

InterPARES Summer School Symposium – San Benedetto, Friday 7 July
Artificial Intelligence to support the ongoing availability and accessibility of trustworthy public records - an overview of InterPARES Trust AI

AI for Multimedia understanding from computer vision to deep learning

CONTENT

- **Introduction**
- **From Machine Learning to Deep Learning**
- **Image understanding**
- **Deep Learning and Archival Science**
- **Managing Public Datasets**
- **The PergaNet study as a use case**
- **Other applications**



INTRODUCTION

What is worth to better
«understand» about ancient
documents?

- Shapes
- Signa
- Position of texts
- Defects and missing parts
- Writing style / abbreviations
- Signatures
- notes (even on the back)



How we can study these
features ?



STUDY INTRODUCTION

- **Ancient written private and public documents on parchment** were an important communication mean to humankind and have, for that motive, an **invaluable historical value to our culture heritage** and at the same time they are **authentic deeds among people**, valid till nowadays.
- Therefore, it becomes imperative that they are **preserved** and perpetuated in order to extend their life span in the interests of population and **future generations**.
- Currently, **digitization of historical parchments is extraordinarily convenient as it allows easy access to the documents** from remote positions and removes the need for possible adverse physical management.



AN OUTSTANDING HERITAGE

- The **Archivio di Stato of Milan** (ASMi), among the others, preserves more than **130.000** documents on parchment, covering nearly a thousand years of history.
- The main collections of parchment documents are the **Museo diplomatico** and the **Pergamene per fondi**.
- The *Museo diplomatico* holds the most ancient Italian original private deed, **dating back to the year 721** and a large number of documents preceding the year 1000.
- The *Pergamene per fondi* holds **imperial and royal charters, papal bulls, chancery documents, and private deeds** coming from a huge variety of ecclesiastical and secular institutions of northern Italy (12th-18th centuries).



[illegible]

The “AI in the Middle Age” study:

Parchments Study and Arrangement by Appearance-Based Recognition

TEAM

Archivio di stato di Milano:

Benedetto Luigi Compagnoni – *former Director*,

Carmela Santoro - *Archivist, Deputy director*,

Archivio di Stato di Novara

Davide Bruno De Franco - *Archivist, Director*

Archivio di Stato di Ascoli Piceno

Emanuele Tedeschi, *director*

University of Macerata & Marche Polytechnic University:

Emanuele Frontoni, *Professor - Computer Science and Computer Vision*

Pierluigi Feliciati - *Professor, Records and Information Science*

Marina Paolanti - *PhD, Assistant professor*

Andrea Felicetti - *PhD, Postdoc*

University of Pavia:

Luca Fois - *PhD, Postdoc, Dipartimento di studi umanistici*

Washington University:

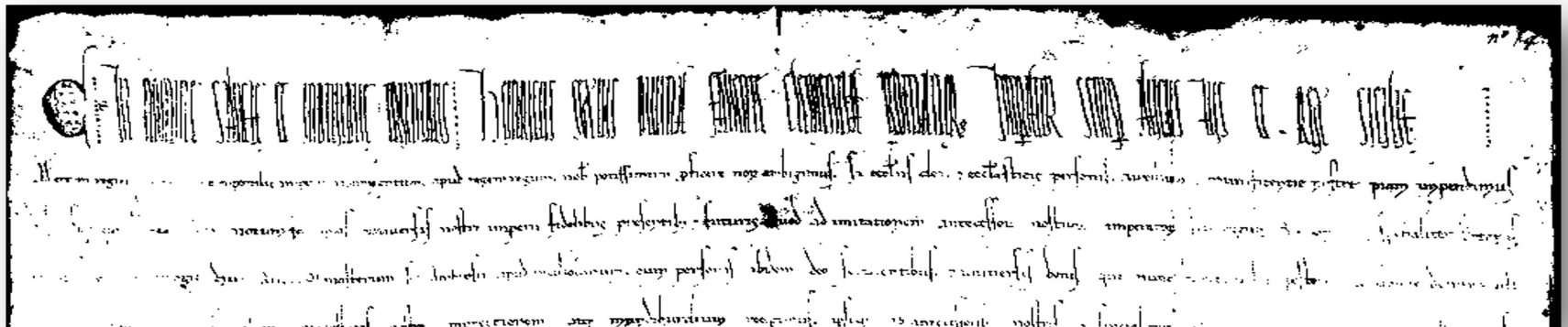
Joseph T. Tennis - *Professor, Associate Dean for Faculty Affairs, Executive Director of Administrative Services*

InterPARES
TrustAI



NEW PATHS OF RESEARCH

- Though studied and edited since the XVIII century, **this extensive materials have never been investigated systematically in order to identify common features** recurrent in homogeneous groups of documents.
- **Processing automatically a large number of scanned documents** can lead to new understandings of general cross-cutting issues, not related to a single fund.



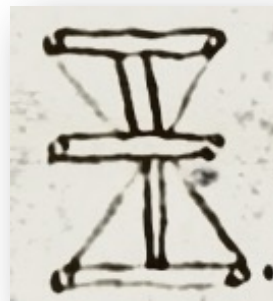
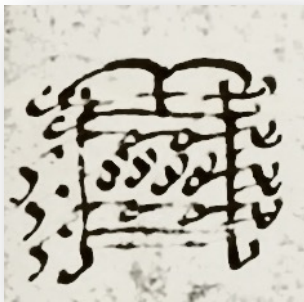
AI & DEEP LEARNING FOR THE PARCHMENT DOCUMENTS

- One of these issues is **defining the number and the activity of the notaries in the city of Milan** and its contado.
- The AI deep learning approach can **automatically investigate a large number of private documents** and give a substantial contribution to the knowledge of **Milanese notaries in the Middle Ages**.
- The basic feature chosen to be identified for the research is **the signum, an authentication element** put by notaries.
- The **signum appears in about all private deeds** and has **specific and easily recognisable characteristics** (shape, position, etc.)



