corners, and junctions, to encode the extracted features, a newly proposed scheme called bagged-vector of locally aggregated descriptors (VLAD) was employed.

Recently, there has been a notable surge in research interest in data-driven feature learning for nearly all pattern recognition tasks, and the same holds true for the writer identification **[110–113]**.

In **[114]**, a deep convolutional neural network (CNN) with SIFT key-points is employed to characterize the writer when CNN features are encoded through multiVLAD and normalized using the L2 norm.

Likewise, authors in **[115]** employed a deep residual CNN to learn effective feature representations using surrogate classes obtained through clustering on training set samples.

The study was extended in **[116]** by introducing a re-ranking method that relies on k- nearest neighbor relationships and significantly improves identification performance on the ICDAR 2107 dataset. Transfer learning on pretrained CNNs has also remained a popular choice for this problem and was employed in **[117]** to identify writers from twelfth century Bible images. Pretrained CNN models have also been evaluated for multiple tasks, including dating and identification of writing styles from historical documents in **[118]**.

In addition to technical contributions, several competitions have also been organized around this problem to assess the performance of existing methods on challenging historical documents. Among these, a recent competition on Image Retrieval for Historical Handwritten Documents **[97]** was organized in conjunction with the international conference on document analysis and recognition "ICDAR 2019". Similarly, The ICDAR2017 Competition on Historical Document Writer Identification **[95]** aimed at evaluating the effectiveness of different feature extraction and machine learning methods for writer identification in historical documents.

These related works demonstrate that the wide range of techniques and approaches devoted for writer identification task are varied in their performance, complexity and resource requirements, and their suitability depends on the particular application and context in which they are employed. In our examination of existing research, we identified a significant gap in the available methodologies. It became evident that studies often fall into one of two categories. On one hand, some rely on traditional manual methods, which, while well established, tend to yield results that lack precision and accuracy, limiting their effectiveness in addressing complex challenges. On the other hand, certain studies turn to cutting-edge machine learning techniques, such as deep learning, which can achieve remarkable results but tend to become excessively specialized and prone to overfitting, particularly when data is limited. Recognizing these limitations, our approach sought to bridge this gap by combining the complementary strengths of both paradigms to create a more robust and versatile solution capable of addressing a broader spectrum of challenges ensuring superior performance compared to exclusive reliance on either.

## 4.4 Dating documents related works

A multitude of research endeavors have risen to confront this challenge of dating historical documents. These approaches encompass a broad spectrum, spanning from conventional machine learning techniques to state-of-the-art deep learning methodologies.

The authors in **[89]** presented a methodology for analyzing the MPS Historical Documents dataset. The adopted approach involved the computation of hinge and fraglet features. It entailed training a global regression model to identify the date's region and a local regression model to estimate the precise date within that region.

A multi-step process has been proposed in **[119],** including encompassed denoising images and detecting edges. Shape context features were extracted from these images, and a codebook was developed using the k-nearest neighbors (KNN) and support vector regression (SVR) algorithms.

Furthermore, in **[120]**, they computed k-contour and stroke fragment features (KCF and KSF); these features were then used to train a codebook that incorporated both of them.

In a subsequent study **[121]**, authors focused on selecting distinctive features such as fork and high-curvature points, which were invariant to scale in handwritten documents. Descriptors for these features were computed, and a model was trained using a Multi-Label Self-Organizing Map (MLSOM).

The same authors in **[122]** constructed a codebook by training a self-organizing time map to capture temporal information from polar stroke features. This codebook was employed with an SVM classifier.

Additionally, **[123]** introduced a novel approach based on the combination of features derived from multiple textures, such as local binary patterns (LBP), Gabor filters, and the gray-level co-occurrence matrix (GLCM).

The methodology described in **[124]** comprised the extraction of features such as GLCM, UV, and IR features from ink pixels; these features were combined and trained using a kernel discriminant learning method for ordinary regression.

The proposed approach in **[125]** entailed image binarization to extract textual and grapheme features using Hinge and CO3. These features were combined and trained using support vector regression.

In **[126]**, document features such as Gabor, HOG, and ResNet-18 were extracted and subsequently fused using the sparse fusion method. The resulting set of features was used for training a multiclass classifier.

The mentioned studies effectively employed traditional hand-crafted techniques to extract pertinent features from historical documents; however, despite their efforts, achieving high accuracy remained challenging. This challenge prompted an exploration of deep learning methods.

As highlighted in **[127]**, where a deep learning approach was explored, the initial step involved feature extraction using the pre-trained GoogLeNet architecture from cropped patches. These extracted features were subsequently utilized in the training of both support vector regression (SVR) and Gaussian process (GP) models.

Additionally, in **[94]**, the training dataset was used to crop patches of size 256 × 256 with a stride of 42 × 42. These patches were employed to train five distinct convolutional neural

networks (CNNs), each comprising 50 layers. The final prediction was reached by averaging the values obtained from these CNNs and patches.

Another approach in **[94]** involved the processing of the dataset by extracting features using the RootSIFT descriptor; these features were encoded using the VLAD encoding method and subsequently employed to train an SVM classifier.

In **[128]**, patches of a similar size were randomly extracted from the dataset and employed for model training using transfer learning techniques.

Furthermore, **[129]** focused on training a CNN specifically on text lines extracted from the training dataset; the document-level output was determined as the median of the text-line-level predictions.

In **[130]**, feature extraction using transfer learning was combined with training on Gaussian processes. Also **[131]** Try different features to determine the effect of data augmentation on dating historical documents.

After a comprehensive review of relevant literature, it becomes evident that diverse methodologies aimed at tackling the challenge of dating historical documents have emerged, predominantly relying on image-based approaches. The former category encompasses traditional techniques, which, while valuable, tend to yield relatively lower accuracy. Conversely, methods heavily reliant on CNNs achieve superior accuracy but are frequently plagued by overfitting issues. Despite the notable achievements of existing methodologies in the realm of historical document dating, a clear imperative for a more robust and high-performing system persists.

## 4.5 Script type recognition related works

Notable advancements in historical document script classification have emerged from recent competitions such as ICFHR 2016 **[93]**, ICDAR 2017 **[94]**, and the recent ICDAR 2021 **[95]**. These competitions have served as arenas for researchers to demonstrate their innovative methodologies and techniques for classifying scripts in historical documents, predominantly using the ClaMM dataset.

ICFHR 2016 **[93]** was dedicated to the classification of ancient handwritten documents, attracting a diverse range of techniques and approaches from participants. Traditional feature extraction methods and machine learning algorithms were employed by several participants, such as FRDC-OCR, NNML, and DeepScript.

The FRDC-OCR system utilizes convolutional neural networks (CNN) to classify scripts. It operates by extracting fixed-size patches from the input image, ensuring comprehensive coverage by evenly selecting patches. Each patch is classified using a trained CNN, providing a feature vector and recognition results with confidence. The system then calculates the average feature vector and average confidences of all patches, which serve as the final results for the entire image, offering an overall representation of script classification.

The NNML system utilizes Deep CNNs with residual learning and batch normalization. Training involves sub-windows of size 256x256 extracted from the provided images, with the CNNs trained to classify sub-windows of size 227x227 into twelve script classes. During testing, the large document image is divided into overlapping sub-windows of size 227x227 for classification. The training protocol incorporates stochastic gradient descent optimization with various parameters and on-the-fly preprocessing techniques.

The DeepScript system adopts a modified VGG-based neural network architecture for script classification. It comprises stacked convolutional layers with increasing filter sizes, followed by fully-connected dense layers and a softmax layer for predictions. Training involves stochastic gradient descent with dropout regularization and ReLU activation. The models were trained for thirty epochs with a decreasing learning rate, and the weights were saved after each epoch.

ICDAR 2017 **[94]** introduced a script identification competition focused on historical and ancient scripts, where participants employed deep learning models, particularly convolutional neural networks (CNNs), for script classification. Several participants incorporated data augmentation techniques to enhance their model's performance. Notable methods include the T-DeepCNN System and CK2.

The T-DeepCNN System utilizes Deep CNNs with residual learning and batch normalization for script classification. Training involves downsampling images, extracting subwindows, and applying preprocessing steps such as random cropping and linear scaling. Stochastic transformations and data augmentation techniques are used during training to improve system performance. The network architecture is based on a 50-layer model with reduced convolution filters. The training protocol employs stochastic Gradient Descent Optimization with specific parameters.

On the other hand, CK2 is an approach for script classification that draws inspiration from successful writer identification methods. It utilizes RootSIFT descriptors and a bag-of-visualwords approach. Descriptors are encoded into global descriptors using VLAD encoding. Multiple background models are employed to compute VLAD descriptors, and PCA whitening is applied for decorrelation. Linear SVMs are used for the final classification step.

In the recent ICDAR 2021 competition **[95]**, significant progress was made in the field of script classification in historical documents. Researchers employed cutting-edge techniques, particularly deep learning architectures. Notable participants include PERO 2 and The North LTU. PERO 2 utilized a CNN for text line classification, employing global average pooling to handle varying line lengths. The model was trained using the Lmin loss function and incorporated multiple checkpoints for improved performance. Page-level outputs were obtained by calculating the mean vector of text lines exceeding 64 pixels. For date tasks, the page-level output was derived from the median of text line-level outputs for lines longer than 128 pixels. On the other hand, The North LTU adopted a method that involved cropping and

downsampling the images. The datasets were merged and split into training and validation sets.

## 4.6 Analysis of existing methodologies

Recent studies in the field have unveiled diverse methods, ranging from traditional machine learning to advanced deep learning techniques. While traditional methods have been foundational in the field and have provided valuable insights into historical document classification, they also have certain limitations. One of the primary challenges is the manual effort required in feature engineering, which can be time-consuming and may not capture all the nuanced patterns present in the data. Additionally, traditional methods may struggle with scalability and adaptability to different types of documents or writing styles, as they rely heavily on predefined features and rules. Deep learning methods also pose challenges, such as the requirement for substantial computational resources and data annotation efforts. Additionally, the interpretability of deep learning models in the context of historical document analysis remains an ongoing research area. Nevertheless, the potential of deep learning in revolutionizing historical document analysis is vast. Remarkably, to the best of our knowledge, no prior study has ventured into the exploration of the potential of Vision Transformers (ViT) within this domain, marking a significant knowledge gap. We posit that this study not only aims to fill this critical gap but also endeavors to unlock new horizons for researchers, beckoning them to delve into the realm of Vision Transformers for historical document classification.

These key challenges in recent studies can be summarized in those main points:

- **Bad documents quality:** Many historical documents suffer from poor image quality due to degradation over time. Our system incorporates advanced image preprocessing techniques to enhance the quality of these documents, making it easier to extract meaningful features for classification.

- **Complex feature-extraction methods:** Previous methods often relied on complex and time-consuming feature extraction processes, leading to inefficiencies. In contrast, our system employs efficient feature extraction algorithms that can accurately capture the distinctive characteristics of historical documents without excessive computational overhead.

- **Lack of enough samples:** The availability of sufficient training samples is crucial for developing accurate classification models. To overcome this challenge, we have utilized very efficient data augmentation techniques and leveraged datasets from multiple sources to ensure a diverse and representative training set for our system.

- **Inaccurate classification algorithms:** Some previous studies may have encountered issues with inaccurate classification results due to limitations in the algorithms used. In response, we have implemented state-of-the-art classification algorithms with

improved accuracy and robustness, ensuring reliable classification outcomes for historical documents.

## 4.7 Conclusion

This chapter provides an insightful overview of the current methods and approaches documented in the literature, which are relevant to our research. By reviewing these state-of-the-art studies, we have gained a comprehensive understanding of the existing landscape, establishing a solid foundation for our own contribution. In the subsequent chapter, we will delve into the specifics of our novel approach for the classification of historical documents. Building upon the insights gleaned from existing methodologies, our novel approach aims to address key challenges and limitations while pushing the boundaries of accuracy, efficiency, and applicability in this domain. Through a detailed exploration of our methodology, including the underlying techniques, algorithms, and experimental results, we aim to demonstrate the effectiveness and significance of our contribution in advancing the field of historical document classification.

# Chapter 5

# Contributions

## 5.1 Introduction

We have paved the way for our contribution to the field of historical documents classification through a comprehensive understanding of the material under study, identification of key challenges, and an extensive review of previous studies. In this chapter, we unveil our methodology designed to address the classification problem, encompassing writer identification, dating, and script classification. Our approach is rooted in computer-based methods, specifically image-based techniques leveraging deep learning. We will commence by providing an overview of our system, outlining the major stages it comprises. Subsequently, we will delve into each stage in detail, covering preprocessing techniques, features-extraction and classification algorithms. And finally, we will discuss each of our contribution in each task.

## 5.2 System Overview

Our system for the classification of historical documents is rooted in the field of paleography, which focuses on leveraging writing styles to differentiate between documents. Paleographical experts have highlighted the significance of writing styles in determining various attributes such as date, writer, and script. Our method aims to automate the classification task using paleographical analysis, building upon the existing studies discussed in previous chapters. We conducted a thorough analysis of existing methods to identify gaps and opportunities where we could contribute to the field. Our goal was to design a robust system that enhances the efficiency and accuracy of classification through innovative paleographical techniques. Therefore, we have developed our system to address aforementioned challenges that researchers in the field have faced.

Our system is designed to tackle each of these key challenges through a structured approach comprising several critical steps:

- **Preprocessing:** The first step in our methodology involves preprocessing the historical documents. We employ various denoising techniques to improve the overall quality of the images. We aim to facilitate better feature detection and extraction in subsequent stages.

- **Feature extraction:** Once the documents are preprocessed, we focus on extracting meaningful features that are relevant to our classification goals. We utilize a specialized feature extraction method tailored specifically for historical document

analysis. This method incorporates techniques to capture distinctive patterns and characteristics in the documents.

- **Classification:** With a high-quality dataset and selected samples in hand, we employ advanced classification algorithms to classify the historical documents accurately. Our system utilizes state-of-the-art classifiers to achieve the highest classification scores possible. We aim to improve the accuracy and reliability of the classification outcomes.

Our systematic approach encompasses **preprocessing for image enhancement**, **feature extraction for capturing relevant patterns**, and **robust classification techniques for accurate categorization** of historical documents. Through these steps, our system strives to provide an effective and reliable solution that contributes significantly to this problem. Figure 5.1 is an overview of our system components.

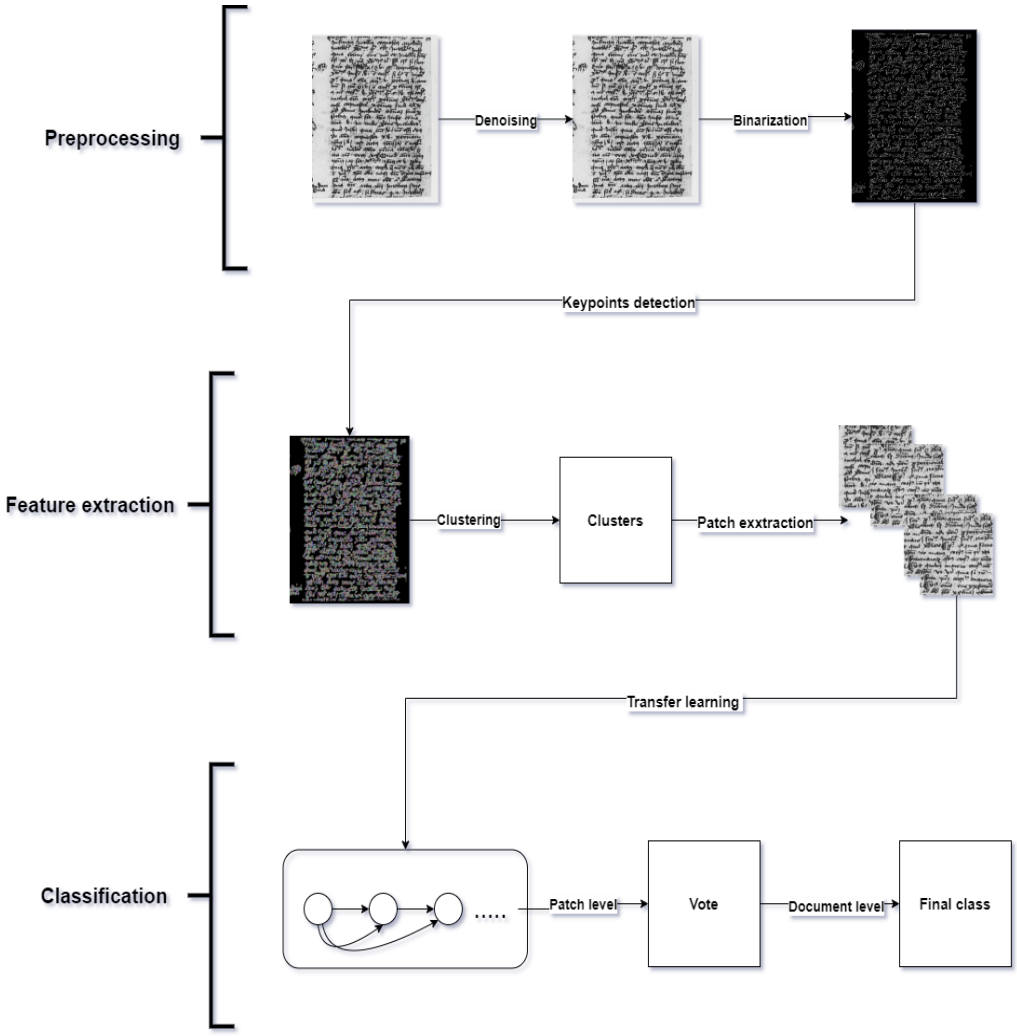



**Figure 5.1** Historical documents classification system design

The first step in our system is preprocessing the data to enhance its quality for subsequent stages. We utilize both denoising and binarization techniques to prepare the image for the

detection of text regions. Next, we use a keypoints detector to extract the coordinates of each keypoint from the binarized image, which represent writing features. The keypoints detector we employ is based on shapes and contours, which are often associated with text. These coordinates are then used to identify regions on the original image, from which small windows around each keypoint are extracted. We avoid using the binarized or denoised images for patch extraction because the original image contains valuable features due to the historical significance of preserving the originality of these writing materials. The size of the patches is chosen based on experiments on different patch sizes and also to the model that will be used for classification, as it depends on the model's input size. The number of patches extracted is also determined by experiments to prevent data redundancy, which can lead to overfitting problems. This step helps convert the dataset into a high-quality and high-quantity dataset, essential for classification, especially when using deep learning models.

For the classification stage, we leverage deep learning models. Using pre-trained models, we classify the patches extracted from the original images. Finally, and to determine the class of a document, we use a majority vote technique to ensures that the final classification decision takes into account the most frequently occurring class among the patches, leading to a more robust and accurate classification outcome.

Following this systematic approach, our system provides a reliable and effective solution for the classification of historical documents, addressing the challenges faced by previous researchers and contributing to advancements in the field. Our choice of techniques at every stage is grounded in extensive experimentation with a multitude of methods and techniques. We conducted these experiments to identify the most effective approaches for each stage of the process, ensuring that our system is optimized for performance and accuracy as we will detail in next sections.

## 5.3 Preprocessing

The examination of our datasets **[89,94,95]** revealed some troubles which may hamper their immediate use. Figure 5.2 showcases some of these problems, including traces of wear and tear on the text layers demonstrating - both inked up as well as clear hindrance. It is important to pre-process these issues, which in turn helps achieve good dataset and leads to better performance of classification algorithms.

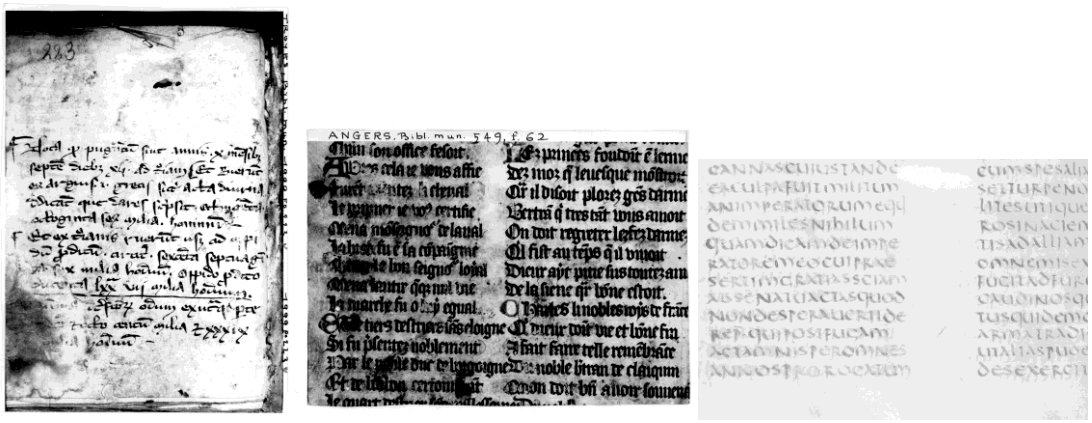
Below, we detail the common problems encountered and the preprocessing techniques used to mitigate them:

- **Wear and Tear:** Historical documents often exhibit physical degradation such as tears, holes, and faded areas, which can obscure text and introduce noise.

- **Ink Traces:** Over time, ink can bleed through pages or leave ghost traces, making it difficult to distinguish the actual text from background noise.

- **Unclear Handwriting:** Handwriting quality varies, with some scripts being poorly legible due to inconsistent pressure, slant, or historical ink quality.

These preprocessing techniques allow us to systematically resolve the issues so that we will be able to appropriately use the methods for building historical document datasets with more features. Respectively, in the next section we will delve into some preprocessing methodologies applied to our system.

