DA 526 (IPML)

DENOISING NOISY DOCUMENTS

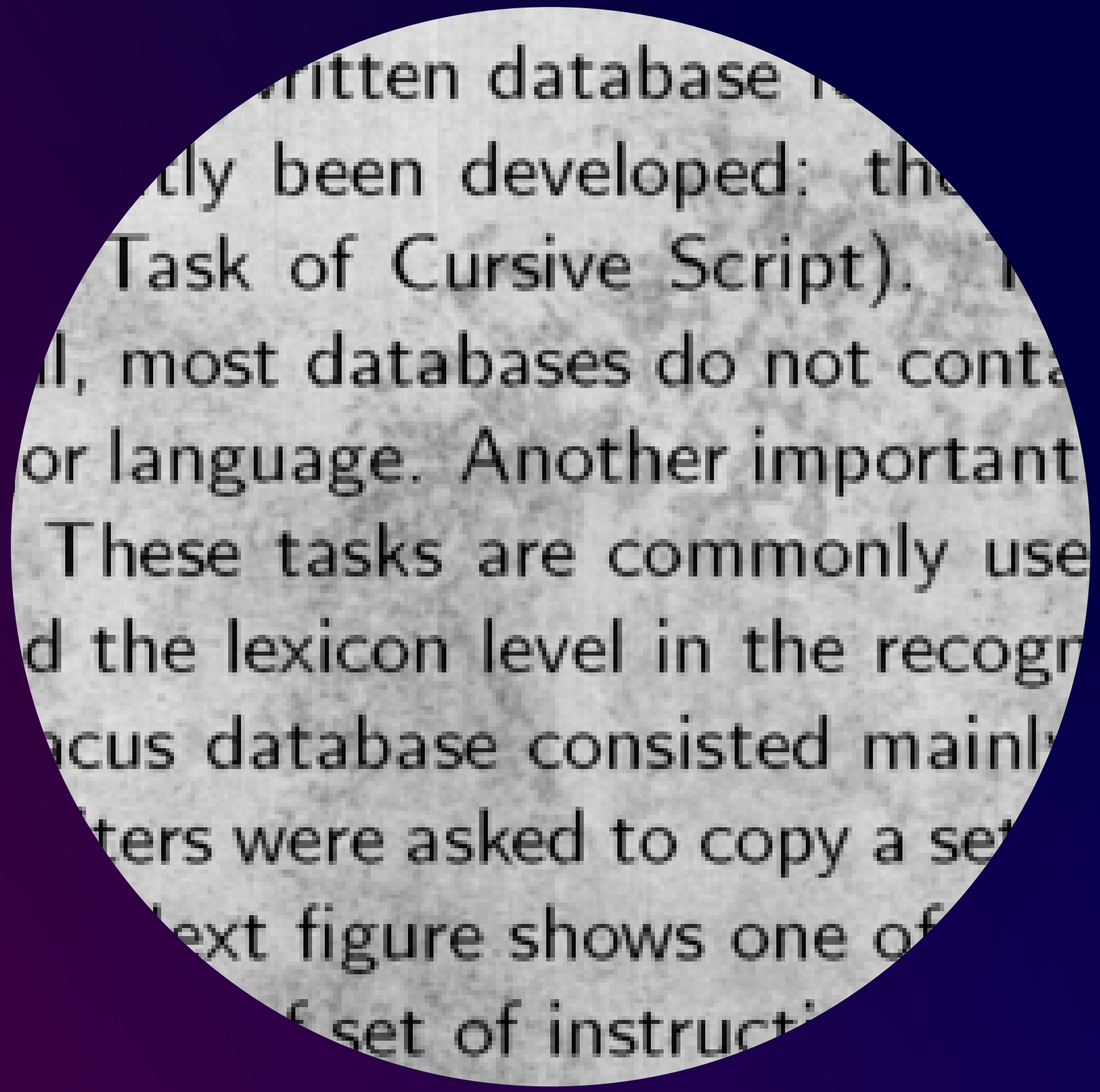
Remove noise from images of scanned text.

PRESENTED BY

Kushal Chakraborty(224101032)
Evanstar Field War(224101018)
Abhik Kumar Khan(224101002)
Vishwas Paikra(224101058)

30/04/2023

Contents



Problem Summary

- Scientific papers, historical documents/artifacts stored as paper, handwritten/typed.
- With time paper accumulates noise through dirt/ fingerprints /stains.
- Several cleaning methods used for preserving, but have risks.
- Digital denoising aimed at creating HiFi recreation of original docs.

https://www.kaggle.com/competitions/denoising-dirty-documents/

Challenge

Goal

Denoising dirty documents .