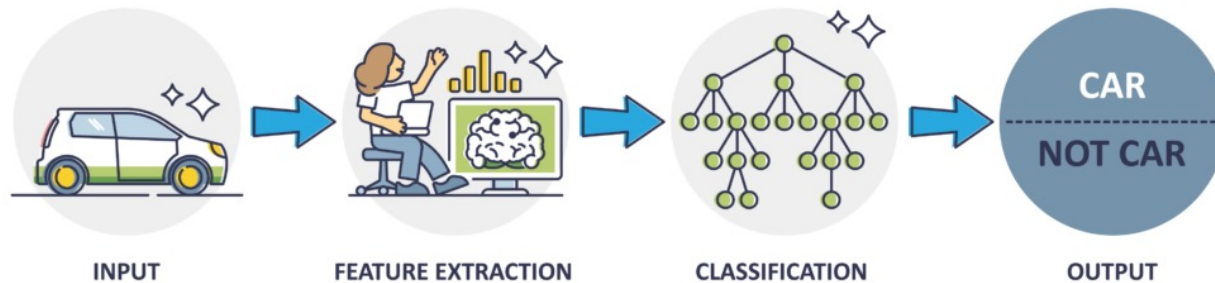
THE SIGNUM: A USEFUL MARKER

- The **signum or Notarial sign** is a specific and personal drawn mark used by a single notary in the protocollo and before his signature.
- Identifying the signum means that every notary could be **recognised and tracked** in a virtually infinite series of documents.
- The AI will contribute either in create a **library of signa: a virtual matricula of notaries** and a basis to investigate their **less visible features**.



FROM ML TO DL

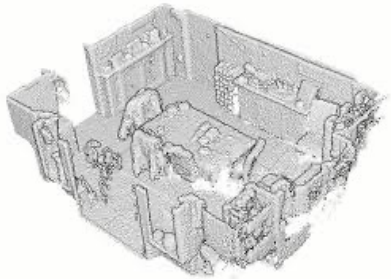
MACHINE LEARNING



DEEP LEARNING



FROM ML TO DL



Field

Task

Leading Methods

CV

Semantic Segmentation

HRNet-OCR | Efficient-Net-L2 | ResNeSt-269 | VMVF

Image Classification

FixEfficientNet | BiT-L | Wide-ResNet-101 | Branching CNN

Object Detection

Efficient-Det-D7x | Rodeo | Patch Refinement | IterDet

NLP

Sentiment Analysis

BERT | T5-3B | NB-weighted-BON + dv-cosine

Language Modeling

Megatron-LM | GPT-3 | GPT-2

Text Classification

XLNet | USE_T + CNN | SGC

Question Answering

T5-11B | SA-Net on Albert | TANDA-RoBERTa

Machine Translation

Efficient-Det-D7x | Rodeo | Patch Refinement | IterDet

RS

Recommender System

Bayesian time SVD++ // flipped w/ Ordered Probit Reg | EASE | H+Vamp Gated

SR

Speech Recognition

ContextNet + Noisy Student | ResNet + BiLSTMs | LiGRU | Large-10h-LV-60k

InterPARES
TrustAI



IMAGE UNDERSTANDING

Classification



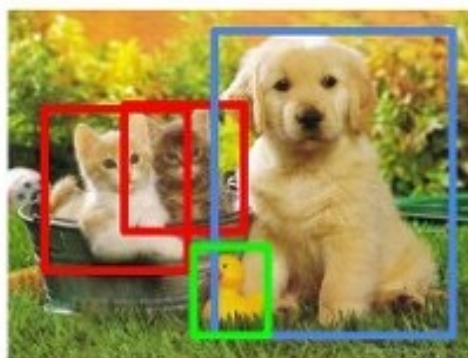
CAT

**Classification
+ Localization**



CAT

Object Detection



CAT, DOG, DUCK

**Instance
Segmentation**



CAT, DOG, DUCK

Single object

Multiple objects

What about archives ?