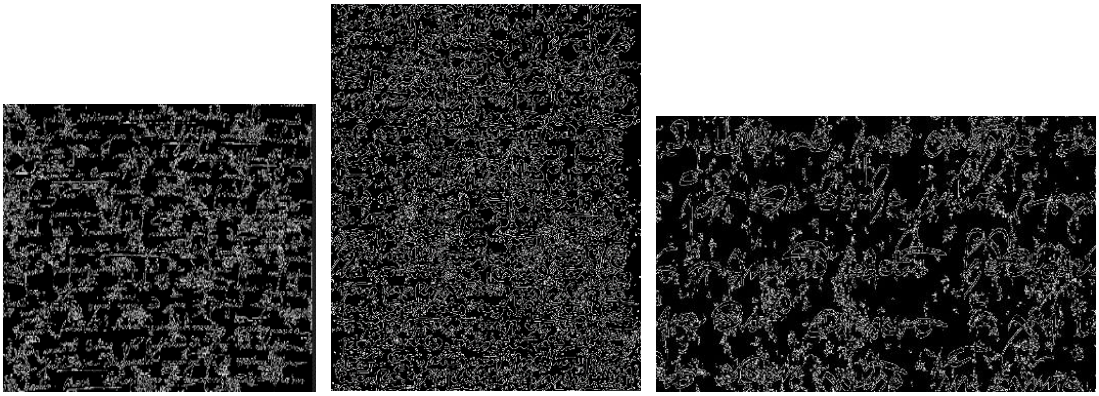


(a) (b) (c)

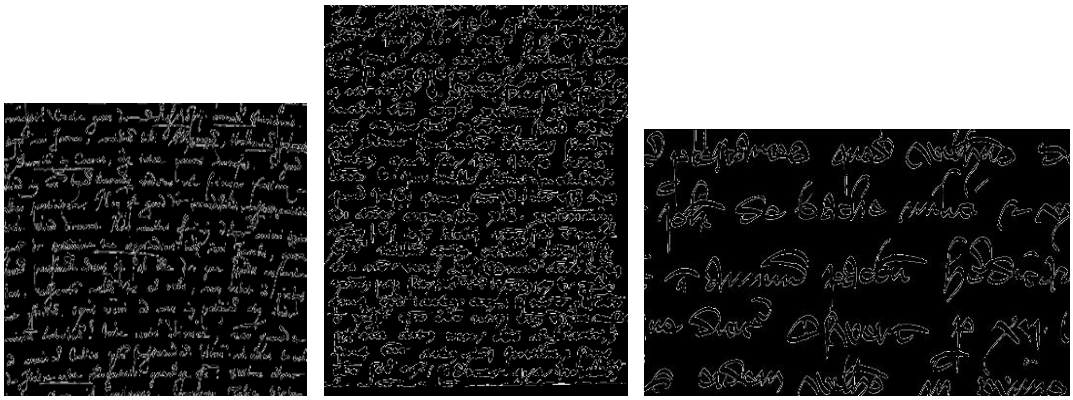
**Figure 5.2** problems with ClaMM samples: (a) worn paper, (b) ink traces, (c) indistinct text.

## 5.3.1 Documents denoising

Historical documents often suffer from significant noise, which poses challenges for subsequent processing tasks like binarization and feature detection (Figure 5.3). In our preprocessing step, one crucial technique used is denoising, aimed at enhancing the quality of the images and effectively removing noise while preserving important features making it better for the detection of regions of interest (Roi). The output of the denoising process is then used as input for binarization algorithms. We have implemented a variety of well-known denoising techniques from the literature, such as Non-Local Means **[132]**, Bilateral Filtering **[133]**, and Total Variation **[134]**. Our choice of denoising technique is based on the output of different techniques and the performance of subsequent binarization on each denoising technique.



(a) binarization on original images.



(b) binarization on denoised images.

**Figure 5.3:** Influence of denoising on the binarization process.

#### 5.3.1.1 Non local means

Non-Local Means (NLM) **[132]** is an advanced denoising algorithm that improves image quality by averaging pixel values based on the similarity of entire patches rather than just neighboring pixels. This method effectively preserves fine details and textures while removing noise, making it highly suitable for applications where the retention of intricate features is crucial. NLM is particularly beneficial in several areas of image processing. In the context of historical document classification, the quality of the images is paramount. NLM is particularly useful here as it:

- **Preserves Textural Details:** Historical documents often contain intricate handwriting styles and detailed decorations, which need to be preserved for accurate classification and analysis.

- **Reduces Noise:** By effectively removing noise, NLM enhances the readability of the text, making subsequent processes such as binarization and feature extraction more reliable.

- **Maintains Structural Integrity:** The algorithm's ability to retain the structural elements of the text ensures that the classification algorithms have a clear and accurate representation of the document's content.

The mathematical foundation of the Non-Local Means algorithm is based on the principle of weighted averaging, where the weight is determined by the similarity between image patches. The NLM filter is defined in Equation (1) as:

$$S\big(P(x), P(y)\big) = e^{-\frac{\|P(x)-P(y)\|^2}{2a^2}} \qquad (1)$$

Where:

- $\boldsymbol{S\big(P(x), P(y)\big)}$: represents the similarity between two patches $P(x)$ $and$ $P(y)$.
- $\boldsymbol{a}$ is a parameter that controls the degree of similarity between patches. A smaller $a$ makes the similarity measure more sensitive to differences between patches, while a larger a makes it less sensitive.

$$I_d(x) = \frac{1}{C(x)} \sum_y S\big(P(x), P(y)\big).I(y) \qquad (2)$$

Where:

- $\boldsymbol{I_d(x)}$ represents the denoised pixel intensity at position $(x, y)$.
- $\boldsymbol{C(x)}$ is a normalization factor that ensures the weights sum up to 1 for each pixel.
- $\sum_y$ represents the summation over all pixels $y$ in the image.
- $\boldsymbol{I(y)}$ is the intensity of the pixel at position $y$.

The algorithm works as follows:

- **Patch Comparison:** For each pixel x, the algorithm compares a patch centered around x with patches centered around all pixels y within a larger search window.

- **Weight Calculation:** The similarity weight between the patches centered at x and y is computed. This weight decreases exponentially with the squared Euclidean distance of the intensity vectors of the patches.

- **Weighted Averaging:** The denoised pixel value $I_d(x)$ is obtained by taking a weighted average of all pixel values within the search window, with the weights determined by the patch similarities.

#### 5.3.1.2 Bilateral Filtering

Bilateral Filtering **[133]** is a non-linear, edge-preserving, and noise-reducing smoothing filter for images. Unlike traditional filters that smooth images uniformly, Bilateral Filtering performs smoothing while preserving edges by considering both the spatial proximity and the intensity difference between pixels. This makes it particularly effective in applications where

edge preservation is crucial. Bilateral Filtering is widely used in various image processing applications.

Bilateral Filtering combines two Gaussian filters: one in the spatial domain and the other in the intensity domain. The filter is defined in Equation (3):

$$I(x) = \frac{1}{W_p} \sum_{y \in \Omega} I(y).\, f_r\big(I(y) - I(x)\big).\, f_s(y - x) \qquad (3)$$

Where :

- $I(x)$ is the filtered intensity value at pixel x.
- $I(y)$ is the intensity value at pixel y in the original image.
- $\Omega$ is the neighbourhood of pixel x.
- $f_r\big(I(y) - I(x)\big)$ is the range kernel that measures the photometric similarity between pixels x and y (Equation(4)):

$$f_r\big(I(y) - I(x)\big) = \exp\left(-\frac{\big(I(y) - I(x)\big)^2}{2\sigma_r^2}\right) \qquad (4)$$

$f_s(y - x)$ is the spatial kernel that measures the geometric closeness between pixels x and y (Equation (5)):

$$f_s(y - x) = \exp\left(-\frac{\|y - x\|^2}{2\sigma_s^2}\right) \qquad (5)$$

$\sigma_r \; and \; \sigma_s$ are the parameters controlling the extent of the range and spatial kernels, respectively.

$W_p$ is the normalization factor defined in Equation (6).

$$W_p = \sum_{y \in \Omega} f_r\big(I(y) - I(x)\big).\, f_s(y - x) \qquad (6)$$

The algorithm works as follows:

- **Spatial Filtering:** Weights are assigned based on the spatial proximity of the pixels. Pixels closer to the target pixel are given higher weights.

- **Range Filtering:** Weights are also assigned based on the similarity of the pixel intensities. Pixels with similar intensities to the target pixel are given higher weights.

- **Weighted Averaging:** The denoised pixel value $I(x)$ is computed by taking a weighted average of all pixel values in the neighborhood, with weights determined by both spatial and range similarities.

#### 5.3.1.3 Total Variation Denoising

Total Variation (TV) Denoising **[134]** operates on the principle of minimizing the total variation of the image, which is a measure of the intensity variation between adjacent pixels. It works by solving an optimization problem that minimizes the total variation of the image while keeping it close to the observed noisy image. The total variation of an image $u$ is given in Equation (7):

$$TV(u) = \int \|\nabla u\| dx \qquad (7)$$

Where $\|\nabla u\|$ is the gradient magnitude of the image, representing the variation in intensity. The denoising process is formulated as an optimization problem defined in Equation (8):

$$min_u \left( \frac{1}{2} \|u - f\|_2^2 + \lambda TV(u) \right) \qquad (8)$$

Where:

- $u$ is the denoised image.
- $f$ is the observed noisy image.
- $\|u - f\|_2^2$ is the fidelity term that ensures the denoised image remains close to the original image.
- $\lambda$ is a regularization parameter that balances the trade-off between noise removal and detail preservation.

## 5.3.2 Documents binarization

Binarization is a crucial image processing technique that converts a grayscale image into a binary image, where the pixels are assigned one of two possible values: black or white. This process is fundamental because:

- **Text Isolation:** It effectively isolates the text from the often noisy and deteriorated background, making the text more readable.

- **Feature Preservation:** Binarization helps in preserving the essential features of the text, which are critical for paleographic analysis.

- **Facilitating Further Processing:** By simplifying the image, binarization facilitates keypoint extraction.

The binarized image is used to identify and extract regions of interest (ROI) from the document. However, it is important to note that while binarized images are instrumental in locating keypoints of text regions, they are not used for the actual processing. The original samples, which contain rich historical features, are retained for classification purposes to ensure that valuable information is not lost.

In our approach, we have experimented with various binarization methods to determine the most effective technique for our dataset. The selection of the optimal method was based on rigorous experimental evaluation.

#### 5.3.2.1 Otsu's Binarization Method

Otsu's method **[135]** is a widely-used global thresholding technique for image binarization. It was introduced by Nobuyuki Otsu in 1979 and is designed to automatically determine the optimal threshold value that separates the foreground (text) from the background in a grayscale image. This method is particularly effective for images where the pixel intensity distribution has two distinct peaks corresponding to the foreground and background.

Our objective is to isolate text regions from the often noisy and deteriorated backgrounds of documents. Otsu's method offers a systematic approach to determine the threshold value without manual intervention, making it suitable for processing large datasets of historical documents.

Otsu's method operates by maximizing the between-class variance, which is the variance between the foreground and background pixel intensities:

- **Histogram Calculation:** Compute the histogram of the grayscale image, which represents the frequency of each pixel intensity level.
- **Probability Distribution:** Calculate the probability of each intensity level in the histogram.

- **Class Probabilities and Means:**
  - **Compute** the cumulative sum of probabilities for the background (class 1) and foreground (class 2) for each possible threshold.
  - **Calculate** the cumulative mean for both the background and foreground for each threshold.
- **Between-Class Variance:** For each threshold, compute the between-class variance, which is the product of the class probabilities and the squared difference between the class means.
- **Optimal Threshold Selection:** The threshold that maximizes the between-class variance is selected as the optimal threshold.

Mathematically, the between-class variance $\sigma_b^2$ is defined in Equation (9):

$$\sigma_b^2(t) = w_1(t)w_2(t)[u_1(t) - u_2(t)]^2 \qquad (9)$$

Where:

- $t$ is the threshold
- $w_1(t)\ and\ w_2(t)$ are the probabilities of the background and foreground classes, respectively.

- $u_1(t) - u_2(t)$ are the means of the background and foreground classes, respectively.

**Advantages**:

- **Automatic Threshold Selection:** Otsu's method does not require manual input for the threshold value, making it highly automated and suitable for large datasets.
- **Simplicity and Efficiency:** The method is straightforward to implement and computationally efficient, which is beneficial for processing numerous images.
- **Effectiveness for Bimodal Histograms:** It performs exceptionally well for images with a bimodal intensity distribution, which is common in many document images.

**Drawbacks:**

- **Sensitivity to Noise:** Otsu's method can be sensitive to noise and variations in the background, which may result in suboptimal thresholding for images with complex or uneven backgrounds.
- **Assumption of Bimodal Distribution:** The method assumes that the histogram of the image is bimodal. For images with more complex intensity distributions, Otsu's method may not perform as well.
- **Global Thresholding Limitation:** As a global thresholding technique, it may not handle local variations in lighting or background texture effectively, which are often present in historical documents.

#### 5.3.2.2 Canny Edge Detection

Canny edge detection **[136]** is a popular and powerful edge detection algorithm developed by John F. Canny in 1986. It is widely used in image processing to identify sharp changes in intensity, which correspond to edges in an image. Edges represent significant structural features and can be used for various image analysis tasks, including object recognition, image segmentation, and, in our case, text region extraction in historical documents. In the context of historical document analysis, identifying the edges of text characters is crucial for accurate binarization and subsequent feature extraction.

Historical documents often contain noise, varying background textures, and degradation, making edge detection a challenging task. Canny edge detection helps in precisely identifying the boundaries of text regions, even in the presence of such challenges. This makes it particularly useful for isolating text from background noise and improving the accuracy of keypoint detection and feature extraction in paleographic analysis. Canny edge detection involves several steps to accurately detect edges while minimizing the impact of noise:

- **Noise Reduction:** Apply a Gaussian filter to the image to smooth it and reduce noise. The Gaussian filter helps in minimizing the impact of small variations in the image that might otherwise be mistaken for edges.
- **Gradient Calculation:** Compute the intensity gradients of the image using finite difference approximations. This involves calculating the gradient magnitude and direction at each pixel. Typically, Sobel operators are used to approximate the derivatives.
- **Non-Maximum Suppression:** Thin out the edges by suppressing non-maximum gradient values. This step involves retaining only the local maxima of the gradient magnitude, which represent potential edges, and setting other values to zero. This results in a set of thin edges.
- **Double Thresholding:** Apply two thresholds to the gradient magnitude: a high threshold and a low threshold. Pixels with gradient magnitudes above the high threshold are considered strong edges, while those between the low and high thresholds are considered weak edges.
- **Edge Tracking by Hysteresis:** Track edges by connecting strong edges and weak edges. Weak edges that are connected to strong edges are preserved, while isolated weak edges are discarded. This step ensures that the detected edges are continuous and connected, resulting in more accurate edge maps.

Mathematically, the edge detection process can be described in Equations (10) and (11):

$$M(x,y) = \sqrt{Gx^2 + Gy^2} \qquad (10)$$

$$\theta(x,y) = \arctan\left(\frac{Gx}{Gy}\right) \qquad (11)$$

Where:

- $Gx$ represents the gradient in the horizontal direction (e.g., the Sobel operator applied in the x-direction).
- $Gy$ represents the gradient in the vertical direction (e.g., the Sobel operator applied in the y-direction).

**Advantages:**

- **Robustness to Noise:** The initial Gaussian smoothing step reduces the impact of noise, making the algorithm robust in noisy environments.

- **Accurate Edge Localization:** By considering both gradient magnitude and direction, Canny edge detection provides precise localization of edges.
- **Thinning of Edges:** The non-maximum suppression step results in thin edges, which are more useful for subsequent processing steps like feature extraction.

**Drawbacks:**

- **Parameter Sensitivity:** The algorithm's performance depends on the choice of parameters, such as the Gaussian filter size and the thresholds for gradient magnitudes. Incorrect parameter settings can lead to either missed edges or excessive noise.
- **Computational Complexity:** The multiple steps involved in the process can be computationally intensive, especially for large images.
- **Assumption of Continuous Edges:** The algorithm assumes that edges are continuous and connected, which might not always be the case in degraded or highly textured historical documents.

## 5.4 Feature extraction

Feature extraction is a critical step in the process of classifying historical documents. After binarizing the documents, we can utilize them to detect text regions within the documents. Handwritten texts exhibit various shapes and curves, which can differ significantly from one document to another. These variations are essential features for classification. To detect these regions, we use the binarized image as input for a feature detector algorithm. By extracting keypoints and then using them to extract patches from the original documents, we can enhance the quality of the dataset. This process involves identifying regions of interest and augmenting the data by extracting multiple patches from a single document. We employed different feature detection methods and made our choice based on the number of features detected and the speed of each method.

### 5.4.1 FAST (Features from Accelerated Segment Test)

FAST (Features from Accelerated Segment Test) **[137]** is a high-speed corner detection algorithm used to identify keypoints within an image. It is designed for real-time applications, where computational efficiency is crucial. The FAST algorithm excels in quickly detecting corners by examining a circle of pixels around a candidate pixel and determining if a significant number of contiguous pixels are brighter or darker than the candidate pixel. In the context of historical document analysis, FAST is particularly useful for detecting text regions efficiently. Historical documents often contain a vast amount of handwritten text, and quickly identifying keypoints can significantly streamline the feature

extraction process. The FAST algorithm operates based on a circle of 16 pixels around a candidate pixel, *p*. The mathematical process can be broken down into the following steps:

- **Pixel Comparison:** For each candidate pixel, a circle of 16 pixels around it is considered.
    - If at least 9 contiguous pixels in this circle are all brighter or darker than the candidate pixel, it is considered a keypoint.
- **Thresholding:** A threshold is applied to determine whether the difference in intensity between the candidate pixel and the surrounding pixels is significant enough to be considered a feature.
- **Non-Maximum Suppression:** Keypoints are refined by suppressing non-maximum points to ensure only the strongest keypoints are retained.

**Advantages:**

- **High Speed and Efficiency:** FAST is one of the quickest corner detection algorithms, making it ideal for real-time applications and large datasets.
- **Simplicity:** The algorithm is straightforward to implement and understand, which facilitates quick integration into various image processing pipelines.
- **High Repeatability:** FAST has a high degree of repeatability, meaning it consistently detects the same corners under similar conditions.

**Drawbacks:**

- **Lack of Scale and Rotation Invariance:** FAST is not inherently invariant to changes in scale or rotation, which can limit its effectiveness in certain applications where the scale and orientation of the text vary significantly.
- **Sensitivity to Noise:** The algorithm can be sensitive to noise in the image, potentially leading to the detection of false corners in noisy regions.
- **Limited Descriptive Power:** While FAST is excellent at detecting corners quickly, it does not provide a detailed description of the detected keypoints, unlike other methods such as SIFT.

### 5.4.2 SIFT (Scale-Invariant Feature Transform)

SIFT (Scale-Invariant Feature Transform) **[138]** is a highly robust feature detection and description algorithm widely used in computer vision. Developed by David Lowe in 1999, SIFT detects keypoints in an image and generates descriptors that are invariant to scale, rotation, and, to some extent, changes in illumination. This makes SIFT particularly effective

for identifying and matching features across different images. Historical documents can vary greatly in scale, orientation, and lighting due to differences in digitization processes or physical conditions of the documents. SIFT's ability to consistently identify and describe features under these varying conditions allows for reliable detection of text regions and extraction of relevant patches. These features can then be used to train models for classification, ensuring that the important historical and paleographic characteristics are preserved and utilized effectively.

- **Scale-Space Extrema Detection:** SIFT detects keypoints by constructing a scale-space representation of the image.
- **Keypoint Localization:** Keypoints are localized by identifying extrema in the DoG images. Each pixel is compared to its 26 neighbors in a 3x3 region at the current and adjacent scales. Pixels that are local maxima or minima are selected as keypoint candidates.
- **Orientation Assignment:** Each keypoint is assigned a consistent orientation based on the local image gradient directions. The orientation is determined by computing the gradient magnitudes and orientations in a neighborhood around the keypoint and forming a histogram of gradient orientations.
- **Keypoint Descriptor:** A descriptor is created for each keypoint by sampling the local image gradients around the keypoint and constructing a histogram of gradient orientations. The gradients are weighted by a Gaussian window and aggregated into a 128-dimensional vector, providing a distinctive descriptor for each keypoint.

**Advantages**

- **Scale and Rotation Invariance:** SIFT is designed to be invariant to scale and rotation, making it highly reliable for matching features across images with different scales and orientations.
- **Robust Descriptors:** The 128-dimensional descriptors are highly distinctive and robust to changes in illumination and minor geometric distortions.
- **Widely Adopted:** SIFT has been extensively used and validated in various applications, ensuring its reliability and effectiveness.

**Drawbacks**

- **Computational Complexity:** SIFT is computationally intensive, which can be a drawback for real-time applications or large datasets.

- **Patent Restrictions:** Historically, SIFT was patented, which restricted its use in some commercial applications. However, the patent has since expired.
- **Sensitivity to Noise:** While robust to many changes, SIFT can still be sensitive to significant noise, which may affect keypoint detection and descriptor accuracy.

### 5.4.3 Harris Corner Detection

Harris Corner Detection **[139]** is a well-established algorithm for detecting corners within an image, developed by Chris Harris and Mike Stephens in 1988. Unlike edge detectors, which identify changes in intensity along a single direction, the Harris algorithm identifies points where the intensity changes in multiple directions, making it effective for detecting corners and other interest points. For historical document classification, detecting stable and reliable keypoints is crucial. The Harris Corner Detector is effective at identifying these points, which can be used to extract meaningful features from binarized images. These features are important for distinguishing between different handwriting styles and enhancing the robustness of the classification process.

- **Gradient Computation:** The algorithm begins by computing the image gradients in the x and y directions using Sobel operators in Equation (12):

$$I_x = \frac{\partial I}{\partial x} \ , I_y = \frac{\partial I}{\partial y} \qquad (12)$$

- **Structure Tensor:** The gradients are used to compute the structure tensor (also known as the second moment matrix) for each pixel as shown in Equation (13):

$$M = \sum_{x,y} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \qquad (13)$$

This matrix captures the intensity changes in the neighborhood of each pixel.

- **Corner Response Function:** The corner response function $R$ is computed from the eigenvalues of the structure tensor. Instead of computing the eigenvalues directly, Harris proposed a simpler measure (Equation (14)):

$$R = det(M) - k(trace(M)) \qquad (14)$$

Where:

  - det(M) is the determinant of the structure tensor M, which captures the product of the eigenvalues of M. It quantifies the local change in intensity in two perpendicular directions.
  - trace(M) is the trace of M, which is the sum of its diagonal elements. It quantifies the overall intensity change.

- k is a constant parameter that adjusts the sensitivity to corners.

- **Non-Maximum Suppression:** The final step involves applying non-maximum suppression to the corner response function to identify the strongest corners. Local maxima in the corner response map are selected as keypoints.

**Advantages**

- **Robustness:** Harris corners are stable and repeatable, making them reliable for matching features between images.
- **Efficiency:** The algorithm is computationally efficient, suitable for applications requiring quick keypoint detection.
- **Effective for Corners:** Harris is particularly effective at detecting corners, which are often abundant in handwritten text regions.

**Drawbacks**

- **Lack of Scale Invariance:** Harris is not inherently scale-invariant, which can be a limitation when dealing with images of varying resolutions.
- **Fixed Orientation:** The algorithm does not provide orientation information for the detected keypoints, which can limit its effectiveness in some applications.
- **Sensitivity to Noise:** Like many feature detectors, Harris can be affected by noise, leading to false corner detections.

### 5.4.4 Clustering

After extracting keypoints from historical documents, the sheer volume of detected keypoints necessitates a method to group them into meaningful clusters. Clustering helps in organizing the keypoints, representing different areas of the document, and facilitating the extraction of significant patches. We use the k-means algorithm for this purpose, which is well-suited for partitioning keypoints into clusters.

