Text Detection and Recognition for Text Extraction of Stained Scanned Printed Document Images

1st Marie Emmanuelle Tacsay
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
qmettacsay@tip.edu.ph

2nd Daniel Kimbell Yu
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
qdkmyu@tip.edu.ph

3rd Roman Richard
Department of Computer Engineering
Technological Institute of the Philippines
Quezon City, Philippines
rrichard.cpe@tip.edu.ph

Abstract* - Printed documents are one of the most important physical records which can have content that hold significant information. As these documents are vulnerable to degradation of its quality such as stains and spilled ink., which would greatly affect the content of such documents. Retrieving these information would be important and crucial for keeping information would maintain the record of such events or knowledge. This study was conducted for the purpose of retrieving such information that may be important from degrading printed documents and help to retrieve these information from disappearing. The mentioned factors of why printed documents would be damaged are the noise factors that this study included to be denoised through autoencoders. The denoising of images was done through three models namely, Autoencoder, UNet, and Noise2Void, but autoencoders showed great performance when it came to denoising images from among the three. With this, preprocessing of images through denoising models was done and other preprocessing methods such as binarization, edge detection, as well as, gray-scale done and was later fed into the Optical Character Recognition (OCR) model. These steps were necessary due to the quality of the images after denoising having their quality dropped when being regenerated by the autoencoder. The OCR models would be used to both detect and extract the information that was on the document and print it into readable text.. The researchers tested three models: Tesseract OCR, Easy OCR, and Keras OCR, with Tesseract OCR being the best in terms of performance among the three. Tesseract OCR observed both low CER and WER compared to both models that it was being compared to.

Keywords* - Denoising, Deep Learning, OCR, Scanned Documents, Autoencoder, Tesseract, CNN, Text Extraction, Text Recognition

I. INTRODUCTION

With successful applications dating back to the early 1990's, Document Image Analysis and Recognition (DIAR), which primarily dealt with analysis and recognition of documents, with focus on scanned documents [1]. and studies recorded from years prior, the field has also experienced advancements, alongside the advancing technology that is present in modern time, specifically with the use of Artificial Neural Networks (ANN) and the influential introduction of Back propagation, as well as the first widely accepted architecture known as Convolutional Neural Network(CNN), creating new waves for further research on ANN, which has also benefited the field [2]. These different deep learning techniques are used not only in the field of DIAR, but is also present in a closely related field, known as Handwritten Text Recognition (HTR), as a means for character recognition [2].

Document images contain various useful information in the form of text, graphics and pictures that depict the records of various contents that impose significant information such as contracts, historical records, manuscripts, academic contents, information about the identity of an individual, record of legal cases, etc [4]. These printed documents are commonly stored as images which then can be retrieved using tools or softwares [5]. With this being said, there are several factors which affect the condition of recognition of these texts, such conditions include: low quality paper, color deformation, low contrast of the text, different shapes of alphabets due to the use of fonts that are no longer standard, and so on. These factors, along with the possible fading of the ink prove to challenge the recovery of the contents of the information within these documents and cause failures in even state-of-the-art Optical Character Recognition (OCR) methodologies [3].

Scanned images of documents should be first considered of high quality wherein OCR would give a high accuracy of text detection. With this denoising images would be implemented and this technique have attracted the interest of researchers which still remains challenging and open for improvement [6]. According to Alshathri et. al (2021), Autoencoders are neural networks that are also known in applying denoising images which is justified since autoencoders are typically used in unsupervised data. Through denoising of images using autoencoders up to detection and recognition of text using OCR, text extraction would be

possible which can be of use to retrieved documents.

II. METHODOLOGY

A. Dataset

The dataset is composed of 504 images which are used in this study and a mixture of both document images or scanned documents already being published online and documents on-hand for both training and testing dataset. The document images have its noise such as bleeding of ink, marks of folded paper, fading ink, etc., which this study would be needed for training denoising autoencoder model to enhance its quality for text detection and extraction.

B. Image Pre-processing

Every image that the researchers have used had undergone preprocessing before placing it to the model for the process of denoising the images. The images were read in greyscale which allow the model to extract the necessary features of the images and process the data properly. Also, the images were reshaped into the closest size of the original size of images (360x612). After these methods, the images will be normalized so that the shape of each array representing the images would fit together in the model.