Task

Do a comparative study of traditional computer vision techniques vs deep learning networks when denoising dirty documents.

Dataset

- UC Irvine Noisy Office Dataset.
- 200 train and 200 test images.

https://drive.google.com/drive/u/0/folders/1CT3fvzjVeHbAss9tFkwN7vM48gSnIc6W

DA 526 (IPML)

Approach

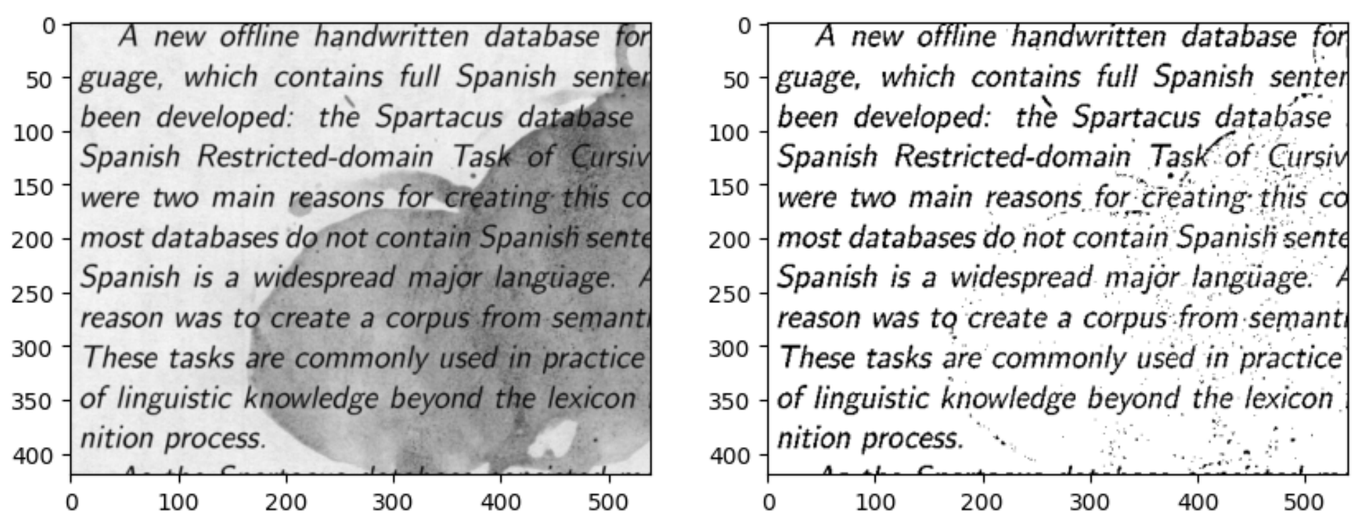
Several Approaches from traditional Computer Vision Techniques to Neural Networks.

Computer Vision — MEDIAN FILTERING, EDGE DETECTION, DILATION & EROSION, ADAPTIVE FILTERING.

ML Techniques — LINEAR REGRESSION.

DNN — AUTOENCODERS.