Clustering keypoints is crucial for several reasons:

- **Manageability:** With potentially thousands of keypoints detected per document, clustering reduces the complexity by grouping similar keypoints together.
- **Relevance:** Clustering helps in identifying and focusing on the most relevant regions of the document, which are likely to contain significant features for classification.

- **Efficiency:** By working with clusters rather than individual keypoints, the subsequent steps in the processing pipeline become more efficient and effective.

How Clustering Works

- **Keypoint Extraction:** Detect keypoints from the document using feature detectors such as FAST, SIFT, or Harris.
- **Cluster Formation Using K-means**: Apply the k-means algorithm to group the keypoints into $k$ clusters.

The k-means clustering algorithm aims to partition $n$ keypoints into $k$ clusters, where each keypoint belongs to the cluster with the nearest mean. The algorithm operates as follows:

- **Initialization:** Select $k$ initial cluster centroids randomly.
- **Assignment Step:** Assign each keypoint to the nearest centroid based on Euclidean distance.
- **Update Step:** Recompute the centroids as the mean of all keypoints assigned to each cluster.
- **Iteration:** Repeat the assignment and update steps until the centroids no longer change significantly or a maximum number of iterations is reached.

Mathematically, the objective of k-means is to minimize the following within-cluster sum of squares (WCSS) Equation (15):

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|^2 \qquad (15)$$

Where:

- k is the number of clusters.
- $C_i$ is the set of keypoints in cluster $i$.
- $u_i$ is the centroid of cluster $i$.
- x represents a keypoint.

## 5.4.5 Patch extraction

Having meticulously identified keypoints through the Harris corner detector and judiciously employed the k-means clustering technique to eliminate redundancy, the subsequent endeavor is to extract regions of paramount interest enveloping these keypoints. This intricate operation entails the capture of image patches, each of dimensions $n \times n$ thoughtfully centered around every chosen keypoint (Figure 5.4).

For each keypoint, extract an image patch of dimensions $n \times n$ centered around the keypoint. The size $n$ is chosen based on the resolution of the document and the level of detail required.

**Patch Extraction Algorithm:**

- Initialize an empty list to store patches.
- For each cluster centroid, calculate the coordinates of the $n \times n$ patch centered on the centroid.
- Extract the patch from the original document image.
- Append the patch to the list.

**Benefits of Patch Extraction**

- **Localized Feature Capture:** By focusing on small regions around keypoints, we capture fine-grained details of handwriting and document features.
- **Improved Classification Accuracy:** The extracted patches provide high-quality, localized information that enhances the classifier's ability to distinguish between different writers and document types.
- **Data Augmentation:** Extracting multiple patches from each document increases the volume and diversity of the dataset, leading to more robust and generalizable models.
- **Reduction of Redundancy:** Clustering keypoints before patch extraction ensures that patches are representative and non-redundant, maximizing the utility of each extracted patch.

Patch extraction is a pivotal step in our system, transforming keypoint data into high-quality, localized patches that enrich the dataset. By meticulously extracting and augmenting these regions of interest, we enhance the dataset's quality and diversity, laying a solid foundation for effective and accurate classification of historical documents. This strategic augmentation is a testament to our commitment to excellence and meticulousness in refining the dataset for subsequent phases of analysis and modelling.

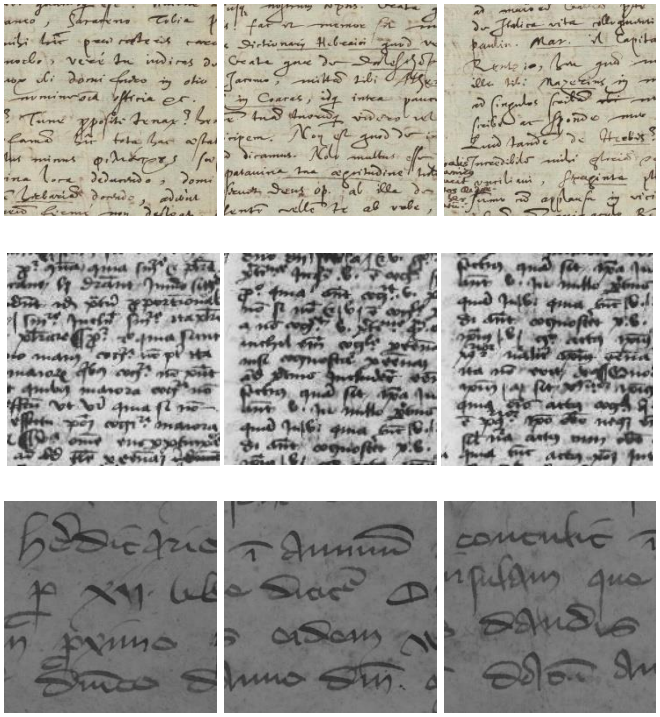


**Figure 5.4** Example of patches extracted from different samples

## 5.5 Classification

Classifying historical documents presents a unique set of challenges that require robust dataset preparation and powerful classification techniques. Deep learning, particularly through the use of Convolutional Neural Networks and Vision Transformers, represents the state-of-the-art algorithms for handling complex feature sets found in such documents. The success of deep learning models hinges significantly on the quality of the dataset, which underscores the importance of our preprocessing and feature detection steps. These steps ensure that our dataset is rich with significant features that can be effectively learned by our classifiers. This section will delve into the utilization of deep learning for classification, including automated feature extraction, the architecture of CNNs, and the emergence of Vision Transformers in our system.

### 5.5.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) **[140]** are specialized neural networks designed for processing structured grid data such as images. They are particularly effective for image classification tasks due to their ability to capture spatial hierarchies through convolutional operations. A typical CNN consists of the following layers as Figure 5.6 shows :

**a. Convolutional Layers**

The core building block of a CNN is the convolutional layer, which applies convolution operations to the input image to produce feature maps. Mathematically, the convolution operation is defined in Equation (16):

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \quad (16)$$

For discrete data (such as images), this is expressed in Equation (17):

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m] \cdot g[n-m] \quad (17)$$

In the context of CNNs, the discrete convolution between an image *I* and a filter (or kernel) K is given by Equation (18):

$$(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \quad (18)$$

Where:

- $\boldsymbol{I}$ is the input image,
- $\boldsymbol{K}$ is the filter,
- $(\boldsymbol{i}, \boldsymbol{j})$ are the coordinates of the output feature map.

### b. Activation Function (ReLU)

The Rectified Linear Unit (ReLU) is a non-linear activation function applied after the convolution operation to introduce non-linearity into the model. The ReLU function is defined in Equation (19):

$$f(x) = max(0, x) \quad (19)$$

### c. Pooling Layers

Pooling layers reduce the spatial dimensions of the feature maps, which decreases the computational load and helps with overfitting. The most common type is max pooling, which takes the maximum value from a patch of the feature map $y = max(x)$.

For a pooling window of size 2×2, the operation can be represented using Equation (20):

$$y(i,j) = max_{\ p,q \in \{0,1\}} x(2i+p, 2j+q) \quad (20)$$

### d. Fully Connected Layers

After several convolutional and pooling layers, the feature maps are flattened into a one-dimensional vector and passed through fully connected layers. Each neuron in a fully connected layer is connected to every neuron in the previous layer.

### e. Output Layer

The final layer is the output layer, which produces the classification results. For a classification problem with C classes, the output layer typically uses a softmax function to convert the raw scores into probabilities, it expressed in Equation (21).

$$softmax(z_i) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} \quad (21)$$

**Where $z_i$** is the raw score for class $i$

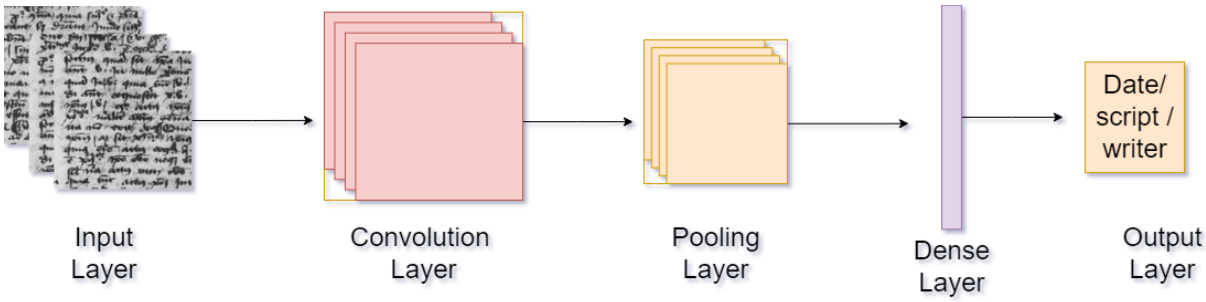



**Figure 5.5** Convolution neural networks architecture

## 5.5.2 Vision Transformers

Vision Transformers (ViTs) **[88]** are a type of neural network architecture specifically designed for image classification tasks, inspired by the success of transformers in natural language processing **[87]**. ViTs treat an image as a sequence of patches, similar to how transformers process sequences of words in text. This approach leverages the self-attention mechanism to capture global context and dependencies between different parts of the image, which is particularly useful for understanding complex and diverse visual patterns (Figure 5.6). Historical documents often contain intricate and varied handwriting styles, complex backgrounds, and noise due to age and wear. The self-attention mechanism in ViTs allows for capturing long-range dependencies and global relationships within the image, making them well-suited for identifying subtle and detailed features in historical documents.

### a. Patch Embedding

The input image is divided into fixed-size patches, each of which is flattened and linearly embedded into a vector. If the image has a size of H×W and the patch size is P×P, the number of patches will be calculated using Equation (22).

$$N = \frac{H \times W}{PXP}. \qquad (22)$$

Let x be the input image. The patch embeddings are given by Equation (23):

$$x_p == [x_1, x_2, \dots, x_N].E \qquad (23)$$

Where:

- $x_i$ is the i-th patch
- E is the linear embedding matrix.

### b. Adding Position Embeddings

To retain positional information, learnable position embeddings are added to the patch embeddings (Equation (24)).

$$z_0 = \left[x_p^1; x_p^2; \dots x_p^N\right] + E_{pos} \qquad (24)$$

**Where $E_{pos}$** is the position embedding.

### c. Transformer Encoder

The sequence of embedded patches is fed into a standard transformer encoder, which consists of multiple layers of multi-head self-attention and feed-forward networks.

- **Multi-Head Self-Attention (Equation (25)):**

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (25)$$

  - **Where Q,K,V** are the query, key, and value matrices, respectively
  - $d_k$ is the dimension of the key vectors.

- **Feed-Forward Network (Equation (26)):**

  Each encoder layer has a position-wise feed-forward network:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \qquad (26)$$

  - Where $W_1$ , $W_2$ are weight matrices
  - $b_1$ and $b_2$ are biases.

  The transformer encoder output is expressed in Equation (27):

$$z_L = TransformerEncoder(z_0) \qquad (27)$$

  **Where L** is the number of encoder layers.

### d. Classification Head

The output from the transformer encoder is processed by a classification head, typically a fully connected layer followed by a softmax function to produce class probabilities (Equation (28)).

$$y = softmax(W_{cls}\, z_L + b_{cls}) \qquad (28)$$

**Where $W_{cls}$ and $b_{cls}$** are the weights and biases of the classification head.

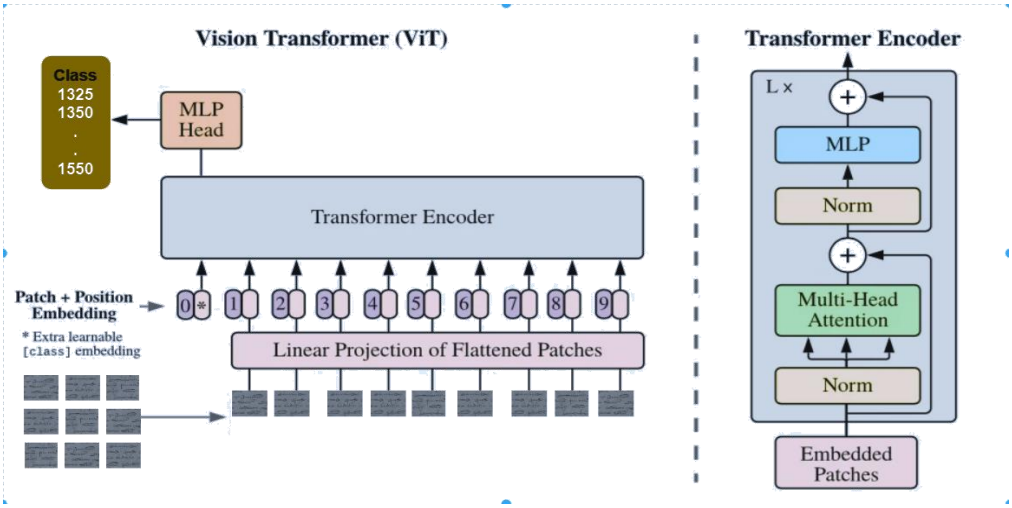



**Figure 5.6** Vision Transformer architecture (ViT).

### 5.5.3 Transfer Learning

Transfer learning is a machine learning technique where a model developed for one task is reused as the starting point for a model on a second task taking a pre-trained model—typically trained on a large and general dataset like ImageNet—and fine-tuning it on a smaller, task-specific dataset. This approach is particularly beneficial when the target dataset is small or specialized where acquiring a large amount of labeled data is challenging. Training deep neural networks for complex tasks, such as the classification of historical documents, often requires a vast amount of data to achieve good generalization and performance. However, historical document datasets are typically limited in size due to the rarity and uniqueness of these documents. Transfer learning addresses this limitation by leveraging pre-trained models that have already learned rich feature representations from large datasets. These pre-trained models can then be fine-tuned on our specific task, significantly reducing the amount of data and computational resources needed, the adjustment process involves:

- **Remove** the top fully connected layer.
- **Add** new fully connected layers tailored to our classification task.
- **Fine-tune** the network using our dataset, ensuring that the early layers retain their learned features while the later layers adapt to our specific needs.

In our approach, we have explored several pre-trained models from both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) categories. The selected models include:

- **ResNet:** ResNet **[141]** revolutionized deep learning by addressing the vanishing gradient problem through the introduction of residual connections. These connections allow gradients to flow directly through the network layers, enabling the training of much deeper networks. We fine-tune the ResNet model on our historical document dataset, adjusting the final layers to classify the specific features relevant to different documents.

- **DenseNet:** DenseNet **[142]**, connects each layer to every other layer in a feed-forward fashion. This dense connectivity pattern improves the flow of information and gradients throughout the network, leading to better performance and efficiency. DenseNet's architecture is well-suited for extracting features from complex and detailed images like historical documents. The dense connections help capture fine-grained details that are crucial for distinguishing different handwriting styles and document structure.

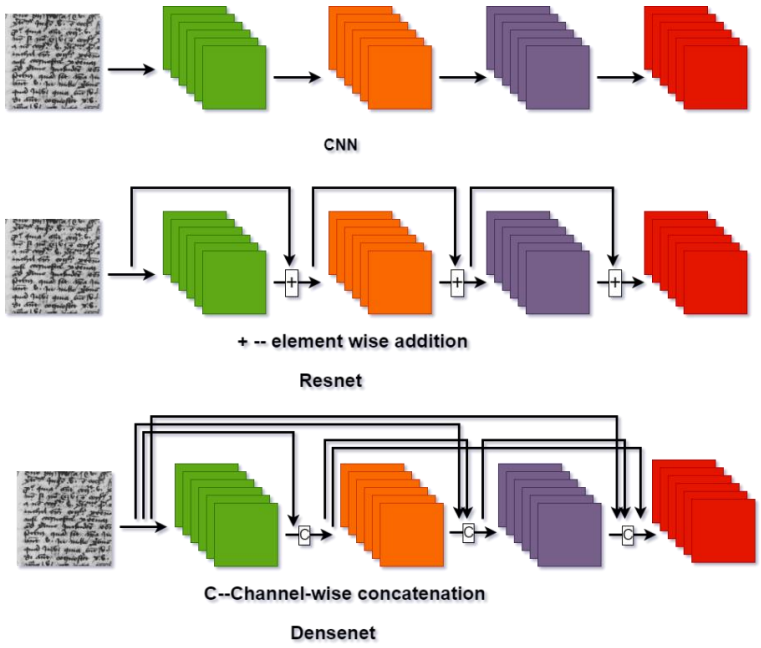



**Figure 5.7** Traditional CNN, Resnet, Densenet architectures.

- **Swin Transformer:** The Swin Transformer **[143]**, incorporates a hierarchical architecture with shifted windows (Figure 5.8), allowing it to capture both local and global features efficiently. This structure makes it highly effective for image classification tasks. Swin Transformer's ability to handle varying scales and its efficiency in capturing global context make it ideal for our problem. fine-tuning a pre-trained Swin Transformer can adapt to recognize the unique patterns and features present in historical documents.

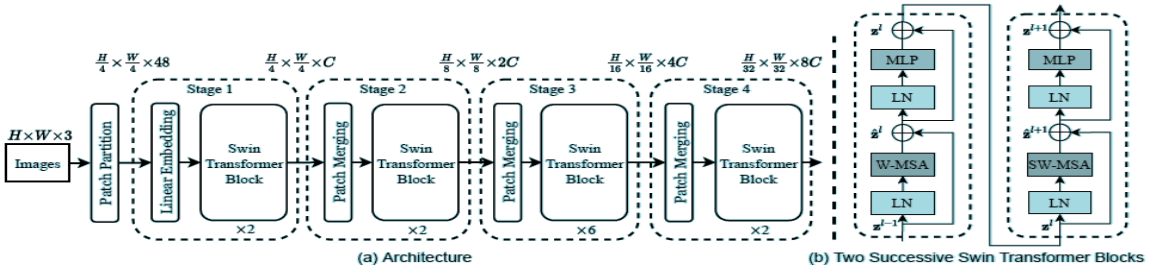



**Figure 5.8** Shifted Windows Transformer architecture (Swin).

- **DeiT (Data-efficient Image Transformer):** DeiT **[144]**, is a Vision Transformer model optimized for data efficiency. It achieves competitive performance with fewer data and less computational resources compared to traditional ViTs. DeiT's data efficiency is particularly advantageous for our limited historical document dataset (Figure 5.9).

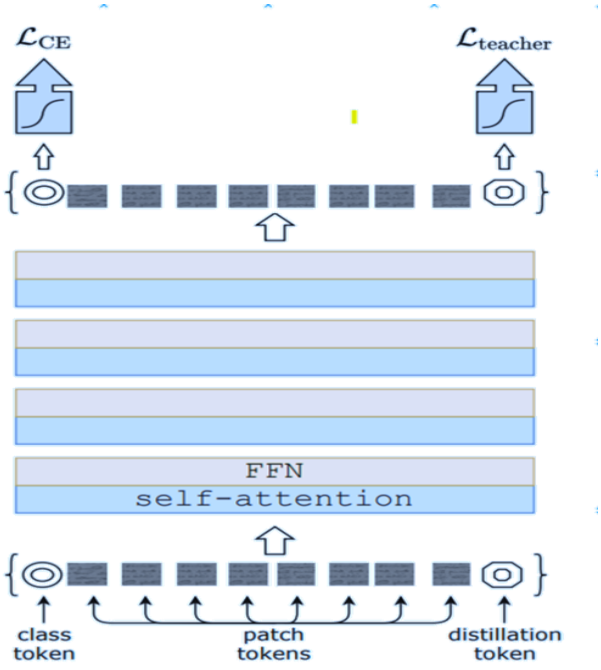



**Figure 5.9** Data Efficient Image Transformer architecture (DeiT)

Transfer learning provides a robust framework for tackling the challenges associated with the classification of historical documents. leveraging pre-trained models like ResNet, DenseNet, Swin Transformer, and DeiT, we can efficiently utilize smaller datasets to achieve high performance. Fine-tuning these models allows us to adapt their powerful feature extraction capabilities to the unique characteristics of historical documents, enabling accurate and effective classification.

### 5.5.4 Majority Vote

The majority vote is a crucial step in our historical document classification system, which ensures that our final output reflects the classification at the document level rather than the individual patch level. Given that our system extracts small patches from each document and performs training and classification on these patches, a method is needed to aggregate the results from these patches to make a reliable document-level prediction (Figure 5.10). Majority voting serves this purpose effectively:

- **Collect** all predicted class labels from the patches of a single document.
- **Count** the frequency of each class label among the patches.
- **Identify** the class label with the highest frequency among the patch-level predictions.

- **Assign** this class label as the final document-level prediction.

Given a document D divided into n patches $P_1, P_{2,...} P_n$ , let $C(P_i)$ represent the predicted class label for patch $P_i$. The majority vote C(D) for the document is determined using Equation (29):

$$C(D) = \arg max_{c \in C} \sum_{i=1}^{n} I(C(P_i) = c) \qquad (29)$$

Where:

- C is the set of all possible class labels.
- $I(C(P_i) = c)$ is an indicator function that returns 1 if the predicted class for patch $P_i$ is c, 0 otherwise

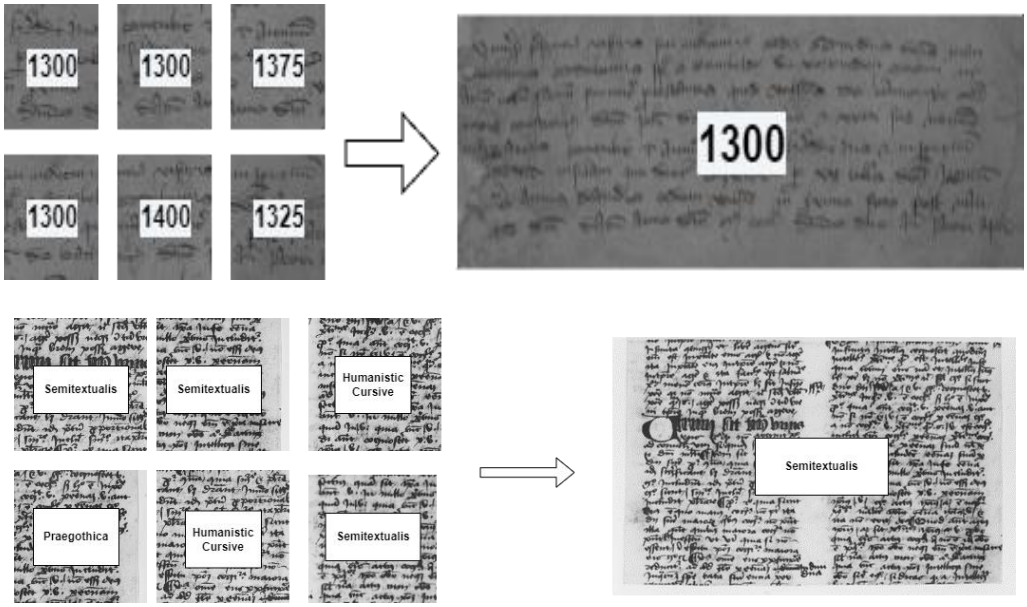



**Figure 5.10** Majority vote technique.

## 5.6 Contribution to writer identification task

Our publication **[21]** represents an extension of the system's application to the realm of writer identification for historical documents, involving three primary steps as outlined in Figure 5.11.