Furthermore, to prepare for the image for Optical Character Recognition (OCR), several methods were used to preprocess the image, those being: Sharpening of the image, Gray Scale, Binarization, and Edge Detection. These methods were utilized to ensure that the results of OCR models would be optimized. To discuss more on the different steps, from observations, the researchers had noticed, after denoising of the image, the image lowered in quality, and thus utilized a kernel to sharpen the image, proceeding from this the images were subjected to gray scale to ensure visibility. With this being said, Binarization was used to darken text that had faded from the denoising of the image, and finally edge detection was used to obtained the version of the image that had the least amount of noise, this method is utilized, as so the model is able to identify the different letters and the spaces between in between in letter.

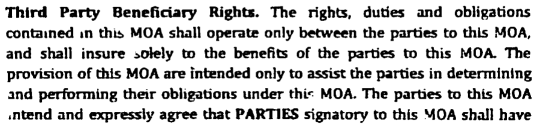

Fig. 1. Sample Image of Binarization

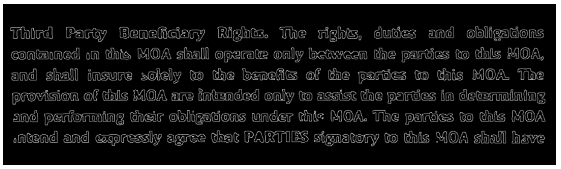

Fig. 2. Sample Image of Edge Detection

C. Visualization

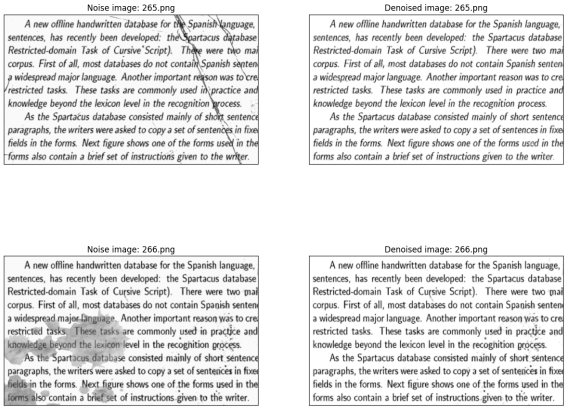

Fig. 3. Sample Image of Denoised Documents

As aforementioned, the dataset consisted of images that contained either markings or stains, and as such was subjected to go through denoising using autoencoders. The autoencoder had undergone training prior to learning and the images below are the result of the denoising.

D. Optical Character Recognition (OCR)

Optical Character Recognition or otherwise known as OCR, for short, refers to the process in which the conversion of image into a machine-readable text format, examples of such documents are not limited to documents but can also refer to signs, billboards, and logos. OCR can be seen in different applications, ranging in various languages, and not being limited to only English. OCR can be broken down into four different steps to achieve it: 1) Image Acquisition, 2)Text Recognition, 3) Text Recognition, and 4) Post Processing. The three (3) models that were used in the study were as follows: Tesseract OCR, Easy OCR and Keras OCR. When comparing the three different models.

Tesseract Technology supports short-short memory neural networks, which can be utilized for the recognition of english text [8] Tesseract is an open-source project which is owned by google. The reason why Tesseract is more accurate than other OCR engines, is due to it checking for any missing words or mispelled words and if there are any present and then correcting it [10]. Easy OCR utilized python and pytorch deep learning library this OCR uses the Character Region Awareness for Text (CRAFT) algorithm for

detection, and Convolutional Recurrent Neural Network (CRNN) for recognition, this OCR consists of three parts: decoding, feature extraction and sequence labeling [9].

E. Denoising Models

As one of the pre-processing of the image in order for the OCR to detect and extract the text, autoencoder was used to denoised the images. Autoencoders are one of the deep learning models that are commonly used in various fields in technology in which this study will use them as denoiser of images. Autoencoder is often used as an image denoiser of various images which generates new cleaned images for a better visualization of the data [7]. Autoencoders have layers where the images will be processed to generate new and close-to-original images. However, not only autoencoders are the one that can be used as denoiser models because CNN models could show some good performance when it comes to denoising images. This study would only not use autoencoders as the denoising model for the purpose of preprocessing of images because CNN would also be explored, compared and used in order to show comparison on the performance of both autoencoders and CNN models when it comes to denoising images.