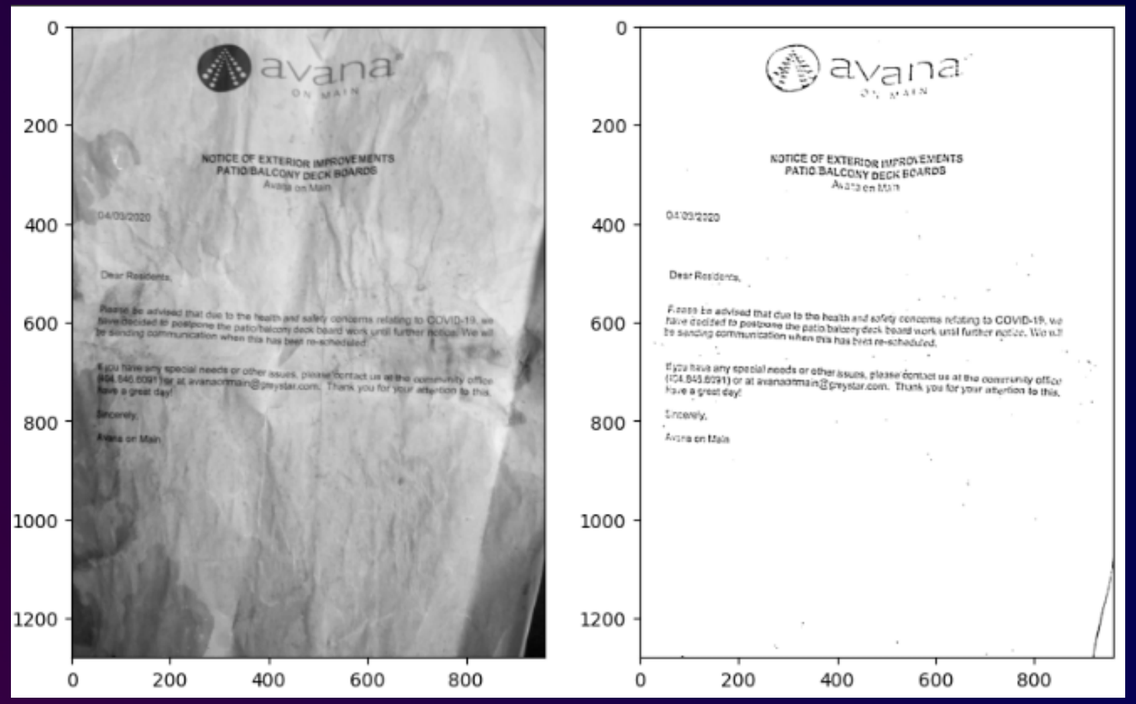
Image Thresholding



Result: (1) Dirty image (2) Denoised image by Thresholding

Thresholding is a simple image processing technique used to segment an image into regions based on intensity values.

Image Thresholding



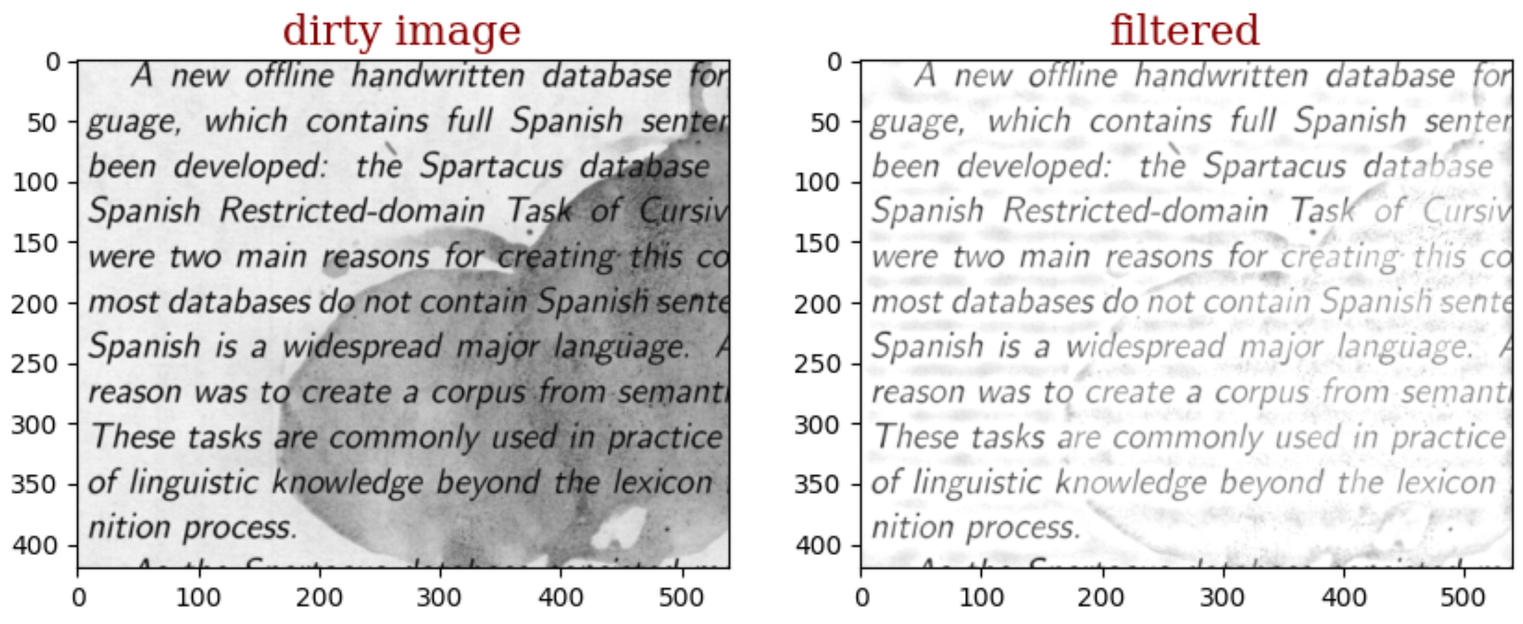
Limitations

- Limited noise removal capability
- Threshold selection
- Loss of image information