- **Firstly**, we applied denoising and binarization techniques to clean the data, using non-local means denoising for noise removal and Canny edge detection for binarization.
- **Next**, we identified key-points using the Harris corner detector after several experimentations with other key-point detectors.
- **Then**, to avoid the problem of grouped key-points, and to ensure a well distribution of the region of interest in the images, we proceeded to group the nearest key-points into clusters using k-means algorithm.

- **Then** we select each cluster's centroid, which will be considered as the center of the patch windows that will be generated around it as a region of interest.
- **These** extracted windows (patches) of a specific size from the original images are viewed as the automatic extracted features that feed the next phase of deep learning model training.
- **Finally**, we employed a majority vote to determine the writer of the document.

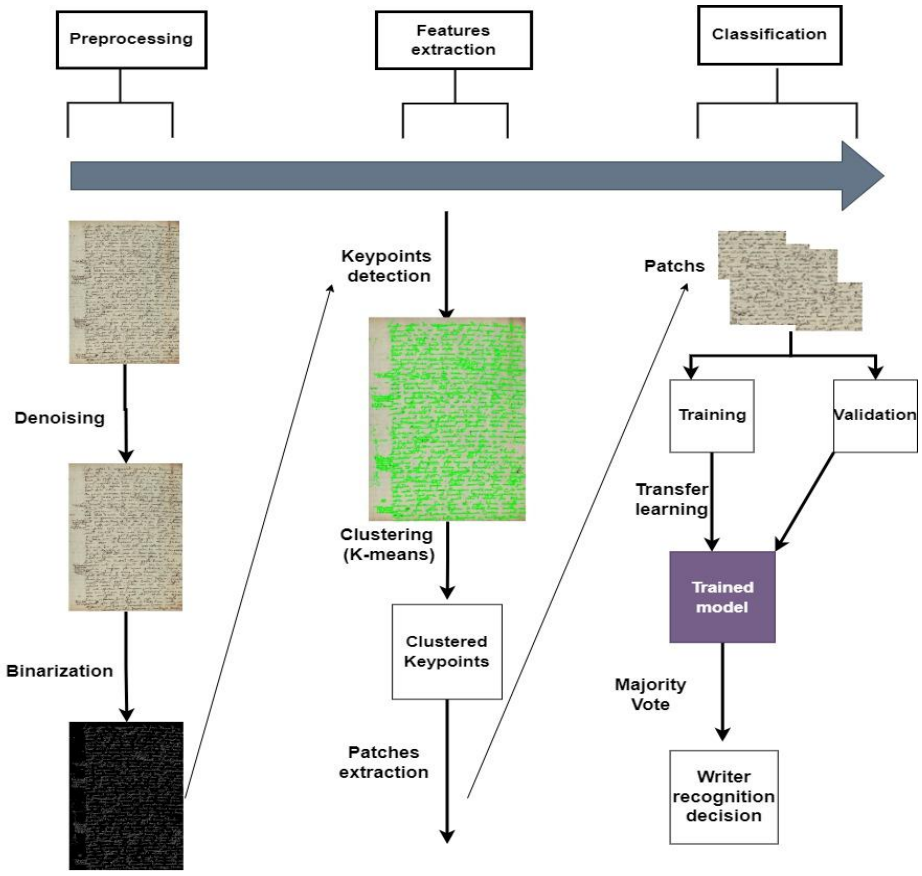



**Figure 5.11** Key steps in the proposed system.

The key highlights of this study are as follows.

• We have proposed a novel approach that combines the machine-learned features with deep learning model in a single efficient framework that provide high accuracy in the automated writer identification of the historical document, which can potentially reduce the workload of historians by improving the speed and accuracy of the historical documents analysis.

• we have provided a state-of-the-art of the recent methods that deal with the same problems,

• We employed effective pre-processing techniques to obtain clean and denoised data. In addition, we addressed the issue of a limited number of writing samples by efficiently augmenting our data.

• We used the automatic key-points detection combined with K-means clustering technique to avoid the manual selection of the region of interest, and to ensure the good distribution of these regions in the entire image, in addition to minimizing the redundancies and the useless information extracted from the images that could affect the training process and the whole classification system.

• We have presented a comparative study and an analysis of the proposed system's performance with the state of the art, and provided several scenarios of experimentation during each phase of the framework to justify the adopted choices.

## 5.7 Contribution to Script recognition task

The crux of our investigation in publication **[24]** revolves around Convolutional Neural Networks (CNNs) and its variants such as Rensnet and Densenet. Renowned for their prowess in various computer vision tasks and adeptness in pattern recognition, seeking to unveil their respective strengths in the intricate realm of historical script analysis. In our study, we rigorously evaluate the performance of these architectures to determine which one is best suited for the task of historical document script classification. Our methodological approach, illustrated in Figure 5.12 initiates with document preprocessing—an essential phase aimed at improving document quality for subsequent analysis. Notably, denoising and binarization procedures play a central role in this preprocessing stage, employing non-local means (NLM) and the Canny-edge threshold method to eliminate noise and standardize visual attributes. These enhancements lay the groundwork for feature detection using the Harris-corner detector (HCD), a crucial step in extracting uniform patches of significance, a prerequisite for the application of deep learning models. In the classification process, we leverage a variety of models sourced from diverse architectures, including CNNs such as ResNet and DenseNet. Ultimately, for the definitive script classification of an entire document, we adopt a majority voting mechanism—an approach that reinforces the accuracy and reliability of our classifications. Our objective is to discern the model most apt for our system. The assessment of our system hinges on the ClaMM dataset recognized as a benchmark in historical document. The results gleaned from our experiments vividly illustrate the remarkable capabilities of our deep learning-based system.

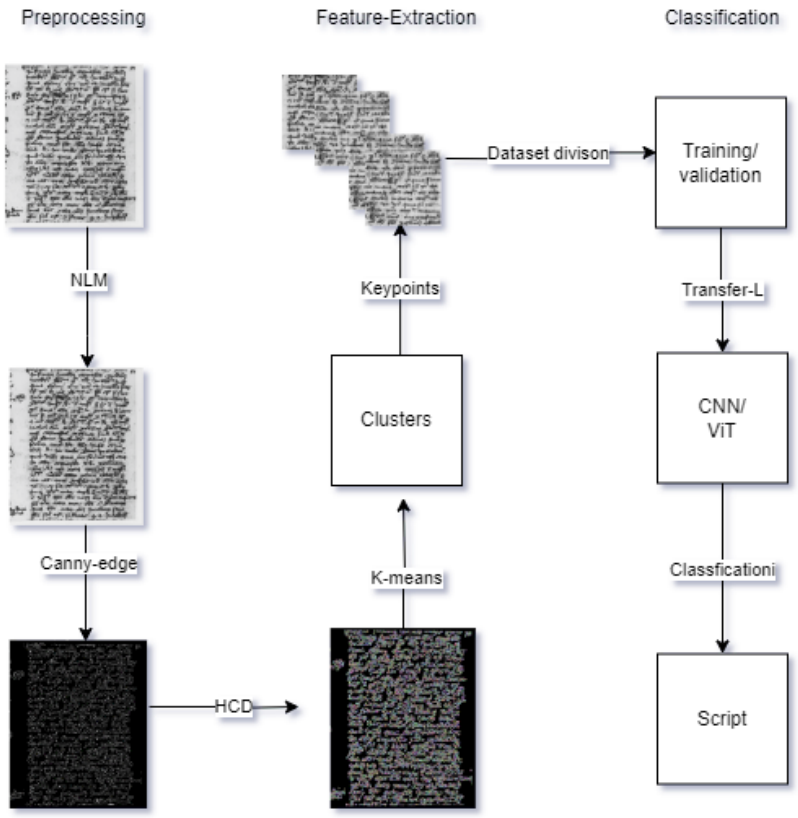



**Figure 5.12** Script classification system design

The main highlights of this research can be summarized as follows:

- We introduce an innovative approach that integrates machine-learned features with a deep learning model, utilizing an intelligent feature selection method. This method significantly contributes to the accurate classification of scripts in old documents.

- Employing effective pre-processing techniques, we ensure the acquisition of clean and denoised data for improved analysis.

- To address the challenge of a limited number of samples, we implement efficient data augmentation strategies, enhancing the diversity and richness of our dataset.

- Automatic key-points detection, combined with the K-means clustering technique, is employed to eliminate manual selection of the region of interest. This ensures a well-distributed selection of regions across the entire image while minimizing redundancies and irrelevant information that could impact the training process and overall classification system.

- Multiple experimentation scenarios are provided for each phase of the framework, offering a thorough justification for the choices made during the development process.

Additionally, In our study **[25]** we have successfully developed a robust system dedicated to the classification of historical scripts (Figure 5.13), employing a thoughtful combination of multiple steps and techniques. Each choice of technique was made based on thorough

experimentation, ensuring the effectiveness of the overall system. The process begins with preprocessing, utilizing Non-local means and Canny-edge techniques. The output serves as input to Harris corner detection for keypoint identification, followed by clustering using k-means to extract patches of relevant information. These patches are then trained on two different architectures, Vision Transformers (Vits) and Convolutional Neural Networks (CNN), enabling a comprehensive comparison between these two approaches. The results reveal that the Vision Transformer model Swin-S has outperformed the CNN architecture and other traditional machine learning methods, establishing its suitability for the classification of historical scripts.

## 5.8 Contribution to Dating task

In our papers on dating historical documents **[22,23]** we explored the functionalities of ViT architecture as well as its derivatives such as Data Efficient Image Transformer (DeiT) and Shifted windows transformer (Swin), emphasizing their self-attention mechanisms and highlighting their uncharted potential in the domain of historical documents dating. The proposed system encompasses three steps, when the choices of the adopted methodologies for each single one has been made after a rigorous experimental analysis. As outlined in Figure 5.13, initially as a preprocessing step, we employ denoising and binarization techniques to enhance data quality. For noise reduction, we utilized the non-local means denoising method (NLM), and for binarization, we relied on Canny edge detection (CED). Subsequently, in features-detection we detect keypoints within the data using the Harris corner detector (HCD), and refine the selection by employing the k-means clustering algorithm to identify specific centroids, then, we extract windows of a specified size around the cluster's centroids from the original images. Lastly, for deep-learned features and classification we fine-tune pre-trained vision transformer models through transfer learning (TL) conducting various experiments with different architectures (ViT, Swin, DeiT) and hyper-parameters (optimizer, model, patch size, etc.), additionally, a majority vote mechanism is applied to determine the dating of the entire document. We believe that this works presents a pioneering contribution to the application of vision transformers in historical documents dating, introducing a novel hybrid approach that encompasses a multi-faceted methodology setting a new benchmark in accuracy and versatility. Its significance lies in its potential to revolutionize historical document dating, not only by enhancing precision but also by providing a versatile framework adaptable to diverse historical contexts and document types. Beyond its immediate applications, this research paves the way for a deeper understanding of historical narratives, cultures, and the invaluable insights that lie within the annals of our past.

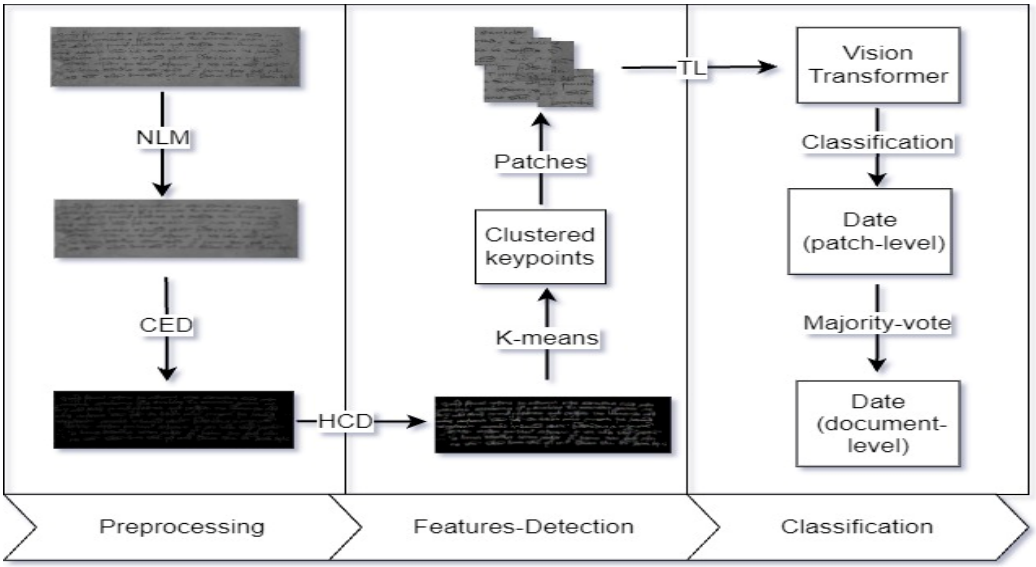



**Figure 5.13** Dating System Design.

The key highlights of this study are as follows;

- We have proposed a practical dating system that allow us to explore the effectiveness of vision transformers in historical documents dating condemning various experiments to define the best system in term of model architecture and hyper-parameter finetuning, and presented a recent state of the art containing the most recent related work in the field.

- We have explored a hybrid approach that combines hand-crafted and deep-learned features to efficiently detect features from the historical document documents, and employed effective preprocessing techniques to obtain clean and denoised data for accurate features extraction, also, we address the issue of limited number of samples by augmenting our data through efficient patch extraction process.

- We have presented a comparative study and an analysis of the proposed system's performance with the state of the art, and provided several scenarios of experimentation during each phase of the framework to justify the adopted choices.

## 5.9 Conclusion

In this chapter, we have provided a comprehensive overview of our methodology, which leverages image-based techniques and deep learning for the classification of historical documents. We meticulously outlined each component of our system, starting from the preprocessing stage, where essential steps were taken to enhance the quality of the images. We then delved into feature extraction, detailing how critical features were identified and extracted to represent the intricate patterns and characteristics of the documents. Finally, we discussed the classification process, where the extracted features were used to train our deep learning models to accurately classify the documents. This methodical approach ensures that each stage contributes effectively to the overall performance of our system. In the next chapter, we will explore the results and discuss the outcomes of various experiments conducted on our system.

# Chapter 6

# Experiments & Results

## 6.1 Introduction

After providing a detailed overview and explanation of our system architecture, this chapter presents a comprehensive evaluation of our system's performance across various tasks, datasets, and evaluation metrics. Each task is treated as a distinct experiment, with detailed results and discussions provided for each. The primary objectives of these experiments are to optimize and fine-tune our system parameters and hyperparameters for optimal results, and to compare our system with state-of-the-art methods in the literature. Each experiment has a specific objective, such as demonstrating the importance of certain techniques ("Preprocessing Impact Experiment"), determining the optimal hyperparameters for better training ("Hyperparameter Experiment"), or identifying the best parameters for our preprocessing steps ("Patch Size Experiment" and "Model Selection Experiment"). The evaluation metrics will vary depending on the task, ensuring that each experiment is assessed appropriately for its unique goals.

## 6.2 Datasets & Evaluation metrics

For our experimental study, we selected widely-used datasets for each task. These datasets are commonly utilized by researchers in the field, serving as benchmark resources for evaluating similar systems. This widespread adoption motivated our choice to use these datasets as standards to assess our system’s performance.

**For the writer identification task**, we used the ICDAR 2017 competition dataset **[95]**, which includes 3,600 document samples from 720 distinct writers, each contributing five samples. To validate our system, we employed k-fold cross-validation with a split of 80/20, dividing the dataset into five equal subsets (k=5). In each run, four folds were used for training, and one-fold was reserved for testing.

For the quantitative evaluation of our system, we utilized several common metrics to assess performance accurately and comprehensively. The primary metrics we focused on are mean average precision (mAP) and Top-N accuracy.

- **Mean Average Precision (mAP):** mAP is a measure that takes into account both precision and recall across all classes (writers) in the dataset. It provides a single metric that reflects the system's ability to correctly identify and rank relevant instances. The mAP is computed as follows:

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} APk$$

$$APK = the\ AP\ of\ class\ k$$

$$n = the\ number\ of\ classes$$

- **Top-N Accuracy:** Top-N accuracy measures the proportion of test samples that are correctly identified within the top N most probable answers. It is particularly useful in scenarios where providing multiple possible matches can still be considered a success, such as in writer identification where there might be several plausible authors for a given document. Top-N accuracy is defined as:

$$Top - N\ Accuracy = \frac{Number\ of\ correctly\ identified\ samples\ in\ the\ top\ N\ results}{Total\ number\ of\ test\ samples}$$

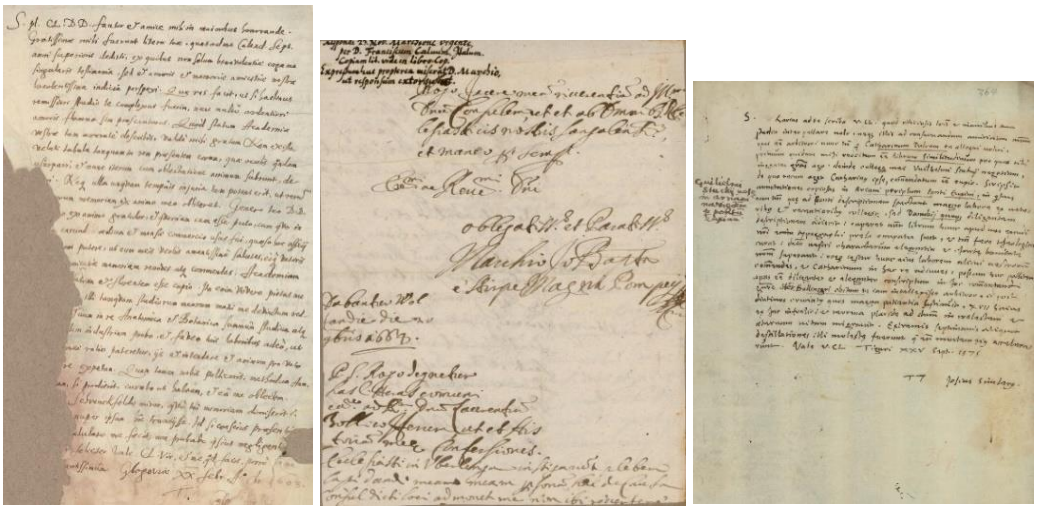


**Figure 6.1** Original samples from historical-wi dataset [95].

**For script type classification**, we used the CLAMM dataset, created for the Competition on the Classification of Medieval Handwritings in Latin Script **[94]**. This dataset is essential for identifying and classifying medieval handwritten scripts (Figure 6.2). Table 6.1 details the composition of the dataset, encompassing 12 distinct script classes.

To ensure rigorous testing, we divided our preprocessed dataset, allocating 80% of the patches for training and reserving 20% for validation. Drawing inspiration from the evaluation methodology utilized in the ICDAR 2017 competition for a similar task, we prioritize the metric of global accuracy (GA). This metric is defined by the equation:

$$GA = \frac{1}{N}\sum_{i=1}^{N} \delta(Dli, GTLi)$$

Where:

- $N$ is the total number of images in the test dataset.

- $\delta(Dli, GTLi)$ is the Kronecker delta function, which is equal to 1 if DLi= GTLi and 0 otherwise.

***Table 6.1*** *Script number of samples per class*

| **Script name** | **Uncial** | **Half-uncial** | **Caroline** | **Humanistic** | **Humanistic Cursive** | **Praegothica** |
|---|---|---|---|---|---|---|
| **Class** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Samples** | 356 | 395 | 200 | 207 | 217 | 341 |
| **Script name** | **Southern Textualis** | **Semitextualis** | **Textualis** | **Hybrida** | **Semihybrida** | **Cursiva** |
| **Class** | 7 | 8 | 9 | 10 | 11 | 12 |
| **Samples** | 411 | 265 | 293 | 202 | 442 | 211 |

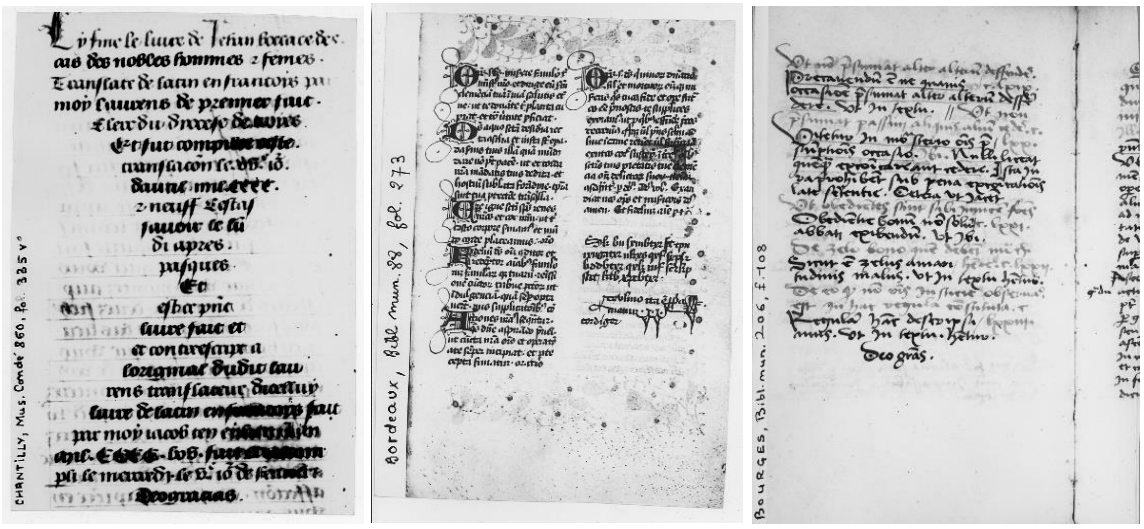



**Figure 6.2** Script samples from ClAMM dataset.

**For the dating task**, we used MPS dataset **[89]** which consists of a collection of ancient charters that date back to the period between 1325 and 1550 (Figure 6.3), each charter belongs to one of the 11 predefined classes as shown in Table 6.2 These classes are typically based on different historical periods, regions, or other relevant categorizations.

**Mean Absolute Error (MAE):** MAE quantifies the average deviation between the estimated date and the actual date of the documents. A lower MAE indicates better accuracy in predicting the dates. The MAE is computed as follows:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - y\hat{}\right|$$

Where:

- n represents the number of samples.
- $y_j$ is the actual observed value (ground truth) for the j-th data point.
- $y$^ represents the predicted or estimated value for the j-th data point.