F. Performance Evaluation Metrics

To measure the performance of autoencoders as a means for denoising images, the following metrics were used as benchmarks, these metrics being: Mean Squared Error (MSE) Loss and Mean Absolute Error (MAE). These performance metrics would be the evaluation metrics in the process of denoising images.

Mean Squared Error (MSE) Loss is a loss function that quantifies the magnitude of the error between the predicted and actual value through taking the squared difference between the prediction and the target values. This metric is calculated through the following formula below:

$$MSE = \left(\frac{1}{n}\right) \bullet \Sigma(actual - forecast)^2$$

Whereas, the Mean Absolute Error (MAE) also refers to the magnitude of the error between the predicted and actual values; however, it takes the average of the absolute errors of the predictions and observations as the measurements of the magnitude of errors of a group of data or values. It can be calculated through the following formula below:

$$MAE = \left(\frac{1}{n}\right) \bullet \Sigma|actual - forecast|$$

When measuring the performance of the OCR models, the researchers took into note different metrics to quantify the performance of the model. These metrics primarily being: Confidence Level, Character Error Rate (CER) and Word Error Rate (WER), but upon further studying on the topic found that using confidence level as a metric proved to be unreliable when detecting the different text, and thus CER and WER was decided on by the researchers to be the preferred performance metric.

Character Error Rate (CER) is a metric which measures the accuracy of predicted text through – substitutions, deletions. This differs from word-level errors, this metric is useful in surfacing mispronunciations and erroneous phonemes. This metric is calculated using the following formula

$$CER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ words\ in\ Reference}$$

Similarly to CER, Word Error Rate (WER), is a metric that also measures the text based on three different types of errors - substitutions, deletions, and insertions. Though CER and WER both use these three types of errors as their basis, Word-level error surface mispredicted words, This metric is useful to visualize the common word-level failures that exist in the model, as so we are able to flesh out the weaknesses present. WER is be calculated using the following formula

$$WER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ words\ in\ Reference}$$

With this being said, when analyzing these metrics for OCR, a lower CER and WER is preferred within a model. These metrics in this study were calculated by comparing the ground truth vs the predicted text of the model, and thus having lower CER and WER would be desired by the researchers.

G. Model Deployment

For our web app deployment, a web application named *"Streamlit",* The web application would be the combination of both methods proposed in the research which are: Denoising and Text Extraction. The web application utilizes the best performing models that the researchers have tested.

Fig 4.GUI of the Application

The model demonstration is also viewable with the link below:

GROUP11_MODELDEPLOYMENT_VID…

III. RESULTS AND DISCUSSION

Before being subjected to Optical Character Recognition, three different models were tested and compared when denoising the images. These models being: Autoencoders , UNet, and Noise2Void. The following are the results obtained from the different denoising models. Firstly the three models were used with the visualization of images.

