UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE (SPECS)

IDENTIFICATION OF BURNED-IN TEXT DATA IN LOW RESOLUTION MEDICAL IMAGING MODALITIES USING DEEP LEARNING TECHNOLOGY

BY EFOSA OSAGIE

Under the Supervision of

Dr. Wei Ji and Dr. Na Helian

Submitted to the University of Hertfordshire in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

July 2024

Dedication

I dedicate this PhD thesis to my late father, Mr George Osagie. He always wished I had achieved the highest academic degree and stood by me until his last breath. I wish you were here with me, but God knows best.

Declaration

I hereby declare that the contents of this PhD thesis are original and have not been submitted wholly or partly for consideration for any other degree or qualification in this or any other university. This PhD thesis is my entire work; references were provided where appropriate. This submission contains less than 60,000 words, including appendices, bibliography, tables and equations, and fewer than 150 figures.

Efosa Osagie

July 2024

Acknowledgements

This PhD would not have succeeded without the help of many people who provided me with their support, time, and resources. I would like to first and foremost give thanks to God, the Inscrutable Creator and Governor of the whole Universe.

I am also highly indebted to my supervisors, Dr Wei Ji and Dr Na Helian, who have found time now and then to review my work and make recommendations. Without their indispensable guidance, this research work would not have been completed. Over the course of four years, I exchanged hundreds of emails and several meetings, which shaped me as a better researcher. Every supervision meeting was always impactful. I remain eternally grateful.

My thanks and appreciation also go to my family for their understanding, continuous prayers and support during this phase of my academic life. They remain a strong pillar of support in all my endeavours.

Finally, I would also like to thank Engr. Uyiosa Egbe, who initially suggested that I pursue this PhD, his words of motivation and belief never withered. Similarly, Engr. Haruna Idoko was always there to provide mental support during challenging times when the journey seemed challenging. I am eternally very grateful to the medical imaging and diagnostics department of Pison Laboratories, Equity Medical Laboratories, and Union Medical Diagnostics in Abuja, Nigeria, for their cooperation during the data collection, support, and resourcefulness.

My endless thanks to you all.

Abstract

Medical Image character recognition (MICR) has become a useful application of optical character recognition (OCR) models due to the advancement in computing resources, large databases of medical image records and the need for an efficient information retrieval system for various needs. However, with the unique nature of medical image modalities (MIM) such as X-rays, Ultrasounds and Magnetic Resonance Imaging (MRI), where patients' demographics and clinical examination data exist as burned-in text on the pixel content in small font sizes with overall image low-resolution, application of traditional OCR on these low-quality image results to a poor accuracy. The traditional OCRs cannot recognise these burned-in texts under these conditions, as they are designed for mainly bi-level text with resolutions of 150DPI and above and scanned documents with a minimum of 300DPI. In contrast, these MIM have a low resolution of 96 dpi.

To solve these challenges, this thesis explores the application of deep learning techniques in the aspects of deterministic modelling, semantic similarity learning, and generative modelling to solve the problems in MICR, which are low resolution, small text, small sample size and background interference. This thesis developed an ensemble of Convolutional Neural Networks (CNN) inspired by the classical Lenet-5 architecture to recognise burned-in text at the character level. Experimental results show promising results when compared with the state of the art. Furthermore, to increase the character recognition rate of the CNN models when dealing with visually similar characters (VSC), this thesis proposed and designed a channel attention-based Siamese network to efficiently apply metric learning and few shot techniques on recognising VSC while training on small sample size per class. The evaluation showed that the Siamese network could discriminate between VSC in MIM compared to regular multi-class classifiers.

To deal with the small sample size problem caused by privacy concerns when acquiring MIM for deep learning tasks, this thesis proposed, deployed, and evaluated a conditional variational autoencoder (CVAE) to generate synthetic image data. The evaluation shows improvement in the accuracy of deterministic models when trained with augmented images generated by the proposed CVAE model.

To ensure the generability of this thesis's findings, two datasets were used for the evaluation: an open-source medical image dataset and a privately collected medical image dataset whose collection was approved by the University of Hertfordshire’s ethics committee. An accurate MICR solution can improve health data analytics by allowing a more accessible and accurate extraction of data from MIM. This can assist in analysing image data to identify patterns, thereby improving patient care and diagnosis.

Table of Contents

Dedication i

Declaration ii

Acknowledgements iii

Abstract iv

Table of Contents vi

List of Figure xii

List of Tables xiv

1.0 Introduction 1

1.1 Overview 1

1.2 Problem Statement 4

1.3 Research Aim 7

1.4 Contributions 8

1.5 Publications related to this thesis 11

1.6 Ethical Consideration 11

1.7 Outline 12

2.0 Background 14

2.1 Medical Imaging Applications: Clinical challenges and issues 16

2.2 Preliminary Definitions 19

2.3 Optical Character Recognition (OCR) 21

2.3.1 Segmentation-based OCR Approach 25

2.3.2 Segmentation-free OCR 26

2.4 Machine Learning 27

2.4.1 Supervised Learning 27

2.4.2 Unsupervised Learning 30

2.4.3 Machine Learning Optimization and Hyperparameters 33

2.5 Deep Learning 38

2.5.1 Multilayer Perceptron (MLP) 39

2.5.2 Convolutional Neural Networks 40

2.5.3 Convolutional Recurrent Neural Network (CRNN) 43

2.5.4 Attention Mechanism 45

2.5.5 Few-Shot Learning Method 46

2.5.6 Generative Models 47

2.6 Chapter Summary 50

3.0 Literature Review 51

3.1 Nature and Acquisition of Medical Image Modalities (MIM). 54

3.2 Character-level Recognition - A Review and Justification 55

3.3 Machine Learning-based Techniques for Burnt-in Textual Data Recognition . 56

3.4 Overview of previous works in the use of ML in recognising burned-in text in MIM 59

3.5 Open issues and challenges in Burned-in Textual Data Recognition 65

3.6 Literature Review Conclusion 68

3.7 Evaluation Metrics 71

3.7.1 OCR Models Performance Evaluation 73

3.7.2 MIM Recognition Pipeline Evaluation.. 74

3.8 Research Objectives and Research Questions 75

3.8.1 Research Objectives 75

3.8.2 Research Questions.. 75

4.0 Research Methodology 77

4.1 Overview of the experimental pipeline 77

4.2 Experimental Research Design 78

4.3 Dataset Description 78

4.4 Resolution of Character Image Patches 81

4.5 Challenges, Technical Factors and Ethical Considerations 82

4.6 Chapter Summary.. 83

5.0 Ensemble Learning for Medical Image Character Recognition based on Enhanced Lenet-5 84

5.1 Introduction 85

5.2 Related Works 88

5.3 Contributions.. 92

5.4 Proposed Model.. 93

5.4.1 Network Hyperparameter Optimisation 94

5.4.2 Designed Model 98

5.4.3 Modifications on Lenet-5 101

5.4.4 Visual description of the models' architecture 105

5.4.5 Majority Voting Algorithm 107

5.5 Experimental Set-Up.. 109

5.5.1 Data Preparation and Training Strategy 109

5.6 Results and Evaluation 110

5.6.1 Models' Results and Optimal Number (n) of Ensemble Members 111

5.6.2 Analysis of hyperparameters.. 117

5.6.3 Classification report for all character classes 118

5.6.4 Result Comparison with past works 122

5.7 Chapter Summary.. 125

6.0 Medical Image Character Recognition using Attention-based Siamese Networks for Visually Similar Characters with Low Resolution.. 126

6.1 Introduction 127

6.2 Related Works 129

6.3 Contributions.. 130

6.4 Proposed Method 131

6.4.1 Model architecture 131

6.4.2 Model + Attention Mechanism.. 134

6.5 Experimental Set-Up.. 135

6.5.1 Dataset Description 135

6.5.2 Training Strategy 135

6.6 Results and Analysis... 139

6.6.1 Performance Analysis on AUC - ROC Curve.................................... 142

6.6.2 Performance Analysis Using Feature Map Visualisation ... 144
6.6.3 Performance Analysis Using Confusion Matrix ... 145
6.6.4 Interpretation of confusion matrix and class-wise evaluation metrics 148
6.6.5 Classwise Evaluation Metrics on the SIAM-MICR + Attention ... 150
6.6.6 Quantitative Analysis with Related Works with Background Interference 155
6.7 Chapter Summary ... 156
7.0 Generating Synthetic Training Data to Improve Character Recognition Accuracy using a Conditional variational autoencoder ... 158
7.1 Introduction ... 158
7.2 Background ... 159
7.3 Conditional Variational Autoencoders (CVAE) and Limitations of VAE ... 165
7.4 Justification of CVAE instead of GAN for this research problem ... 168
7.5 Models’ Architecture ... 171
7.6 Results and Evaluation ... 173
7.6.1 Latent variable investigation ... 174
7.6.2 Generated Character Images – Samples ... 178
7.6.3 Qualitative Analysis - Augmenting Datasets with Synthetic Images . 179
7.6.4 Evaluation of individual classes' Improvements ... 181
7.6.5 Investigation of Impact classes with small sample size on latent space 182
7.7 Chapter Summary ... 184
8.0 Conclusion ... 185

8.1 Overview 185

8.2 Research Findings 185

8.3 Research Limitations 188

8.4 Research Contributions to Knowledge 189

8.5 Research Significance 190

8.6 Future Outlook 192

Appendix A – Data Collection Approval (UH Ethics Committee) 193

Appendix B – Data Request to Data Collection Location 1 194

Appendix C – Data Approval to Data Collection Location 2 195

Appendix D – Data Request and Approval to Data Collection Location 2 196

Appendix E - Risk Assessment for Data Collection 197

Appendix F- Results from Commercial OCRs 198

Bibliography 199

List of Figure

Figure 1. 1 : Xray Image (Wang (2002)) 3

Figure 1. 2 : Using Tesseract to check burned-in text (source: Author) 5

Figure 2. 1 : Some real-world applications of Computer Vision 15

Figure 2. 2 : General OCR Workflow 22

Figure 2. 3 : Plot of the data points with two Clusters 31

Figure 3. 1: Study's Evaluation Methods 73

Figure 4. 1: Experimental Pipeline 77

Figure 4. 2: Resolution of character "M" using the PIL Python library 81

Figure 5. 1 : X-ray image (The Cancer Imaging Archive (TCIA) Public Access) 85

Figure 5. 2: Flow chart of the BO hyperparameter optimisation process 96

Figure 5. 3: Proposed Enhanced MICR model 100

Figure 5. 4: Visual description of model's enhancement. 105

Figure 5. 5: Majority voting ensemble approach used 108

Figure 5. 6: Plot for the optimal number of ensemble members. 114

Figure 5. 7: Lenet-5 learning curves on MEDPIX 116

Figure 5. 8: Learning curves for the MICR model 131

Figure 5.9: Hyperparemeters' importance 132

Figure 5. 10: Recognition of low-resolution MIM sample 139

Figure 6. 1: SIAM-MICR 148

Figure 6. 2: SIAM-MICR + Attention. 150

Figure 6. 3: Pairing of images for training 151

Figure 6.4: Optimal layer of insertion of attention module 156

Figure 6. 5: ROC AUC on MEDPIX 157

Figure 6. 6: ROC AUC on PRIVATEDT 158

Figure 6. 7: SIAM-MICR's Second layer output for Character "Z." 159

Figure 6. 8: SIAM-MICR + Attention's Second layer output after channel attention module insertion -Character "Z." 159