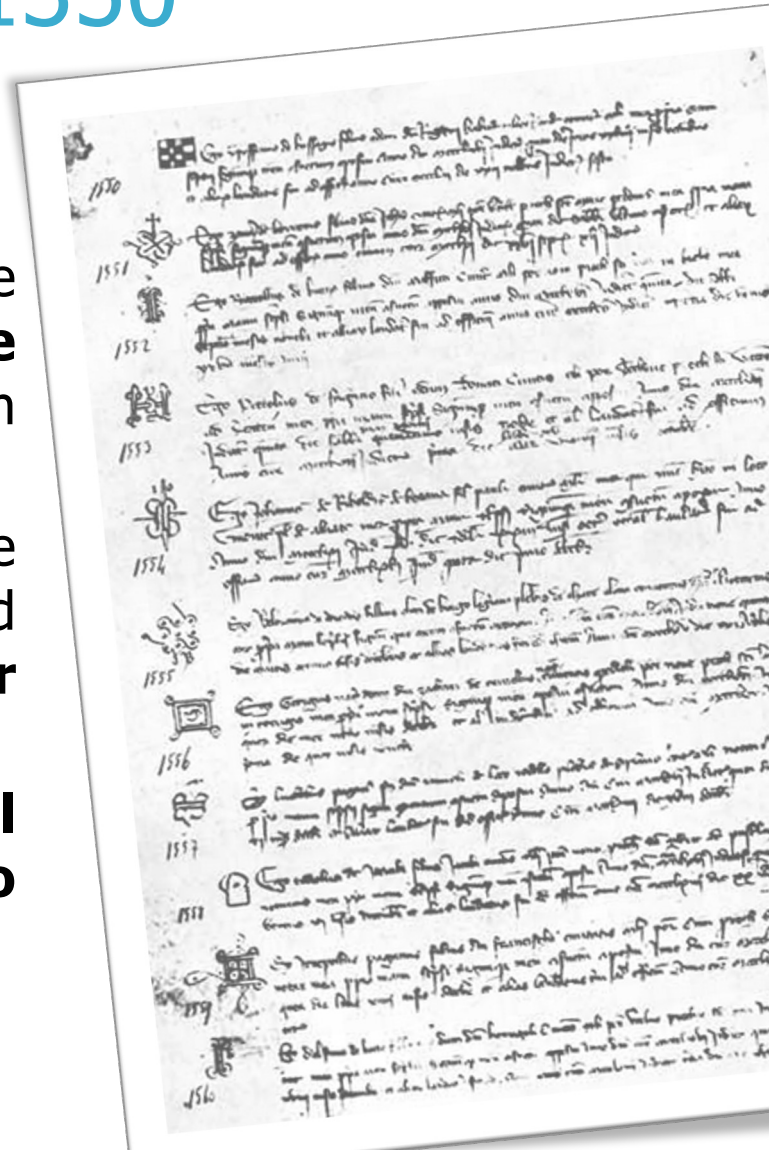


MANAGING PUBLIC DATASETS

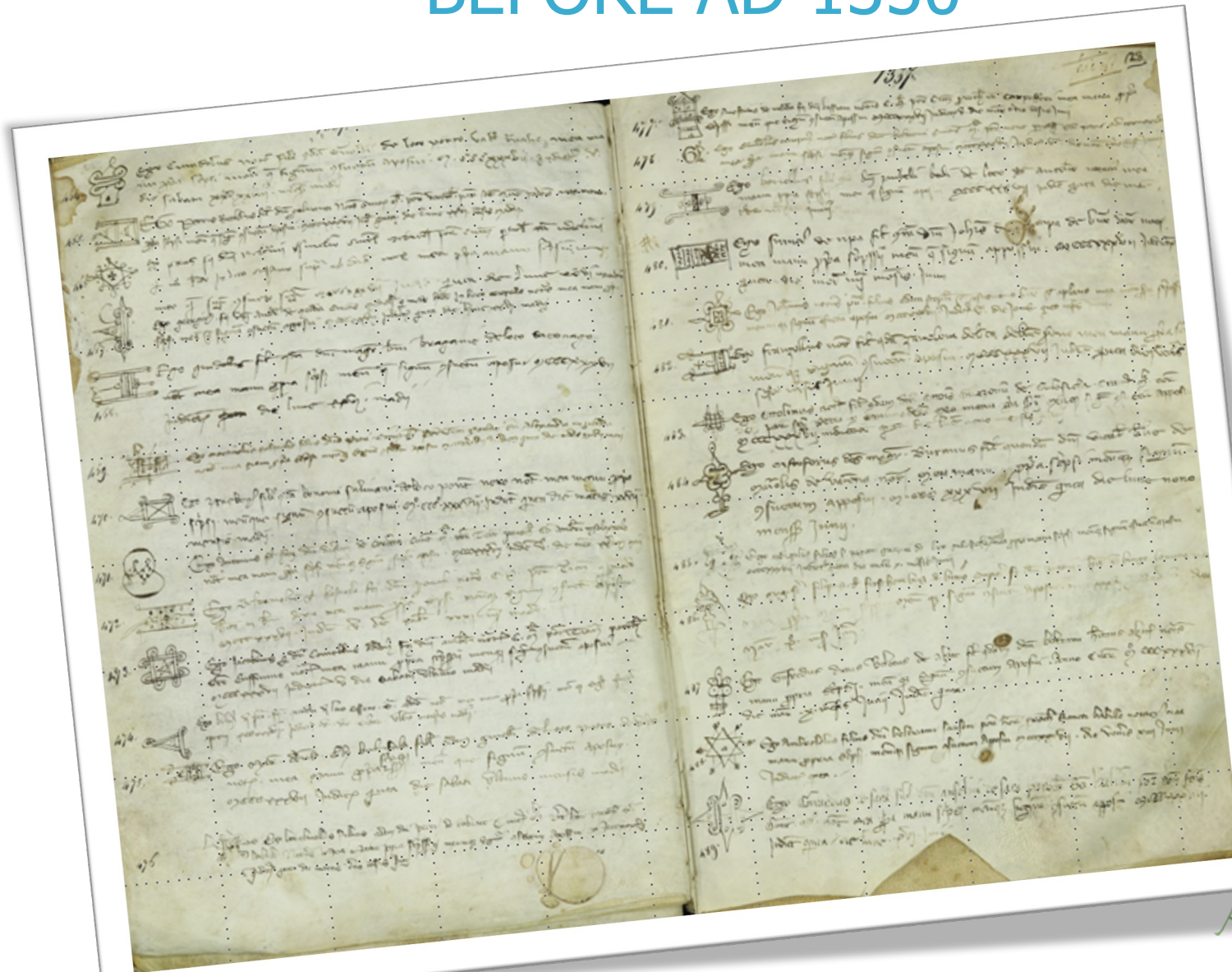


BUILDING UP A NOTARIAL MATRICULA BEFORE AD 1350

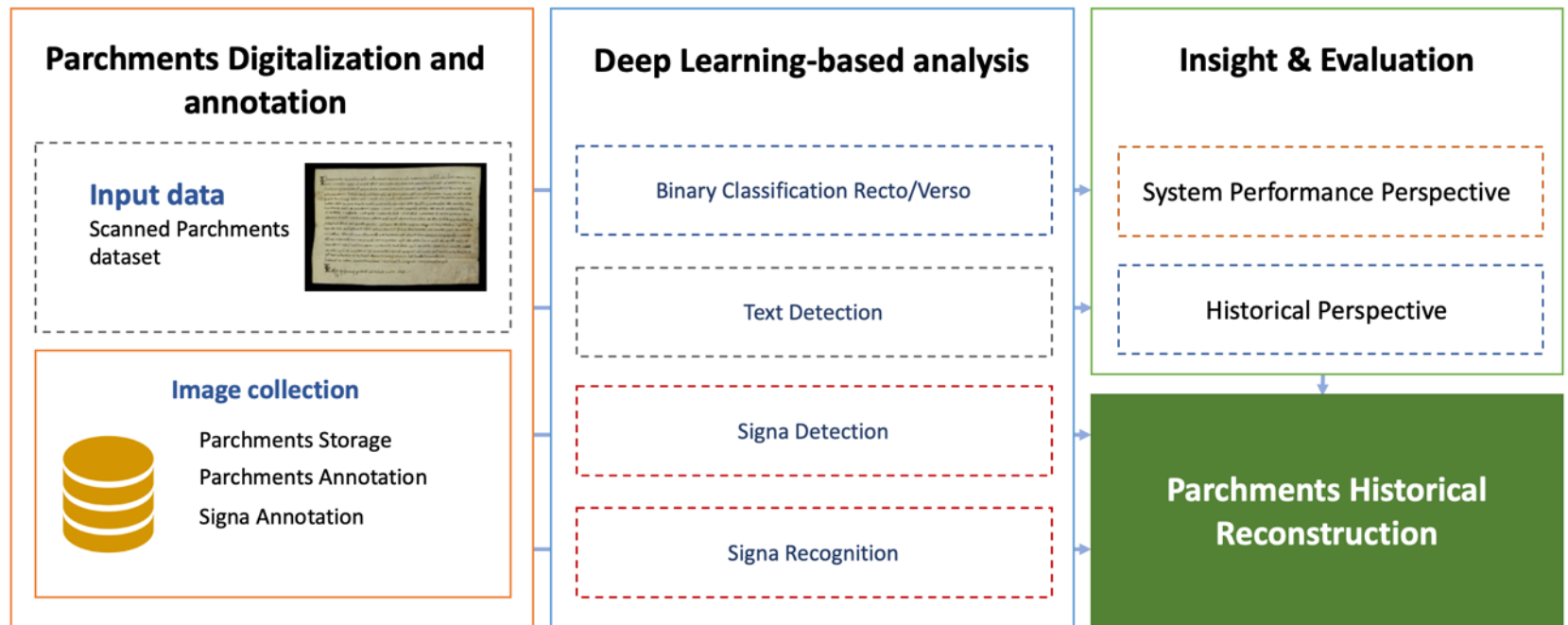
- One of the aims of the project is to use the **AI to build up a Matricula of Milanese notaries** of the twelfth and thirteenth centuries (in separate steps).
- The progressive investigation of all the parchment documents of the ASMi will lead to a **complete list of notaries and their identification**.
- The process will also generate **a list of all the surviving documents related to every single notary**.



