university of groningen

University Medical Center Groningen

University of Groningen

Artificial Intelligence in Historical Document Analysis

Dhali, Maruf A.

DOI:
10.33612/diss.869247881

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2024

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Dhali, M. A. (2024). *Artificial Intelligence in Historical Document Analysis: Pattern recognition and machine learning techniques in the study of ancient manuscripts with a focus on the Dead Sea Scrolls*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. https://doi.org/10.33612/diss.869247881

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

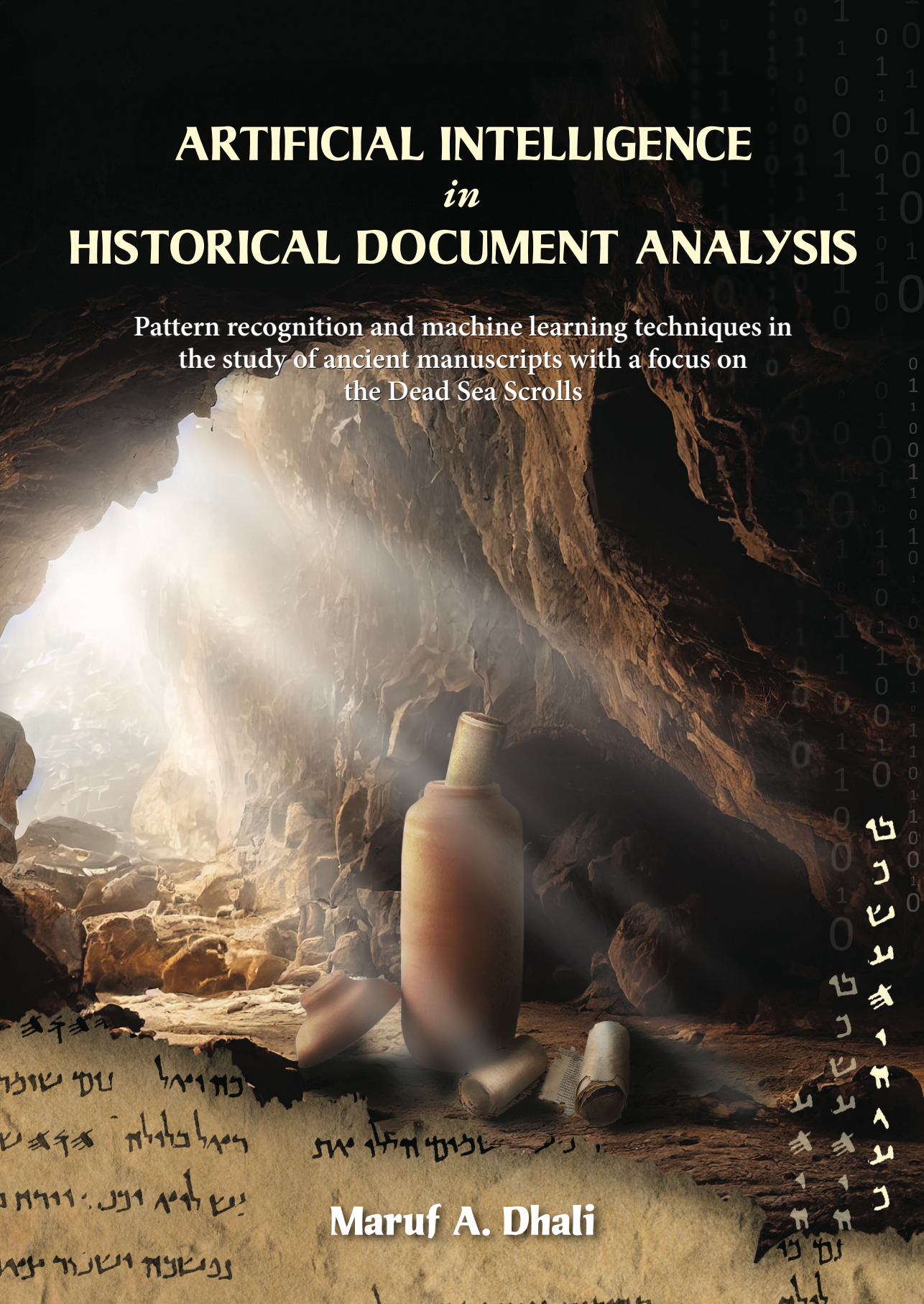
Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 01-02-2024

ARTIFICIAL INTELLIGENCE *in* HISTORICAL DOCUMENT ANALYSIS

Pattern recognition and machine learning techniques in the study of ancient manuscripts with a focus on the Dead Sea Scrolls



Maruf A. Dhali

ARTIFICIAL INTELLIGENCE
in
HISTORICAL DOCUMENT ANALYSIS

Pattern recognition and machine learning techniques in the study of ancient manuscripts with a focus on the Dead Sea Scrolls

MARUF A. DHALI
মারুফ আহমেদ ঢালী

rijksuniversiteit
groningen

FUNDING

The research has been carried out under the ERC Starting Grant of the European Research Council (EU Horizon 2020): The Hands that Wrote the Bible: Digital Palaeography and Scribal Culture of the Dead Sea Scrolls (HandsandBible № 640497).