Figure 6. 9: Confusion Matrix for MEDPIX (SIAM-MICR + Attention) 161

Figure 6. 10: Confusion Matrix for PRIVATEDT (SIAM-MICR + Attention)..………162

Figure 6.11: Comparison with related works in accuracy and parameter size…….172

Figure 7. 1: VAE generative modelling process …..…..………..…..…..…..…..….. 176

Figure 7. 2: CVAE generative modelling process…..…..…..…..…..…...…..…..…..181

Figure 7. 3: MICR Model…..…..…..…..…..…..…..…..…..…..…..…..…..…..…..…..186

Figure 7. 4: Proposed CVAE Architecture after Bayesian Optimisation with a Dropout of 0.3 ……..…..…..……..…..…..…..…..…...…...…..…..…..…..…..…..….. 188

Figure 7. 5: Optimal Latent Variables Investigation .. …..…..…..…..…..…..…..…..175

Figure 7. 6: Scatter plot of latent space ..…..…..…..…..…..…..……...…..…… 176

Figure 7. 7: Minimum reconstruction loss curve ..…..…..…..…..…..…..………178

Figure 7. 8: Generated character images ……..…..…..…...…..…..…..…..…... 178

Figure 7. 9: Accuracy vs Sample Size Augmentation for both datasets…………..180

Figure 7.10: Similarity comparison of images from small classes..………………....183

Figure F1: Google Document AI results on sample medical image ………………....198

Figure F2: Amazon AWS Textract results on sample medical image ……….……....198

List of Tables

Table 3. 1: Overview of previous works in the use of ML in recognising burned-in text in MIM …………………………………………………………………………………… 75

Table 5. 1: Layerswise Summary of the proposed CNN model……………………..….114

Table 5. 2: Modifications carried out on the Lenet-5 base model …………………..…116

Table 5. 3: Results for different ensemble members' configuration…………….……126

Table 5. 4: Classification report on all classes in MEDPIX ….………….………..…..133

Table 5. 5: Classification report on all classes in PRIVATEDT ………………………134

Table 5. 6: Comparison with Related Works on MEDPIX ….…………………………137

Table 6.1: Character frequency of MEDPIX for Siamese training ………………..……152

Table 6.2: Character frequency of PRIVATEDT for Siamese training ……………..152

Table 6. 3. Comparison of Model’s Accuracy with/without channel attention……..153

Table 6. 4: Classification report for all character classes for MEDPIX...................165

Table 6. 5. Improvement in VSC recognition compared to the multi-class MICR model (MEDPIX) 166

Table 6. 6: Classification report for all character classes for PRIVATEDT.............168

Table 6. 7. Improvement in VSC recognition compared to the multi-class MICR model (PRIVATEDT)...169

Table 6. 8: Comparison with related works on MEDPIX on 62-way 25-shot learning averaged over 30 runs...170

Table 7. 1: Latent variables and Reconstruction loss for the CVAE model............189

Table 7. 2: Accuracy of MICR model on augmented datasets—averaged on 20 runs. …………………………………………………………………………………….… 194

Table 7.3. Investigation of Individual Classes' Improvements ………………………197

1.0 Introduction

1.1 Overview

Major advancements in computational power, hardware, artificial intelligence, image processing, and pattern recognition technologies have been applied to various medical image modalities (MIM), such as X-rays and ultrasounds. A closer understanding of the default features of these MIMs shows they incorporate patients' demographics and information from medical examinations, and these exist as burned-in text data on these images; that is, the text is embedded in the pixel content of the image. The burned-in text is helpful for various information retrieval purposes, and therefore, it is essential to have efficient and accurate means to identify them for further processing needs. However, the MIM have a complicated nature due to its acquisition and acquisition device method, where the high quality of the imaging is given up, allowing for storage and transmission needs. Hence, these images have poor quality and low - resolution, making the burned-in text appear very small. These complexities of the MIM and the burned-in text affect the accuracy of traditional optical character recognition (OCR) systems when used for medical image character recognition.

The state-of-the-art OCR systems include open-source, for instance, Tesseract, OCRopus, and others for commercial use, such as ABBYY and Transym OCR, and they work in a similar mode of segmentation and recognition (Reul et al., 2018). A comprehensive analysis of Tesseract showed it is regarded as the best open source in critical comparison with other systems (Patel et al., 2018), particularly in line finding, extracting features and text classification methods (Smith, 2007). However, Tesseract has a very low accuracy level when applied in recognising burned-in textual data on low-resolution MIM due to the small font size of the textual data and the complex background interference. To provide a more comprehensive analysis, some commercial OCR systems were applied to recognise burned-in textual data on low-resolution MIM and results from popular systems, Google Document AI, Microsoft Azure Cognitive Services for Vision and Amazon AWS Textract, show an inability to recognise the burned-in text accurately or accurately. I subscribed to these paid services and attempted to extract the text from the low-resolution MIM, and the results were not accurate, which further reveals the challenges currently in this domain. The evidence of the result is provided in Appendix F.

This PhD study focuses on only the issue of low resolution and small font size problems in MIM and aims to provide innovative solutions to these challenges. Further attention is given to the problems of small sample sizes in medical image datasets and visually similar character images. A visual representation of the low resolution and small font size problem in MIM is shown in Figure 1.1 below:

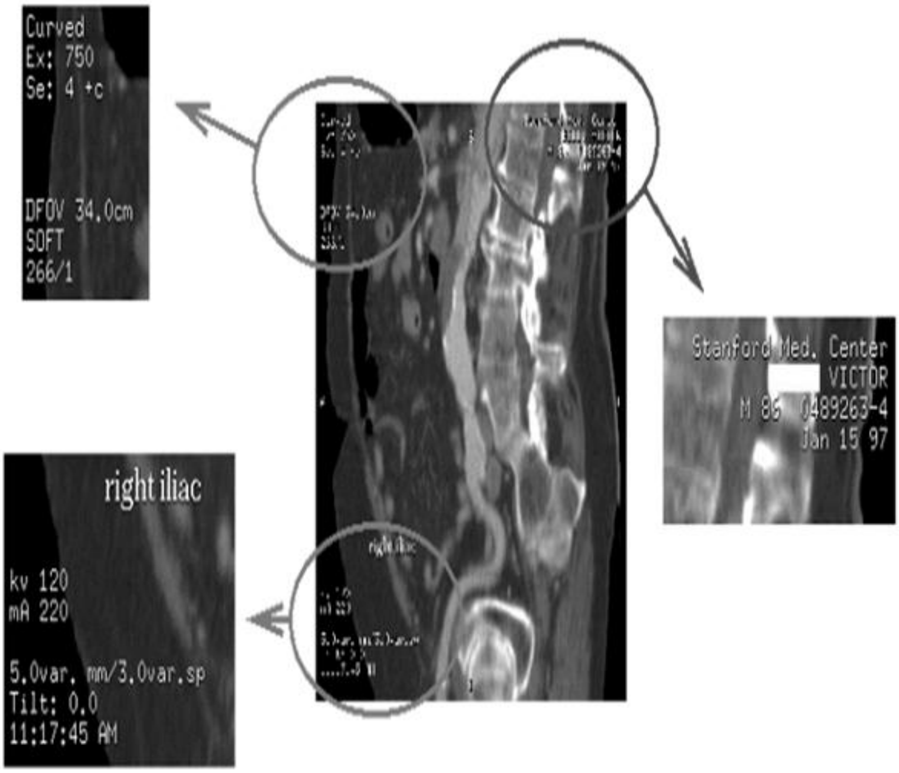


Figure 1. 1 : Xray Image (Wang (2002))

Extracting the burned-in text region for recognition from Figure 1.1 would more closely encounter the low-resolution problem resulting from the small font size and low resolution, as shown in the extracted burned-in text on the left and right panels.

From the varying conventional MIM shown in the Figures above, it would be noticed that the burned-in text characters have a small font size, resulting in the low resolution of that region if extracted for post-processing needs. When the part with the text is further extracted and enlarged to recognise the text data, the low resolution of the image region is extensively revealed. This fuzzy and small font size of these burned texts occurs in most MIM, and it is a major challenge for traditional OCR methods to recognise the characters accurately. The problem of recognising small font sizes of burned-in text due to the overall low resolution of MIM remains an unsolved and challenging problem, and past authors have pointed this out while proposing different solutions. Unlike printed text on paper, these burned-in texts are stored as data in the pixel structure of the acquired medical image (Reul et al., 2016). Recognition of these burned textual data using traditional methods has been difficult because their recognition is affected by the image's low resolution, and the character recognition accuracy is usually poor when the traditional methods are applied.

This PhD thesis proposes varied solutions based on advanced deep learning algorithms to tackle the problem in different aspects: (a) Leveraging the ensemble model advantage in recognition of these burned-in texts in MIM, (b) Tackling the issue of the visually similar characters using proposed semantic similarity learning methods, which classical classifiers find it difficult to achieve and (c) Proposing a data augmentation technique to improve the recognition accuracy, as MIM data samples are small in size due to privacy policies in the health domain, and acquisition cost. Chapters 4, 5, and 6 will extensively present these solutions with experimental validations.

1.2 Problem Statement

Medical imaging acquisition devices, during capturing, usually save modalities with very low resolution to reduce required storage infrastructure, usually at the cost of losing vital pixel information and clarity (Thambawita et al., 2021), which are relevant during information retrieval processes. The recognition of these burned-in texts in these modalities poses several challenges to modern OCR systems due to their small font size and low resolution. The burned-in text is rendered at a low resolution of an average value of less than 100 DPI and has a small font size. This creates a problem during textual recognition as the characters end up being connected together with an overall low quality and, hence, becomes a challenge for traditional OCR solutions.

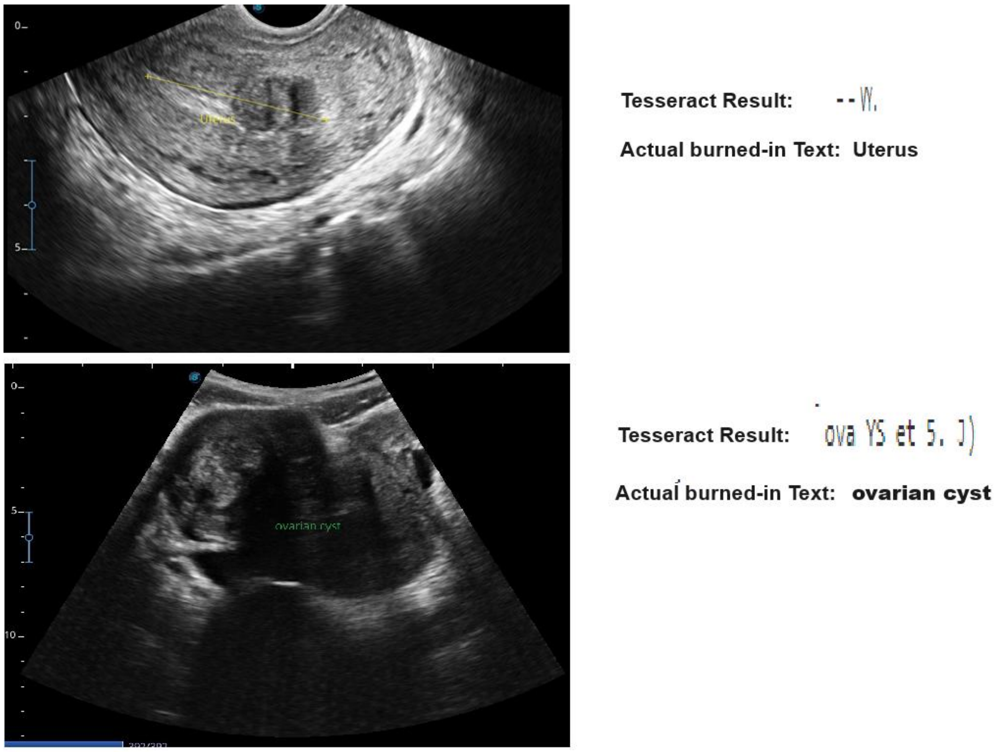

Uterus
Tesseract Result: - - W.
Actual burned-in Text: Uterus
ovarian cyst
Tesseract Result: ova YS et 5. J)
Actual burned-in Text: ovarian cyst