BUILDING UP A NOTARIAL MATRICULA BEFORE AD 1350



PERGANET

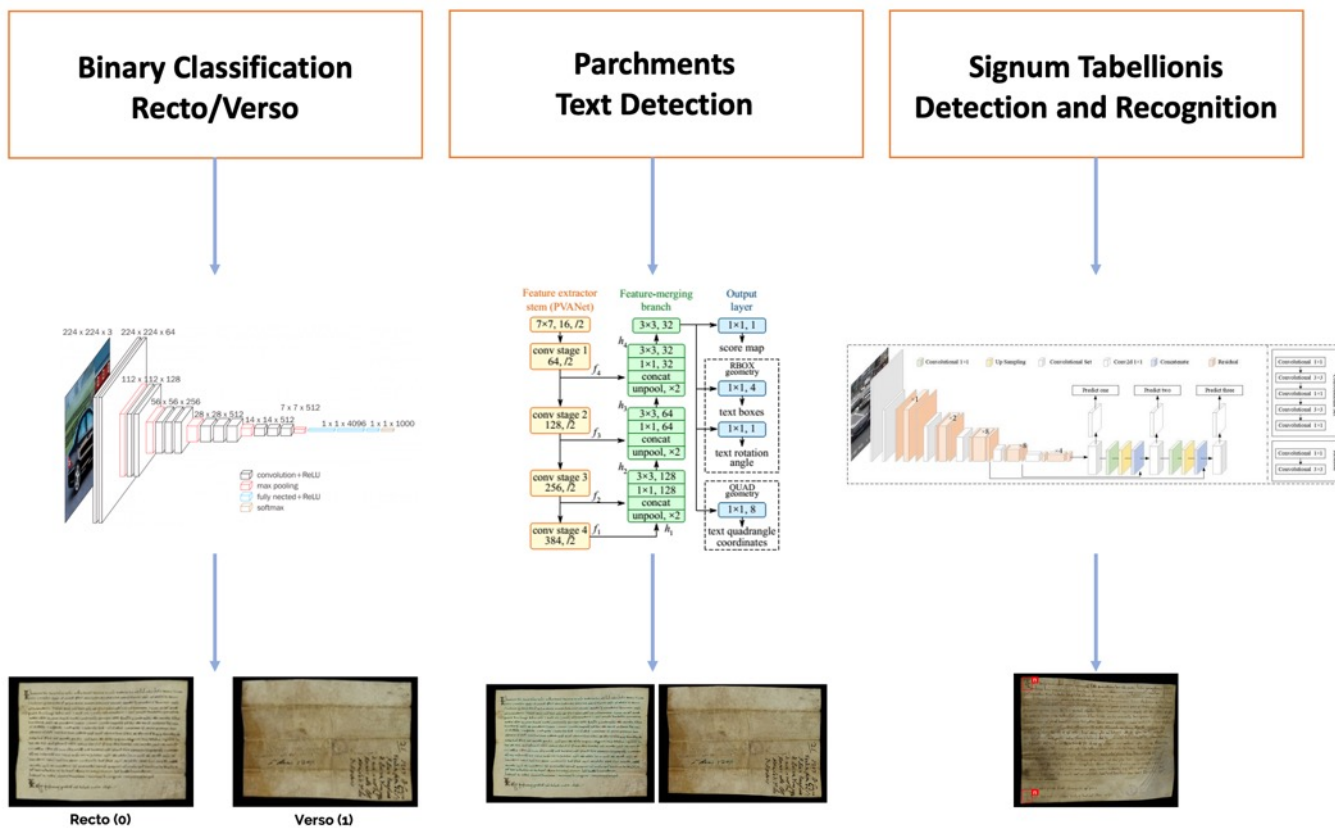


DATASET COLLECTION

- Data collection is moving forward and actually we have the following **archives cooperation together and collecting data for the study**:
 - **Archivio di Stato di Milano**
(220 documents already processed, nearly 1200 scanned and ready to be processed)
 - **Archivio di Stato di Novara**
(91 documents ready to be processed)
 - **Archivio di Stato di Ascoli Piceno**
(27 documents ready to be processed)

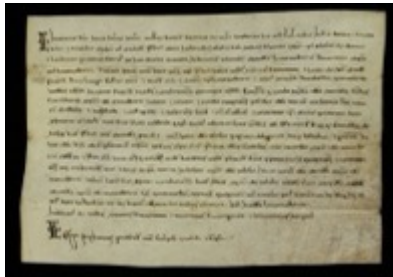


DEEP LEARNING PIPELINE



BINARY CLASSIFICATION RECTO/VERSO

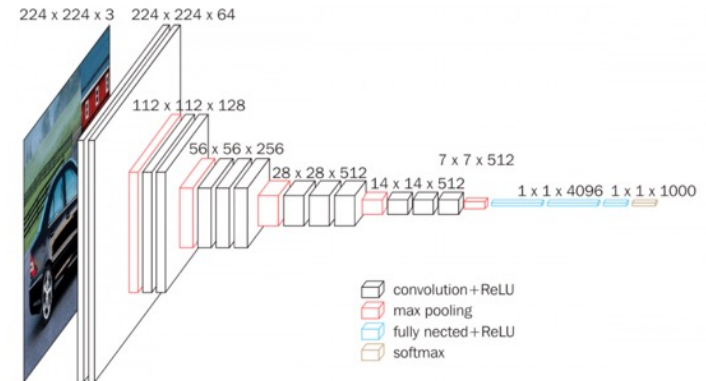
VGG16 DEEP NEURAL NETWORK



Recto



Verso



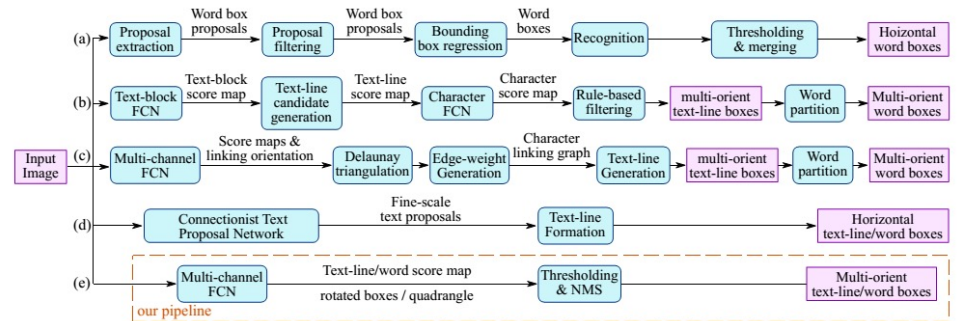
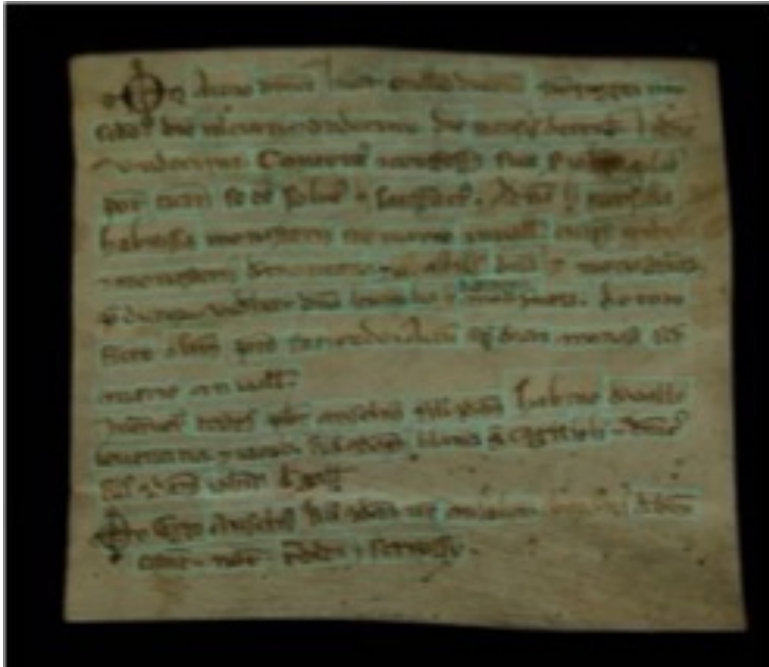
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.