Result: Adaptive Thresholding on Actual document

Colab link

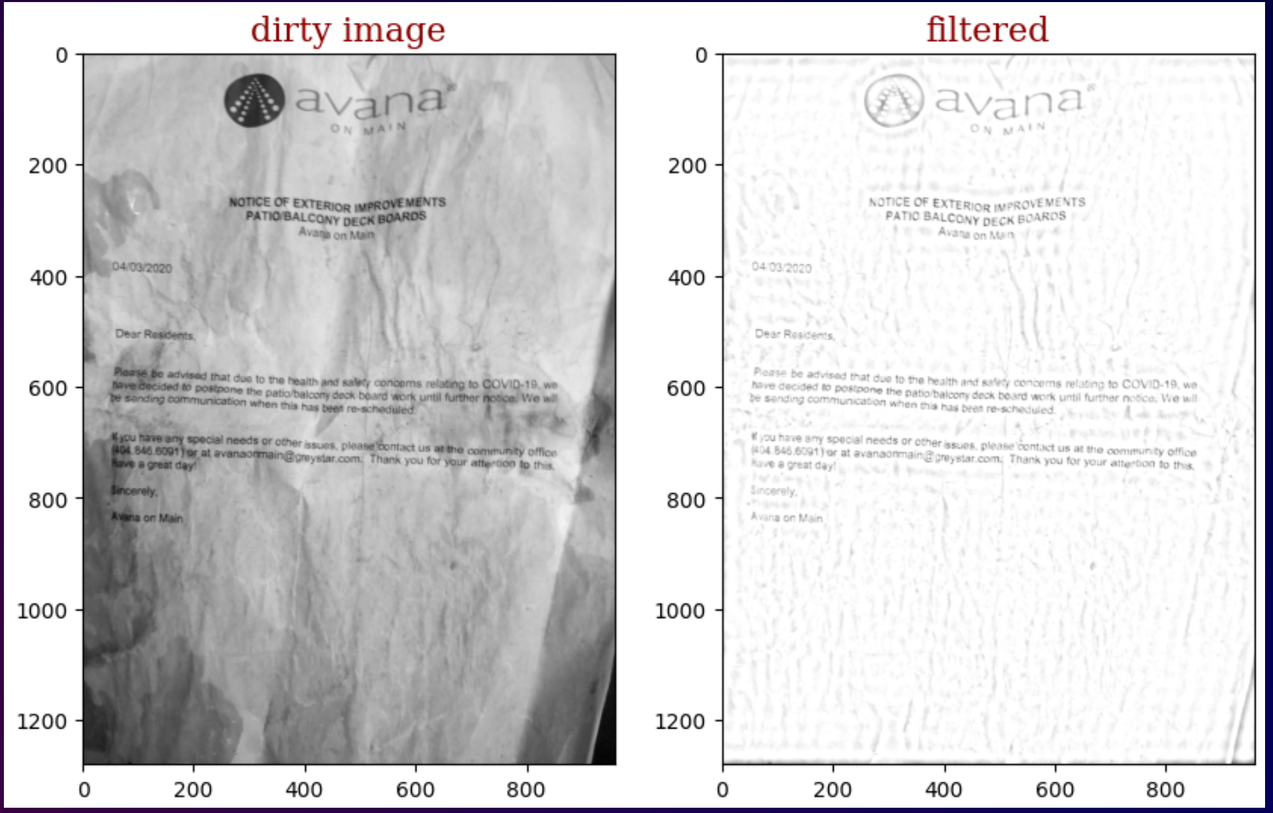
High pass filtering



- A high-pass filter is applied to the transformed image by iterating over each frequency component and setting to zero if the frequency component is within a circular region of radius 0.03.