Figure 1. 2 : Using Tesseract to check burned-in text (source: Author)

Figure 1.2 above shows the recognition results from the latest version of one of the most accurate and reliable OCR engines, Tessaract (Badla, 2014), and its poor result in recognising burned-in textual data. The results seem extremely poor, as seen in the figure. From a detailed literature study of various traditional OCR solutions with consideration on their application to the recognition of burned-in texts in MIM, most of these systems are based on the character-segmentation approach, in which words are segmented into characters and recognition is done at a character level (Due-Trier et al., 1996). However, in MIM, where burned-in text data exist in small font sizes and low resolution, incorrect character segmentation leads to poor recognition rates by these existing OCRs. Other recent OCR techniques follow a holistic word recognition method because they do not identify at the character level but use global features like

T-junctions, B-loop, ascenders, and descenders information for identifying the entire word in cases where font size may be too small with a low resolution in the input image, achieving up to a recognition accuracy of 65% (Lavrenko et al., 2014). In the case of MIM, this holistic word recognition approach does not solve the problem of burned-in text recognition, as this method has the major drawback of being limited only to a small vocabulary and only useful with static small lexicon cases (Cote et al., 1998). MIM may contain private and diagnostic data unique to each patient involved; therefore, a unique solution is required to recognise these burned-in texts for post-processing actions. More recent deep-learning approaches have been unable to recognise these small texts in low-resolution MIM accurately when considering the performance of OCRs in other related domains (Xu et al., 2021 & Monteiro et al., 2017) as the image's complex background and noise have negatively affected their performance.

Critically considering the recent study by Xu et al. (2021), though the authors suggested their method effectively solved the low-resolution and background interference problem, there is no specific indication about the exact DPI they worked on that could be regarded as "low-resolution". There is also no comprehensive information on any background interference of the concerned image in their paper. Their precision was 80% for the synthetic character dataset and 70% for a medical image dataset used; however, considering a past work by Sangiacomo et al. (2022), who used OCR in a related domain, they suggested that an accuracy of at least 90% is sufficient for semantic analysis and data entry. This means that Xu et al. 's (2021) work still has room to be improved and be more effective. Additionally, as mentioned earlier, Xu et al. (2021) carried out an evaluation on a synthetic character dataset, Mjsynth, and a small medical image dataset, respectively. Their conclusion would be more convincing if their work were evaluated on a larger non-synthetic dataset.

This current research aims to clarify the magnitude of the low resolution in terms of DPI and propose, implement, and validate solutions accordingly. Furthermore, considering the current performance of OCR in other domains, further improvement can be made not only in character recognition but also in tackling the problem of visually similar characters and small sample size problems.

This is a relevant gap to which this PhD thesis aims to contribute by proposing techniques to improve recognition rates in medical image character recognition. A solution to this problem would increase performance in information retrieval systems needed for diagnostics and health management requirements.

1.3 Research Aim

The research aims to propose and apply deep learning techniques in recognising burned-in textual data on low-resolution MIM with background interference. This PhD thesis proposes different advanced deep learning-based algorithms to tackle the associated problems of accurately recognising these textual data on a character-by-character basis. The best approach, though difficult, is character-by-character. It removes the limitations and difficulty of using a vocabulary. It allows the practical application of the proposed solution in any location and device, as long as the burned-in text is constituted of characters. Furthermore, to support this choice, past works on character recognition in printed text and historical documents show higher accuracy and more generalisation than word-based recognition, as word-based recognition usually requires various post-corrections (Islam & Iacob, 2023; Drobac & Lindén, 2020).

To achieve this aim, this research will provide a critical analysis of the state-of-the-art techniques in OCR and medical imaging, as seen in the literature, for recognising burned-in textual data in MIM to reveal significant research gaps. Furthermore, this research will propose specialised deep learning techniques to solve these identified research gaps and validate these proposed techniques using both open-source and privately collected data. Ethical guidelines will be followed as set out by the University’s ethical committee.

The objectives and research questions are discussed extensively in section 3.8 after a comprehensive literature review is provided.

1.4 Contributions

The contributions of this research are well presented in Chapters 3, 5, 6 and 7. The chapters follow a common theme of optical character recognition, burned-in textual data recognition in MM, improved recognition accuracy for visually similar characters (VSC) in real-world applications and generative modelling for data augmentation for MICR. These chapters are adaptations of academic publications from this PhD thesis except Chapters 7 and 8, a version of the literature review on existing machine learning practices in burned-in text recognition from MIM, which is an integral part of the contents in Chapter 3 (Osagie et al., 2024a) has been published in a journal; Parts of Chapter 5 (Osagie et al., 2023) has been presented in a conference, and Chapter 6 (Osagie et al., 2024b) has also been accepted for a Springer Nature conference in Europe. This section summarises each chapter and focuses on the contributions outlined below.

1. Presented a critical review of the existing machine learning practices in burned-in text recognition regarding their challenges and open issues.

   In Chapter 3, this research reviewed the significance of burned-in textual data recognition in MIM and recent works regarding the ML approach, challenges, and open issues for further investigation. The chapter describes the significant problems in this research area, such as low resolution, background interference of textual data, small dataset size and VSC recognition. Finally, the chapter suggests applying more advanced deep-learning algorithms as possible solutions (Osagie et al., 2024a). The chapter provides an understanding of the gaps in the literature that exist in MICR-based ML and DL-based solutions.

2. Proposed, implemented and validated an enhanced CNN model and a majority voting algorithm for burned-in text data recognition in low-resolution medical imaging modalities having background interference.

   Chapter 5 presents two vital contributions to improving the performance of CNNs for MICR. With a focus on solving the issues identified in Chapter 3, an enhanced CNN model for MICR is proposed in Chapter 5. The Lenet-5 architecture inspires this proposed Model, and justification is provided for the choice of this base model. This Chapter further designs a majority voting ensemble of enhanced CNN models to optimise this new technique to recognise VSC (Osagie et al., 2023). Bayesian optimisation is used to optimise the hyperparameters. Multiple evaluations are done using open-source and original datasets collected by this research from a data collection study conducted around December 2022 – February 2023, with the University Ethics Committee’s approval (Appendix A).

3. Proposed, implemented and validated a channel attention-based Siamese Neural Network to recognise visually similar characters with small sample sizes in burned-in text on medical imaging modalities.

   In Chapter 6, the issue of VSC recognition and small dataset size for training due to the data privacy issue of collecting medical image datasets is addressed explicitly by proposing an attention-based Siamese Network to accurately recognise VSC by efficiently learning the semantic similarities between the extracted embeddings from sample images. The semantic similarities and attention-focused feature extraction layer enable the proposed model to discriminate between different character classes efficiently, with only small sample sizes (Osagie et al., 2024b). Bayesian optimisation is used to determine optimal network hyperparameters, similar to what is done in Chapter 5.

4. Proposed, implemented and validated a specially designed Conditional Variational Autoencoder (CVAE) as a practical data augmentation technique to improve the performance of deterministic models in burned-in text data recognition in medical imaging modalities.

   In Chapter 7, the issue of small dataset size is further addressed by proposing a specially designed CVAE that can be used to synthesise new images of characters with the same constraints of low-resolution with background interference. The experiments in Chapter 6 show that training deterministic models with different subsets of augmented training data generated by the CVAE model achieve better performance compared to models trained with the original data alone.

1.5 Publications related to this thesis.

The following publications are related to the chapters of this thesis:

- ***Chapter 3**:* Osagie, E., Ji, W. and Helian, N. (2024a) Burnt-in Text Recognition from Medical Imaging Modalities: Existing Machine Learning Practices, Journal of Advanced Computational Intelligence, and Intelligent Informatics, 28 (1), pp. 103–110. DOI:10.20965/jaciii.2024.p0103.
- ***Chapter 5**:* Osagie, E., Ji, W. and Helian, N. (2023) Ensemble Learning for Medical Image Character Recognition based on Enhanced Lenet-5, in: *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Eindhoven, Netherlands: IEEE, pp. 1–8. doi: 10.1109/CIBCB56990.2023.10264911.
- ***Chapter 6** (Accepted):* Osagie, E., Ji, W. and Helian, N. (2024b) Medical Image Character Recognition using Attention-based Siamese Networks for Visually Similar Characters with Low Resolution. In Lecture Notes in Networks and Systems (pp. 3–12). Springer Nature Switzerland. This was presented at the 2024 Third International Conference on Innovations in Computing Research in Athens, Greece

1.6 Ethical Consideration

This research involved using privately collected medical images approved by the university's ethics committee, protocol number SPECS/PGR/UH/05141. The rules and regulations set out were followed, as shown in appendix A—E, which show the data collection approval, request, and risk assessment.

1.7 Outline

This PhD thesis is separated into the following sections:

- Chapter 2 provides the comprehensive theoretical framework required to understand this thesis. It introduces computer and machine vision concepts: optical character recognition, image preprocessing, machine learning, and deep representation learning.
- Chapter 3 provides a comprehensive literature survey of the related and past works in recognising these burned-in textual data, including machine learning practices. The research objectives and questions are presented based on the gaps identified from the literature, with the aim of providing practical solutions to the identified problems.
- Chapter 4 provides the methodology utilised in this PhD research, including an overview of the experimental pipeline and data collection. It also discusses the technical challenges in this research and how they were addressed.
- Chapter 5 focuses on the problem of burned-in text data recognition at the character level in low-resolution MIM with background interference. It proposes a new CNN model and an ensemble classifier model to tackle it, optimised based on a hybrid Bayesian hyperparameters optimisation technique.
- Chapter 6 proposes a channel attention-based semantic similarity learning technique to solve the problem of recognising visually similar characters and a small sample size per class in low-resolution MIM.
- Chapter 7 proposes an innovative generative model to increase the available character datasets by including synthetic character images for training. This model aims to increase the performance of deterministic models trained with this augmented data.

- Chapter 8 concludes this thesis by summarising the contributions and presenting future work as regards medical image character recognition and deep learning techniques.

2.0 Background

Computer vision has rapidly developed into a vast area of application, from collecting raw visual data to more advanced techniques of pattern recognition, automated feature extraction, and representation learning of visual content (Wiley & Lucas, 2018). The modern concept of computer vision combines techniques, ideas and methods of digital image processing, pattern recognition, computer graphics and artificial intelligence to extract features and information from input images. The output is a comprehensive and usable understanding of the image in a particular domain. The human eye can see and interpret images easily due to its complex biological structure. It can adjust the amount of light it lets in, focus on objects near and far, and produce better interpretations of incomplete and/or visually similar objects. However, the practical efficiency of computer vision models is still far from that of the human eye, and this has led to research for a better understanding of 2D and 3D shapes and appearances of objects in imagery. As research in computer vision progresses, it has become widely applied in real-world applications. Some of these applications Include:

- **Optical Character Recognition (OCR):** Handwritten and printed text recognition (Figure 2.1a), automatic plate number recognition, archiving, and automatic data entry.
- **Medical Imaging:** Tumours, cancer detection in computed tomography (CT) images (Figure 2.1b), and smart operating facilities for surgical procedures to improve precision.
- **Retail and sales:** Customer tracking in cashierless stores and automated warehouse inventory management (Figure 2.1c).