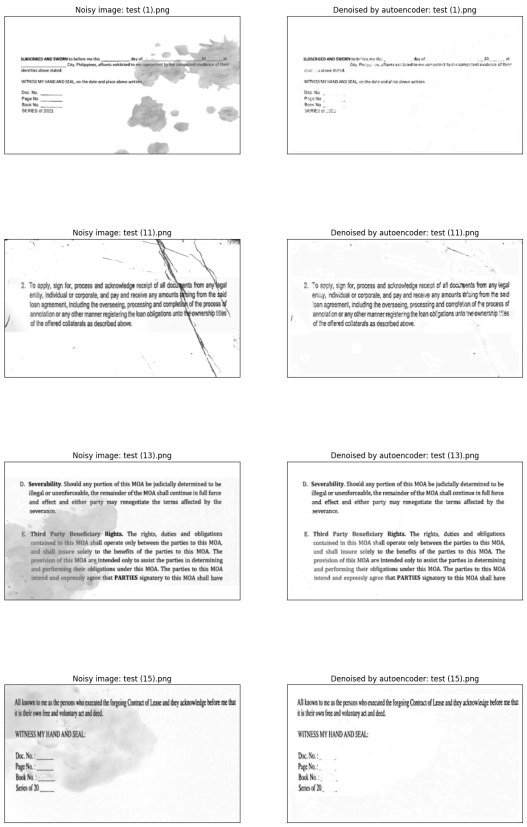

Fig 5. Autoencoder Denoising Visualization

As seen in the image above, the autoencoder model was able to effectively clean the image. The autoencoder was able to fully clear out the dark stain on the image making it easier to read.

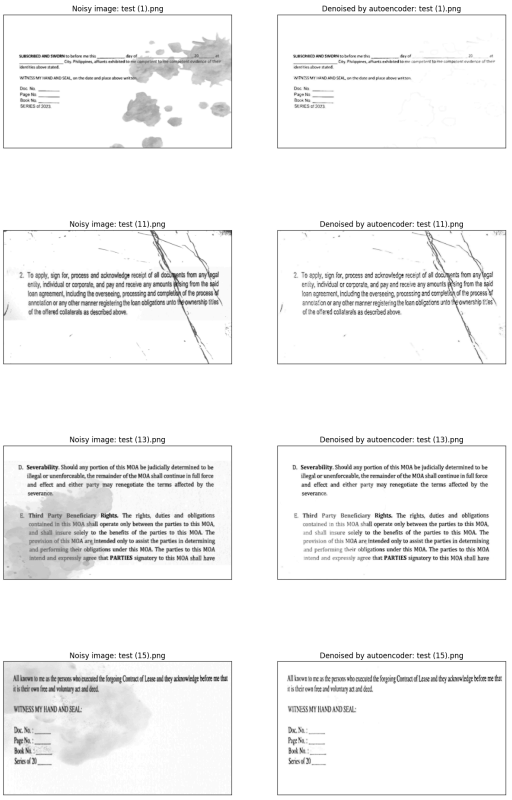

Fig 6. UNet Denoising Visualization

As one of the CNN models, UNet proved its quality when it comes to denoising images which can be seen from the images shown above. Most of the document images were cleaned because of the model and its ability to denoise or clean images.

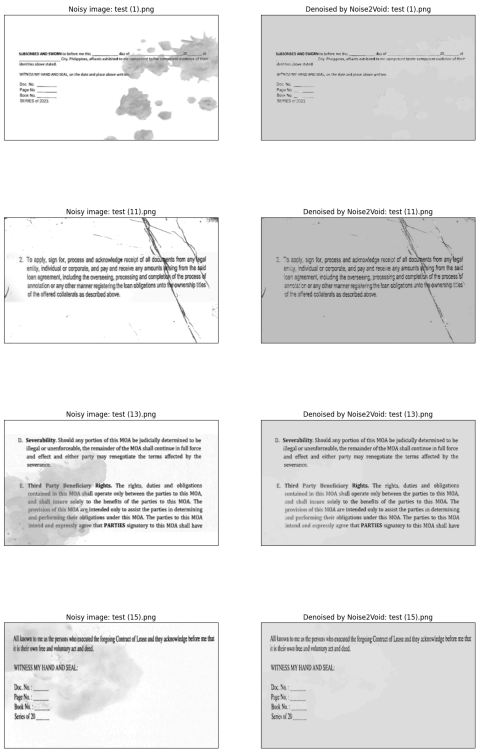

Fig 7. Noise2Void Denoising Visualization

As one of the CNN models, Noise2Void also proved its quality when it comes to denoising images which can be seen from the images shown on the left lower corner of the page. The only difference of this model to the previous ones is that the generated result of images is darker compared to the others. The images look to have gray color background and also have slightly obvious stane which is an indication that the model wasn't able to perform its best in denoising the data.