- **Cumulative Score (CS):** CS represents the percentage of test images for which the absolute error in estimating the production year $Ne$ is less than or equal to a specified threshold $\alpha$. The CS is computed as follows:

$$CS(\alpha) = \frac{N_{e \leq \alpha}}{N} \times 100\%$$

Where:

- $N_{e \leq \alpha}$ represents the number of relevant samples with scores less than
- or equal to α.
- N is the total number of samples in the dataset.

***Table 6.2*** *Number of samples in each period of the MPS dataset.*

| **Period** | **1300** | 1325 | 1350 | 1375 | 1400 | 1425 | 1450 | 1475 | 1500 | 1525 | 1550 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Samples** | 106 | 164 | 199 | 386 | 311 | 323 | 501 | 423 | 372 | 241 | 241 |

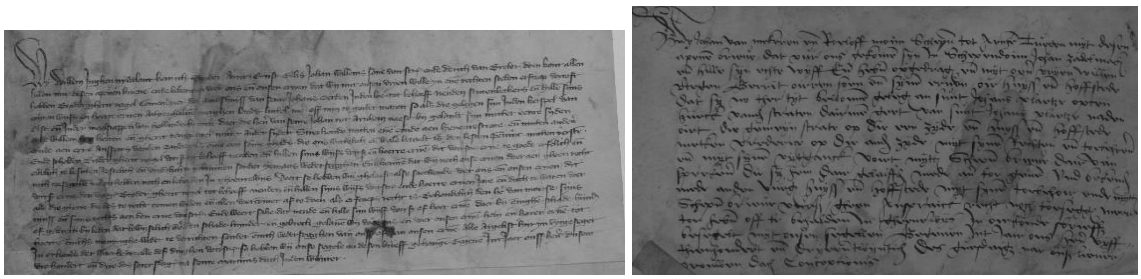
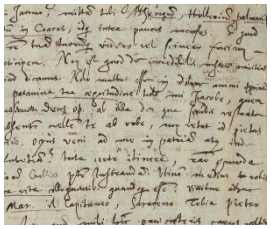
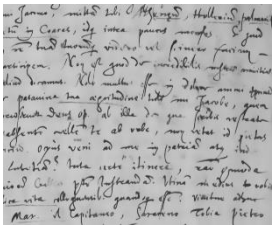
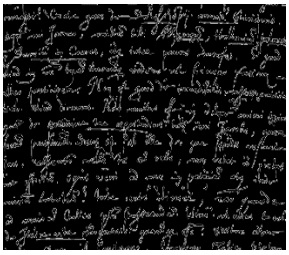
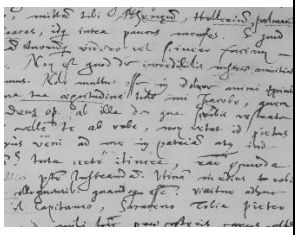
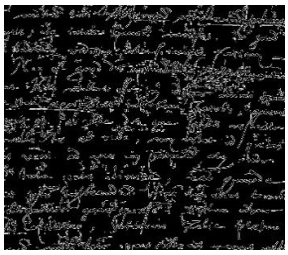
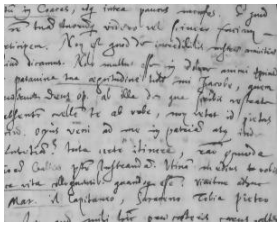
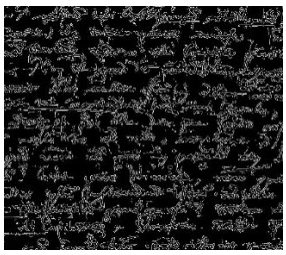
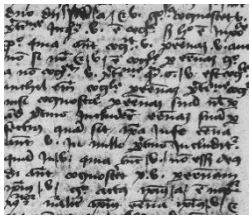
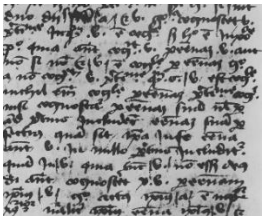
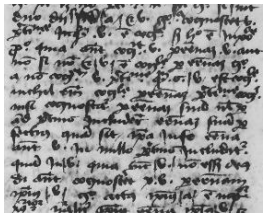
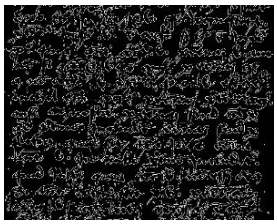
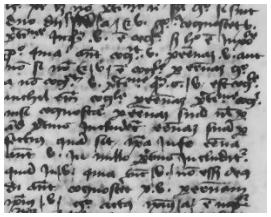
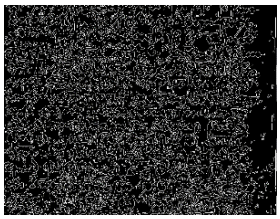
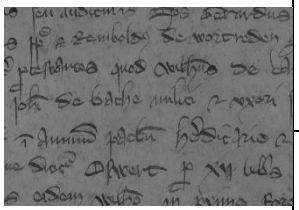
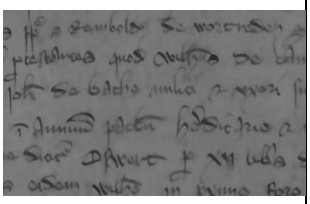
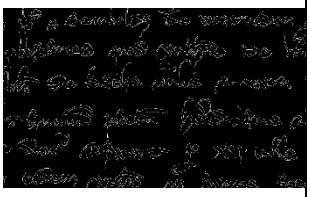
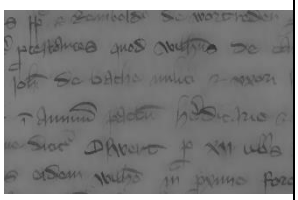
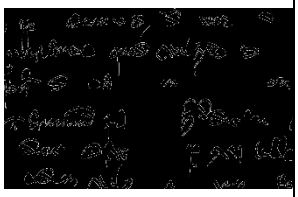
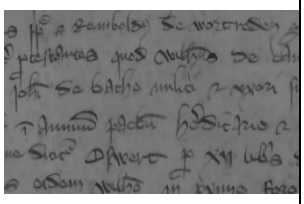
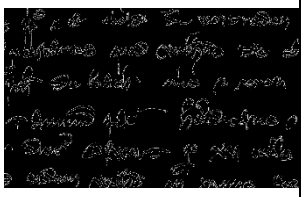


**Figure 6.3** Script samples from MPS dataset.

## 6.3 Denoising techniques experiment

For denoising technique choice in our system, in our experiments, we employed the most accurate denoising methods on cropped samples from our datasets: Non-Local Means **[132]**, Bilateral Filtering **[133]**, and Total Variation Denoising **[134]**. In Table 6.3 we present the outcomes achieved with each technique. To ascertain which methods exhibited superior performance, we consistently applied the same binarization method to all the denoised images and disclosed a sample of the outcomes in the same Table 6.3.

***Table 6.3*** *Outcomes of different denoising techniques*

| **Original Sample** | **Non-Local Means Denoising** | **Bilateral Filtering Denoising** | **Total Variation Denoising** |
|---|---|---|---|
|  **Writer Task** |   |   |   |
|  **Script Task** |   |   |   |
|  **Dating Task** |   |   |   |

The denoising results in Table 6.3 from our experiment reveal key insights into the performance of different techniques—Non-Local Means, Bilateral Filtering, and Total Variation Denoising—in preparing images for further processing, particularly in the binarization stage.

## 6.3.1. Visual Quality and Denoising Efficiency

- **Non-Local Means Denoising:** This technique demonstrated superior performance in preserving details while effectively removing noise, as confirmed by visual inspection. Non-Local Means tends to retain texture and fine details, critical for historical document images where essential information might be embedded in subtle variations. This quality preservation was evident after binarization, where Non-Local Means images yielded clean, high-contrast results with minimal artifacts. The effectiveness of this method across different types of documents in our dataset indicates its robustness and adaptability, making it suitable for consistent application throughout the study.

- **Bilateral Filtering:** Although Bilateral Filtering effectively reduces noise while retaining some edges, it struggled with finer details compared to Non-Local Means. Post-binarization, images processed with Bilateral Filtering sometimes displayed slight blurring around detailed regions, which could obscure valuable features in tasks such as script classification and writer identification. While it performs adequately, the method's limitations become more apparent in highly detailed areas where texture and subtle strokes are critical.

- **Total Variation Denoising:** Total Variation Denoising provided significant noise reduction but tended to oversimplify the images by smoothing out fine details. In binarized results, this simplification sometimes led to the loss of nuanced elements, which could be detrimental to accuracy in classification tasks that rely on intricate patterns. Although useful in cases where extreme noise reduction is required, Total Variation Denoising is less suited to our goals given the fine detail in document samples.

## 6.3.2 Impact on Binarization

- **Binarization Results:** When binarization was applied to each denoised output, Non-Local Means consistently yielded the best results. This outcome reinforces its effectiveness in both noise reduction and detail preservation, as binarization revealed clean separations between text and background. The other techniques, particularly Total Variation Denoising, led to less distinct binarized images due to the overly smooth textures. For historical documents, maintaining text clarity post-binarization is critical since many downstream tasks (e.g., writer identification, script classification) depend on high-contrast, detailed features.

- The clear, artifact-free binarization of Non-Local Means images suggests it enhances the quality of features fed into machine learning models, improving classification performance across tasks.

### 6.3.3 Implications for Specific Tasks

- **Writer Identification:** This task requires highly detailed images to capture subtle differences in handwriting. Non-Local Means, by preserving these details post-binarization, enhances the discriminative power of the features used in this task, potentially improving Top-N accuracy and mAP metrics. Bilateral Filtering and Total Variation Denoising, however, could compromise accuracy in writer identification by blurring or simplifying unique handwriting patterns.

- **Script Type Classification:** Given that script classification benefits from clear patterns and textures, the Non-Local Means approach again proves ideal. Its ability to retain structural elements ensures that the classifier can better differentiate script styles. The more simplified outputs from Total Variation Denoising could obscure stylistic nuances, leading to lower global accuracy in this task.

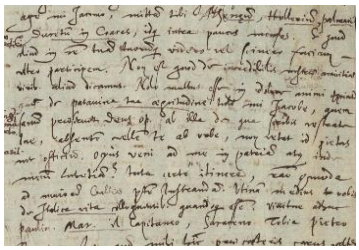
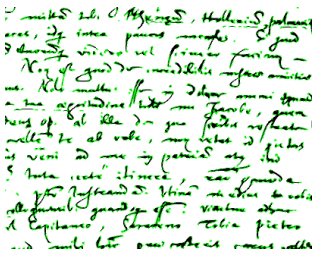
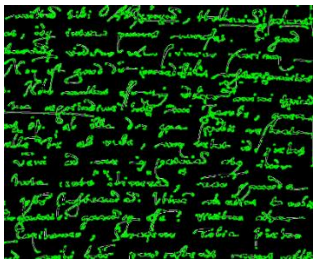
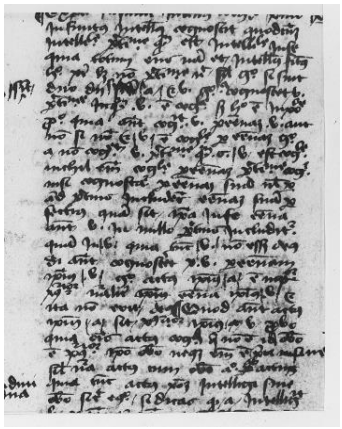
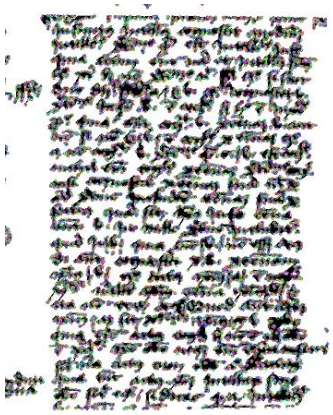
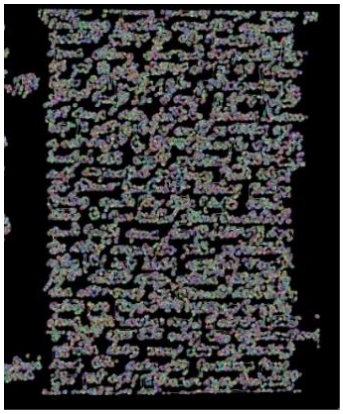
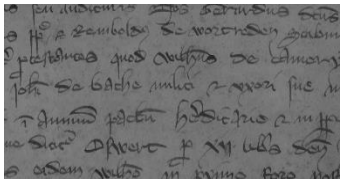
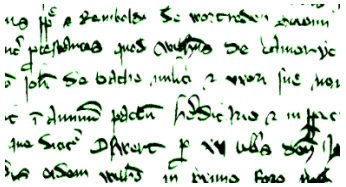
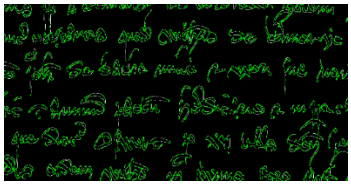
- **Dating Task:** For accurate dating, slight variations in script style and ink degradation may hold valuable temporal clues, preserving these characteristics, Non-Local Means prepares the data in a way that enables better estimation of document dates, as reflected in lower MAE and higher CS values in post-binarization results. In contrast, excessive smoothing from other methods may remove essential age-related artifacts, reducing dating accuracy.

Based on our experiments, Non-Local Means **[132]** emerged as the optimal denoising method for historical document classification. It strikes a balance between effective noise reduction and detail preservation, essential for accurate binarization and feature extraction in downstream tasks. Applying this method consistently across all tasks ensures that the images retain their defining characteristics, maximizing system performance across writer identification, script classification, and dating tasks. This choice aligns with our goal of developing a robust classification system that accommodates the complexities and variations inherent in historical document data.

## 6.4 Binarization techniques experiments

In this experiment, we tested two binarization techniques—Otsu **[135]** and Canny edge **[136]** detection—to determine which method was more suitable for our feature detection approach. Below is a detailed analysis of the results based on the number and quality of detected key points and a visual evaluation of the resulting key-point distributions.

***Table 6.4*** *Outcomes of Binarization methods on different samples*

| Original Sample | Otsu Threshold | Canny Edge detection |
|---|---|---|
|  **Writer Task** |  |  |
|  **Script Task** |  |  |
|  **Dating Task** |  |  |

## 6.4.1 Comparison of Binarization Techniques

- **Otsu Method:** The Otsu method, a popular global thresholding technique, produced generally good binarization results, effectively separating foreground and background in the images. However, when analyzed for feature detection performance, images binarized with Otsu displayed fewer detected key points. This limitation could be attributed to the fact that Otsu's thresholding, being global, might overlook finer edge details, especially in images with variable lighting or complex patterns, resulting in fewer distinct features available for detection.
- **Canny Edge Detector:** In contrast, the Canny edge detector, which uses gradient-based edge detection, excelled at capturing fine details in the images. This technique provided a more nuanced separation of edges, which allowed the Harris corner detector to identify a greater number of key points. The Canny edge method enhanced

the quantity and also the quality of detected key points, creating a more detailed and comprehensive feature map essential for our classification tasks.

### 6.4.2 Impact on Feature Detection and Key-Point Distribution

- **Harris Corner Detection Results:** The Harris corner detector's output on Canny edge binarized images showed a significant increase in key points compared to Otsu-binarized images. The improved edge clarity in Canny edge images allowed the Harris detector to identify high-quality corner features, resulting in a denser and more meaningful distribution of key points. This distribution provides a stronger foundation for feature extraction, improving the potential accuracy of downstream classification tasks by offering more distinguishing data points.
- **Quality of Detected Key Points:** The quality of detected key points in the Canny edge images was visibly superior, with well-defined and accurately localized corners that were evenly distributed across relevant regions. In contrast, the Otsu-binarized images, though still usable, had a more limited key-point spread and sometimes produced spurious points due to its broader thresholding approach, which affected the precision of detected features.

### 6.4.3 Implications for Classification Tasks

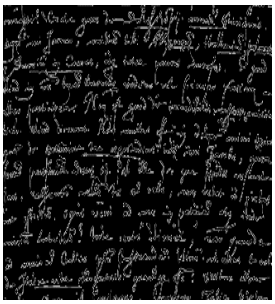
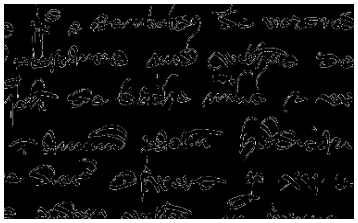
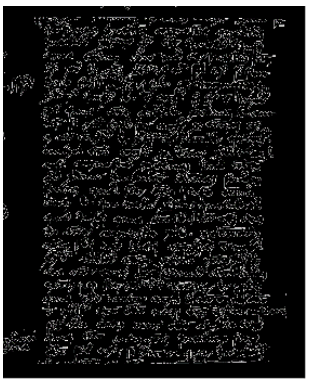
- **Writer Identification:** For writer identification, where small differences in handwriting style are crucial, the detailed edge representation provided by Canny edge binarization is advantageous. More key points equate to a richer set of features that can differentiate subtle variations in individual handwriting. This enhancement in feature quality is likely to increase Top-N accuracy and mAP for writer identification.
- **Script Type Classification:** In script classification, distinguishing different script styles depends on the detection of unique structural elements. The Canny edge technique's ability to capture fine details facilitates more accurate corner detection, allowing the classification model to discern variations in stroke patterns and script morphology. This feature density from Canny edge binarization directly contributes to achieving higher global accuracy (GA) in script classification.
- **Dating Task:** Accurate dating of historical documents relies on detecting minor yet informative artifacts, which might fade with age. The Canny edge method's ability to reveal subtle textural details supports more accurate feature extraction in age estimation. The increased number of key points enhances the dating model's capacity to differentiate between subtle historical attributes, potentially reducing MAE and improving CS values.

The comparative analysis of the Otsu and Canny edge binarization techniques clearly indicates that the Canny edge detector is more effective for our system's requirements, providing higher-quality and more numerous key points, Canny edge binarization ensures a rich feature set for keypoints detection, thus enabling better performance across classification tasks. This method's success in enhancing both the quantity and distribution of key points makes it the optimal choice for our experiments. Given its strong performance in all relevant metrics, Canny edge detection is adopted as the standard binarization technique for the remainder of the study, ensuring a more robust and comprehensive feature extraction process that benefits all downstream tasks.

## 6.5 keypoints detection techniques experiment

In this experiment, we assessed three keypoint detection methods—Scale-Invariant Feature Transform (SIFT) **[137]**, Features from Accelerated Segment Test (FAST) **[138]**, and the Harris Corner Detector (HCD) **[139]**—to determine which best supports the classification tasks involving historical document images. The criteria for selecting the optimal method were primarily based on the number of keypoints detected. Table 6.5 provides a detailed comparison of the results.

***Table 6.5*** *Comparison between different feature detection methods*

| | Detector | Keypoints number |
|---|---|---|
|  | **HCD** | **61872** |
| | FAST | 7128 |
| | SIFT | 3970 |
|  | **HCD** | **84933** |
| | FAST | 7603 |
| | SIFT | 3934 |
|  | **HCD** | **619313** |
| | FAST | 10356 |
| | SIFT | 10859 |

### 6.5.1 Comparison of Keypoint Detection Methods

- **Harris Corner Detector (HCD):** HCD stands out as the method with the highest keypoint detection capability, identifying significantly more keypoints than both SIFT and FAST across all tests. The superior keypoint density offered by HCD makes it

highly effective for extracting detailed features from historical documents, which are rich in intricate visual patterns. Although HCD has longer processing times than FAST and SIFT, the number of detected keypoints—up to 619,313 in the highest case—demonstrates its strong suitability for applications where detail is crucial.

- **FAST:** FAST is extremely efficient in terms of processing speed, outperforming both SIFT and HCD with minimal time required for keypoint detection. However, the downside is that it detects relatively few keypoints compared to HCD. With a keypoint count significantly lower than HCD, FAST might miss some important features in the documents, limiting its effectiveness for detailed feature extraction.
- **SIFT:** SIFT performs moderately, detecting more keypoints than FAST but significantly fewer than HCD. SIFT's processing times are also longer than FAST but shorter than HCD in most cases. Despite its ability to detect more keypoints than FAST, the keypoint density remains insufficient compared to HCD for highly detailed document analysis.

### 6.5.2 Implications for Classification Tasks

- **Writer Identification:** In writer identification, capturing fine-grained details of the handwriting is essential. HCD's high keypoint count provides a richer dataset for analyzing handwriting features, potentially improving performance metrics such as Top-N accuracy and mean average precision (mAP). The lower keypoint density of FAST and SIFT may limit their effectiveness in accurately distinguishing between individual writing styles.
- **Script Type Classification:** For script type classification, which relies on detecting specific stylistic elements in the text, the high keypoint density from HCD is advantageous. With more keypoints, HCD can capture distinctive patterns in script style, contributing to improved global accuracy (GA). While SIFT offers moderate keypoint detection, its performance may still fall short of the detail provided by HCD. FAST's efficiency may come at the cost of lower accuracy due to the lower number of detected keypoints.
- **Dating Task Classification:** Accurate dating of historical documents often depends on detecting subtle features indicative of age. HCD's high keypoint density ensures a more comprehensive representation of the document's texture and aging characteristics, helping reduce the Mean Absolute Error (MAE) and improve the Cumulative Score (CS). SIFT and FAST may overlook minor details, which could impact date prediction accuracy.

The Harris Corner Detector (HCD) emerged as the most effective feature detection method for our needs, with the highest keypoint density, offering a more detailed representation of document features that is vital for classification tasks. Despite its higher computational cost, HCD's benefits in accuracy and detail outweigh the processing time drawback, making it the preferred choice for our study.

## 6.6 Hyper-parameters experiment

In this section, we analyze the impact of different hyper-parameter configurations on the performance of our system across various tasks, including writer identification, script type classification, and dating task classification. We evaluated the model's performance using a

variety of combinations of epochs, learning rates, optimizers, and batch sizes. The experiments demonstrate that hyperparameter optimization is task-dependent, with each task benefiting from different configurations:

**In Writer identification,** the evaluation metrics include Top-1, Top-5, and Top-10 accuracy, as well as mean Average Precision (mAP). The results are summarized in Table 6.6 where we observe that Adam optimizer with a lower learning rate (0.001) and longer training (20 epochs) achieved the best performance.