- **Manufacturing and industrial systems:** Industrial anomaly detection for defective and non-defective machine parts (Figure 2.1d).

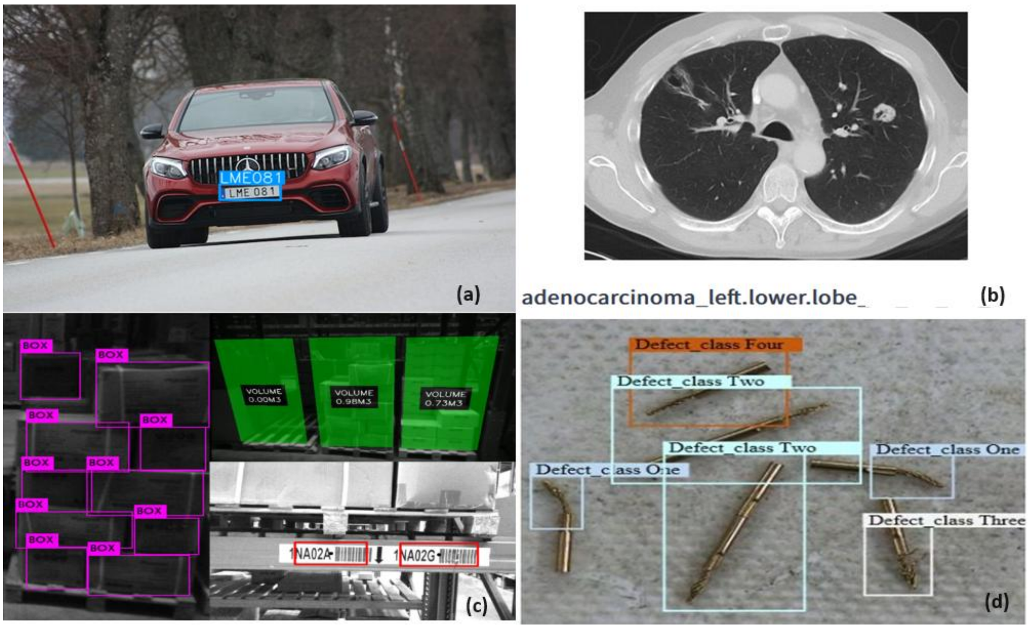


Figure 2. 1 : Some real-world applications of Computer Vision

(a) Visual character recognition for reading number plates (ANPR)[1], (b) Lung cancer classification model using CT images[2] (c) Inventory counting[3] (d) Real-time defect detection[4].

However, this study focuses specifically on applying OCR in medical imaging to provide accessibility and automate the recognition of burnt-in textual data embedded in the pixel content under the constraints of low-resolution and background interference.

This chapter provides adequate background on the computer and machine vision concepts important to OCR and MICR to enable easy understanding of this thesis. It

---

[1] https://viso.ai/computer-vision/optical-character-recognition-ocr/
[2] https://viso.ai/applications/computer-vision-in-healthcare/
[3] https://xosight.com/2020/12/30/the-future-of-inventory-management/
[4] https://viso.ai/applications/computer-vision-in-manufacturing/

comprises sub-sections: medical imaging application, preliminary definitions in image processing and medical imaging, OCR for text extraction, machine learning approaches, and deep representation learning.

2.1 Medical Imaging Applications: Clinical challenges and issues

With the vast growth in computing power, artificial intelligence has been rapidly applied to develop useful models for various medical imaging modalities. Some vital applications in medical imaging are:

- Prescribing targeted treatments: Computer vision techniques in medical images remove the dependence on quantitative methods and allow medical personnel to decide on more effective personalised treatments that precisely target the specific illness. This is seen in instances such as accurately identifying and segmenting cancerous tissues (Bai et al., 2023).
- Predictive medicine: Computer vision techniques are helpful in identifying existing conditions and can also provide useful insights into developing conditions, such as the risk of a cardiac attack and neurological decline. Past work has shown that combining MRI with clinical reports can help spot signs of lesions and shrinkage in Alzheimer's disease (Moscoso et al., 2019).
- Diagnostics medicine: Here, computer vision techniques in medical imaging can detect illnesses such as tumours faster and more accurately, as conventional mammogram scans have an error rate of 20% in detecting breast cancer. Compared to Google AI's AI-powered model, which has an error rate of 1% (Liu et al., 2018),
- Clinical data entry: This is useful for medical diagnostics and a robust electronic health management system. An efficient and highly accurate OCR system enables the scanning text on MIM to convert into readily accessible forms such

as synthetic speech and plain text, thereby improving diagnosis speed and mobility (Hom et al., 2022). OCR solutions provide an effective mechanism to convert medical imaging to allow the application of text analysis techniques, such as natural language processing, to yield highly actionable data insights (Hom et al., 2022).

However, despite the rapid integration of computer vision techniques in medical images, significant challenges and problems remain. In terms of clinics and medical centres, these are:

- Visual impairment, a decreased ability to see to a degree, has caused many problems that are not fixable by usual means, such as glasses. According to a report from the global blindness and visual impairment data in 2015, there were an estimated 253 million people with visual impairment worldwide, out of which 36 million were blind and a further 217 million had various cases of moderate to severe visual impairment (MSVI) include the problem with seeing in low contrast and brightness condition (Ackland et al., 2017). Vision is essential for seeing objects and dark adaptations, contrast sensitivity, balance, and colour perceptions. This visual impairment has led to many errors in clinical data entry, especially from medical images, due to human errors. An efficient and highly accurate OCR system can provide health workers who are visually impaired with the capacity to scan text on medical images and modalities and then convert it into easily accessible forms such as synthetic speech and plain text. It can help them recognise abnormality without the help of third-party verification, thereby enabling improvement of diagnosis speed and mobility. With the rapid entry into electronic health record (EHR) systems, which replaced the old paper-based storage and retrieval processes, which were designed to make patients' management more accurate, safer, and more accessible, the EHR system involves a large number of documentation, medical images, investigation reports and prescription, there is always the difficulty tracking files and keeping inventory (Dash et al., 2019). Clinical data entry for MIM is a significant and challenging task that health workers face daily, and the significance of an efficient medical image character recognition system would significantly improve the speed, accuracy, and management of medical data entry systems.

- Medical image colourisation, including the large variability in image characteristics and the need for robust and accurate colourisation methods (Pinto-Coelho, 2023). This has limited the application of computer vision techniques in medical imaging, and some medical centres may need to adjust to manual methods in analysing these images. This has resulted in a need for enhanced modelling techniques to tackle the colour complexity and variability of these images to allow automatic extraction and representation learning of relevant features.
- Health workers' workload - In this digital age, the high volume of medical imaging examinations has increased health workers' workload. This increased workload can lead to burn-out, excess fatigue, and an increased error rate (McDonald et al., 2015), especially in manually recognising burned-in text in MIM, which is usually in low resolutions and has background interference. Hence, it is vital to recognise this burned-in text for extraction and fusion with EHR to get these benefits.

The above sub-section has discussed the background of this research and its significance from a clinical perspective. It shows that issues of visual impairment, image colourisation and variability, and increased health workers' workload remain motivations for research into automated solutions for clinical data entry and imaging analysis.

2.2 Preliminary Definitions

The common terminologies in computer vision, image analysis, and OCR are explained here to enable understanding of their use in the content of this thesis:

- **Pixel**: A digital computer represents an image as a sequence of tiny dots called pixels (abbreviated px). Depending on the application and the digitiser used, a pixel's colour/grey shade is entered as an integer dimension between 0-256. In OCR, however, the pixel's reference is either 0 or 1, white or black, accordingly. The spatial location of a pixel P is often denoted by its offset from the top left corner of a binary image. P_{ij}, means the magnitude of the pixel on the i^{th} row and the j^{th} column (Wiley & Lucas, 2018). Pixel is the determinant of object sharpness and location in an image. Pixel optimisation is helpful for object detection, segmentation, and recognition (Wiley & Lucas, 2018).
- **Binary Image:** Binary images have only two values, 0 and 1, but they often use the range of values 0 and 255 to represent the colour of black and white. They are referred to as bi-level images.
- **RGB (TrueColor) Image:** An RGB image, sometimes referred to as a TrueColor image type, is basically an array of colour pixels, where each pixel is associated with three values of the image's colour components (red, blue, and green) at a specified spatial location.

- **DPI:** This signifies the resolution of a digital image, which is notably measured by the number of dots (pixels) per inch.
- **Medical imaging acquisition devices:** These are photo-electronic image acquisition devices used in diagnostic medicine. During the image acquisition process, the imaging part is automatically combined with patient text data, and both are merged into the same pixel structure, resulting in the text data appearing as burned-in text data and not printed on the image.
- **Medical Imaging Modalities:** This medical imaging technique utilises a specialised imaging acquisition device to visualise the human internal organs and reflect them as images. These Imaging modalities are often classified by the technique in which images are generated, such as ultrasound, radiation such as X-rays, and MRI.
- **Convolution:** This is a widely used technique in the imaging process. It is a mathematical operation on two arrays of numbers that outputs a third array of numbers with the same dimension. The mathematical formulation of 2D convolution is shown below.

$$y\,[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} h[m,n] \cdot x[i-m, j-n] \qquad (2.1)$$

In Equation 2. 1 , $\boldsymbol{x}$ is the input image matrix to be convolved with the kernel h to output a third and new matrix $y,$ which denotes the final image. Matrices are referenced here as [column, row]. m and n represents the shift with kernel h. i and j represents the coordinates of the output in the resulting matrix $y\,[i,j]$. Zero-padding is when border pixels with all values zeros are added to the edges of the input feature map. This ensures that the border pixels receive convolutions and contribute to the feature extraction process. The convolution function is generally used in image analysis and processing to apply a certain function whose discrete pre-computation is stored in an array (usually called a mask) onto the discrete greyshades function of the input image. Convolution is used often since the discrete function is pre-determined, and its utilisation involves multiplication and summation operations. Convolution involving one-dimensional data is called 1D convolution, and three-dimensional data is called 3D convolution.

- **Image Resolution:** Resolution indicates the number of pixels displayed per inch for an image. It describes the image's level of detail—higher resolution means more clarity, and lower resolution means less clarity. The main difference and relation between DPI and image resolution is straightforward: the higher the DPI, the more detail can be shown in an image, which means a higher resolution. The DPI is used to measure the resolution of the image both on screen and in print, whereas the image’s resolution is used to describe the overall quality of the image (Mohamed & Yousif, 2010). Finally, DPI and resolution are both significant elements in describing an image. Rakhshan (2014) supports this relation in medical imaging, as the DPI describes the image resolution measurement, and the image resolution explains the overall clarity of the image, with instances of X-rays and other digital radiography images.

2.3 Optical Character Recognition (OCR)

OCR is a process that extracts textual data from an input image into a machine-readable and accessible format. Much information is stored in printed media, including images such as newspapers, paper prints, legal documents, scanned documents, and acquired images such as camera, medical, and industrial images. Due to advancements in modern technology, there is a need for an accessible machine format for the textual data contained in these images. Manually accessing this large volume of images to extract the textual data is challenging as it is highly time and resource-consuming (Adnan & Akbar, 2019). The data entry errors will also be high through the manual extraction due to the possibility of human errors. These led to the advancement in computer vision techniques to solve this problem. OCR solves the problem by automating textual data extraction, improving operational efficiency, and reducing human errors in data entry. The OCR workflow is shown in Figure 2.2, and the major steps for the OCR are explained briefly below.