DA 526 (IPML)

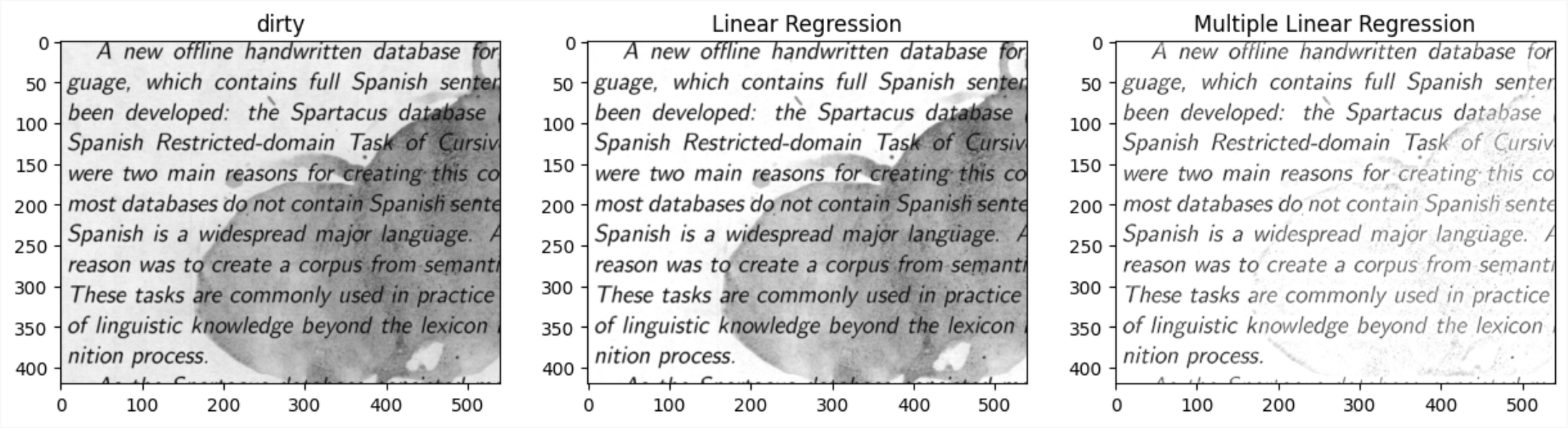
High pass filtering



- It has been observed that high pass filter can causes ringing effect due to sharp circle cut off.
- Hence the need arises for ML techniques.

https://colab.research.google.com/drive/17xeHducQZ1w1Zzwntsq5U4c0QhRSsI6v?usp=sharing