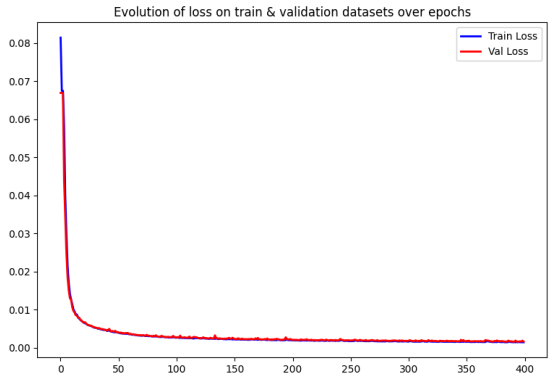

Fig 8. MSE Loss Function of Autoencoder

MSE Loss Function is one of the most performance metrics of denoising models since loss function indicates that the model was able to retain the significant features of the data while processing and denoising images inputted.

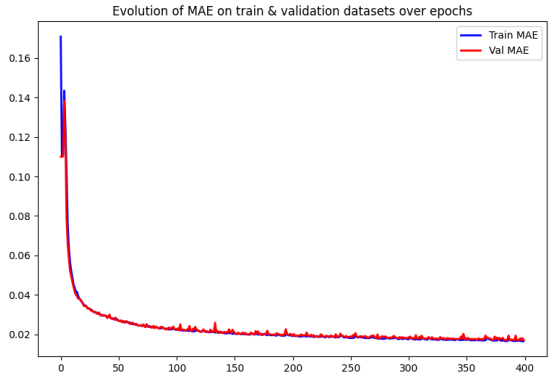

Fig 9. MAE Function Graph of Autoencoder

As seen in the graph above, the data trained was able to adapt to the train data which will later on be used to denoise unseen data. The graph above shows an ideal visualization of a mode's performance when it comes to denoising images. Also, because of these graphs, it is evident that generating cleaner images or results is possible.

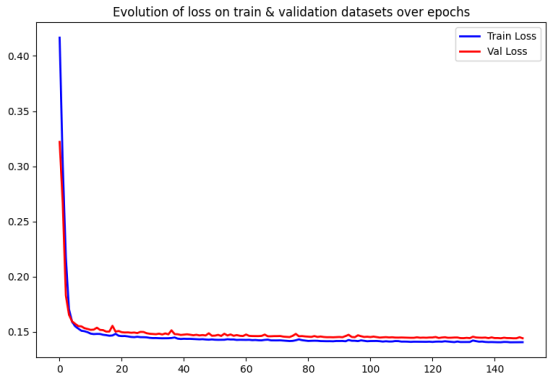

Fig 10. MSE Loss Function Graph of UNet

The figure above shows the MSE Loss Function Graph of UNet in which, according to it, the model showed a promising performance when it comes to denoising images. The graph shows an almost ideal curve or flow between the train and validation losses which is also an indication of a good performance of the model in denoising images.

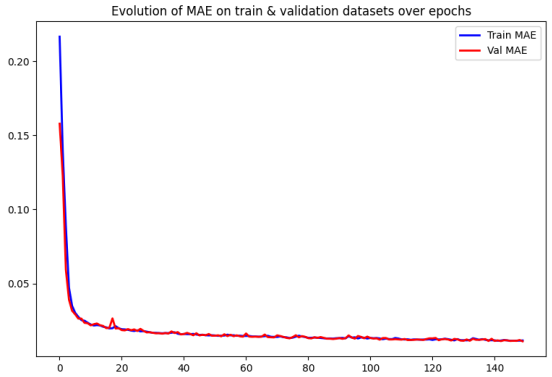

Fig 11.: MAE Function Graph of UNet

The graph above shows the MAE Function of the UNet Model in terms of denoising the images. Just like the previous model, the performance of the UNet reflects on the values generated as MAE function which in this case is an indication of good performance in terms of denoising images.