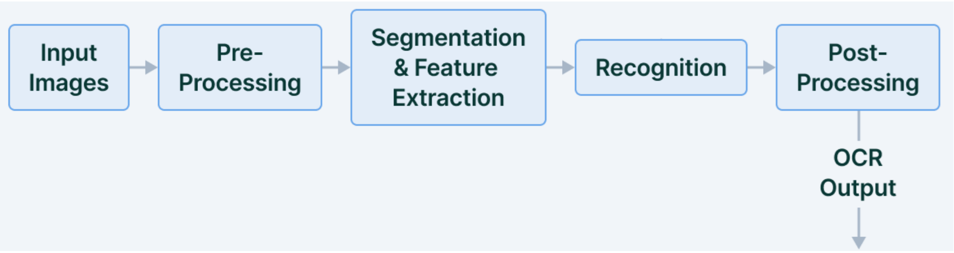


Figure 2. 2 : General OCR Workflow[5]

- **Pre-processing:** The OCR performs noise removal and other cleaning processes to increase the overall quality of the input image. Some popular pre-processing techniques in OCR include binarisation, noise removal, thinning, skew correction, and skeletonisation. Binarisation converts a coloured image with three channels into a bi-level image with only black and white pixels using the popular method of thresholding conditions, as shown in the algorithm below.

---

[5] https://www.v7labs.com/blog/ocr-guide

Algorithm 1: Thresholding

```
Input:  Input.png
Output: Output.png
def threshold (Input):
    thresholdValue = SomeValue
    If  (CurrentPixelValue > thresholdValue)
        CurrentPixelValue = 255
    else
        CurrentPixelValue = 0
```

The major challenge is finding the optimal value of the threshold, and various techniques, such as the local maxima and minima, OTSU binarisation and region-based adaptive thresholding, have been proposed to determine the value. Skew correction is to correct the image projection, which may be skewed from scanning or acquisition. Noise can be introduced into images if scanned from photographic materials, which may have fine grains or damage if acquired directly in a digital format and/or if the image is transmitted electronically. This noise will reduce the OCR's accuracy; hence, removing or reducing as much as possible is always a good pre-processing step. In this step, noise removal aims to smoothen the image to improve quality by removing non-uniform pixels using average, median, and adaptive filtering techniques. The most practical filtering is adaptive filtering, improving degraded images' quality.

Skeletonisation helps to uniformise the stroke width of textual data due to the different writing styles and font sizes. It is closely related to thinning. Most programming frameworks, such as MATLAB and OpenCV, provide utilities to carry out these image pre-processing steps automatically.

- **Segmentation, Feature extraction and Recognition:**
  The segmentation, extracting features, and recognition stages are used to extract the most relevant information from the input and then to recognise the characters in the textual data (Singh & Budhiraja, 2011). The feature selection is highly contributory to the accuracy of character recognition. This is pattern recognition, where the modelling is more complex. Pattern matching isolates a character image patch via segmentation and compares it with a similar prototype image patch. This works well when the entire textual data has a font and scale similar to the prototype stored. It may also involve comparing extracted features, such as open and closed loops in characters, edges, line thickness, intersection, and edges, from the input image and the stored prototype using nearest neighbour algorithms.
- **Post-processing:** Post-processing involves approaches such as error detection and error correction and conversion of the extracted text into another required format (an example is annotated pdf). The essence of this stage is to provide human assistance to correct errors quickly. This can be done using the lexical approach, candidate generation and candidate ranking to find the most appropriate candidate words to replace the erroneous words (Nguyen et al., (2022).

To further understand how the OCR workflow is approached, the next subsections will present a comprehensive summary of the OCR approaches, which modern-day OCR engines are based on: segmentation-based and segmentation-free OCR.

2.3.1 Segmentation-based OCR Approach

The segmentation-based OCR approach identifies the individual characters that will be used for recognition and strongly relies on the accuracy of the individual character segmentation process. This has remained the state-of-the-art approach for most OCR engines, which led to vast research for handwritten and printed character segmentation methods. Segmentation-based OCR can either be template matching or an over-segmentation technique (Qaroush et al., 2022). Template matching involves extracting connected characters and matching them with possible templates based on the nearest neighbour. These templates are representative samples of each character, and character-character matching is done based on pixel-by-pixel matching. The match is found when the number of matched pixels exceeds a predetermined value. The basic similarity measure used in practical applications for template matching in segmentation-based OCR is a cross-correlation function, presented in (2.2).

$$X_{(x,y)} = \frac{\Sigma_{x,y}\left[I_{(x,y)}-\bar{\mathrm{I}}_{u,v}\right]\cdot[(\mathrm{T}(\mathrm{x}-\mathrm{u},\mathrm{y}-\mathrm{v})-\bar{\mathrm{T}}]}{\sqrt{\Sigma_{x,y}(I_{(x,y)}-\bar{\mathrm{I}}_{u,v})^2\cdot\Sigma_{x,y}(\mathrm{T}(\mathrm{x}-\mathrm{u},\mathrm{y}-\mathrm{v})-\bar{\mathrm{T}})^2}} \tag{2.2}$$

Where *I* is the input image, and *T* is the template, $\bar{T}$ is the mean of the template. $\bar{I}_{u,v}$ is the mean of $I_{(x,y)}$ and is the region under the template. Briechle & Hanebeck (2001) adequately explained the cross-correlation function equation, and Hashemi et al. (2016) further supported this equation for template matching using a cross-correlation equation with a robust decision-making algorithm.

The over-segmentation approach applies to situations where correct character segmentation is not possible due to overlapping characters. Character candidates are found using imperfect segmentation, and matching is based on standard pattern recognition techniques (Support Vector Machine, Bayes classifiers, or neural networks). This method attempts to solve the over-segmentation problem where the OCR engine may not differentiate two close characters using projection-based, feature-based, or skeleton analysis-based methods.

2.3.2 Segmentation-free OCR

The segmentation-free OCR uses a more holistic approach, integrating feature extraction and contextual information in the text recognition stage. It can recognise a single character by considering the state and its surrounding context. It depended on extracting the entire word, part-of-word, or sentence line; features are extracted at a word or sentence level, and recognition is done at the word or sentence level. A trained classifier is then designed to carry out the recognition, thus avoiding the need for character segmentation. A popular, well-known segmentation model is the Hidden Markov Model (HMM) (Agazzi & Kuo,1993), which is very similar to recurrent neural networks (Baucum et al., 2020). HMM is a very powerful statistical modelling tool for OCR due to its high usage in temporal pattern recognition. More extensive information regarding the HMM segmentation-free approach and basic algorithms is covered in a background study by Rabiner(1989). This PhD thesis focuses on deep learning approaches and, hence, will not elaborate on the mathematical foundations of the HMM.

2.4 Machine Learning

Machine Learning (ML) is an aspect of artificial intelligence that develops algorithms that can learn relevant representations from available data. It can be referred to as the automatic detection of patterns in existing data. ML enables machines to carry out complex functions without being explicitly programmed. Recently, it has become popular due to the large amount of data continuously generated around us. Such ML applications include search engines, loan approval applications, fraud detection systems, medical abnormality detection devices, and many others. ML algorithms can be classified into supervised and unsupervised learning. These are explained briefly in the following subsections.

2.4.1 Supervised Learning

Supervised learning is an ML approach that develops an algorithm to learn the input-output relationship information of data based on a given set of paired input-output data samples. To learn this input-output relationship of input, $\boldsymbol{X}$, to output $\boldsymbol{Y}$, that is $\int : \boldsymbol{X} \rightarrow \boldsymbol{Y}$, the model is trained with a labelled dataset that has input-output pairs, $\boldsymbol{D}$.

$$\boldsymbol{D} = \{(x_1, y_1), \dots, (x_n, y_n)\} \qquad (2.3)$$

Where each x_i is the feature vector of the input, and y_i is the corresponding output (Cunningham et al., 2008).

The overall goal is typically to develop a final deterministic model that takes input $\boldsymbol{x} \in \boldsymbol{X}$ and predicts its matching output $\boldsymbol{y} \in \boldsymbol{Y}$. In computer vision and, most specifically, OCR, the inputs are the image’s pixels, and the outputs are the characters the image represents. The aim is to enable the model to generalise over samples outside the training samples, assuming all the training samples are independent with identical distributions.

A loss function must be defined to find the best approximate function that can define the input-output relationship. The loss function quantifies the difference between the predicted value $\hat{y}_i = f(x_i)$ and the actual target value y_i, where $f(x_i)$ is the model function (or hypothesis) that make predictions. For instance, in a regression task for stock prices based on historical data, the loss function evaluates the model's prediction based on a sample from the training dataset by quantifying the error margin between the model's price prediction and the actual price on the dataset. In practice, the cost function is often used interchangeably with the loss function, but they can have slightly different connotations depending on the context and the specific domain. However, these terms differ because the loss function concerns a single training iteration. In contrast, the cost function, an objective function, is the average loss function of all the training iterations done on the entire dataset. Loss function can be categorised based on the task being done, which are broadly regression and classification tasks. For the regression task, which involves the prediction of continuous output values, the Mean Square Error (MSE) denotes the loss function. It is given by the mathematical equation in (2.4).

$$MSE = \frac{1}{n} . \sum_{i=1}^{n} (y_i \; - \hat{y}_i)^2 \qquad (2.4)$$

Where n is the number of samples in the dataset, $\hat{y}_i$ is the model's prediction for the *i-th* sample and y_i is the actual target value for the *i-th* sample. MSE is a standard loss function and optimises the minimising of the squared differences between the predicted and target values of the training samples. Another loss function for the regression task is the Mean Absolute Error (MAE), which calculates the average absolute distance between the predicted and the target values. It does not square the difference and can be defined mathematically in (2.5).

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i \; - \; \hat{y}_i| \tag{2.5}$$

n is the total number of instances , $\hat{y}_i$ is the model's prediction for the *i-th* sample and y_i is the actual target value for the *i-th* sample. MSE is more sensitive to large errors and is useful when you want to heavily penalize large errors. On the other hand, MAE is less sensitive to outliers and provides a direct measure of the errors' average magnitude. MAE has the same units as the target variable, making it easier to interpret. MSE are positive values. Since absolute values are always non-negative, MAE is also always non-negative.

The binary cross-entropy (BCE) is used in binary classification tasks for performance measurement, where the prediction is an output with a probability value between 0 and 1. A variant of this loss for multi-class classification is the categorical cross-entropy. Binary cross entropy loss is calculated from the negative value of the summation of the logarithm value of the probabilities of the predictions made by the model against the total number of samples in the dataset. It is used to train artificial neural networks to predict the likelihood of a data sample belonging to a class and leverage the sigmoid activation function internally. The sigmoid function ensures the output of the input-output relationship to a value between 0 and 1. The BCE loss only takes one channel with a number ranging between 0 and 1 and is used only when there are two classes. BCE is mathematically given in (2.6).

$$BCE = -\frac{1}{n} \sum_{i=1}^{n} (y_i \; \cdot \log \hat{y}_i \; + \; (1 - y_i) \; \cdot \; \log(1 - \; \hat{y}_i)) \tag{2.6}$$

Where y_i represents the actual class label for the *i-th* sample and $\hat{y}_i$ is the predicted probability value for the *i-th* sample of the positive class, which ranges from 0 to 1.

Most ML models learn functions that are controlled by a set of parameters θ . Therefore, there is a need to determine the best values for the parameters θ^* that minimises the loss over all the training samples in the dataset. This research involves multi-class and binary classification, which will be discussed more specifically in the coming chapters.