Linear Regression

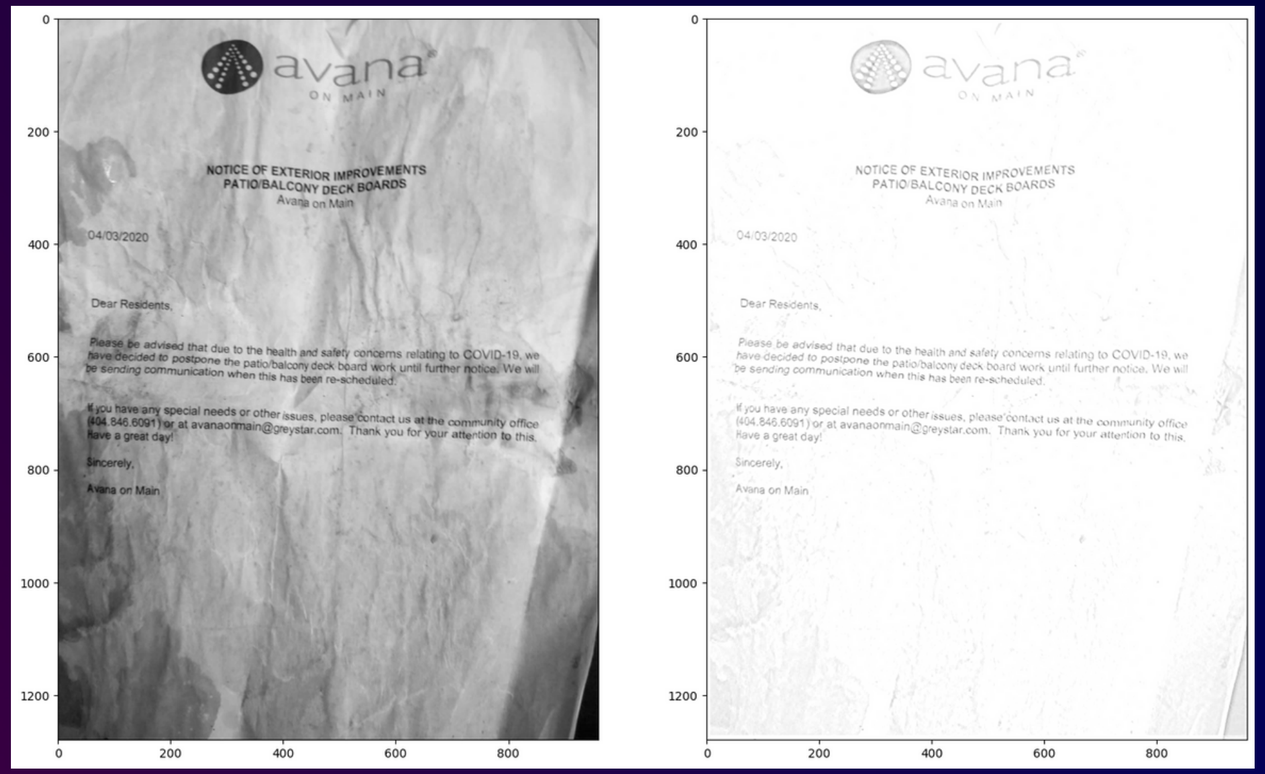


Result: (1) Dirty image (2) Denoised image by Linear Regression (3) Denoised image by Multiple Linear Regression

https://colab.research.google.com/drive/1bCRAJa6FXzWrCLsal9C1Ge_PwAA34iRc?usp=sharing

DA 526 (IPML)

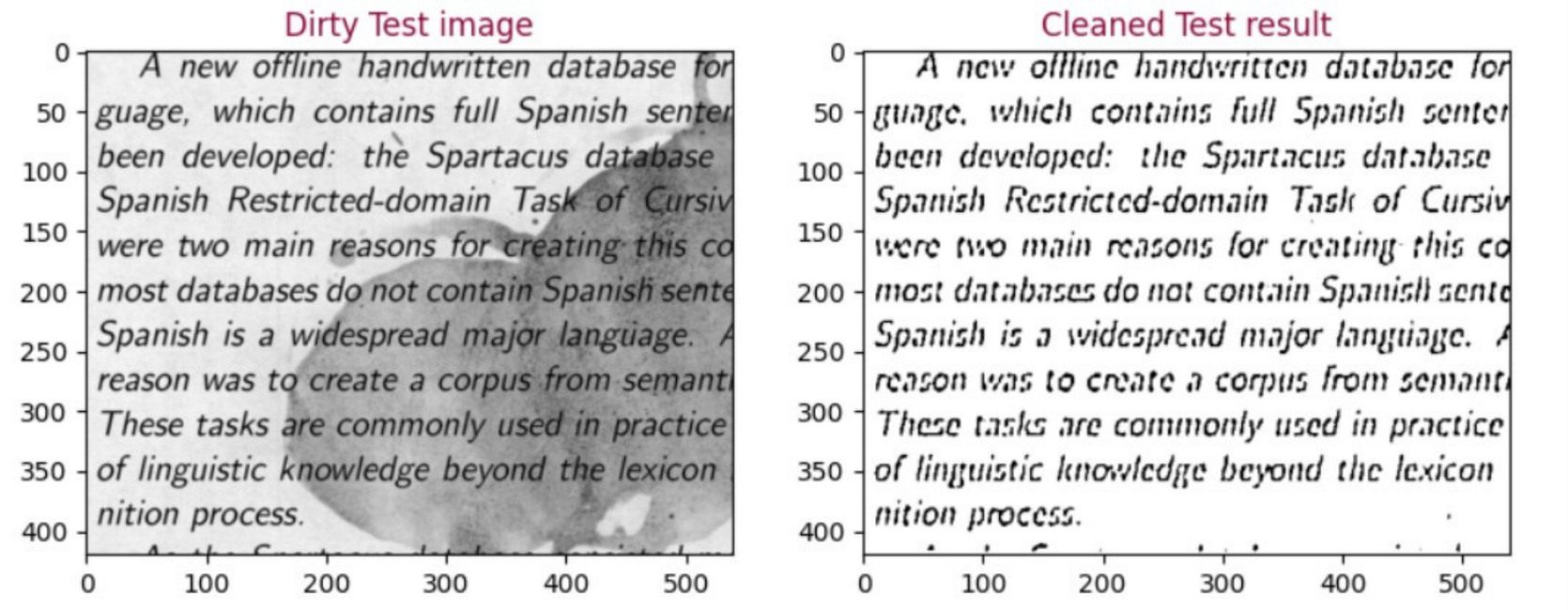
Linear Regression



- Linear Regression: one feature
- Multiple Linear Regression: multiple features
- Generation of Features
- Two models for different Patterns of stains
- More features improved results
- Features increased computing

