AUTHORS

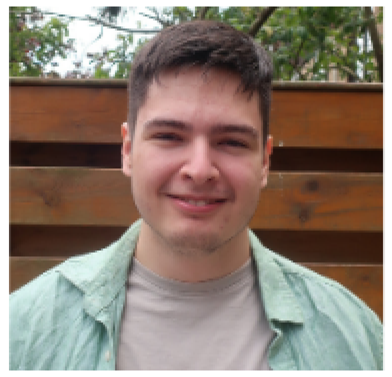
Crăciun Marius-Cosmin

SUPERVISORS

Dr. Zekeriya Erkin, Devriş İşler

Watermarking numerical datasets used for ML

Watermarking numerical datasets in the wavelet domain



Contact: M.C.Craciun@student.tudelft.nl

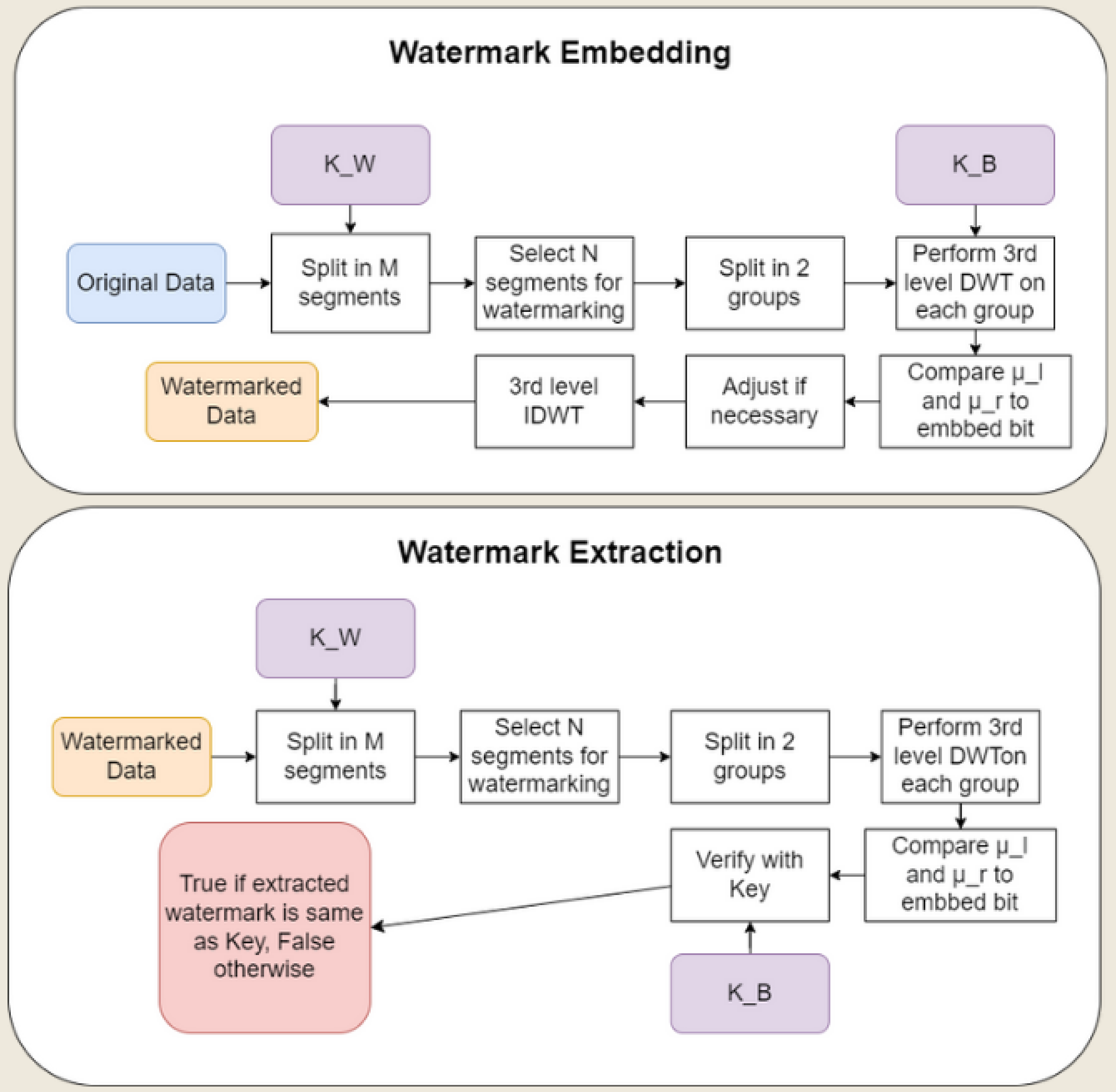
01. Introduction

- value of data increases as ML and AI develop
- collecting data is expensive
- high interest in proving ownership of this asset