2.4.2 Unsupervised Learning

In unsupervised learning, there is no clear target output that the model is being trained to predict; that is, the dataset, ***D***, is unlabelled. See (2.7).

$$\boldsymbol{D} = \{(x_1), \dots, (x_n)\} \tag{2.7}$$

The overall goal of unsupervised learning is to build representation of relevant features from the dataset, such as pixel structure, image patterns, or characteristics (Ghahramani, 2004). Such algorithms include clustering and data compression. The unsupervised learning algorithms relevant to this research are known as generative models. The generative models aim to learn the true data distribution itself to generate new data points with some variations. This can be useful for generating samples similar to those in the training set by sampling from the estimated distribution learned by the model. As a typical instance of an unsupervised learning algorithm, K-means clustering is a commonly used clustering algorithm. Kmeans clustering is an iterative unsupervised learning technique that aims to divide a dataset into *K* pre-determined defined, separable clusters without any overlap, where each data point can only belong to a single cluster. The distance between data points in a cluster is minimised, and this keeps the clusters away from each other. Data points are assigned such that the sum of the square distance between them and the cluster's centroid is maintained at a minimum. This ensures more consistency within the cluster and highly similar data points within a particular cluster. A visual representation of the Kmeans clustering is shown in Figure 2.3.

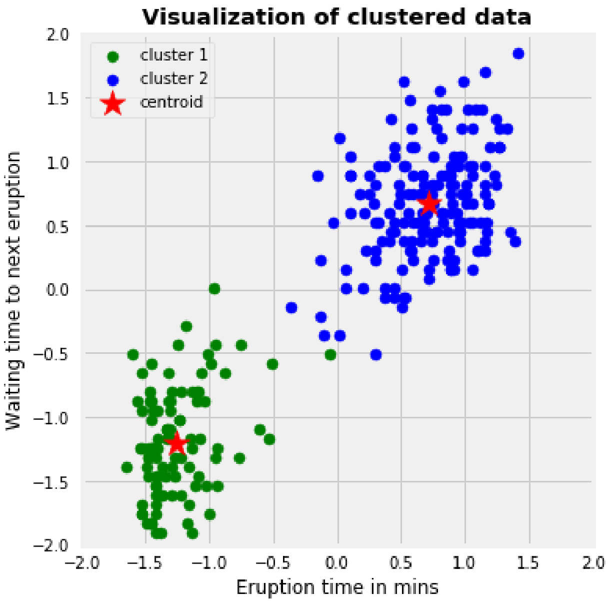


Figure 2. 3 : Plot of the data points with two Clusters

K-means is usually applied for tasks such as compressing images, document clustering, and image segmentation. The K-means procedures can be summarised as outlined below.

1. Determine the specific number of required clusters
2. Randomly shuffle the dataset and select K points to represent the centroids without replacement.
3. Initiate the selected centroids.

4. Carry out iterations until the values of the centroids do not change; that is, the convergence is completed.
5. Compute the summation of the squared distance between the centroids and data points. Assign each data point to the nearest cluster.
6. Compute the centroids by averaging each cluster's data points.

Since Kmeans clustering uses distance-based measurements to determine the similarity between data points and the centroids, the Euclidean distance between two points, p and q, in a multi-dimensional space is given as (2.8).

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (2.8)$$

The same subscript, i, enables for component-wise comparison; that is, each component of p is matched with the corresponding component of q. For instance, p_1 matched with q_1 . If each cluster's centroid is denoted, then by $\boldsymbol{c_i}$**,** then, each data point, x, is assigned to a cluster based on (2.9).

$$\arg\min_{c_i \in C} dist(c_i, x)^2 \qquad (2.9)$$

$dist()$ is the Euclidean distance in (2.9). Due to Kmeans' iterative characteristics and the random initialisation of centroids, varying initialisations will lead to varied clusters. The algorithm may be stuck in a local optimum and not converge to the global optimum. It is suggested that iterations are always performed based on different values of centroids and that the values of the iterations with the lowest sum of squared distance be chosen.

2.4.3 Machine Learning Optimization and Hyperparameters

Machine Learning Optimization is the process of iteratively improving the accuracy of a machine learning model, lowering the degree of error by approximating the underlying function or relationship between input and output data. A major goal of training a machine learning algorithm is to minimise the degree of error between the predicted output and the true output. Gradient descent is an iterative first-order optimisation algorithm used commonly in ML. Gradient descent numerically finds the minima of multivariate functions. It minimises a function $J(x)$ by altering $x.$ $J(x)$ is referred to as the objective function to be minimised. There are two requirements before gradient descent can be applied to optimise an objective function: differentiable and convex. If the objective function is differentiable, it has a derivative for each point in the domain. For a function to be convex, its second derivative must be bigger than zero (2.10).

$$\frac{d^2J(x)}{dx^2} > 0 \qquad (2.10)$$

The gradient descent is the first derivative at a selected point for a univariate function. In the case of a multivariate function, the gradient descent is a vector of derivatives in each main direction along variable axes. A gradient for an n-dimensional function $J(x)$ at a given point *p* is defined mathematically as follows (2.11):

$$\nabla \int(p) = \frac{\partial J}{\partial x_1}(p) \ldots \frac{\partial J}{\partial x_n}(p) \qquad (2.11)$$

Gradient descent iteratively calculates the next point using gradient values at the current position, scaling it by a learning rate and subtracting the value to minimise the function. This process is shown mathematically in (2.12).

$$\theta = \theta' - \alpha \nabla \mathrm{J}(\theta') \qquad (2.12)$$

θ is the vector of the parameter being optimised from its initial value θ', α is the learning rate, which controls the step size and influences the model's performance. $\mathrm{J}(\theta)$ is the cost function, measuring how well the model fits the data. $\nabla\mathrm{J}(\theta)$ is the gradient of the cost function with respect to θ. In summary, gradient descent methods are as follows:

1. Initialise a starting point.
2. Calculate the gradient value at this point.
3. Make a time-scaled move in the opposite direction to the gradient.
4. Redo points 2 and 3 until the maximum number of iterations is reached or the step size is smaller than the tolerance.

Many variations of the basic gradient descent algorithm update rule exist, such as Adam, Stochastic gradient and AdamGrad. Most ML algorithms are controlled by parameters the model cannot learn or set for itself, such as batch size. These parameters are known as hyperparameters. The usual way to choose values for hyperparameters is to randomly test these values and select the ones that provide a promising result during model evaluation.

This PhD thesis study uses Bayesian optimisation (BO) to select the best hyperparameter configuration due to its ability to handle expensive-to-evaluate objective functions, flexibility in computing varied objective functions, and search spaces and find the global optimum with a small number of evaluations (Yang & Shami, 2020). Past works with empirical analysis results show that the BO algorithm outperforms other global optimisation algorithms (Wu et al., 2019) for hyperparameter configuration. Here, the objective function being optimised is the accuracy metric.

BO builds a probability model of the objective function and uses it to select preferred hyperparameters to evaluate the true objective function (Wu et al., 2019). For BO to optimise a function $\int(\boldsymbol{x})$, the function has to have an unknown expression and the cost of finding $\int(\boldsymbol{x})$ for a value of $\boldsymbol{x}$ must be high (Rodemann, & Augustin, 2024). These are usually in the case of neural networks and Deep nets, where there are large possible configurations of hyperparameters (Feurer & Hutter, 2019). If $\int(\boldsymbol{x})$ meets both conditions, BO can be applied to find $\boldsymbol{x}^*$, the global value of $\boldsymbol{x}$ while minimising the iterations on $\int(\boldsymbol{x})$. The objective becomes (2.13).

$$\underset{x}{\text{Max}} \int(x) \tag{2.13}$$

Bayesian optimisation works by integrating samples drawn from the objective function into the model's prediction for $\boldsymbol{x}$, a prioir is defined for $\boldsymbol{x}^*$and samples drawn from the objective function are updated to define a posterior to improve the accuracy of the probable $\boldsymbol{x}^*$. This entire process is done with the aid of the surrogate and the acquisition functions (Diessner et al., 2022). The surrogate function $\boldsymbol{g}(.)$ address the problem of no analytical expression condition for $\int(\boldsymbol{x})$ and the acquisition function $\boldsymbol{u}(.)$ provides guidance on the next value from the objective function to be sampled by balancing exploration and exploitation.

The iterative process to optimise the objective function using BO and these identified functions is as follows:

1. Iterate for t = 1, 2, 3, …. T steps for sampled points, (x, y)Which are added to the set $D_{1:t-1}$.
2. Select the next sampling point for x_t, by doing argmax of the acquisition function.

$$x_t = \ argmax_x u(x|D_{1:t-1}) \tag{2.14}$$

3. Sample the objective function at this point $\boldsymbol{yt} = \int(\boldsymbol{xt})$ and add this sample to the test,

$$D_{1:t} = \{D_{1:t-1}(x_t, y_t)\} \tag{2.15}$$

4. Update the surrogate function $\boldsymbol{g}(.)$ with the newly sampled points $(\boldsymbol{x_t}, \boldsymbol{y_t})$.

It is also termed sequential model-based optimisation because the hyperparameters are added to update the surrogate model sequentially (one by one). It uses a less expensive model-based approximation technique with a surrogate model. This optimisation is used during this research to optimise the hyperparameters of the proposed models. Figure 2.4 shows the BO process.

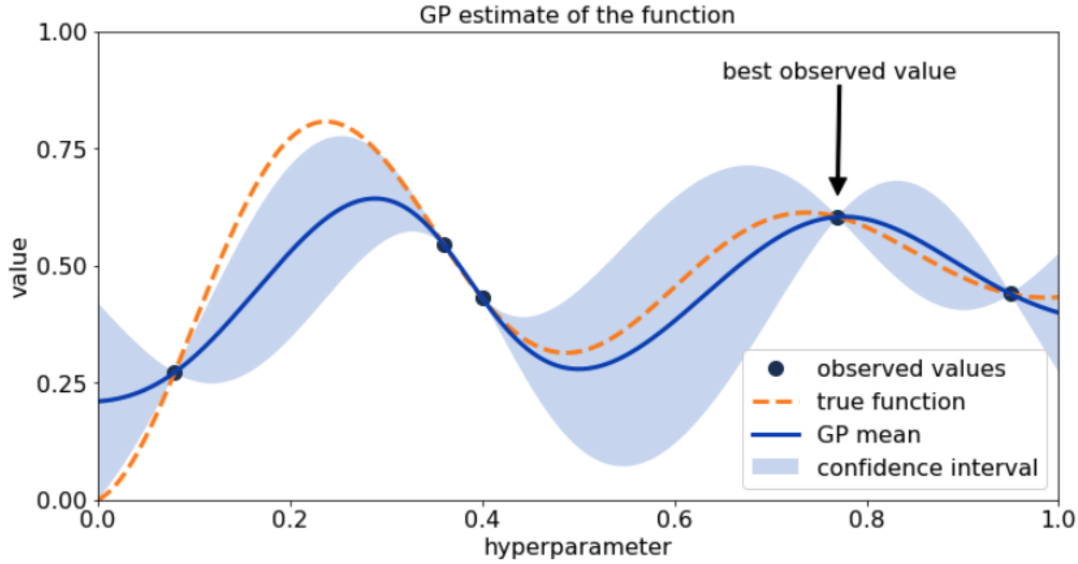


Figure 2. 4: Bayesian Optimisation with a Gaussian process fitted to the observed data from previous steps[6]

As shown in Figure 2.4, the Gaussian Process (GP) is fitted to the observed data. The GP is a flexible class of non-parametric statistical models over function spaces with domains that can be continuous (Cheng et al., 2019). GP can be the maximum entropy probability distribution in the context of statistical inference, given mean m and covariance v constraints, the probability of a distribution with m and v, the GP "most