Result: Multiple Linear Regression prediction on Actual document

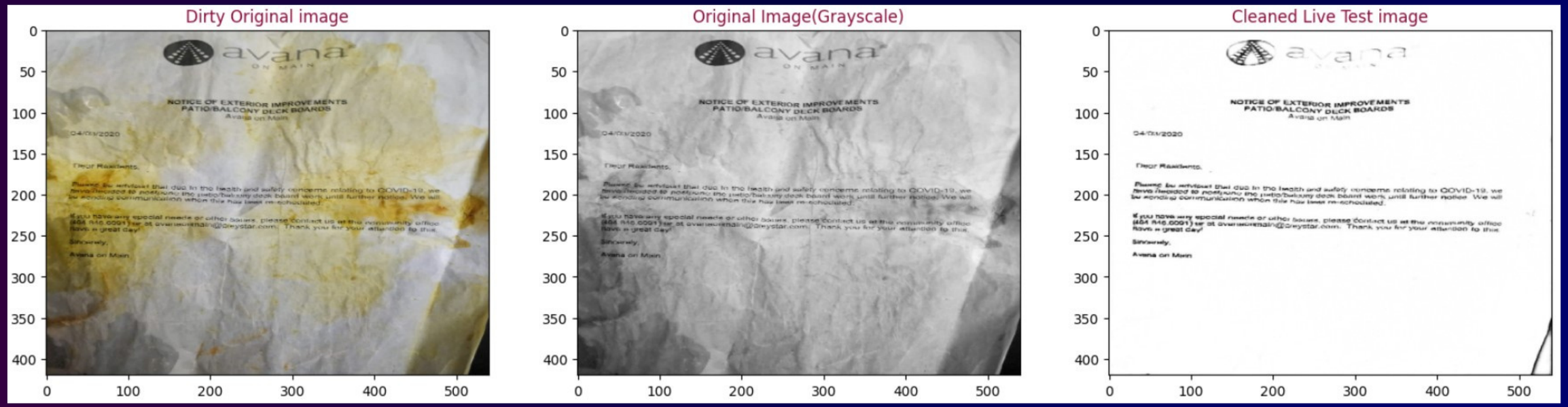
Autoencoder Results



- Autoencoder removes all stains.
- However we loose some features

DA 526 (IPML)

Autoencoder Results



https://colab.research.google.com/drive/1FN5TBXLgdRSz2znfuuOmnI9aGlzScA-C#scrollTo=omaqqhJtaK_C

DA 526 (IPML)

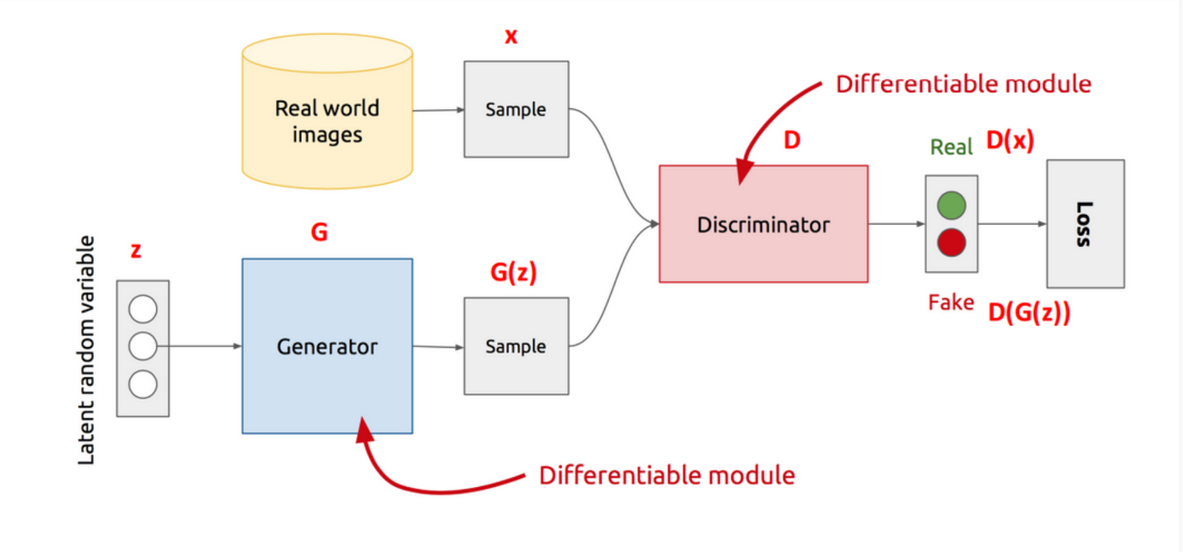
Performance Metrics of each Method

	RMSE	PSNR	UQI
Adaptive Thresholding	35.66	17.08	0.98
High Pass Filtering	15.30	64.43	0.97
Linear Regression	24.77	20.25	0.99
Autoencoder	234.29	0.735	6.409