COLOPHON

The typesetting of this thesis is done using LaTeX. The base typographical template, *classicthesis 4.6* - originally developed by André Miede and Ivo Pletikosić, is modified and updated by the author under GNU General Public License (GPL-2.1).

GRAPHICAL OUTLOOK

Cover design by PixelKreator (www.pixelkreator.com). Cave images, sand textures, and cover fonts are used under a CC BY-NC-ND (4.0) license with the registered project name PixelKreator.

The images of the Dead Sea Scrolls are from IAA, courtesy of the Leon Levy DSS Digital Library. Some illustrative photos (for individual parts' covers) are AI-generated by the author using Midjourney (2024).

ISBN

978-94-6496-017-4

PRINTING

Gildeprint – www.gildeprint.nl

© Maruf A. Dhali, Groningen, The Netherlands, 2024

university of groningen

Artificial Intelligence in Historical Document Analysis

Pattern recognition and machine learning techniques in the study of ancient manuscripts with a focus on the Dead Sea Scrolls

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of the Deans.

This thesis will be defended in public on

Tuesday 23 January 2024 at 16:15 hours

by

Maruf Ahmed Dhali

born on 31 December 1987

Supervisors
Prof. L.R.B. Schomaker
Prof. M. Popović

Assessment Committee
Prof. R. Ingold
Prof. H. Jaeger
Prof. C. Brune

CONTENTS

I Inception

1 Introduction 2
1.1 Background knowledge . . . 4
1.1.1 Dead Sea Scrolls (DSS) . . . 4
1.1.2 Handwriting recognition . . . 6
1.1.3 Multidisciplinary knowledge integration . . . 6
1.1.4 Writer identification . . . 7
1.1.5 Date estimation . . . 8
1.1.6 Pattern recognition techniques . . . 9
1.1.7 Artificial neural networks . . . 10
1.2 Research motivations . . . 12
1.2.1 Identifying the scribes . . . 12
1.2.2 Enhancing handwriting . . . 13
1.2.3 Dual-perspective time axis . . . 15
1.2.4 Ingenuity, adaptation, interpretability, and explainability . . . 15
1.3 Thesis outline . . . 16

2 Initial studies on writer identification 18
2.1 Introduction . . . 19
2.1.1 Challenges in digital palaeography . . . 20
2.2 Data . . . 20
2.2.1 Manuscript images . . . 20
2.2.2 Ground truth . . . 21
2.3 A pilot experiment . . . 23
2.3.1 Writer identification . . . 23
2.4 Results . . . 28
2.5 Discussions . . . 30
2.5.1 Performance evaluation . . . 30
2.5.2 Propositions . . . 31
2.5.3 Conclusions . . . 32

II Image and Handwriting

3 Binarization techniques 34
3.1 Introduction . . . 35
3.1.1 Why binarization is still important . . . 37

3.1.2 Goals . . . 38
3.2 Related works . . . 40
3.3 Methodology . . . 43
3.3.1 Dataset . . . 43
3.3.2 Image fusion . . . 44
3.3.3 Ground truth . . . 44
3.3.4 BiNet . . . 46
3.3.5 Network architecture . . . 47
3.3.6 Transfer learning . . . 49
3.4 Experiments . . . 49
3.4.1 Training . . . 49
3.4.2 Evaluation measures . . . 50
3.5 Results . . . 51
3.6 Conclusions . . . 56
Appendix of Chapter 3 . . . 58

III Writer identification

4 Great Isaiah Scroll & AI 66
4.1 Introduction . . . 67
4.2 Materials and methods . . . 71
4.2.1 Dataset and image preparation . . . 72
4.2.2 Primary analyses: feature-space explorations . . . 73
4.2.3 Secondary analyses: statistical evaluations . . . 75
4.2.4 Tertiary analyses: post-hoc visual analyses . . . 76
4.3 Results . . . 77
4.3.1 Primary analyses . . . 77
4.3.2 Secondary analyses . . . 81
4.3.3 Tertiary analyses . . . 85
4.4 Discussion and conclusions . . . 88
Appendix of Chapter 4 . . . 94
4.5 Supposed scribal idiosyncrasies . . . 94
4.6 Image information . . . 95
4.7 Primary analyses . . . 95
4.7.1 Preprocessing: binarization & alignment correction . . . 95
4.8 Secondary analyses . . . 99
4.8.1 Kohonen map of fraglets . . . 99
4.8.2 Statistical tests on the fraglet feature distances . . . 99
4.8.3 Least-squares fitting of a logistic curve . . . 106
4.9 Tertiary analyses . . . 109