BINARY CLASSIFICATION RECTO/VERSO: RESULTS



PARCHMENTS TEXT DETECTION



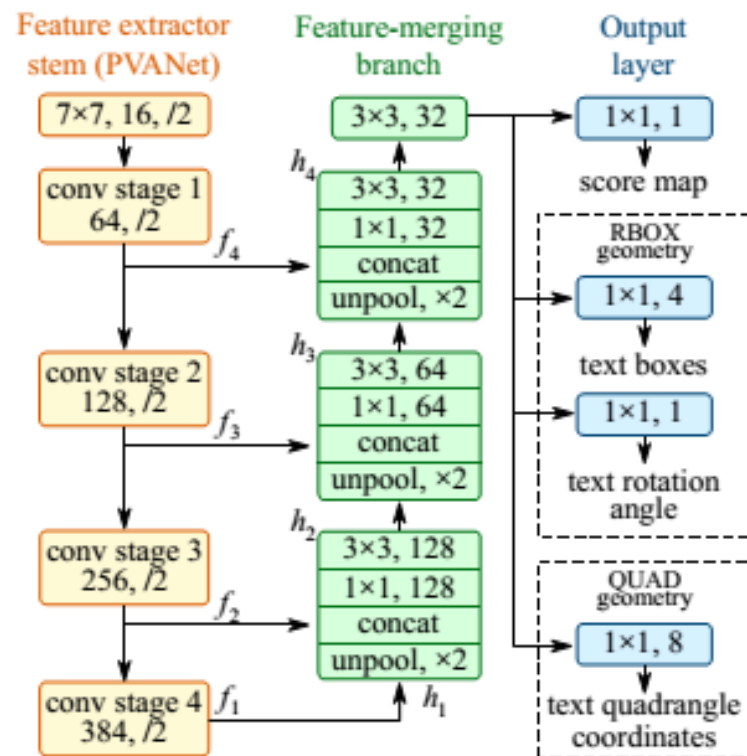
Xinyu Zhou., et all, A. (2017). EAST: An Efficient and Accurate Scene Text Detector. *arXiv:1704.03155v2*.



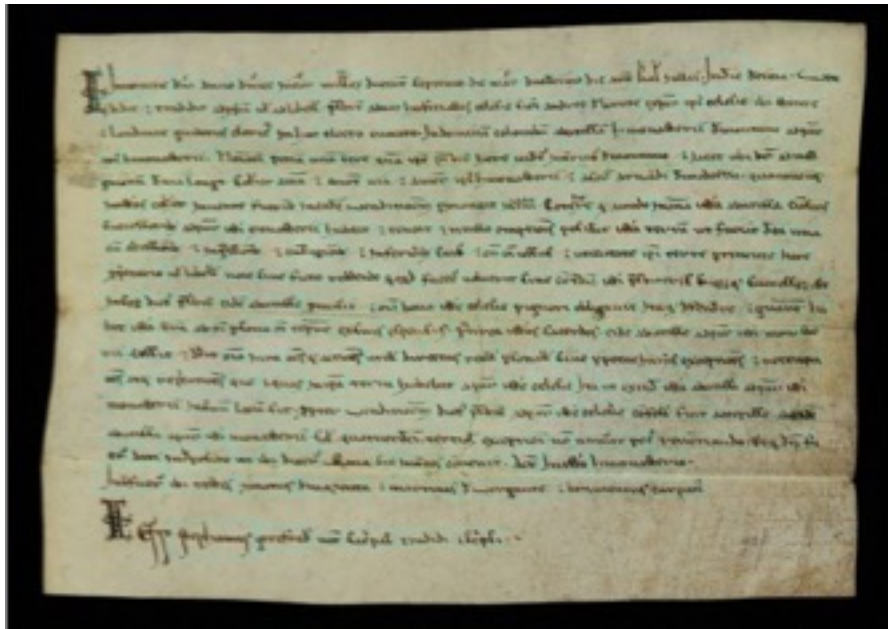
PARCHMENTS TEXT DETECTION: EXPERIMENTS

Word detection

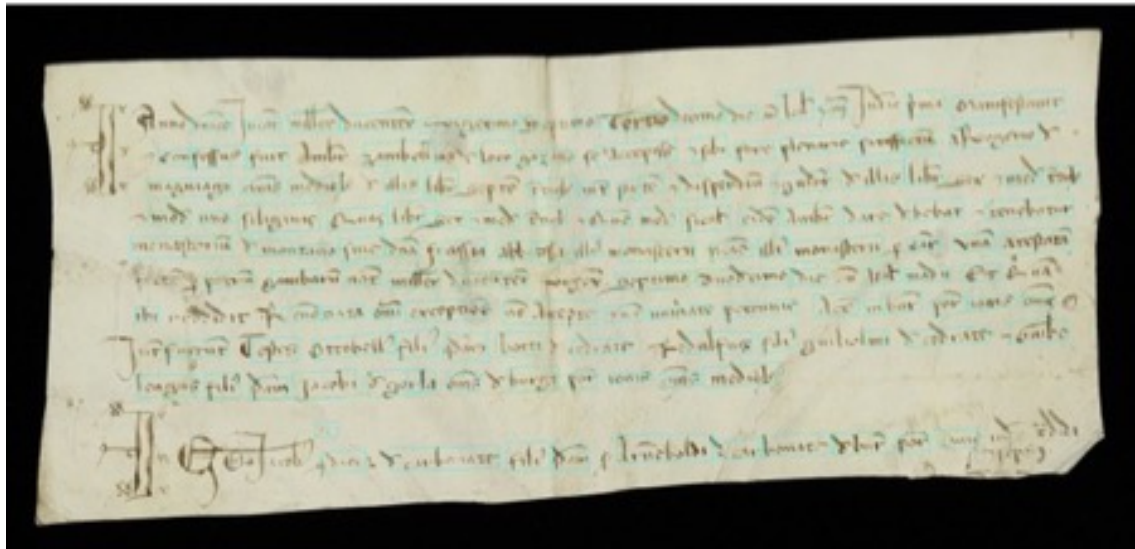
- DNN Model: EAST
- Use of trained model (tested over datasets: ICDAR 2015, MSRA-TD500, COCO-Text)