---

[6] https://medium.com/aimonks/

spread out probability", has the greatest uncertainty due to higher variance. In other words, they are a good choice when modelling a high degree of uncertainty, such as deep learning models. GP is derived from the Gaussian probability distribution. Gaussian distribution estimates the probability of an input vector based on the hyperparameters, mean, and variance; GP generalises this concept, enabling a more flexible prediction and modelling (Hamoudi et al., 2023). Hyperparameters are significant in GP, and they determine high-level characteristics of the prior through the mean and covariance (Noack et al., 2023). Furthermore, this PhD thesis chooses to use GP success due to its flexibility in implementation and robustness and allows modelling of complex functions which may have large variables, the ability to adapt to noisy data, and its nonparametric nature, and there are no assumptions about the underlying distribution of the data. Additionally, GP can capture salient patterns in the input, through expressing complex covariance structures (Stoddard et al., 2019). Training a GP model via Bayesian inference involves computing the marginal likelihood of a given set of hyperparameters that can be used to predict new data points. Instead of defining a fixed set of parameters, the GP can be set by a mean and a kernel function.

$$f(x) \sim GP\ (m(x), k(x, x')) \tag{2.16}$$

Where $f(x)$ represents the function from the GP, $m(x)$ is the mean function which determines the expected value of the function at a given input x. $k(x, x')$ is the kernel function, representing how the function values at varied inputs of x and x' correlate. The mean shows the average behaviour of the function, while the kernel shows how the function's values vary with respect to each other across varying inputs. In this PhD thesis, the Matern Kernel is used with GP due to advantage of a trade-off between smoothness and computational efficiency, and its flexibility. The Matern Kernel equation is its simplest form, with hyperparameter ν and length scale σ , is denoted $K_\nu(d)$, where d is the Euclidean distance between two points. This could be further expressed as $K_\nu(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu}\frac{d}{\sigma})^\nu \cdot K_\nu (\sqrt{2\nu}\frac{d}{\sigma})$. σ and ν are positive parameters, Γ represents the gamma function, K_ν is the modified Bessel function of the second kind. σ controls the rate of change between the points and normalises d.

2.5 Deep Learning

Deep learning (DL) is a sub-division of ML concerned with algorithms with interconnected nodes arranged in layers and inspired in structure and function by the brain. This network of interconnected nodes is known as the artificial neural network (ANN). In the ANN, each layer takes the outputs of the previous one as its input and performs some non-linear transformations on this input, and the result is forwarded to the next layer within the network. When numerous layers are stacked (exceeding 1 hidden layer between the input and output layer), it is called a deep network. The core motivation for the rapid advancement in DL is the availability of modern computational power and the vast availability of data, which is available for training deep learning algorithms to learn more generalised representations. DL can be trained with more data, and their performance increases, unlike many traditional ML algorithms that may reach a level of no growth or decline. This sub-section briefly summarises the main DL techniques relevant to this research. For a more in-depth guide on deep learning, please see Goodfellow et al. (2016).

2.5.1 Multilayer Perceptron (MLP)

The fundamental architecture in neural networks is the multilayer perceptron (MLP). It is also referred to as a feedforward neural network due to its weight updating and information flow, which is only forward. There are no backward connections in the basic MLP. The MLP was developed to tackle the limitations of the linear computation of the basic perceptron, as the MLP can map a non-linear relationship between inputs and outputs. The MLP has an input, one or more hidden, and output layers stacked together. A visual representation is presented in Figure 2.5. The arrows in the Figure denotes a connection between the layers, and the learnable parameter of the model is represented with a weight $\boldsymbol{w_{1 \ldots n}}$.

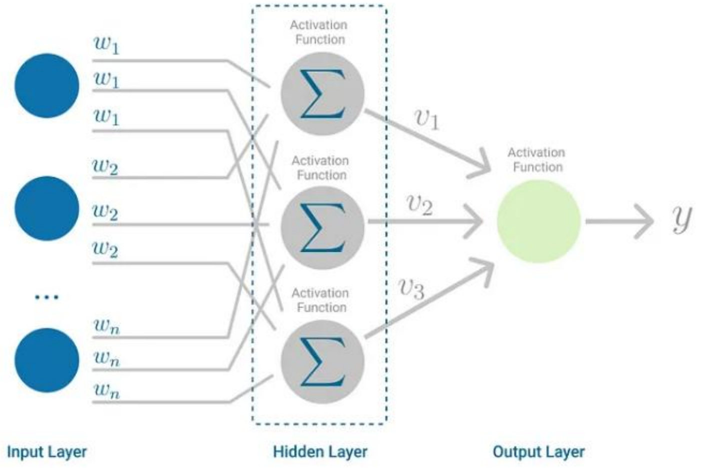


Figure 2.5: Multilayer Perceptron with a single hidden layer[7]

However, since it is a feedforward algorithm, the computed weighted sum in each neuron is propagated to the next layer, and there is no further action afterwards. This leads to inadequacies in adjusting weights and minimising the cost function. The weighted sum WS is shown in (2.17).

---

[7] Towardsdatascience.com

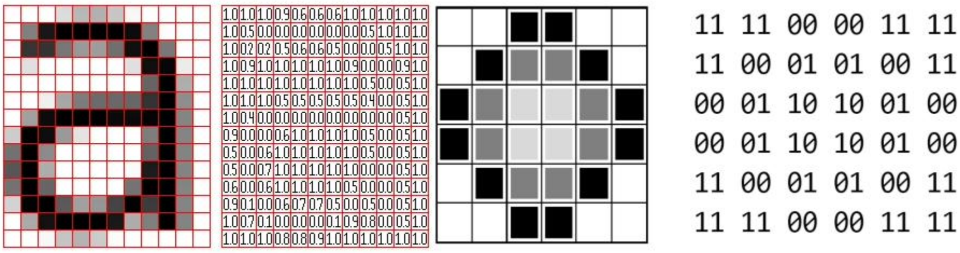
$$WS = \sum_{i=1}^{n} (W_i * X_i) + b \tag{2.17}$$

Where n is the total number of inputs, W_i is the weight for the *i-th* input, X_i is *the i*-th input value and b is the bias. Weights determine the strength of connections between neurons, while biases are added to the neuron’s output to introduce a threshold for activation.

The learning technique, backpropagation, was developed to solve this problem by enabling the MLP to adjust the weights during iterations to minimise the cost function. The activation function (such as ReLU, Sigmoid, and Softmax introduced non-linearity into the MLP, enabling it to map complex input-output relationships. In summary, each MLP layer builds on the feature the proceeding layer learned to learn more complex representations. The lower layers learn local features, and the deeper layers learn more high-level (abstract) features. For instance, in images, the lower layers may learn lines and curves, whereas the deeper layers may learn the shapes of objects.

2.5.2 Convolutional Neural Networks

Convolutional neural networks (CNN) are a variant of the MLP that specialises in processing data such as images or videos with a grid-like topology. Images are representations of data consisting of varied pixel values, which determine each pixel's colour, contrast, and brightness. This is shown in Figure 2.6 for the characters “a” and “o”, where the pixel values are presented on the grid beside the images.



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

11	11	00	00	11	11
11	00	01	01	00	11
00	01	10	10	01	00
00	01	10	10	01	00
11	00	01	01	00	11
11	11	00	00	11	11

Figure 2. 6: Representation of Character images as a pixel grid – “a” and “o” [8]

A CNN architecture comprises three layers: the convolutional layer, a pooling layer, and a dense layer (also called a fully connected layer). In a fully connected layer, every input node is connected to a corresponding output node, whereas in the convolutional layer, no such connection exists between nodes.

A basic CNN architecture with its components is displayed in Figure 2.7 below.

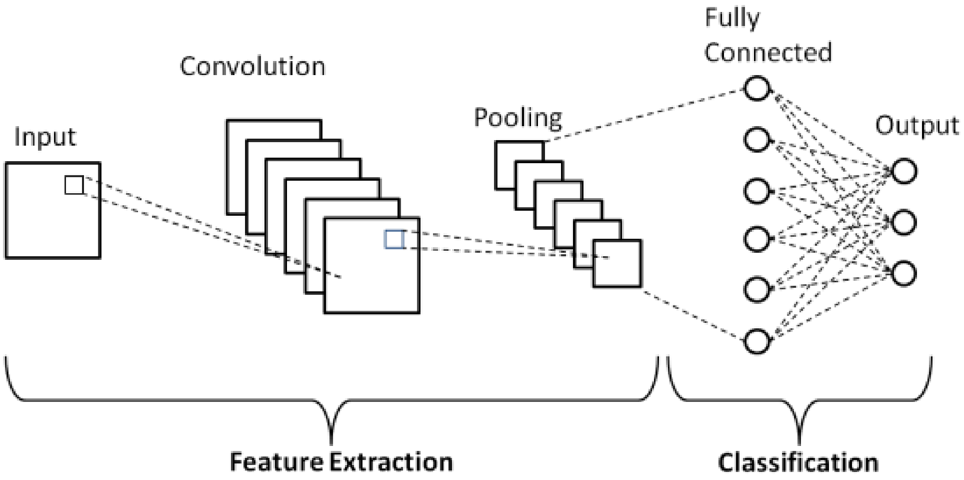


Figure 2. 7: A basic CNN architecture (Phung & Rhee, 2019).

The convolutional layer is the main operational block of the architecture, which performs the convolution and hence constitutes the main computational load of the network. The convolution operation is a dot matrix between two matrices, where one

---

[8] https://towardsdatascience.com/

is the kernel, a learnable set of parameters, and the second matrix is a restricted portion of the receptive field in the image. The kernel is usually smaller than the input image to achieve sparse interaction and learn relevant information from the pixels in the image. During a forward pass, the kernel slides across the height and width of the input image and produces an output representing the receptive field. This output is known as the activation map, which is the kernel's response at each spatial position of the image. The sliding action that produces the activation maps is known as stride. To determine the output volume (feature map) of an input image size W X W X D and kernel size *F*, stride *S* and padding P, is given by (2.18).

$$W_{out} = \frac{W-F+2P}{S} + 1 \qquad (2.18)$$

There is a significant reduction in the number of parameters in a convolutional layer compared to a dense layer since the weights are shared across all spatial locations in the input. The pooling layer replaces the output of the preceding layer by sliding a 2D filter over certain locations to derive a summary statistic of the nearby outputs. This helps reduce the representation's spatial size, which decreases the required computation and weights. For a feature map with dimension n_h x n_w x n_c, where n_c represent the channel depth, the pooling layer is done by (2.19).

$$\frac{(n_h-f+1)}{s} \, . \, \frac{(n_w-f+1)}{s} \, . \, n_c \qquad (2.19)$$

As seen in the MLP architecture, the dense layer contains the entire connection between the neurons in the preceding and subsequent layers.

2.5.3 Convolutional Recurrent Neural Network (CRNN)

Convolutional Recurrent neural networks (CRNN) are a variant of the CNN combined with a recurrent neural network (RNN), with a hidden state and the ability to use feedback loops in processing sequential data, which decides the final output. Hence, CRNN can learn sequential features of input data and predict the next possible data point in the overall data sequence. CRNN can capture long-term dependencies, giving them the advantage of understanding and efficiently modelling contextual and temporal information. It is more suitable for sequential images, text and speech analysis due to its application in sequential data. CRNN can handle varied input and output lengths, and unlike feed-forward neural networks, it can access its internal memory to process inputs in sequences (Yasrab et al., 2020). In summary, the CRNN works as follows:

- Input: A sequence of data
- Convolutional layers: These layers extract features from image inputs, making them particularly effective.
- Recurrent layers: They receive the output from the previous layers and effectively process the sequential data, with each layer maintaining a hidden state that captures the contextual information about previous sequence entries.
- Connections between recurrent and convolutional layers: This reduces the overall network complexity and preserves salient input features along the network.
- Output: This can be a sequence of words or any relevant output related to the input. It is produced by the last layer, which is a fully connected layer.