***Table 6.6*** *Writer task Hyperparameters experiment results using CNN.*

| Epochs | Learning rate | optimizer | Batch-size | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|---|---|---|
| 10 | 0.01 | SGD | 16 | 75.6 | 78.3 | 84.4 | 53.2 |
| 20 | 0.001 | SGD | 32 | 76.6 | 78.9 | 85.2 | 55.3 |
| **20** | **0.001** | **Adam** | **32** | **85.2** | **90.4** | **93.6** | **73.5** |
| 10 | 0.01 | Adam | 16 | 65.3 | 75 | 81.3 | 43.3 |

**In Script classification**, we evaluated the performance of our model using global accuracy (GA%) for both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The results are summarized in Table 6.7. For CNNs, the Adam optimizer with a lower learning rate was optimal, while for ViTs, a higher learning rate with SGD was more effective.

***Table 6.7*** *Script task Hyperparameters experiment results for both CNN & Vit*

| Epochs | Learning rate | optimizer | Batch-size | GA% (CNN) | GA%(Vit) |
|---|---|---|---|---|---|
| 10 | 0.01 | Adam | 16 | 75.36 | 23.42 |
| **20** | **0.001** | **Adam** | **32** | **83.28** | **65.07** |
| 10 | 0.001 | SGD | 16 | 52.40 | 62.93 |
| 20 | 0.01 | SGD | 32 | 70.55 | 77.88 |

**In Dating task,** the performance metrics evaluated are the Mean Absolute Error (MAE) and Cumulative Score (CS) at two thresholds: $\alpha$=0 years and $\alpha$=25 years. The results are summarized in Table 6.8. Using Vit, this task required faster learning with fewer epochs and a higher learning rate using SGD to achieve the lowest error and highest cumulative scores.

***Table 6.8*** *Dating task hyperparameters experiment results using Vit*

| Epochs | Learning rate | optimizer | Batch-size | MAE | CS($\alpha$=0) | CS($\alpha$=25) |
|---|---|---|---|---|---|---|
| 10 | 0.01 | Adam | 32 | 63.32 | 18.22 | 45.63 |
| 20 | 0.001 | Adam | 16 | 45.04 | 31.85 | 66.61 |
| **10** | **0.01** | **SGD** | **32** | **16.33** | **66.15** | **89.85** |
| 20 | 0.001 | SGD | 16 | 20.95 | 61.25 | 82.45 |

These findings underscore the importance of tailoring hyperparameters to the specific goals and characteristics of each task and model architecture.

## 6.7 Impact of preprocessing

Different preprocessing steps, such as resizing, generating random patches, and key-points detection, were evaluated to understand their effect on identification performance across different tasks. This section outlines the influence of these preprocessing approaches on writer identification, script classification, and dating tasks.

The following Table 6.9 summarizes the identification performance under different preprocessing conditions. The evaluation metrics include Top-1 accuracy for writer task, GA for script task, and CS (a=0) for dating task.

***Table 6.9*** *Identification performance with and without preprocessing steps*

| **Experiment** | **Top-1 Writer Task** | **GA Script task** | **CS (a=0) Dating task** |
|---|---|---|---|
| Without Preprocessing (only resizing) | 35.8 | 40.96 | 64.16 |
| Random patches of size 550x550 | 47.6 | 68.14 | 76.30 |
| **Preprocessing +key-points detection + patches of size 550x550** | **85.2** | **83.28** | **79.59** |

**Without Preprocessing (Only Resizing):** This baseline condition involved minimal preprocessing, where only resizing was applied to standardize image dimensions. The performance was noticeably lower across all tasks:

- Top-1 accuracy for the writer task was 35.8%, indicating limited success in writer identification.
- Global accuracy (GA) for the script task was 40.96%, showing weak classification performance.
- Cumulative score (CS) for the dating task was 64.16%, reflecting moderate accuracy in date classification.

This result underscores the importance of additional preprocessing steps, as simple resizing alone does not sufficiently enhance data quality for accurate identification.

**Random Patches of Size 550x550:** Introducing random patches of size 550x550 showed a notable increase in performance across all tasks:

- Top-1 accuracy improved to 47.6%, reflecting an enhancement in writer identification.
- GA rose to 68.14%, indicating better script classification.
- CS (α=0) increased to 76.30%, showing improved date classification accuracy.

Random patches likely improved performance by providing localized details within images, which helped the model focus on specific features within a controlled area.

**Preprocessing + Key-points Detection + Patches:** The combination of key-points detection with patch extraction had the most substantial impact on performance:

- Top-1 accuracy reached 85.2%, demonstrating a significant increase in writer identification accuracy.
- GA improved to 83.28% for the script task, nearly doubling the accuracy compared to the baseline.
- CS (α=0) reached 79.59%, showing further improvement in date classification.

The use of key-points detection likely contributed by highlighting crucial regions of interest within the images, allowing the system to focus on meaningful details and discard irrelevant data. This approach provided the best performance across all metrics, reinforcing that advanced preprocessing steps are critical for accurate identification.

The results clearly indicate that incorporating preprocessing steps, particularly key-points detection with patch extraction, significantly improves identification performance across all tasks. This method maximizes relevant data retention and enhances the model's ability to distinguish critical features, leading to more accurate writer, script, and date classifications.

## 6.8 Impact of patch size

The size of image patches is a significant factor that influences model performance, as it determines the level of detail and context available for feature extraction in each task. This experiment assesses the effect of varying patch sizes on three tasks: writer identification, script classification, and dating classification. The results, presented in Table 6.10, highlight how different patch sizes impact accuracy across these tasks.

***Table 6.10*** *patch size Impact on model performance across different tasks*

| **Patches Size** | **Top-1 Writer Task** | **GA Script task** | **CS (a=0) Dating task** |
|---|---|---|---|
| 150 x 150 | 49.3 | 39.83 | 55.58 |
| 224 x 224 | 65.8 | 49.64 | 61.56 |
| **550 x 550** | **85.2** | **71.16** | **66.15** |
| 750 x 750 | 83.4 | 68.14 | 64.01 |

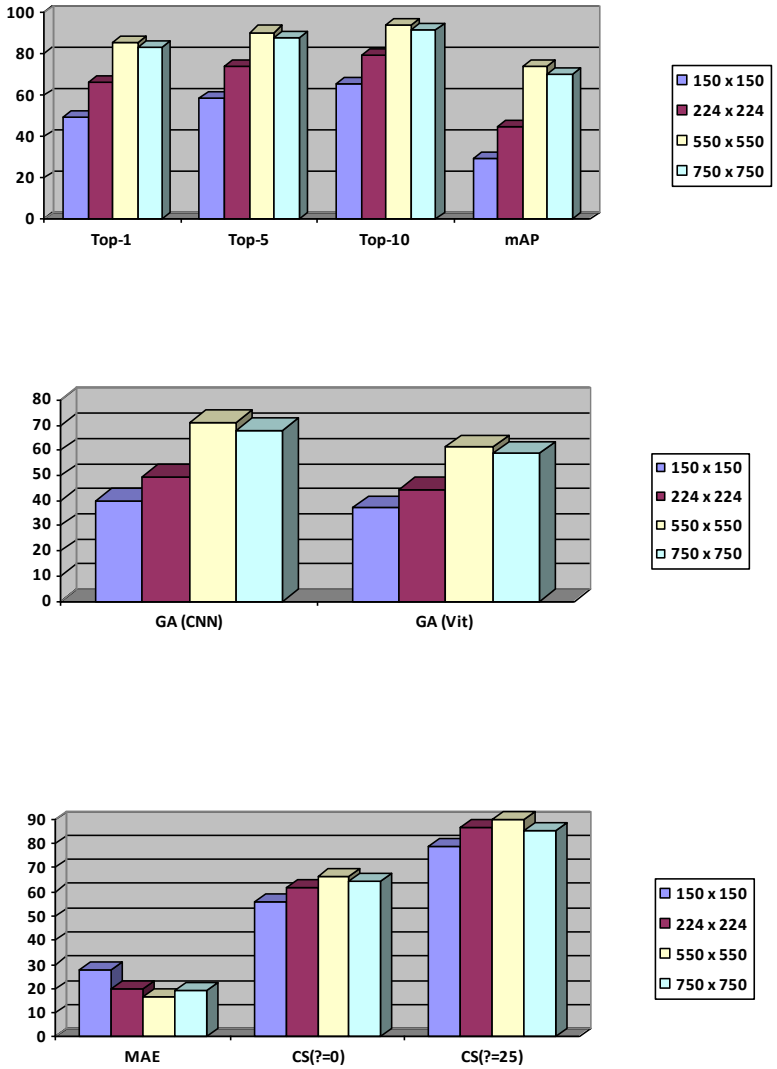



**Figure 6.4** Chart of patch size impact with metrics based on each task (writer, script, date)

## Small Patch Sizes (150 x 150 and 224 x 224):

- **150 x 150:** At this smallest size, all three tasks yielded relatively low accuracies, indicating that the patch size may be insufficient for capturing the necessary details for accurate feature extraction. The Top-1 accuracy for the writer task was 49.3%, and the GA and CS (α=0) were 39.83% and 55.58%, respectively.
- **224 x 224:** Increasing the patch size to 224 x 224 resulted in improved accuracy for all tasks. Top-1 accuracy reached 65.8%, while GA for the script task improved to 49.64%, and CS (α=0) reached 61.56%. This increase suggests that 224 x 224 patches offer more contextual information, helping the model achieve better feature representation.

**Medium Patch Size (550 x 550):** The patch size of 550 x 550 provided the highest performance across all tasks:

- Top-1 accuracy for the writer task was 85.2%, representing a significant improvement in writer identification.
- GA for the script task reached 71.16%, the highest value in this experiment.
- CS (α=0) for the dating task was 66.15%, indicating better performance in identifying historical periods.

The results suggest that this patch size strikes a balance between capturing enough detailed features within a single patch and providing sufficient context for accurate task performance.

**Large Patch Size (750 x 750):** Increasing the patch size to 750 x 750 led to a slight decrease in performance across tasks:

- Top-1 accuracy dropped to 83.4% for the writer task, while GA for the script task was 68.14%.
- CS (α=0) for the dating task decreased to 64.01%.

This decrease in performance suggests that larger patches might incorporate too much contextual information, which can dilute the focus on specific details critical for accurate feature detection, especially in highly detailed images like historical documents.

The results reveal that a patch size of 550 x 550 provides optimal performance across writer, script, and dating tasks, offering the best balance between detail and context. Smaller patches may lack sufficient context for feature extraction, while larger patches can incorporate excess information that may hinder focused feature detection. Therefore, a patch size of 550 x 550 was chosen for subsequent experiments due to its superior accuracy across multiple evaluation metrics.

## 6.9 Impact of number of clusters

In this experiment, we investigate the effect of varying the number of clusters in the k-means clustering algorithm on the performance across different tasks. The k-means clustering algorithm is used to partition the feature space into distinct groups, which can aid in improving feature representation and, subsequently, model performance. The results are shown in graph chart of Figure 6.5.

**Writer Identification (Top-1, Top-5, Top-10, and mAP):**

- As the number of clusters increases from 4 to 16, the model's accuracy improves significantly across all metrics, with Top-1 accuracy increasing from 55.4% to 85.2%, Top-10 accuracy reaching 93.6%, and mAP improving from 36.1 to 73.5.
- The optimal number of clusters for the writer identification task is observed at 16, where the model reaches its highest performance. Increasing clusters beyond 16 (e.g., 20 clusters) does not yield further improvement and instead results in a slight

performance drop, suggesting that higher clustering numbers may lead to over-segmentation or redundancies within feature representation.

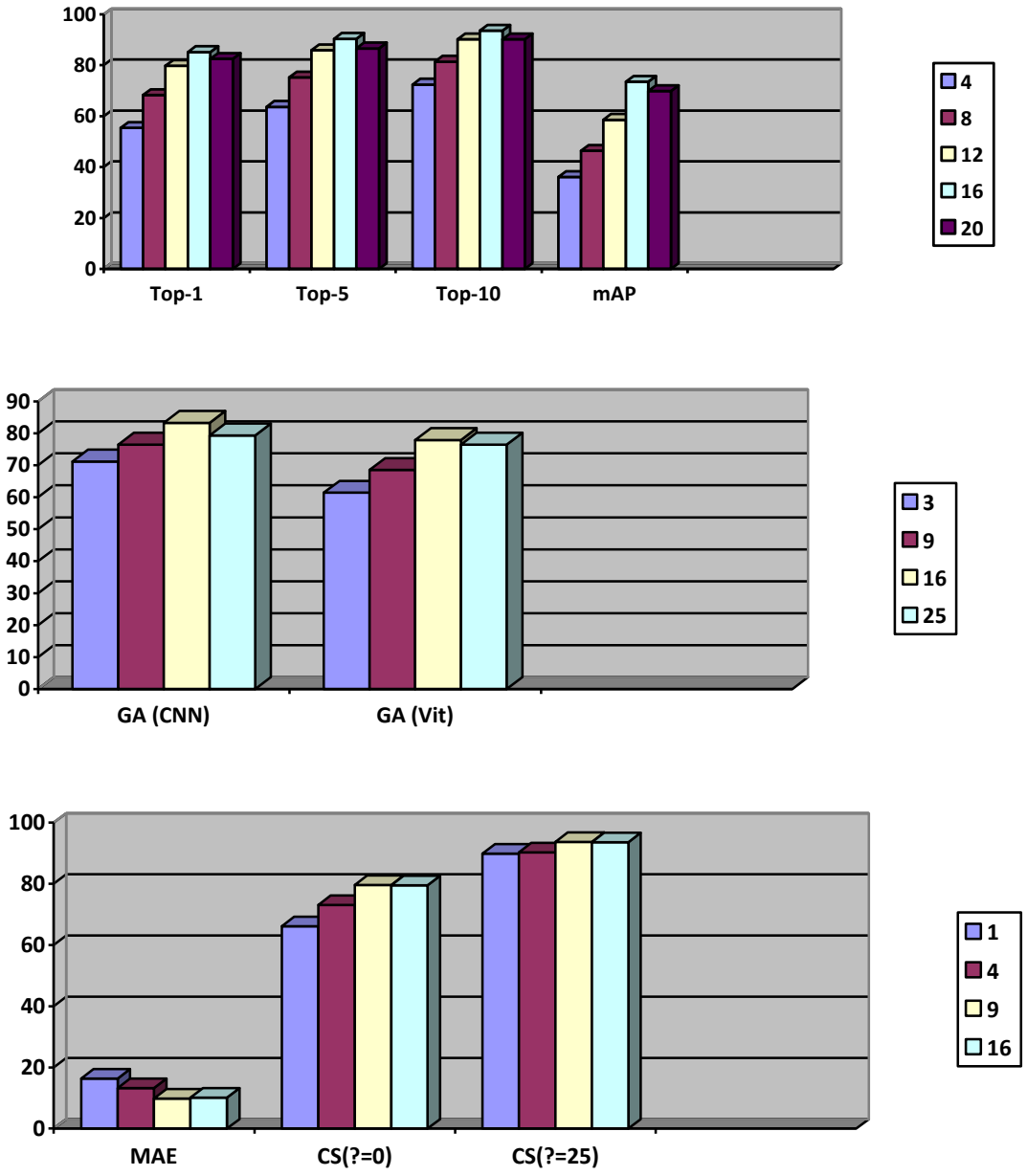



**Figure 6.5** Chart of Clusters number impact with metrics based on each task (writer, script, date)

## Script Classification (GA for CNN and ViT):

- Global Accuracy (GA) shows an upward trend as the number of clusters increases from 3 to 16. For CNN, GA rises from 71.16% to 83.28%, while for ViT, it increases from 61.41% to 77.88%.
- 16 clusters emerges as the ideal number, particularly for CNN, where it achieves the highest GA. Further increasing clusters to 25 does not improve performance and, for ViT, results in a slight decrease in GA. This indicates that, while clustering enhances

feature extraction and boosts performance, excessively high cluster counts can create overly specific groups that might hinder the generalizability of learned features.

**Dating Classification (MAE and CS):**

- The dating task shows optimal performance at 9 clusters, where MAE is minimized to 9.82, and CS (α=0) reaches 79.59, the highest cumulative score observed for this task.
- Increasing clusters from 4 to 9 leads to a significant improvement in both MAE and CS, indicating that finer partitioning benefits date prediction accuracy. However, further increasing clusters to 16 slightly increases the MAE to 10.11 and reduces CS (α=0), suggesting diminishing returns and a potential loss of temporal coherence when over-clustering.

The experiment demonstrates that the number of clusters plays a critical role in optimizing model performance across different tasks. Specifically:

- For writer identification, 16 clusters provides the best accuracy metrics, balancing detail and generalizability.
- For script classification, 16 clusters also yields the highest global accuracy, particularly benefiting CNN architectures.
- For dating classification, 9 clusters achieves optimal results, balancing cluster specificity without sacrificing cumulative scoring accuracy.

These findings suggest that each task benefits from a different optimal number of clusters, tailored to the nature of the features required. For subsequent experiments, 16 clusters will be used for writer and script tasks, while 9 clusters will be selected for the dating task.

## 6.10 Model Selection

The purpose of this experiment is to assess and compare the effectiveness of various pre-trained models on different tasks, using a simplified approach with a single patch per document. This preliminary evaluation provides an initial understanding of which models are best suited for each task, allowing us to select the most promising models for full training on the complete dataset in future stages. The tasks evaluated are writer identification, script classification, and dating estimation.

***Table 6.11*** *Performance on different pre-trained models fir writer task*

| Model | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| VGG16 | 70.4 | 77.8 | 85.1 | 48.9 |
| ResNet50 | 77.3 | 83.8 | 87.2 | 55.8 |
| **Densenet121** | **85.2** | **90.4** | **93.6** | **73.5** |

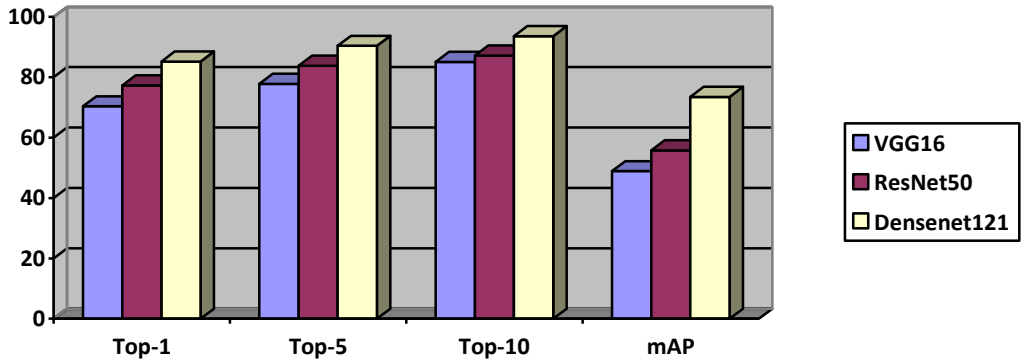



**Figure 6.6** Chart of models performance on writer task

- **Densenet121** outperforms other models, achieving a Top-1 accuracy of 85.2% and a mAP of 73.5. Its Top-5 and Top-10 accuracies also indicate robust performance, with values of 90.4% and 93.6%, respectively.
- **ResNet50** performs moderately well, with a Top-1 accuracy of 77.3% and a mAP of 55.8, though it lags behind Densenet121 across all metrics.
- VGG16 shows the lowest performance, with a Top-1 accuracy of 70.4% and mAP of 48.9. This suggests that VGG16 may struggle with capturing the complex features needed for effective writer differentiation compared to more recent architectures.
- **Conclusion**: Densenet121 emerges as the top-performing model for writer identification, suggesting it has strong feature extraction capabilities suitable for distinguishing between different writers in this dataset.

***Table 6.12*** *Model's experiment results for script task.*

| **Model** | **GA** |
|---|---|
| Vit-Ti | 61.41 |
| Vit-S | 68.43 |
| Swin-S | 72.59 |
| Resnet50 | 71.16 |
| **Densenet121** | **73.02** |

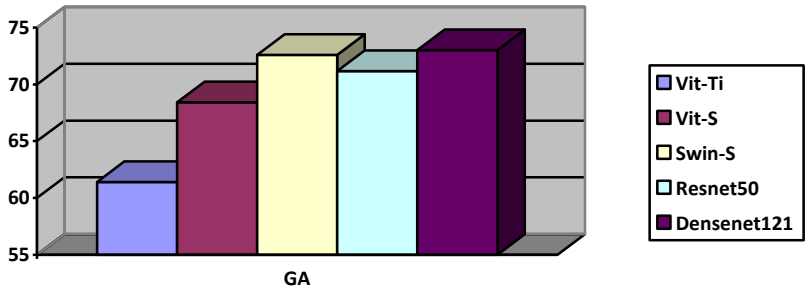



**Figure 6.7** Chart of models performance on Script task

- **Densenet121** again performs best, with the highest Global Accuracy (GA) of 73.02%, closely followed by Swin-S with a GA of 72.59%.
- **ResNet50** and Vit-S also demonstrate solid performance, with GA values of 71.16% and 68.43%, respectively, though they do not match the accuracy achieved by Densenet121 and Swin-S.
- **Vit-Ti** has the lowest GA of 61.41%, indicating it may lack the complexity or capacity to handle the nuances of script classification effectively.
- **Conclusion**: Densenet121 and Swin-S are the best-performing models for script classification, indicating their architectures are well-suited for this task's requirements. These models will likely perform even better when trained on the full dataset.

***Table 6.13*** *Performance on different pre-trained models for dating task.*

| **Model** | **MAE** | **CS(α=0)** | **CS(α=25)** |
|---|---|---|---|
| Vit-Ti | 16.33 | 66.15 | 89.85 |
| Vit-S | 15.28 | 70.59 | 90.01 |
| Vit-B | 17.71 | 64.77 | 88.20 |
| DeiT-S | 14.71 | 71.51 | 91.11 |
| **Swin-S** | **11.47** | **73.66** | **93.41** |

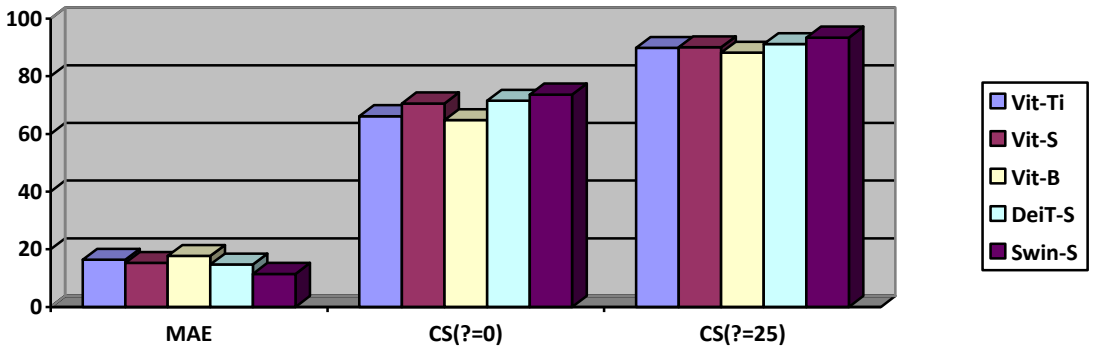



**Figure 6.8** Chart of models performance on Dating task

- **Swin-S** achieves the best results in dating classification, with a Mean Absolute Error (MAE) of 11.47 and the highest Cumulative Scores: CS (α=0) of 73.66 and CS (α=25) of 93.41.
- **DeiT-S** also shows good performance, with an MAE of 14.71 and cumulative scores of 71.51 and 91.11, respectively.
- **Vit-S** performs comparably well, while Vit-B and Vit-Ti have lower cumulative scores, suggesting that these models may be less effective at learning chronological features from the dataset.
- **Conclusion:** Swin-S and DeiT-S emerge as the most suitable models for dating estimation. Swin-S, in particular, provides strong accuracy across all cumulative score metrics, suggesting it has a sophisticated feature extraction approach conducive to dating predictions.