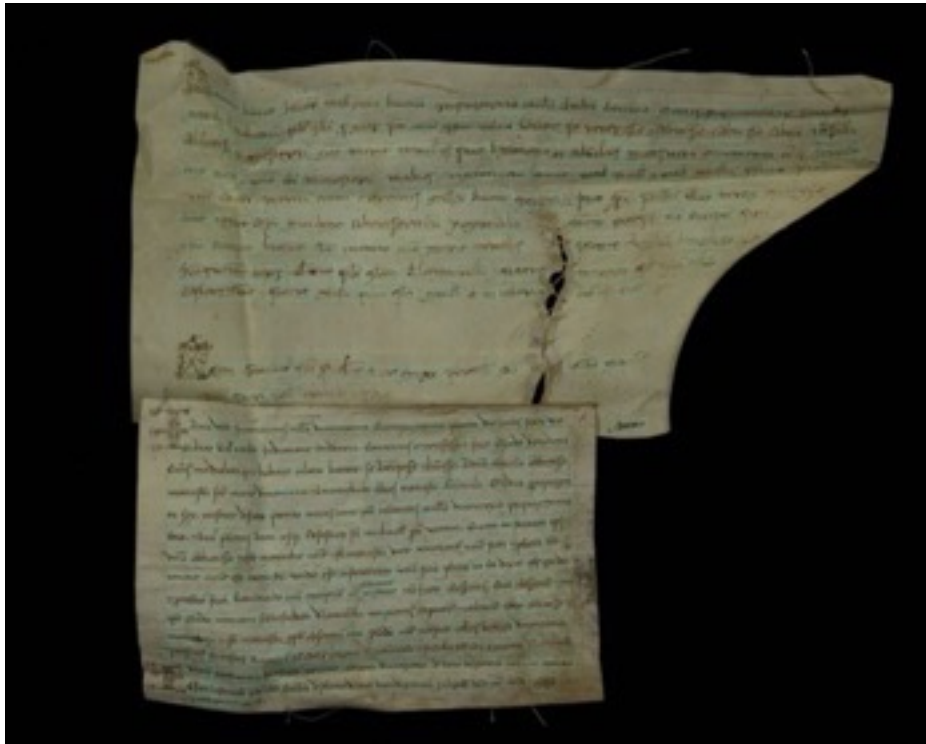
PARCHMENTS TEXT DETECTION: RESULTS



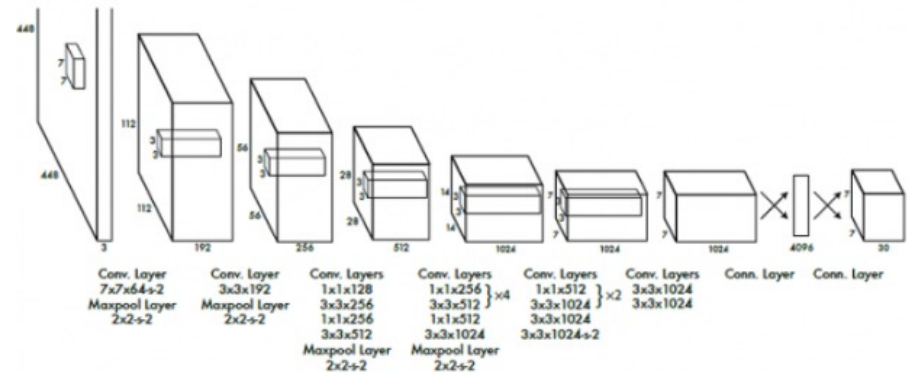
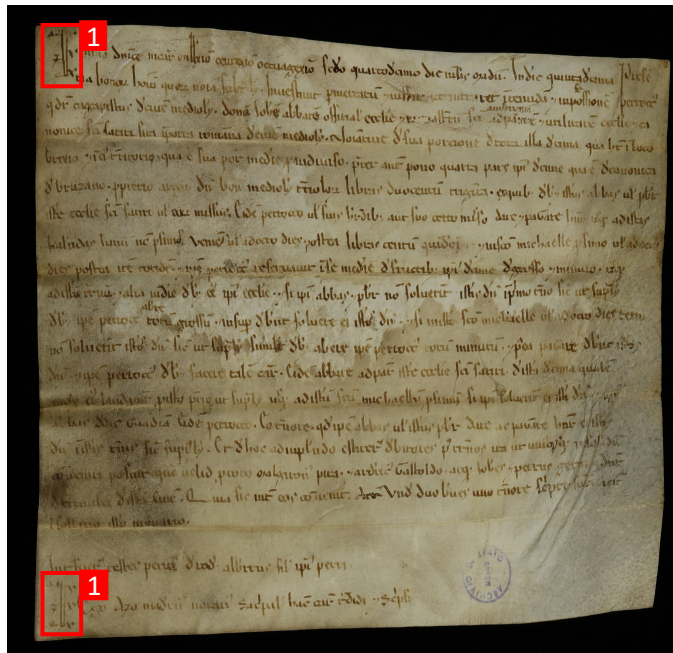
PARCHMENTS TEXT DETECTION