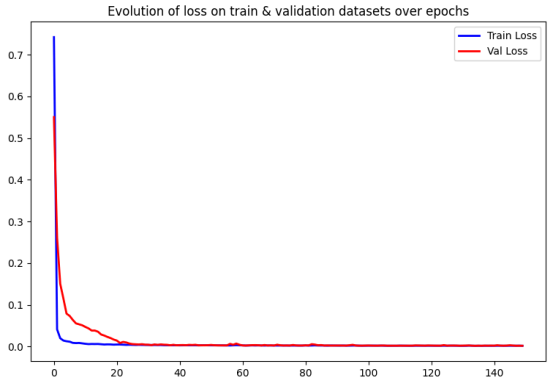

Fig 12. MSE Loss Function Graph of Noise2Void

Since MSE is one of the used performance metrics for models, using this metric would be

appropriate in terms of monitoring performance of the model used in denoising images. MSE Loss Function would be able to determine if the model can maintain the quality of the images while processing them since generating the result would be focused on generating them based on the original data. As what the graph showed, the performance of the CNN model, Noise2Void, seems to generate good results and was able to generate a good performance through the graph.

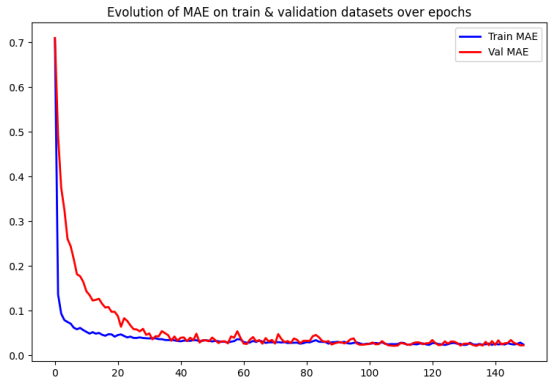

Fig 13. MAE Function Graph of Noise2Void

As for the last performance metric which is MAE, just like MSE, it is also one of the metrics that is commonly used to evaluate the performance of denoising models. As shown in the graph, the model was able to show great performance in terms of denoising images because of the graphs generated. This would indicate that this model also is appropriate and can be used to denoise images.

From all the results generated by the two models mentioned and used in this study, autoencoders still give the best performance compared to others. It is because the denoised images of the autoencoder model are cleaner compared to the other two. In addition, UNet and Noise2Void models still left stains on the images to process and denoised which made the autoencoder be the best to perform when it comes to denoising images. Even though these three models showed good visualization of their performance through graphs, comparing the denoised images generated by these three models, autoencoder had generated cleaner images than the two models used for this study. Thus, using autoencoder for model deployment as well as using it as the denoising model as the preprocessor of the images for text detection and extraction, is more appropriate and guarantees good quality in terms of denoising stained document images.

After comparing and visualizing the results of the denoising models, the images were then put under different Optical Character Recognition (OCR) models for comparison namely: Tesseract OCR, Easy OCR, Keras OCR. Below are the results with the image for reference.

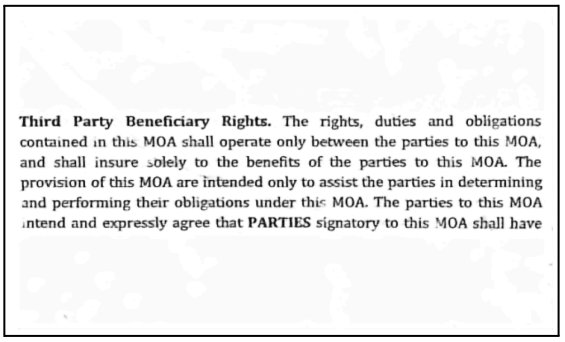

Fig 14. First Denoised image used for OCR

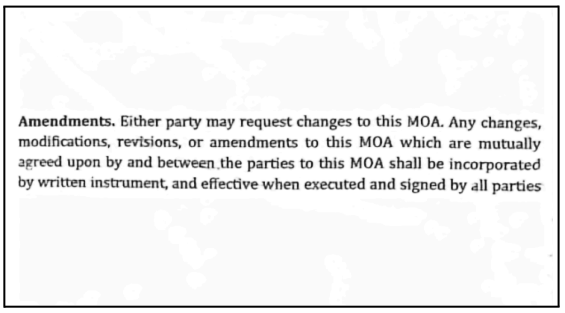

Fig 15. Second Denoised image used for OCR

The images above were subjected to previously mentioned denoising models, and will be used for testing of the performance of the OCR models.

```
Text Detected:
Third Party Beneficiary Rights. The rights, duties and obligations
contained in this MOA shall operate only between the parties to this MOA,
and shall insure solely to the benefits of the parties to this MOA. The
provision of this MOA are intended only to assist the parties in determining
and performing their obligations under this MOA. The parties to this MOA
intend and ogress agree that PARTIES signatory to this MOA shall have
```