02. Background

- **Watermarking :**
  - embedding a small amount of information in the target data
  - commonly used for proving ownership of the data or verifying if the data has been altered
- **Discrete Wavelet Transform :**
  - signal processing technique that tries to reconstruct the signal using short-lived signals (wavelets) from the same family
  - breaks down into a list of coefficients
  - can be done multiple times

03. Watermarking



Watermark Embedding

K_W

K_B

Original Data

Split in M segments

Select N segments for watermarking

Split in 2 groups

Perform 3rd level DWT on each group

Compare μ_l and μ_r to embbed bit

Adjust if necessary

3rd level IDWT

Watermarked Data

Watermark Extraction

K_W

Watermarked Data

Split in M segments

Select N segments for watermarking

Split in 2 groups

Perform 3rd level DWTon each group

Compare μ_l and μ_r to embbed bit

Verify with Key

K_B

True if extracted watermark is same as Key, False otherwise

Figure 1: Watermarking scheme

04. Results

Imperceptibility

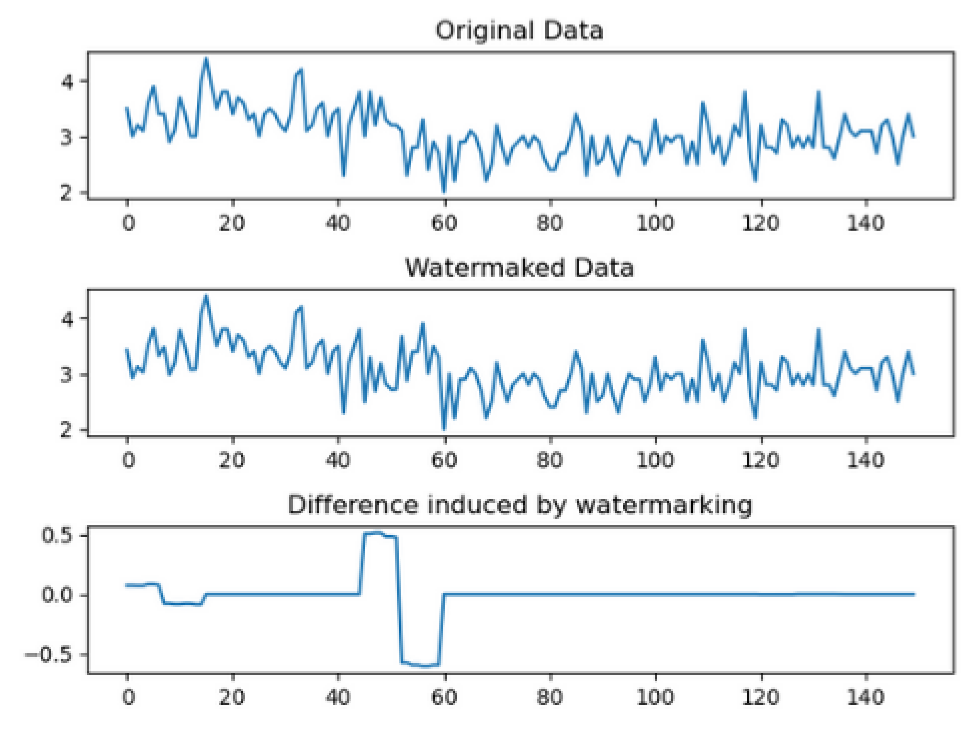


Figure 2: Iris Data before and after WM

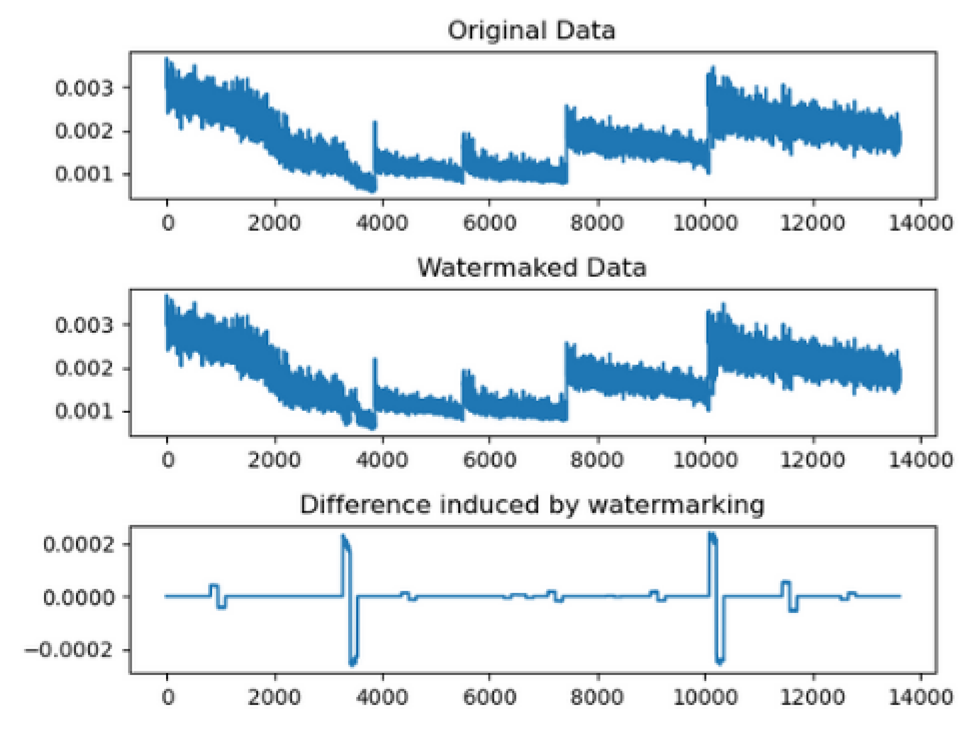


Figure 3: Dry Bean Data Before and after WM

ML Model	Δ Acc Test 50%	Δ Acc Test 100%
Log Regression	0.440744	-3.697356
KNN(n=6)	0.514201	0.538687
KNN(n=20)	0.024486	0.75906
SVM	0.563173	1.22429
Decision tree	-0.905974	0.514202
Random Forest	3.893242	3.648384

Table 1: Accuracy differences between watermarked and original data

- Watermarking induces small changes in data
- Hard to infer which dataset is the original
- Changes in the accuracy performance of the models trained on watermarked data are not significant

Robustness

- The initial watermark is not very robust to attacks
- However, the method is flexible and robustness can be improved at the expense of imperceptibility
- Most resistant to update attacks

- When the change in data is doubled, the resulting change in variance is around 1% => ~ 10 times larger