PARCHMENTS TEXT DETECTION



SIGNA TABELLIONIS DETECTION AND RECOGNITION



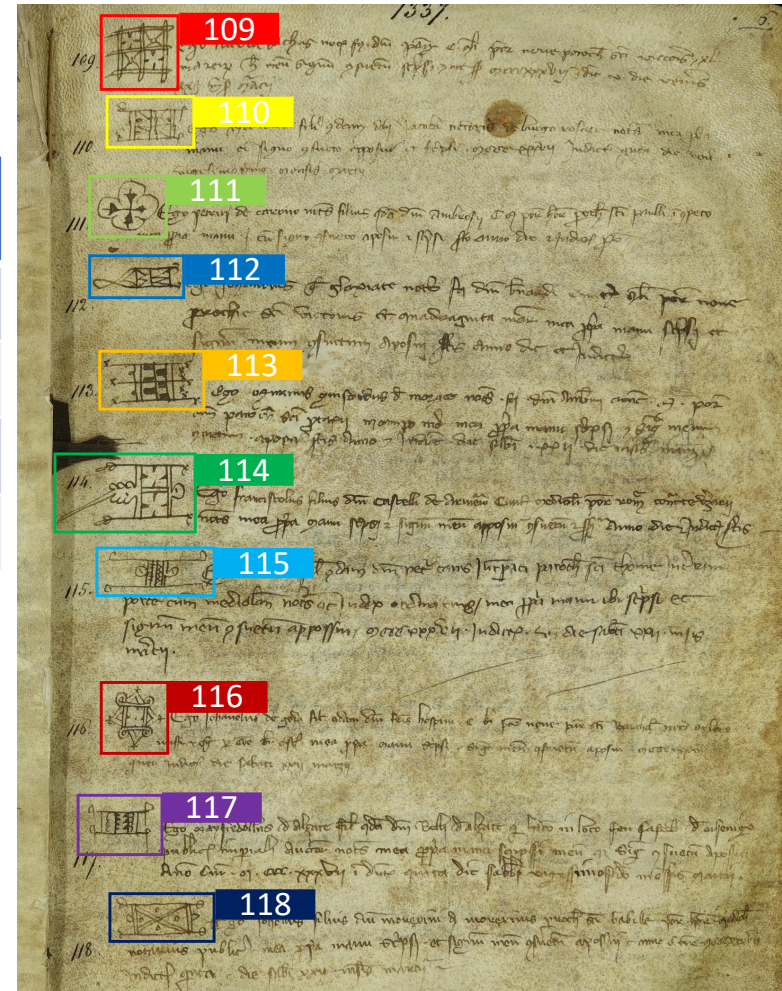
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).



NOTARIAL MATRICULA ANNOTATION

2768 signa

ID	NAME	HISTORICAL INFORMATION			
...	...				
109	notary 109				
110	notary 110				
...	...				



SIGNA TABELLIONIS DETECTION AND RECOGNITION: RESULTS



SIGNA TABELLIONIS DETECTION AND RECOGNITION: RESULTS

signum 0.73

...fuit boni hominis ... unum dñi Johis qd' nullan' dñe medietate facere qd' ingratu. ad partē monariū qd' d' clauuallē. Itē autē unā dñi in p
... facere debet ipse Johis qd' facere debet facere nepotes suos. in abbate ip' monariū ut infus misse. talē autē quale uidet ip' monariū laudauerit. p' p'riū libray octraginta
... dñi. d' omib' paratū qd' geritū qd' buscū. que ipse Johis qd' sui nepotes detinet ut ad suā manū ut p'ius nullanō. ab illa ripa uestalū que ē uertū ip' monariū usq' in strada s'ci martini sic certat
... amolino Johis gariardi in furū qd' amolino azoni fante in subū. qd' si sui nullanū ut alū nullanū d' madremiano habent paratū qd' ip' nullanū teneat p' se ut p' alios suos seniores. a uocē qd' est in
... capite d' paratū qd' Johis usq' in buscū d' monariū. ipse Johis ad quere debet ip' paratū ab illo cui ē qd' facere similiter facere eū p' eodē p'rio. Et insup facere debet ipse Johis eū qd' facere debet facere eū
... ip' sui nepotes in laudamito iudicū. ut monasteriū possit firmare qd' habet clusa in paratū qd' Johis qd' sui nepotib' qd' ē ab alia ripa uestalū a uado in furū. sic fuerit ut illa qd' clusa ab
... monariū lectū ubi ip' monariū uoluerit. ita ut nō tollat eis lotas frē nec ligna. qd' ipsa clusa ingurgauerit paratū ut frā ip' Johis qd' sui nepotib' nō debet querere restaurari a eū
... monariū. Et ipsa clusa fuerit facta ab ipse uado in subū. qd' ingurgauerit paratū ut frā eū restaurari debet dñi ip' monariū. Et insup facere debet eū in ip' monariū sedā mātū qd' p'ue
... actū ut ip' monariū faciat mātū si uoluerit a sua ripa que ē uertū monariū. qd' p' mātū clusa si uoluerit i dñā ripa. Et debet similiter facere eū qd' ip' monariū habet eū mātū mātū eundē. et ad mātū eū
... caritū qd' eū omī eū uoluerit utriusq' uestalū p' accessū uerū qd' nō est. p' frā ip' Johis qd' sui nepotib' qd' p' nullā. aundo usq' ad uicū maiorē qd' redeundo. Et si fuerit opus. ut liceat facere eū monariū
... pollata sup frā ip' Johis qd' sui nepotib' ab una p'ir uicū qd' ab alia. Et ut ipse Johis qd' sui nepotes nō debet eū dicere eū monariū. leuare pontē sup uestalū ubi monasteriū habet ip' ab una p'ir qd' ab alia. Et in
... talis tenore d' adimplendo omīa quātū sup' b'. posuit si fiduissorē Arderiū qd' d' dñe clausē. in pena d' dñi libray quadraginta den. bon. medietatē fuit ad impleri qd' fuerit. Et si ip' ip' monariū
... uno dñe mātū null' antef' ingratu octauo. nōne scilicet dñe dñe dñe. fuit

signum 0.36

signum 0.32

signum 0.76

signum 0.72

... hoc breue scriptū

OTHER APPLICATIONS

Once refined and fully developed, the AI could be used in a **wide range of applications:**

- Recognizing the **peculiar system of writing** of every single notary.
- Training **Handwritten Text Recognition** tools to read text and abbreviations
- Analyzing **archival notes on the back** and retrace previous archival arrangements or uses of groups of documents.
- Recognizing **recurring images or other features** in huge series of documents.
- Identifying **common patterns in manuscript maps or drawings.**
- Opening **original & relevant AI Data Sets.**
- And many others...



InterPARES Summer School Symposium – San Benedetto, Friday 7 July
Artificial Intelligence to support the ongoing availability and accessibility of trustworthy public records - an overview of InterPARES Trust AI

AI for Multimedia understanding from computer vision to deep learning