Fig 16. Text Detected by Tesseract OCR

```
easyocr_result

['Third Party Beneficiary Rights: The rights, duties and obligations contained in this MOA shall
operate only between the parties to this MOA, and shall insure solely to the benefits of the
parties to this MOA The provision of this MOA are intended only t0 assist the parties in
determining and performing their obligations under thic MOA The parties to this MOA intend and
expressly agree that PARTIES signatory to this MOA shall have']
```

Fig. 17. Text Detected by Easy OCR

```
party beneficiary obligations third duties and rights the rights
moa between moa the this this shall only to in opelate contained parties
of moa the the shall solely the this benefits insure and to to parties
determining moa intended of this only the assist in to parties are provision
under obligations moa this the moa this and performing their to parties
parties moa that this have and ntend shall signatory expressly to agree
```

Fig.18. Text Detected by Keras OCR

To determine the accuracy of the result, the CER and WER of the predicted test was taken and compared with the ground truth. Below are the results of the following tests.

TABLE I. CER Results of OCR Model

Model	Image 1	Image 2	Average of CER Results
Tesseract OCR	0.23%	1.03%	0.63%
Easy OCR	1.38%	0.69%	1.03%

Keras OCR	62.67%	34.36%	48.51%

The table above shows the different Character Error Rate (CER) results of the OCR models between two images. From these results we can see that in terms of CER, Tesseract OCR performed best, scoring generally lower than the other models in this case is ideal, scoring 0.63% when averaging the performance between both images. Whereas, Keras OCR performed the worst among the three with an average CER rate of 48.51%.

TABLE II. WER Results of OCR Model

Model	Image 1	Image 2	Average
Tesseract OCR	1.37%	8.70%	5.04%
Easy OCR	8.22%	4.35%	6.29%
Keras OCR	83.56%	60.87%	72.22%

In the given table, showcase the Word Error Rate (WER) results when testing between the ground truth text and the predicted text. As seen in the table, it can be said that similarly to CER, Tesseract generally performed better when averaging the score, scoring only 5.72% WER, whilst Keras OCR remains to be the most unreliable one in terms of WER as well.

TABLE III. Accuracy Results of OCR Model in CER

Model	*Image 1*	*Image 2*	*Average Accuracy*
Tesseract OCR	*99.77%*	*98.97%*	*99.37%*
Easy OCR	*98.62%*	*99.31%*	*98.97%*
Keras OCR	*37,33%*	*65.64%*	*51.49%*

The table above is the tabular version of the accuracies scored by the model, These indicate how accurately the model is able to predict character-level wise, which means that there were fewer mistakes in terms of substitutions, deletion, as well as insertion of characters, and from the given table Tesseract OCR is the best performing model in this regard.

TABLE IV. :Accuracy Results of OCR Model in WER

Model	*Image 1*	*Image 2*	*Average Accuracy*
Tesseract OCR	*98.63%*	*91.30%*	*94.97%*
Easy OCR	*91.78%*	*95.65%*	*93.72%*
Keras OCR	*16.44%*	*39.13%*	*27.79%*

The table above reflects the accuracy scores attained by the different models in terms of word-level. This would in-turn mean that the model was able to closely match the text that was in the image. From the given table TesseractOCR was the best performing model in this regard, similarly to CER.

To summarize, from the given results of the above, the top performing models, when tested by the researchers, for denoising were Autoenders and Tesseract OCR for the text recognition portion these two models showed promising results, the autoencoder model was able to effectively clean the stained images and whilst also maintaining readability, even when compared to the results of the other three (3) models. When comparing the OCR models, Tesseract OCR had the lowest average CER and WER, amongst the three OCR models. This would entail that from among the three, the model had minimal substitution, insertion, and deletion, whilst also being able to match the contents of the image the closest.