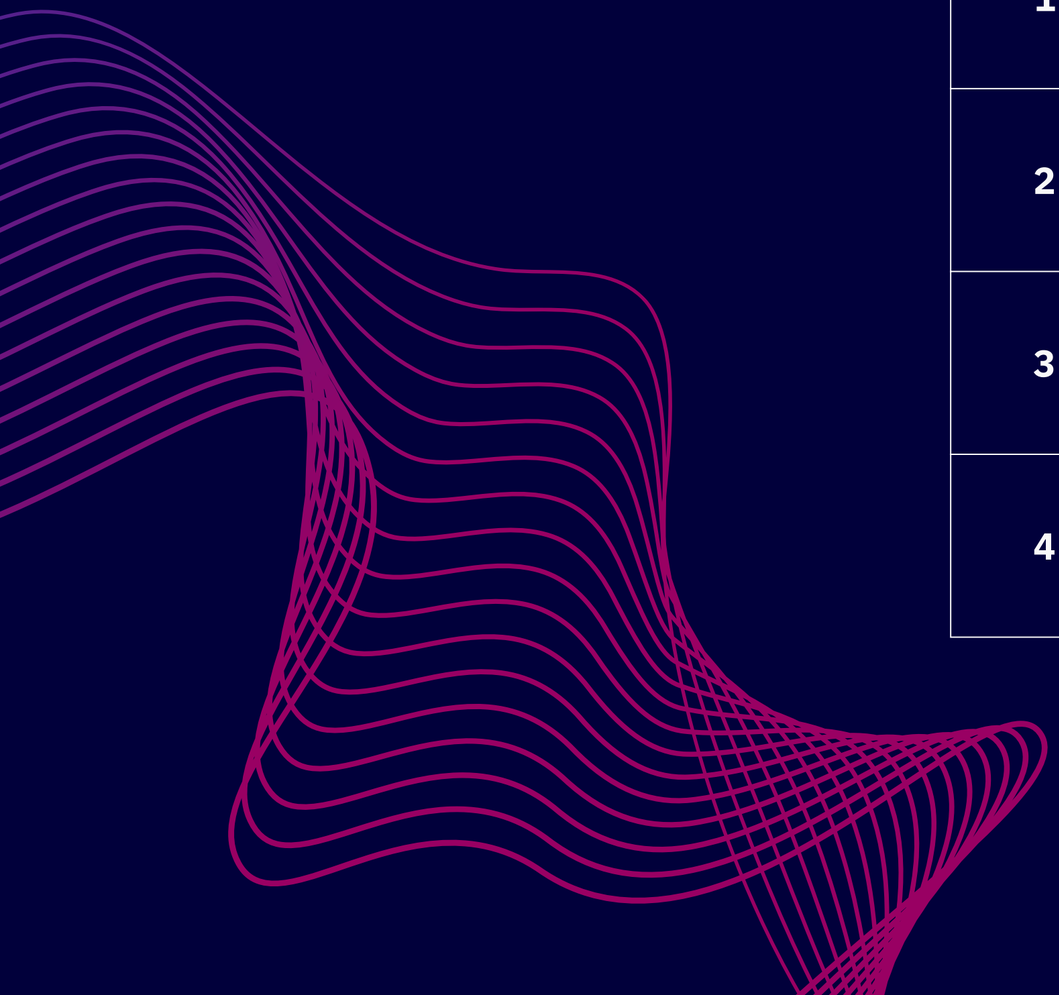
RMSE: Root Mean Square Error
PSNR: Peak Signal to Noise Ratio
UQI: Universal Quality Index

Ongoing Work

- GAN(Generative Adversarial Networks) is known to generate very realistic images.
- Generator tries to fool discriminator.
- Discriminator distinguishes between real and fake samples.
- Loss is sent as a feedback to Generator.
- In the process the model learns the salient features of the input data.



Resources



1	https://www.kaggle.com/competitions/denoising-dirty-documents
2	https://archive.ics.uci.edu/ml/datasets/NoisyOffice#
3	https://towardsdatascience.com/denoising-noisy-documents-6807c34730c4
4	Courses like IITG DA 526, Stanford's CS 229 and CS 231n.

Thanks