IV Style classification and dating

5 Time period classification 112
5.1 Introduction . . . 113
5.2 Methodology . . . 115
5.2.1 Data . . . 115
5.2.2 Ground truths . . . 116
5.2.3 Preprocessing . . . 117
5.2.4 Feature extraction techniques . . . 117
5.2.5 Estimating time periods . . . 121
5.3 Experimental results . . . 122
5.3.1 Measures . . . 122
5.3.2 Sub-codebook size . . . 123
5.3.3 Overall results . . . 123
5.3.4 Cumulative scores . . . 124
5.4 Discussion . . . 124
5.5 Conclusions . . . 127

6 Date estimation using radiocarbon and AI 128
6.1 Introduction . . . 129
6.2 Radiocarbon dating . . . 131
6.3 Integration of multiple dating methods . . . 132
6.3.1 Neural networks for handwritten ink-trace detection . . . 132
6.3.2 Features for style attribution . . . 134
6.3.3 Bayesian ridge regression . . . 134
6.3.4 Testing Enoch . . . 135
6.4 Methodology . . . 136
6.4.1 Data preparation . . . 136
6.4.2 Data augmentation . . . 139
6.4.3 Allographic codebook with neural networks . . . 140
6.4.4 Textural-level features . . . 141
6.4.5 Adjoined feature . . . 141
6.4.6 Date-prediction model . . . 141
6.4.7 Data balancing . . . 148
6.4.8 Balance using augmentation . . . 149
6.4.9 Training options . . . 150
6.5 Results . . . 151
6.5.1 ^{14}C dates and palaeographic estimates . . . 151
6.5.2 Validation of Enoch . . . 152
6.6 Harvest of Enoch's predictions for undated manuscripts . . . 153
6.7 Discussion and conclusions . . . 156
6.7.1 The Enoch approach to dating ancient manuscripts . . . 157

6.8 Online contents . . . 158
Appendix of Chapter 6 . . . 160
6.9 The dating problem of the DSS . . . 160
6.9.1 Too few date-bearing manuscripts . . . 160
6.9.2 Weak workarounds . . . 161
6.9.3 The way out of the gap . . . 161
6.10 Radiocarbon dating of the DSS . . . 162
6.10.1 Selection of samples . . . 163
6.10.2 Soxhlet treatment and AAA pretreatment . . . 164
6.10.3 AMS measurements . . . 164
6.10.4 AMS dating results . . . 165
6.10.5 Result not to be used for palaeography: 4Q185 . . . 171
6.10.6 Technically rejected results . . . 171
6.10.7 Analytical chemistry . . . 171
6.10.8 Comparing radiocarbon results and palaeographic estimates . . . 172
6.10.9 Mur19 and 4Q52 results not used for the training . . . 174
6.11 List of images for different tests . . . 177
6.12 Additional information, plots, and results . . . 180
6.12.1 Radiocarbon sample information . . . 180
6.12.2 Data-sheet radiocarbon runs . . . 180
6.12.3 Worksheet of comparative data for ^{14}C dates and palaeography 180
6.12.4 Comparative plots for different information sources . . . 181
6.13 Enoch's predictions for 135 undated manuscripts . . . 182
6.13.1 Physical and image quality of the data . . . 182
6.13.2 How to read a prediction plot . . . 183
6.13.3 First evaluation . . . 187
6.13.4 Second evaluation . . . 188
6.14 Deep learning methods for image-based dating . . . 198
6.15 OxCal plots: ^{14}C determinations and calibrated dates . . . 202

v Discussion and conclusion

7 Discussion and conclusion 207
7.1 Addressing the research questions . . . 209
7.2 Future research . . . 211
7.2.1 Character reconstruction . . . 211
7.2.2 Scaling up writer identification test . . . 212
7.2.3 Neural network-based dating models . . . 212
7.2.4 Full page text recognition and transcription . . . 212
7.2.5 Material analysis and localization . . . 213
7.3 Conclusion . . . 213

VI Epilogue

8 Appendix-A: Material analysis 215
8.1 Introduction . . . 216
8.1.1 Dead Sea Scrolls collection . . . 218
8.2 Methodology . . . 220
8.2.1 Classification using Fourier transform . . . 221
8.2.2 Hierarchical k-means clustering . . . 225
8.2.3 Convolutional neural networks . . . 229
8.3 Results . . . 232
8.3.1 Classification using Fourier transform . . . 232
8.3.2 Hierarchical k-means clustering . . . 233
8.3.3 Convolutional neural networks . . . 233
8.4 Discussions . . . 234
8.5 Conclusions . . . 237
Supplement of Chapter 8 . . . 238

9 Appendix-B: Data augmentation 242
9.1 Introduction . . . 243
9.2 Related Works . . . 244
9.3 Methods . . . 246
9.4 Results . . . 257
9.4.1 Sub-codebook size . . . 258
9.4.2 Augmentation . . . 261
9.4.3 FRC . . . 262
9.5 Discussion . . . 265
9.5.1 Future research . . . 267
Supplement of Chapter 9 . . . 269

Bibliography 270

Summary 296

Nederlandse samenvatting 299

Abbreviations 302

List of publications 304

Media coverage 306

Acknowledgments 307