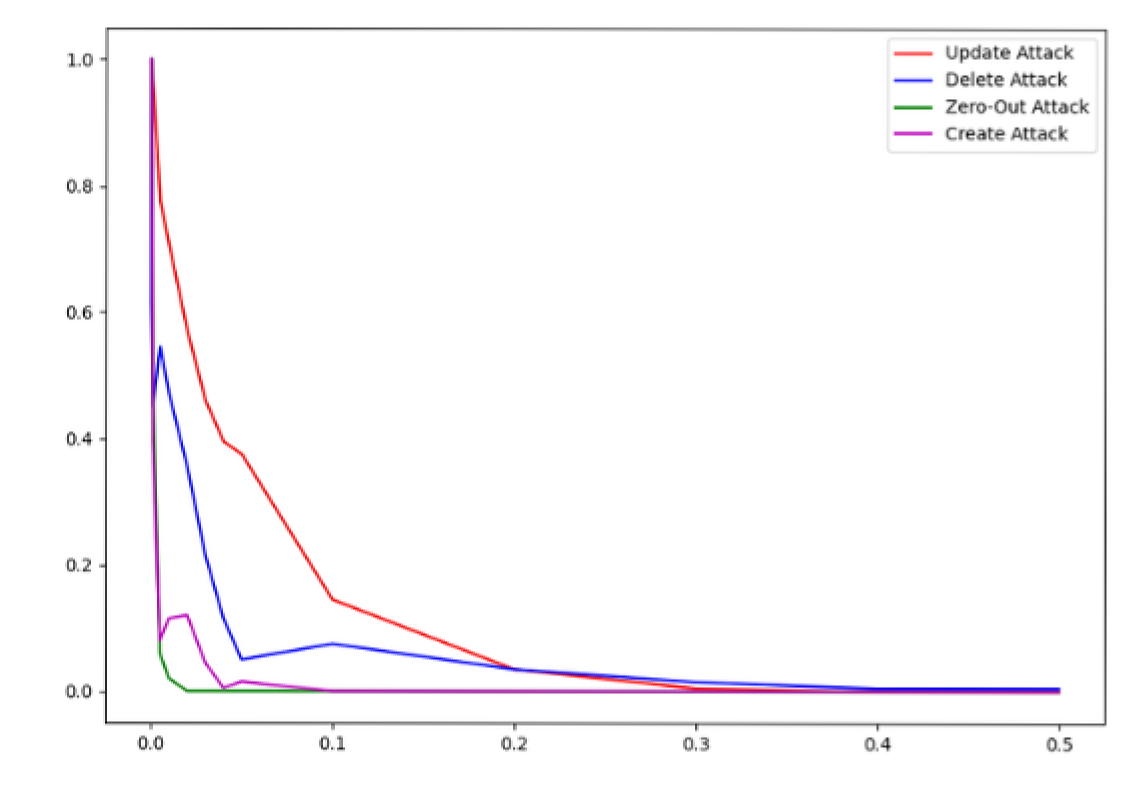


Figure 4: Extraction rate after attacks

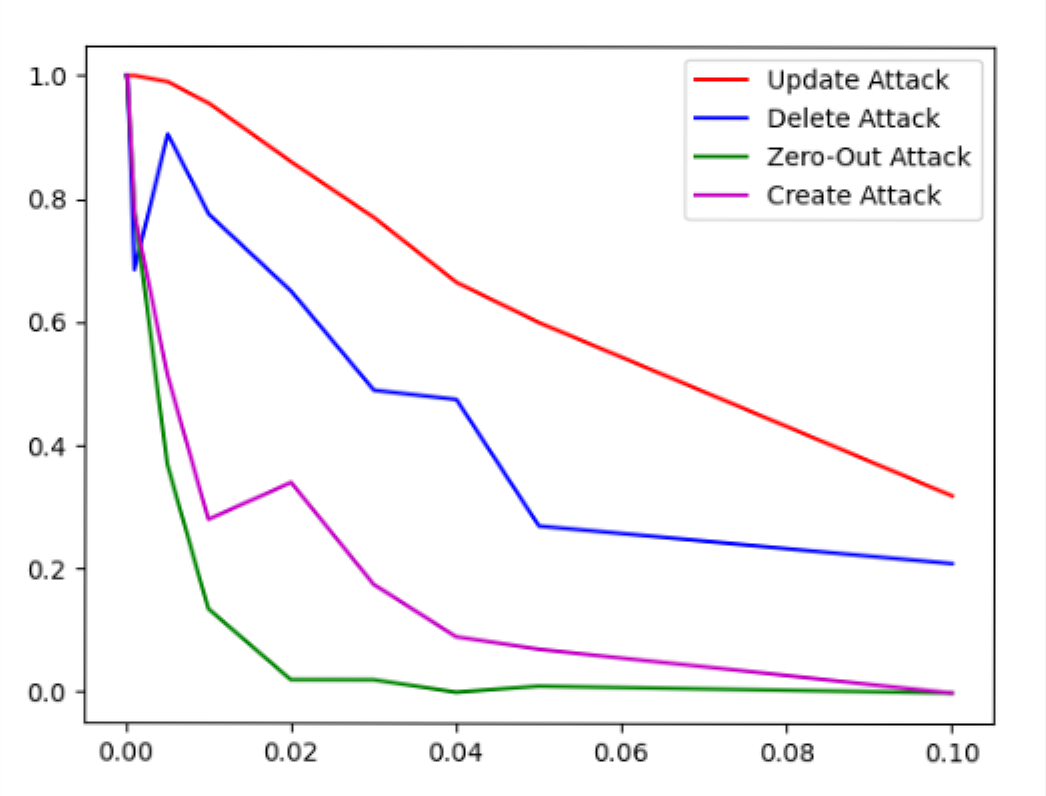


Figure 5: Extraction rates when robustness is improved

05. Future Work

- Sorting the data could make the method resistant to reordering
- Multiple attributes could be utilized for watermarking in order to increase the amount of data that can be embedded
- A majority voting system during extraction might improve the robustesness of the technique

06. Conclusions

- Watermarking did not introduce relevant distortion to the data
- The proposed method is flexible, offering a tradeoff between robustness and imperceptibility
- Training ML models on watermarked data does not affect the quality of the model in terms of accuracy
- Not robust enough for professional use

Related literature

Arezou Soltani Panah, Ron Van Schyndel, Timos Sellis, and Elisa Bertino. On the properties of non-media digital watermarking: a review of state of the art techniques. IEEE Access, 4:2670–2704, 2016