IV. CONCLUSION

From the different tests that were performed on different models, it can be concluded that from these tests, AutoEncoders and Tesseract OCR respectively performed the best for denoising and text recognition. AutoEncoders were able to generally clean the document, whilst maintaining the text printed onto the document with minimal erasures. Whereas Tesseract OCR yielded the lowest CER and WER Rate, which would mean that from among the three, it had the least amount of substitutions and erasure when compared to the actual text.

REFERENCES:

[1] F. Lombardi and S. Marinai, “Deep Learning for Historical Document Analysis and Recognition—A Survey,” *Journal of Imaging*, vol. 6, no. 10, p. 110, Oct. 2020, doi: https://doi.org/10.3390/jimaging6100110.

[2] Diplom-Ingenieur, H. Scheidl, and R. Sablatnig, “Handwritten Text Recognition in Historical Documents DIPLOMARBEIT zur Erlangung des akademischen Grades Visual Computing eingereicht von.” Accessed: Apr. 28, 2024. [Online]. Available: https://scholar.archive.org/work/6poyaajafrfitkgiplszvkifqu/access/wayback/https://repositum.tuwien.at/bitstream/20.500.12708/5409/2/Scheidl%20Harald%20-%202018%20-%20

Handwritten%20text%20recognition%20in%20historical%20documents.pdf

[3] C. Biswas, P. S. Mukherjee, K. Ghosh, U. Bhattacharya and S. K. Parui, "A Hybrid Deep Architecture for Robust Recognition of Text Lines of Degraded Printed Documents," 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018, pp. 3174-3179, doi: 10.1109/ICPR.2018.8545409. keywords: {Image segmentation;Optical character recognition software;Text recognition;Databases;Image recognition;Degradation;Engines},

[4] D. Ghai and N. Jain, "Text Extraction from Document Images- A Review," *International Journal of Computer Applications*, vol. 84, no. 3, p. 40, 2013, Accessed: Apr. 28, 2024. [Online]. Available: https://www.academia.edu/50741354/Text_Extraction_from_Document_Images_A_Review?sm=b

[5] Gangeh, M. J., Tiyyagura, S. R., Dasaratha, S. V., Motahari, H., & Duffy, N. P. (n.d.). *Document enhancement system using auto-encoders*. OpenReview. https://openreview.net/forum?id=S1Mnzp9qLB

[6] M. Ibrahim *et al.*, "Denoising Letter Images from Scanned Invoices Using Stacked Autoencoders," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1371–1386, 2022, doi: https://doi.org/10.32604/cmc.2022.022458.

[7] L. Gondara, "Medical Image Denoising Using Convolutional Denoising Autoencoders," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 241-246, doi: 10.1109/ICDMW.2016.0041. keywords: {Noise reduction;Training;Noise measurement;Noise level;Convolutional codes;Image denoising;Biomedical imaging;Image denoising;denoising autoencoder;convolutional autoencoder},

[8] J. Li, "Research on English Automatic Recognition System Based on OCR Technology," 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT), Jiaxing, China, 2023, pp. 1061-1066, doi: 10.1109/ACAIT60137.2023.10528556. keywords: {Handwriting recognition;Image recognition;Text recognition;Optical character recognition;Neural networks;Information retrieval;Convolutional neural networks;OCR technology;text recognition;Tesseract;image preprocessing},

[9] C. Jeeva, T. Porselvi, B. Krithika, R. Shreya, G. S. Priyaa and K. Sivasankari, "Intelligent Image Text Reader using Easy OCR, NRCLex & NLTK," 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICPECTS56089.2022.10047136. keywords: {Emotion recognition;Vocabulary;Social networking (online);Text recognition;Anxiety disorders;Speech recognition;Real-time systems;Easy OCR;Text-to-Speech;Translation;Emotion Detection using NRCLex (lexicon-based method)},

[10] S. Sundara Pandiyan and C. Kelvin, "A Translator for Indian Sign Boards to English using Tesseract and SEQ2SEQ Model," 2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM), Mathura, India, 2021, pp. 1-4, doi: 10.1109/SASM51857.2021.9841215. keywords: {Machine learning;Optical character recognition software;Smart manufacturing;tesseract OCR engine;seq2seq model and Tamil OCR.},