Full dataset training with these models is expected to further enhance performance, leveraging the high-quality feature extraction observed in these initial trials.

## 6.11 Final system results

This section presents the final evaluation of our classification systems for writer identification, script classification, and dating tasks, using all optimal parameters and hyperparameters identified from prior experiments. Here, we compare patch-level and document-level performance, the latter obtained through majority voting across patches. The

results, shown in Tables 6.14, 6.15, and 6.16, reveal notable insights into model performance and improvements with document-level aggregation.

***Table 6.14*** *Final system results for writer task (Patch-level vs. Document-level performance)*

| Experiment | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| **Patch-level** | **77.8** | **82.2** | **88.3** | **61.6** |
| **Document-level (majority vote)** | **85.2** | **90.4** | **93.6** | **73.5** |

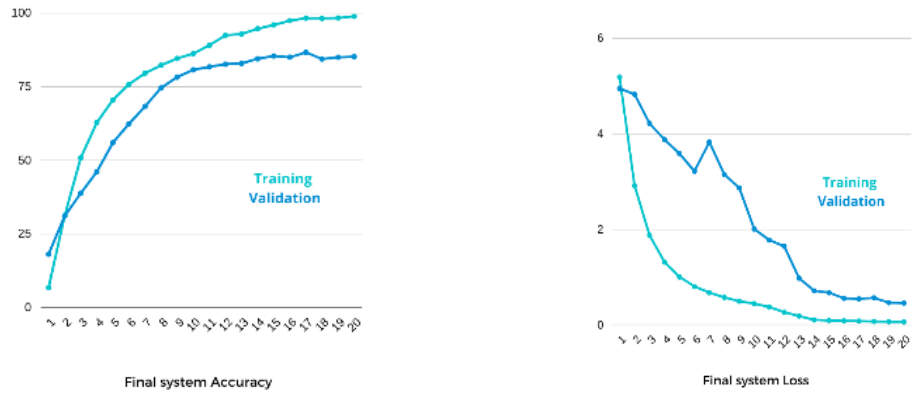



**Figure 6.9** Final system Loss and Accuracy for writer task.

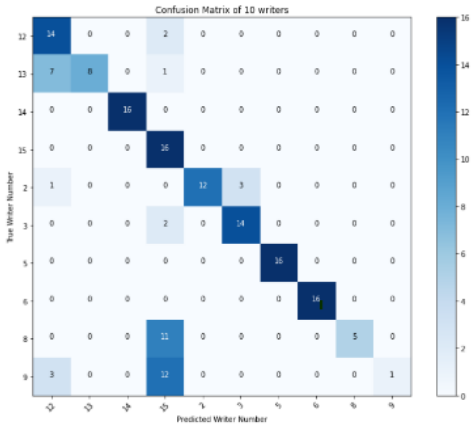



**Figure 6.10** Confusion Matrix of 10 classes for writer task.

- Patch-level performance achieves a Top-1 accuracy of 77.8% and mAP of 61.6, while document-level aggregation improves Top-1 accuracy to 85.2% and mAP to 73.5%.
- The increase in performance through majority voting at the document level demonstrates the model's improved reliability when incorporating multiple patches, suggesting that the system benefits from aggregated context across different parts of the document.

Figures 6.9 and 6.10 depict the final training loss and accuracy trends and a confusion matrix focusing on 10 writer classes. The matrix highlights performance patterns:

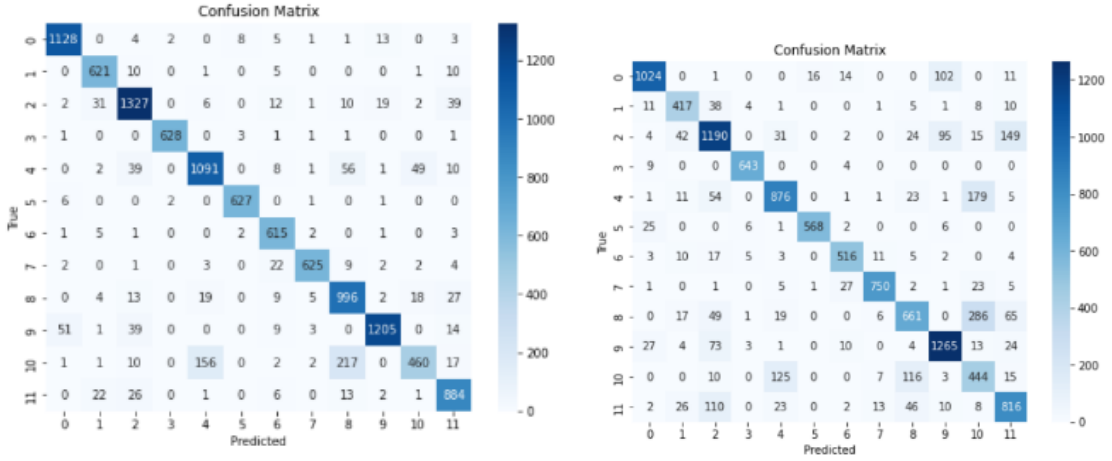
- **Accurate predictions:** Writers like writer 14 were consistently recognized across all samples.

- **Partial recognition:** Some writers, such as writer 2, were only partially recognized, suggesting some ambiguities in the system's classifications.
- **Low recognition:** For instances like writer 9, recognition was largely unsuccessful, indicating areas for further refinement in distinguishing similar styles.

Quantitative metrics for these specific classes — Precision: 87.47%, Recall: 73.75%, and F1 Score: 71.82% — provide a more granular look at the model's effectiveness in specific cases, showing strengths and highlighting areas needing improvement. The system achieved high training accuracy (98%) and validation accuracy (85%), with a validation loss of 0.06, indicating strong generalization while avoiding overfitting.

***Table 6.15*** *Final system results for Script task (Document-level performance)*

| **Model** | **Vit (Swin-s)** | CNN (Densenet-121) |
|---|---|---|
| **Performance** | **91.99** | 89.2 |




(a) (b)

**Figure 6.11** Confusion Matrix for script task, (a) Vit, (b) CNN.

Swin-S (Vision Transformer) achieves an accuracy of 91.99%, outperforming DenseNet121 (89.2%). This result underscores the effectiveness of Vision Transformers in tasks that require precise pattern recognition, particularly on script datasets.

Figure 6.11 visualizes the model's loss and accuracy, alongside a confusion matrix showcasing detailed performance for different script classes. Performance varies by script:

- **High Accuracy Scripts:** Scripts like Uncial, Humanistic, and Praegothica exceed 97% accuracy, showing the model's high proficiency in identifying these distinct styles.
- **Moderate Accuracy Scripts:** Scripts such as Half-uncial and Humanistic Cursive achieve accuracies between 88% and 94%, though with some room for improvement.
- **Low Accuracy Scripts:** Caroline and Southern Textualis scripts show accuracies between 86% and 89%, reflecting some difficulty in distinguishing similar styles.

- **Very Low Accuracy Scripts:** Hybrida Semihybrida and certain other classes, particularly Class 12, exhibit lower accuracies, with significant gaps in accurate prediction, suggesting that these scripts may require further feature refinement or additional data for better recognition.

***Table 6.16*** *Final system results for Dating task (Patch-level vs. Document-level performance)*

| **Results** | **MAE** | **CS(α=0)** | **CS(α=25)** |
|---|---|---|---|
| **Patch-level** | **7.42** | **83.21** | **95.80** |
| **Document-level (majority vote)** | **3.70** | **88.06** | **97.47** |

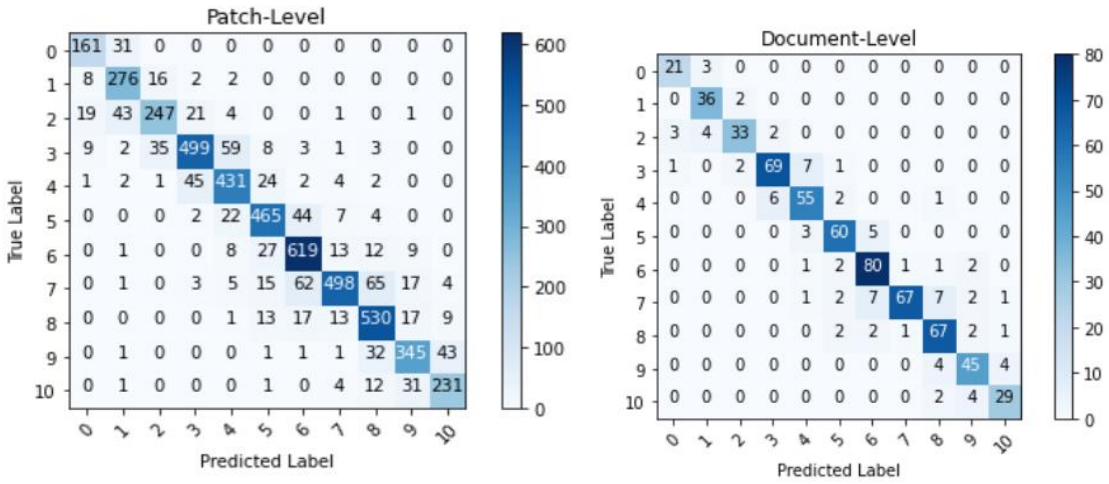



**Figure 6.12** Confusion matrix fir dating tasks at both patch and document levels).

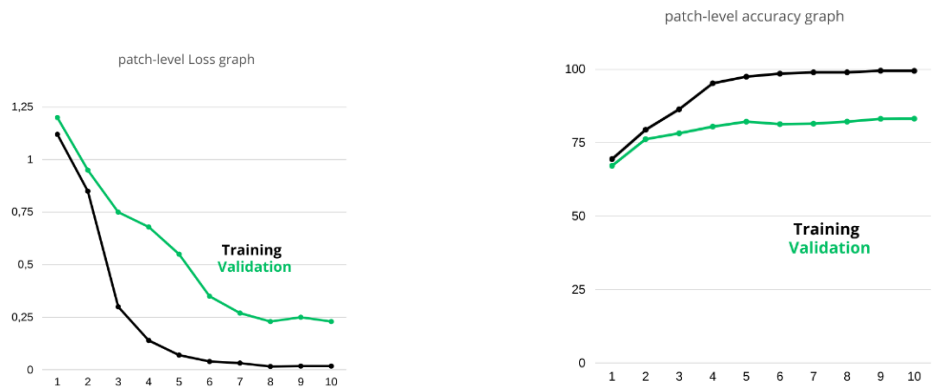



**Figure 6.13** Graphs illustrating the final system's loss and accuracy at the patch-level for dating task.

- **Document-level** (majority vote) accuracy greatly improves Mean Absolute Error (MAE) and Cumulative Scores (CS).

- **Patch-level** results show an MAE of 7.42 and CS (α=0) of 83.21%, while document-level performance improves these metrics to an MAE of 3.70 and CS (α=0) of 88.06%.
- **Figure 6.12** provides a confusion matrix for validation set performance, with observations of accurate classifications for specific classes (e.g., Class 5 with 92% accuracy).
- **Misclassifications** generally occur in adjacent classes, suggesting that even when errors occur, the predicted dates are close to the actual dates, indicating the system's proximity in predictions.

Figure 6.13 illustrates the model's training progression at the patch level:

- **Reduction in Loss:** The decreasing trend in loss signifies the model's ability to refine its predictions over time.
- **Increase in Accuracy:** This consistent improvement demonstrates the model's increasing precision, effectively mitigating overfitting risks common in historical document dating tasks.

## 6.12 Performance comparison

To assess the effectiveness of our proposed classification system across writer identification, script classification, and dating tasks, we compared our results with those of several prior studies. The performance comparison, summarized in Tables 6.17, 6.18, and 6.19, demonstrates the advantages of our approach over existing methods

***Table 6.17*** *Performance comparison our system for writer task with other techniques*

| Method | Top-1 | Top-10 | mAP |
|---|---|---|---|
| **Proposed Method** | **77.8** | **88.3** | **61.6** |
| | **85.2** | **93.6** | **73.5** |
| moment-based distance features matching [106] | 78.75 | 91.08 | 58.62 |
| Combined basic image features [145] | 77.39 | - | 56.82 |
| OBIF [145] | 76.4 | - | 55.6 |
| SRS-LBP [145] | 67.0 | 80.1 | 45.9 |
| SIFT+NBNN [145] | 67.1 | 80.2 | 46.9 |
| Co-Hinge Featuers [146] | 76.1 | 85.8 | 54.2 |
| AngU-Net + R-SIFT [146] | 48 | 92 | 42.8 |
| TwoPathwriter [147] | 77.1 | / | 33.5 |

Our proposed method outperformed other techniques in Top-1, Top-10 accuracy, and mAP:

- **Top-1 Accuracy:** Our document-level model achieved 85.2% Top-1 accuracy, which is higher than the moment-based distance features matching (78.75%) and Co-Hinge Features (76.4%).
- **Top-10 Accuracy:** We reached 93.6% Top-10 accuracy, outperforming TwoPathwriter (77.1%) and AngU-Net + R-SIFT (92%).
- **mAP:** Our method achieved an mAP of 73.5%, surpassing techniques like Combined Basic Image Features (56.82%) and Co-Hinge Features (55.6%).

These results show that our approach is more effective at distinguishing between writers, suggesting the robustness of our feature extraction and classification techniques.

***Table 6.18*** *Performance comparison our system for script task with other techniques*

| **Method** | **Global Accuracy** |
|---|---|
| **Proposed Method (CNN)** | **89.2** |
| **Proposed Method (Vit)** | **91.99** |
| NNML [93] | 83.80 |
| FCDR-OCR [93] | 79.80 |
| DeepScript [93] | 76.79 |
| S2-CCMV [94] | 56.55 |
| FAU [93] | 83.90 |
| The North LTU [96] | 73.96 |
| T-DeepCNN [94] | 85.20 |
| P-CNN [94] | 73.25 |
| CK2 [94] | 78.05 |

Our system also showed superior performance in script classification:

- **Accuracy:** Our Vision Transformer-based (ViT) method achieved 91.99% accuracy, while the CNN-based variant reached 89.2%, outperforming all compared techniques. Notably, FAU achieved 83.90%, and T-DeepCNN reached 85.20%.

- **Comparison with Prior Methods:** Techniques like DeepScript (76.79%) and S2-CCMV (56.55%) exhibited significantly lower accuracy, emphasizing the effectiveness of our proposed approach, especially with ViT.

The higher accuracy of our method in script classification tasks highlights its ability to capture and classify intricate patterns in historical script data better than traditional and prior deep learning-based approaches.

***Table 6.19*** *Performance comparison our system for dating task with other techniques*

| **Method** | **MAE** | **CS(α=25)** |
|---|---|---|
| **Proposed Method** | **3.70** | **97.47** |
| [89] 2014 | 35.4 | 25 |
| [120] 2016 | 7.9 | 93.2 |
| [121] 2016 | 9.1 | 91.4 |
| [122] 2016 | 7.8 | 93.3 |
| [123] 2018 | 20.3 | - |

Our system demonstrated strong performance in document dating:

- **Mean Absolute Error (MAE):** Our model achieved an MAE of 3.70, a substantial improvement over the best prior methods. For example, studies from 2016 by [120] and [122] reported MAEs of 7.9 and 7.8, respectively.
- **Cumulative Score** (CS) at α=25: We achieved a CS(α=25) of 97.47%, which is significantly higher than the results from studies like [120] (93.2%) and [122] (93.3%).

The lower MAE and higher cumulative score indicate that our dating model produces more precise age estimates, demonstrating the advantages of our model's design, which effectively minimizes the error margin compared to earlier methods.

Our proposed classification system exhibits superior performance across all tasks, consistently outperforming previous techniques. The high Top-1 accuracy in writer identification, high accuracy in script classification, and low MAE in dating reflect our model's strong generalization and robustness.

## 6.13 Conclusion

In this chapter, we have conducted a series of experiments to evaluate and refine our system. Through these experiments, we fine-tuned the parameters and hyperparameters of various components, leading to the optimization of our system's performance. Our observations highlight the critical impact that these parameters have on the overall effectiveness of the system. After refining these aspects, we presented the final results of our system, demonstrating its capabilities. Finally, we compared our system's performance with similar studies in the literature, providing context and validation for our approach. These experiments enhanced the performance of our system and underscored the importance of careful parameter selection in achieving optimal results.

# Conclusion & Perspectives

In concluding this study, titled "Contribution to Historical Documents Classification," we have explored the multifaceted concepts surrounding the classification of historical documents from various perspectives.

We began by defining what historical documents are and underscoring the importance of studying them. This foundational understanding set the stage for grasping the significance of classifying these invaluable documents, which serve as crucial links to our past.

Next, we examined the classification of historical documents from the viewpoints of historians and paleographers, identifying the unique challenges and problems they face in their work. These challenges highlighted the need for more advanced, systematic approaches to classification.

We then explored the existing methodologies for classifying historical documents, including physical, paleographical, and computer-based methods. Within the computer-based approaches, we distinguished between image-based and text-based methods, providing a comprehensive overview of the current landscape in document classification.

In reviewing the literature, we encountered various studies and their associated challenges, which informed the direction of our contribution. Our work culminated in the development of an end-to-end system for the classification of historical documents.

Our system is designed to tackle each of these key challenges through a structured approach comprising several critical steps:

- **Preprocessing:** The first step in our methodology involves preprocessing the historical documents. We employ various denoising techniques to improve the overall quality of the images. We aim to facilitate better feature detection and extraction in subsequent stages.

- **Feature extraction:** Once the documents are preprocessed, we focus on extracting meaningful features that are relevant to our classification goals. We utilize a specialized feature extraction method tailored specifically for historical document analysis. This method incorporates techniques to capture distinctive patterns and characteristics in the documents.

- **Classification:** With a high-quality dataset and selected samples in hand, we employ advanced classification algorithms to classify the historical documents accurately. Our system utilizes state-of-the-art classifiers to achieve the highest classification scores possible. We aim to improve the accuracy and reliability of the classification outcomes.

Our systematic approach encompasses **preprocessing for image enhancement**, **feature extraction for capturing relevant patterns**, and **robust classification techniques for accurate categorization** of historical documents. Through these steps, our system provide an effective and reliable solution that contributes significantly to this problem. The results of our contribution demonstrate the effectiveness of our approach in improving the accuracy and reliability of document classification. However, our work does not end here.

Looking forward, we aim to conduct further research to enhance the accuracy of our system.

We also plan to design and develop a user-friendly tool for the classification of historical documents, which will be of practical use to historians and researchers.

Finally, we intend to integrate semantic analysis with our visual methods, creating a more robust and comprehensive system that combines the strengths of both approaches to further improve classification performance.

This study has laid the groundwork for significant advancements in the field of historical document classification, and we look forward to continuing this important work.

# References


[1] Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. Ieee Access, 8, 75264-75278.

[2] Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine, 7, 509744.

[3] Wei, S., Huang, P., Li, R., Liu, Z., & Zou, Y. (2021). Exploring the application of artificial intelligence in sports training: a case study approach. Complexity, 2021(1), 4658937.

[4] Kansteiner, W. (2022). Digital doping for historians: can history, memory, and historical theory be rendered artificially intelligent?. History and Theory, 61(4), 119-133.

[5] Terras, M. (2022). Inviting ai into the archives: The reception of handwritten recognition technology into historical document transcription. Digital Humanities Research| Volume 2, 179.

[6] Lasker, D. J., & Olszowy-Schlanger, J. (2013). The importance of documents for the study of Early Modern Eastern European Karaism. Documents hébreux et arabes. Mélanges en l'honneur de Colette Sirat, 403-412.

[7] Fedi M (2009) Accelerator mass spectrometry for 14C dating. In: Colombini MP, Modugno F (eds)Organic mass spectrometry in art and archaeology. Wiley, Chichester, pp 459-482

[8] Abdel-Maksoud, G., Emam, H., & Ragab, N. M. (2020). From traditional to laser cleaning techniques of parchment documents: A review. Advanced Research in Conservation Science, 1(1), 52-76.

[9] Andrist, P. (2021). Toward a new generation of databases and database applications for describing ancient documents. Digital Scholarship in the Humanities, 36(Supplement_2), ii9-ii16.

[10] Sulaiman, A., Omar, K., & Nasrudin, M. F. (2017, November). A database for degraded Arabic historical documents. In 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI) (pp. 1-6). IEEE.

[11] Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., & Ingold, R. (2016, October). Diva-hisdb: A precisely annotated large dataset of challenging medieval documents. In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 471-476). IEEE.

[12] Aljalbout, S., & Falquet, G. (2018). A semantic model for historical documents. arXiv preprint arXiv:1802.00295.

[13] Yahia, M. H. N. (2011). Content-based retrieval of Arabic historical documents using latent semantic indexing (Master's thesis, King Fahd University of Petroleum and Minerals (Saudi Arabia)).

[14] Reynolds, T., Dhali, M. A., & Schomaker, L. (2022). Image-based material analysis of ancient historical documents. arXiv preprint arXiv:2203.01042.

[15] Wolf, L., Potikha, L., Dershowitz, N., Shweka, R., & Choueka, Y. (2011, September). Computerized paleography: tools for historical documents. In 2011 18th IEEE International Conference on Image Processing (pp. 3545-3548). IEEE.

[16] Chazalon, J., Carlinet, E., Chen, Y., Perret, J., Duménieu, B., Mallet, C., ... & Král, P. (2021,

September). ICDAR 2021 competition on historical map segmentation. In International Conference on Document Analysis and Recognition (pp. 693-707). Cham: Springer International Publishing.

[17] Mehri, M., Héroux, P., Mullot, R., Moreux, J. P., Coüasnon, B., & Barrett, B. (2019, September). ICDAR2019 competition on historical book analysis-HBA2019. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1488-1493). IEEE.

[18] Antonacopoulos, A., Clausner, C., Papadopoulos, C., & Pletschacher, S. (2013, August). ICDAR 2013 competition on historical book recognition (HBR 2013). In 2013 12th International Conference on Document Analysis and Recognition (pp. 1459-1463). IEEE.

[19] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., ... & de Freitas, N. (2023). Machine learning for ancient languages: A survey. Computational Linguistics, 49(3), 703-747.

[20] Lombardi, F., & Marinai, S. (2020). Deep learning for historical document analysis and recognition—a survey. Journal of Imaging, 6(10), 110.

[21] Bennour, A., Boudraa, M., Siddiqi, I. et al. A deep learning framework for historical documents writer identification using data-driven features. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-18187-y

[22] Boudraa, M., Bennour, A. (2024). Combination of Local Features and Deep Learning to Historical documents Dating. In: Bennour, A., Bouridane, A., Chaari, L. (eds) Intelligent Systems and Pattern Recognition. ISPR 2023. Communications in Computer and Information Science, vol 1940. Springer, Cham. https://doi.org/10.1007/978-3-031-46335-8_11

[23] Boudraa, M., Bennour, A., Al-Sarem, M., Ghabban, F., & Bakhsh, O. A. (2024). Contribution to historical document dating: A hybrid approach employing hand-crafted features with vision transformers. Digital Signal Processing, 149, 104477.

[24] Boudraa, M., Bennour, A., Mekhaznia, T., Alqarafi, A., Marie, R. R., Al-Sarem, M., & Dogra, A. (2024). Revolutionizing Historical Document Analysis: A Deep Learning Approach with Intelligent Feature Extraction for Script Classification. Acta Informatica Pragensia, 13(2), 251-272.

[25] Bennour, A., Boudraa, M., & Ghabban, F. HeritageScript: A cutting-edge approach to historical document script classification with CNN and vision transformer architectures. Intelligent Decision Technologies, (Preprint), 1-24.

[26] Merriam-Webster. (n.d.). Historical. In Merriam-Webster.com dictionary. Retrieved July 8, 2024, from https://www.merriam-webster.com/dictionary/historical

[27] Merriam-Webster. (n.d.). Document. In Merriam-Webster.com dictionary. Retrieved July 8, 2024, from https://www.merriam-webster.com/dictionary/document

[28] Cappon, L. J. (1956). Historical documents as Archives: Some Definitions and Their Application. The American Archivist, 19(2), 101–110. http://www.jstor.org/stable/40289525

[29] Wakelin, D. (2017). Paleography. The Encyclopedia of Medieval Literature in Britain, 1-6.

[30] Griffin, S. M. (2023). Epigraphy and paleography: bringing records from the distant past to the present. International Journal on Digital Libraries, 24(2), 77-85.

[31] Ciula, A. (2017). Digital palaeography: What is digital about it?. Digital Scholarship in the Humanities, 32(suppl_2), ii89-ii105.

[32] Webmaster. (n.d.). IWCP 2024. https://www.csmc.uni-hamburg.de/iwcp2024.html

[33] Lombardi, F., & Marinai, S. (2020). Deep learning for historical document analysis and recognition—a survey. Journal of Imaging, 6(10), 110.

[34] Spectroscopy Matters. (n.d.-b). https://www.horiba.com/int/scientific/resources/spectroscopy-matters

[35] Hajji, L., Boukir, A., Assouik, J., Lakhiari, H., Kerbal, A., Doumenq, P., ... & De Carvalho, M. L. (2015). Conservation of Moroccan document papers aged 150, 200 and 800 years. Analysis by infrared spectroscopy (ATR-FTIR), X-ray diffraction (XRD), and scanning electron microscopy energy dispersive spectrometry (SEM–EDS). Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 136, 1038-1046.

[36] Nesměrák, K., & Němcová, I. (2012). Dating of historical documents using spectrometric methods: a mini-review. Analytical letters, 45(4), 330-344.

[37] Sistach, M. C., Ferrer, N., & Romero, M. T. (1998). Fourier transform infrared spectroscopy applied to the analysis of ancient documents.

[38] Pessanha, S., Manso, M., & Carvalho, M. L. (2012). Application of spectroscopic techniques to the study of illuminated documents: a survey. Spectrochimica Acta Part B: Atomic Spectroscopy, 71, 54-61.

[39] A Comprehensive Guide to the Light Microscope - How to Use a Light Microscope. (n.d.). https://www.microscopeworld.com/p-4463-a-comprehensive-guide-to-the-light-microscope-how-to-use-a-light-microscope.aspx

[40] Clark, R. J., & Gibbs, P. J. (1998). Analysis of 16th century Qazwını documents by Raman microscopy and remote laser Raman microscopy. Journal of archaeological science, 25(7), 621-629.

[41] Hamed, S. A. E. K. M., & Ali Hassan, R. R. (2019). Microscopic characterization of acidic paper documents. Pigment & Resin Technology, 48(2), 119-128.

[42] Hajdas, I., Ascough, P., Garnett, M.H. et al. Radiocarbon dating. Nat Rev Methods Primers 1, 62 (2021). https://doi.org/10.1038/s43586-021-00058-7

[43] Aghaei, A., Bahrololoumi, F., Hajdas, I., Jafarian, R., Kordavani, L., & Marx, M. (2023). Radiocarbon Dating Of Documents Kept In The Central Library Of The University Of Tehran. Radiocarbon, 65(2), 307-321.

[44] Brock, F. (2013). Radiocarbon dating of historical parchments. Radiocarbon, 55(2), 353-363.

[45] Marx, M. J., & Jocham, T. J. (2019). Radiocarbon (14C) dating of Qurʾān documents. In Qurʾān Quotations Preserved on Papyrus Documents, 7th-10th Centuries (pp. 188-221). Brill.

[46] Nicolis, O., & Gonzalez, C. (2021). Wavelet-based fractal and multifractal analysis for detecting mineral deposits using multispectral images taken by drones. In Methods and Applications in Petroleum and Mineral Exploration and Engineering Geology (pp. 295-307). Elsevier.

[47] Rahiche, A., Hedjam, R., Al-Maadeed, S., & Cheriet, M. (2020). Historical documents dating using multispectral imaging and ordinal classification. Journal of Cultural Heritage, 45, 71-80.

[48] Hollaus, F., Diem, M., Fiel, S., Kleber, F., & Sablatnig, R. (2015, September). Investigation of ancient documents based on multispectral imaging. In Proceedings of the 2015 ACM Symposium on Document Engineering (pp. 93-96).

[49] Mazzocato, S., Cimino, D., & Daffara, C. (2024). Integrated microprofilometry and multispectral

imaging for full-field analysis of ancient documents. Journal of Cultural Heritage, 66, 110-116.

[50] paleography. (n.d.). In Merriam-Webster Dictionary. https://www.merriam-webster.com/dictionary/paleography

[51] Longacre, D. (2018). Reconsidering the Date of the En-Gedi Leviticus Scroll (EGLev): Exploring the Limitations of the Comparative-Typological Paleographic Method. Textus, 27(1), 44-84.

[52] Merriam-Webster. (n.d.). Codicology. In Merriam-Webster.com dictionary. Retrieved July 14, 2024, from https://www.merriam-webster.com/dictionary/codicology

[53] Herde, P. (2022, August 12). diplomatics. Encyclopedia Britannica. https://www.britannica.com/topic/diplomatics

[54] Tereschenko, E. Y., Vasiliev, A. L., Lyakhovitsky, E. A., Balachenkova, A. P., Tsypkin, D. O., & Yatsishina, E. B. (2021). Electron Microscopy Studies of Documents from Russian National Library. Crystallography Reports, 66, 1143-1149.

[55] info@nireos.com. (2024). What is Hyperspectral Imaging | Nireos. Nireos - Understand the Invisible. https://nireos.com/application/what-is-hyperspectral-imaging/

[56] HMML Palimpsest Project. (n.d.). https://hmml.org/research/palimpsest/

[57] France, Haguenau, Médiathèque de la Vieille-Île - 1. 2. (n.d.-b). arca.irht.cnrs.fr. https://arca.irht.cnrs.fr/ark:/63955/md67wm11bc80

[58] Leydier, Y., Le Bourgeois, F., Emptoz, H. (2004). Serialized k-Means for Adaptative Color Image Segmentation. In: Marinai, S., Dengel, A.R. (eds) Document Analysis Systems VI. DAS 2004. Lecture Notes in Computer Science, vol 3163. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28640-0_24

[59] Tensmeyer, C., & Martinez, T. (2020). Historical document image binarization: A review. SN Computer Science, 1(3), 173.

[60] Otsu N. A threshold selection method from gray-level histograms. Trans Syst Man Cybern. 1979;9(1):62–6.

[61] Sauvola J, Pietikäinen M. Adaptive document image binarization. Pattern Recognit. 2000;33(2):225–36.

[62] Canny J. A computational approach to edge detection. Trans Pattern Anal Mach Intell. 1986;8:679–98.

[63] Sehad A, Chibani Y, Hedjam R, Cheriet M. Gabor flter-based texture for ancient degraded document image binarization. Pattern Anal Appl. 2018;1–22.

[64] Lelore T, Bouchara F. Fair: a fast algorithm for document image restoration. Trans Pattern Anal Mach Intell.2013;35(8):2039–48.

[65] Howe NR. Document binarization with automatic parameter tuning. Int J Doc Anal Recognit (IJDAR). 2013;16(3):247–58

[66] Jindal, A., Ghosh, R. Text line segmentation in indian ancient handwritten documents using faster R-CNN. Multimed Tools Appl 82, 10703–10722 (2023). https://doi.org/10.1007/s11042-022-13709-y

[67] Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Abstracts of the sixth international conference on computer vision IEEE, Bombay, India, pp 839–846. https://doi.org/10.1109/ICCV.1998.71081

[68] Yang GZ, Burger P, Firmin DN, Underwood SR (1996) Structure adaptive anisotropic image filtering. Image Vis Comput 14(2):135–145. https://doi.org/10.1016/0262-8856(95)01047-5

[69] Hou JH (2007) Research on image denoising approach based on wavelet and its statistical characteristics. Dissertation, Huazhong University of Science and Technology

[70] Jiao LC, Hou B, Wang S, Liu F (2008) Image multiscale geometric analysis: theory and applications. Xidian University press, Xi'an

[71] Benesty J, Chen JD, Huang YT (2010) Study of the widely linear wiener filter for noise reduction. In: Abstracts of IEEE international conference on acoustics, speech and signal processing, IEEE, Dallas, TX, USA, pp 205–208. https://doi.org/10.1109/ICASSP.2010.5496033

[72] Zhang KB, Gao XB, Tao DC, Li XL (2012) Multi-scale dictionary for single image super-resolution. In: Abstracts of 2012 IEEE conference on computer vision and pattern recognition. IEEE, Providence, pp 1114–1121. https://doi.org/10.1109/CVPR.2012.6247791

[73] Chen YY, Pock T (2017) Trainable nonlinear reaction diffusion: a flexible framework for fast and effective image restoration. IEEE Trans Pattern Anal Mach Intell 39(6):1256–1272. https://doi.org/10.1109/TPAMI.2016.2596743

[74] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ... & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing, 39(3), 355-368.

[75] He, S., Samara, P., Burgers, J., & Schomaker, L. (2016). Historical document dating based on temporal pattern codebook. Computer Vision and Image Understanding, 152, 167-175. https://doi.org/10.1016/j.cviu.2016.08.008

[76] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1, doi:10.1109/CVPR.2005.177.

[77] Iqbal N, Mumtaz R, Shafi U, Zaidi SMH. 2021. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms. PeerJ Computer Science 7:e536 https://doi.org/10.7717/peerj-cs.536

[78] Pietikäinen, M. (2010). Local binary patterns. Scholarpedia, 5(3), 9775.

[79] Mehrotra, R., Namuduri, K. R., & Ranganathan, N. (1992). Gabor filter-based edge detection. Pattern recognition, 25(12), 1479-1494.

[80] O'shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[81] Alsaleh A, Perkgoz C. 2023. A space and time efficient convolutional neural network for age group estimation from facial images. PeerJ Computer Science 9:e1395 https://doi.org/10.7717/peerj-cs.1395

[82] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[83] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[84] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015).

Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[85] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.

[86] Jian Zheng, Cencen Xu, Ziang Zhang and Xiaohua Li, "Electric load forecasting in smart grids using Long-Short-Term-Memory based Recurrent Neural Network," 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 2017, pp. 1-6, doi: 10.1109/CISS.2017.7926112.

[87] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[88] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[89] 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

[90] Svenskt Diplomatariums huvudkartotek (SDHK), https://sok.riksarkivet.se/SDHK.

[91] Adam K, Baig A, Al-Maadeed S, Bouridane A, El-Menshawy S (2018) KERTAS: Dataset For automatic dating of ancient arabic documents. International Journal on Document Analysis and Recognition (IJDAR) 21:283–290. https://doi.org/10.1007/s10032-018-0312-3

[92] Gervers M (2000a) The DEEDS project and the development of a computerised methodology for dating undated English private charters of the twelfth and thirteenth centuries. In: Gervers M (ed) Dating undated medieval charters. Boydell Press, Woodbridge, pp 13–35

[93] Cloppet F, Eglin V, Kieu VC, Stutzmann D, Vincent N (2016) ICFHR2016 Competition on the classification of medieval handwritings in latin script. In: Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR)

[94] Cloppet F, Eglin V, Helias-Baron M, Kieu VC, Stutzmann D, Vincent N (2017) ICDAR 2017 competition on the classification of medieval handwritings in latin script. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 2017, Kyoto, pp 1371–1376. https://doi.org/10.1109/ICDAR.2017.224

[95] Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Nikos, S., & Gatos, B. (2017, November). Icdar2017 competition on historical document writer identification (historical-wi). In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 1377-1382). IEEE.

[96] Seuret, M., Nicolaou, A., Rodríguez-Salas, D., Weichselbaumer, N., Stutzmann, D., Mayr, M., ... & Christlein, V. (2021, September). ICDAR 2021 competition on historical document classification. In International Conference on Document Analysis and Recognition (pp. 618-634). Cham: Springer International Publishing.

[97] Christlein, V., Nicolaou, A., Seuret, M., Stutzmann, D., & Maier, A. (2019, September). ICDAR 2019 competition on image retrieval for historical handwritten documents. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1505-1509). IEEE.

[98] Seuret, M., Nicolaou, A., Stutzmann, D., Maier, A., & Christlein, V. (2020, September). ICFHR 2020 competition on image retrieval for historical handwritten fragments. In 2020 17th International conference on frontiers in handwriting recognition (ICFHR) (pp. 216-221). IEEE.

[99] Mohammed, H., Marthot-Santaniello, I., & Märgner, V. (2019, September). GRK-Papyri: a dataset of Greek handwriting on papyri for the task of writer identification. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 726-731). IEEE.

[100] Abbas F, Gattal A, Djeddi C, Bensefia A, Jamil A, Saoudi K. Offline writer identification based on CLBP and VLBP. In: Mediterranean Conference on Pattern Recognition and Artificial Intelligence. Switzerland: Springer; 2020. p. 188–99.

[101] Abbas F, Gattal A, Djeddi C, Siddiqi I, Bensefia A, Saoudi K. Texture feature column scheme for single-and multi-script writer identification. IET Biometrics. 2021;10(2):179–93.

[102] Chammas, M., Makhoul, A., & Demerjian, J. (2020, December). Writer identification for historical handwritten documents using a single feature extraction method. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1-6). IEEE.

[103] Bennour, A. (2018, December). Automatic handwriting analysis for writer identification and verification. In Proceedings of the 7th International Conference on Software Engineering and New Technologies (pp. 1-7).

[104] A.Bennour, et All. (2019). Handwriting based writer recognition using implicit shape codebook. Forensic science international 301, 91-100.

[105] Fecker, D., Asit, A., Märgner, V., El-Sana, J., & Fingscheidt, T. (2014, August). Writer identification for historical Arabic documents. In 2014 22nd International conference on pattern recognition (pp. 3050-3055). IEEE.

[106] Asi, A., Abdalhaleem, A., Fecker, D., Märgner, V., & El-Sana, J. (2017). On writer identification for Arabic historical documents. International Journal on Document Analysis and Recognition (IJDAR), 20, 173-187.

[107] Dhali, M. A., He, S., Popović, M., Tigchelaar, E., & Schomaker, L. (2017). A digital palaeographic approach towards writer identification in the dead sea scrolls. In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods-Volume 1: ICPRAM (Vol. 2017, pp. 693-702). Scitepress; Setúbal.

[108] Lai S, Zhu Y, Jin L. Encoding pathlet and SIFT features with bagged VLAD for historical writer identification. IEEE Trans Inform Forensics Security. 2020;15:3553–66.

[109] Bennour, A. (2018, March). Clonal selection classification algorithm applied to arabic writer identification. In Proceedings of the 8th International Conference on Information Systems and Technologies (pp. 1-5).

[110] Chammas M, Makhoul A, Demerjian J, Dannaoui E. A deep learning based system for writer identification in handwritten Arabic historical documents. Multimedia Tools Appl. 2022;81:30769–84.

[111] He S, Schomaker L. GR-RNN: Global-context residual recurrent neural networks for writer identification. Pattern Recognition 2021;117:107975.

[112] Semma A, Hannad Y, Siddiqi I, Djeddi C, El Kettani MEY. Writer identification using deep learning

with FAST key-points and Harris corner detector. Expert Syst Appl. 2021;184:115473.

[113] Rehman, A., Naz, S., Razzak, M. I., & Hameed, I. A. (2019). Automatic visual features for writer identification: a deep learning approach. IEEE access, 7, 17149-17157.

[114] Chammas M, Makhoul A, Demerjian J. Writer identification for historical handwritten documents using a single feature extraction method. In: 19th International Conference on Machine Learning and Applications (ICMLA 2020). USA: IEEE; 2020

[115] Christlein V, Gropp M, Fiel S, Maier A. Unsupervised feature learning for writer identification and writer retrieval. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1. USA: IEEE; 2017. p. 991–7

[116] Jordan S, Seuret M, Král P, Lenc L, Martínek J, Wiermann B, et al. Re-ranking for writer identification and writer retrieval. In: International Workshop on Document Analysis Systems. Switzerland: Springer; 2020. p. 572–86.

[117] Cilia N, De Stefano C, Fontanella F, Marrocco C, Molinara M, DiFreca AS. An end-to-end deep learning system for medieval writer identification. Pattern Recognition Lett. 2020;129:137–43.

[118] Mohammed H, Märgner V, Stiehl HS. Writer identification for historical documents: analysis and optimisation of a classifier as an easy-to-use tool for scholars from the humanities. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE; 2018. p. 534–9.

[119] F. Wahlberg, L. Mårtensson, A. Brun, Large scale style based dating of medieval documents, in: Proc. 3rd International Workshop on Historical Document Imaging and Processing, New York, ACM Press, 2015, pp. 107–114, https://doi.org/10.1145/2809544.2809560.

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

[121] S. He, P. Samara, J. Burgers, L. Schomaker, A multiple-label guided clustering algorithm for historical document dating and localization, IEEE Trans. Image Process. 25 (11) (2016) 5252–5265.

[122] S. He, P. Samara, J. Burgers, L. Schomaker, Historical document dating based on temporal pattern codebook, Comput. Vision Image Underst. 152 (2016) 167–175.

[123] A. Hamid, M. Bibi, I. Siddiqi, M. Moetesum, Historical document dating using textural measures, in: International Conference on Frontiers of Information Technology (FIT), 2018, pp. 235–241, https://doi.org/10.1109/FIT.2018.00048.

[124] A. Rahiche, R. Hedjam, S. Al-Maadeed, M. Cheriet, Historical documents dating using multispectral imaging and ordinal classification, J. Cult. Herit. 45 (2020) 71–80.

[125] M.A. Dhali, C.N. Jansen, J.W. De Wit, L. Schomaker, Feature-extraction methods for historical document dating based on writing style development, Pattern. Recognit. Lett. 131 (2020) 413–420

[126] K. Adam, S. Al-Maadeed, Y. Akbari, Hierarchical fusion using subsets of multi-features for historical arabic document dating, J. Imaging 8 (3) (2022) 60.

[127] Wahlberg, F., Tomas, W., Anders, B. (2016) Historical document production date estimation using deep convolutional neural networks.

[128] A. Hamid, M. Bibi, M. Moetesum, I. Siddiqi, Deep learning based approach for historical document dating, in: International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 967–972, https://doi.org/10.1109/ICDAR.2019.00159

[129] M. Seuret, A. Nicolaou, D. Rodríguez-Salas, N. Weichselbaumer, D. Stutzmann, M. Mayr, V. Christlein, ICDAR 2021 competition on historical document classification, in: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, 2021, pp. 618–634. September 5–10, 2021, Proceedings, Part IV.

[130] Paparigopoulou, A., Pavlopoulos, J., & Konstantinidou, M. (2022). Dating Greek papyri images with machine learning.

[131] Koopmans, L., Dhali, M.A., & Schomaker, L. (2022). The effects of character-level data augmentation on style-based dating of historical documents. arXiv preprint arXiv:2212.07923.

[132] A. Buades, B. Coll, J.M. Morel, Non-local means denoising, Image Process. On Line (2011) 208–212.

[133] S. Paris, P. Kornprobst, J. Tumblin, F. Durand, Bilateral filtering: theory and applications, Found. Trends® Comput. Graphics Vision 4 (1) (2009) 1–73.

[134] C.R. Vogel, M.E. Oman, Iterative methods for total variation denoising, SIAM J Scientif. Comput. 17 (1) (1996) 227–238.

[135] Otsu N. A threshold selection method from gray-level histograms. Trans Syst Man Cybern. 1979;9(1):62–6.

[136] Canny, J. (1986). A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, p. 679-698.

[137] E. Rosten, R. Porter, and T. Drummond. Faster and better: A machine learning approach to corner detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 32:105–119, 2010. 1

[138] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004. 1, 2.

[139] B. Harris, M. Stephens, A combined corner and edge detector, Proceedings of the 4th Alvey Vision Conference, (1988), pp. 147–151.

[140] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[141] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778.

[142] Huang G, et al. Densely Connected Convolutional Networks. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. p. 2261–2269.

[143] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030.

[144] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (pp. 10347-10357). PMLR.

[145] Gattal A, Djeddi C, Siddiqi I, Al-Maadeed S (2018) Writer identifcation on historical documents using oriented basic image features. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, USA. pp 369–73

[146] Vincent Christlein et al (2022) Writer Retrieval and Writer Identi_cation in Greek Papyri, Intertwining Graphonomics with Human Movements: 20th International Conference of the

International Graphonomics Society, IGS 2021, Las Palmas de Gran Canaria, Spain, June 7–9

[147] Mathias Seuret et al (2020) ICFHR 2020 Competition on Image Retrieval for Historical Handwritten Fragments, 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp 216–221