Some drawbacks regarding using CRNN include difficulty in training, complex model architecture, difficulty in interpretability, lack of robustness and computationally expensive and time-consuming (Liu et al., 2023; Xu et al., 2023). Some common cases for CRNN are mainly in social media for sentiment analysis and reinforcement learning. A basic visual representation of CRNN is shown in Figure 2.8 below:

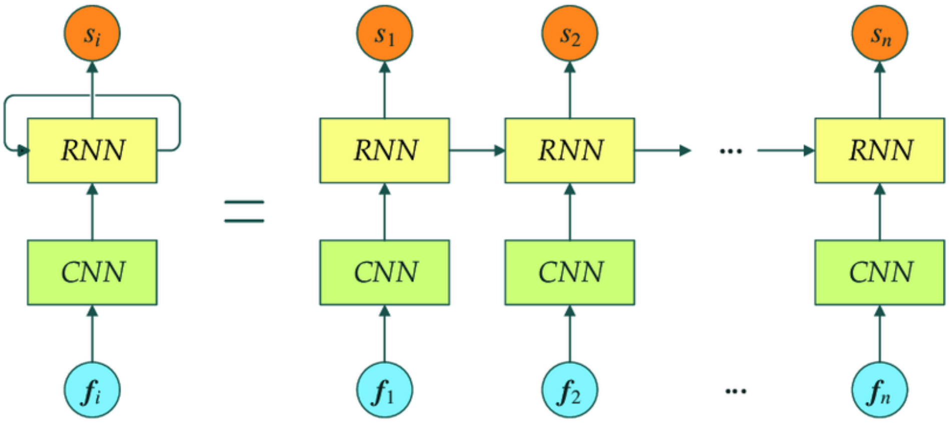


Figure 2. 8: A basic CRNN architecture (Yuan et al., 2019).

As explained in the steps provided, given a sequence of data, CRNN learns to predict its scores. The 2D CNN layers first process each entry, and the output of features as a vector is fed into the RNN. The RNN concatenates the information of the current step with that of the previous step and outputs the current entry's score (Yuan et al., 2019).

However, from a comprehensive medical imaging literature analysis, CNN is considered to be more potent than RNN (Banerjee et al., 2019). CRNN includes less feature compatibility when compared to CNN (Alzubaidi et al., 2021). CNN is ideal for images and video processing and has been widely applied in medical imaging analysis as a powerful modelling technique (Sarvamangala & Kulkarni, 2021), and this justifies the choice as the neural network of choice in this PhD thesis in addition to visual inspection of the burned-in text on the images. Additionally, the problem domain for this research does not involve a vocabulary, considering there is no fixed lexicon for patients' names and other burned-in textual data that may appear. Hence, the data are not sequential, and this PhD thesis aims to design modelling solutions applicable to varied medical images across different locations and without any dependency on a vocabulary.

2.5.4 Attention Mechanism

Attention mechanisms are a neural network layer inserted into DL models to selectively focus their attention to specific regions of input data based on different weights assigned to different regions. This priority-based mechanism improves prediction accuracy, as it emphasises discriminative parts of the data. The attention mechanism generally works by breaking the inputs into smaller regions and deciding which region has more relevance by comparing it to a pre-determined query. For instance, these could be words in sentences or different aspects of an image. It then assigns each part with a score and, based on this score, determines how much attention it gives to each part by assigning weights (Bahdanau et al., 2014). The original attention mechanism was proposed by Bahdanau et al. (2014), but it was mainly for putting emphasis on words in neural machine translation. However, these have been applied across different medical image analysis aspects such as classification, segmentation, and detection, as they enable CNNs to focus more on semantically important regions (Rao et al., 2021; Li et al., 2023). In images, the attention mechanism computes the correlation of feature vectors from input images, and this correlation shows the relationship between the global pixels, and then weights are assigned according to this correlation. Types of attention mechanisms include channel attention (calculates the importance of channel components through the exploitation of the inter-channel relationship of features) and spatial attention (considers the spatial feature information and assigns higher weights to discriminative location information) (Shi et al., 2022). However, to take advantage of both types, hybrid attention can be composed of channel and spatial attention fused or in series or parallel. The hybrid attention fully takes into consideration both the channel and spatial information of feature maps, hence making weights more effective, thereby increasing overall representation capability (Li et al., 2022).

2.5.5 Few-Shot Learning Method

With the issue of limited data samples, the few-shot learning method was developed to enable models to learn and make predictions based on only a few data samples. Few-shot learning leverage generalisation over memorisation (Seo et al., 2021). In Few-shot learning, the goal is to train the model to know the similarities and differences between different classes of samples rather than simply training the model to know what class each sample belongs to (Zhang et al., 2023). A support set containing a few samples of each class is used to train a model using the few-shot learning method, which is ordinarily impossible to use in training a DL model. For instance, 5 samples per class. The basic way the few-shot learning works is as follows:

- Assign a similarity function , simi(x, x’)
- The function measures the similarity between two data samples, x, and x’
- If the samples are the same, the function returns 1
- If the samples are not the same, the function returns 0
- The model is trained to learn the similarity function, which can be used to make predictions for unseen data samples by calculating their predicted similarity scores.

With the issue of dataset accessibility in medical imaging, a few-shot learning method has been applied to solve the issues of data scarcity and enhance medical image analysis for classification (Cai et al., 2020; Singh et al., 2021) and segmentation (Sun et al., 2022; Feng et al., 2023).

2.5.6 Generative Models

Generative models can synthesise new data with the same distribution as their training data samples. These models can either explicitly learn an estimate of this distribution or be trained to sample from the estimate. Since its introduction, generative models have found diverse applications in engineering, medicine, and sciences, such as super-resolution of images, image-to-image translation, image reconstruction, and so much more. Based on the recent advancement of DL, some generative models have gained much research attention. This sub-section will briefly discuss the Variational Autoencoders (Kingma & Welling, 2013) and generative adversarial networks (Goodfellow et al., 2014).

Variational Autoencoders (VAE) (Kingma & Welling, 2013) is a generative model whose training can be regularised to avoid overfitting and ensure that the latent space has good properties to enable the generative process. In contrast to the autoencoder training, where the input is encoded as a single point, the VAE is fed with the input as a distribution over the latent space. During training, a data point from the latent space is sampled, then the sampled data point is decoded, and the reconstruction error is calculated and backpropagated through the network. These encoded distributions are Gaussian distributions that enable training to return the mean and covariance matrix.

The Bayes theorem can be used as an approximator to compute this latent space and look for the best approximations. A summary of this is shown in Figure 2.8 below.

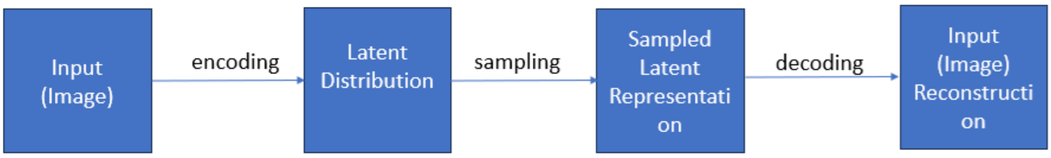


Figure 2. 8: VAE generative modelling process

VAE is faster than other generative models as it can generate samples in a single iteration. However, these generated samples are blurry due to the mean squared error typically used in the reconstruction term (Bredell et al., 2023). VAE consists of an encoder, a decoder, and a loss function. The encoder is a neural network, and the input is a data point $\boldsymbol{x}$. The output is a hidden representation $\boldsymbol{z}$ having weights and biases $\boldsymbol{\theta}$. For instance, for an input of a character image with dimension 28 X 28, the encoder encodes the data points which is a 784-dimension, into a latent representation space $\boldsymbol{z}$. The encoder can be denoted as $\boldsymbol{q_{\theta}(z \mid x)}$, and since the lower-dimensional space is stochastic, the encoder search for the optimal parameters for the encoding, which is a gaussian probability density. Sampling from this distribution can be done to get noisy values of the latent representation space $\boldsymbol{z}$. The decoder takes the latent representation space $\boldsymbol{z}$ as input and outputs the parameters of the distribution with weights and biases $\emptyset$. The decoder can be denoted as $\boldsymbol{P_{\emptyset}(z \mid x)}$. The decoder outputs 784 parameters from $\boldsymbol{z}$, each pixel represented from the distribution. The entire information from the original 784-dimensional vector cannot be transmitted because only a summary of the data points is available to the decoder on the latent space. The reconstruction log-likelihood $logP_{\emptyset}(z \mid x)$ measure the information lost in between encoding and decoding. The VAE's loss function is the negative reconstruction log likelihood, and for a datapoint $\boldsymbol{x_i}$, the loss function $\boldsymbol{l_i}$ can be denoted as (2.20)

$$L_i(\theta, \emptyset; x_i) = - \; KL(q_\theta(z \mid x_i) \mid p(z)) \quad + \; E_{z \sim q_\theta(z \mid x_i)}[logP_\emptyset(x_i \mid z)] \qquad (2.20)$$

KL is a regulariser known as Kullback-Leibler divergence, which measures the information loss by checking the difference in the divergence between two distributions. The first term on the right $E_{z \sim q_\theta(z \mid x_i)}[logP_\emptyset(x_i \mid z)]$ represents the expected reconstruction log-likelihood and it measures the similarity between the original data and the generated data. The second term on the left, $KL((q_\theta(z \mid x_i)) \mid p(z))$ is the KL divergence between the approximate posterior, $q(z|x_i)$, and the prior $p(z)$. KL it acts as a regularization term that ensures that the latent representation z is a sample from the prior distribution $p(z)$. These terms are derived and explained much in depth in the original paper by Kingma & Welling, (2013).

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are a group of neural networks based on unsupervised learning that can generate new samples. A classic GAN comprises two parts: a Generator and a Discriminator. The generator competes with the discriminator, which aims to distinguish between real and generated data. This competition between these two parts is referred to as "adversarial". GANs are widely applied in major aspects of generative modelling, such as image reconstruction and inpainting. Training GAN is difficult, as the goal is to optimise the Generator and Discriminator alternatively during iterations. GAN produces more high-resolution images than the VAE. In GAN, the loss function for the simultaneous optimisation of these two parts is known as the MinMax loss, and it is given mathematically as (2.21).

$$min_G \, max_D \, (G, D) = [\mathbb{E}_{xp_{data}}[\log D(x)] \; + \; \mathbb{E}_{zp_z(z)}[\log(1 - D(g(z)))]] \qquad (2.21)$$

G is the generator network, and **D** is the discriminator network. $\boldsymbol{p_{data}}(\boldsymbol{x})$ is the true data distribution that samples actual data samples $\boldsymbol{x}$. $\boldsymbol{p_z}(\boldsymbol{z})$ is a previous distribution that samples a random noise $\boldsymbol{z}$. $\boldsymbol{D}(\boldsymbol{x})$ is the likelihood of the discriminator to identify original data as real correctly. $\boldsymbol{D}(\boldsymbol{G}(\boldsymbol{z}))$ is the likelihood of the discriminator identifying generated data from **G**, as authentic. There has been much research in the design of various variants of GAN to solve this limitation: Conditional GAN (CGAN), Deep convolutional GAN (DCGAN), and Super Resolution GAN (SRGAN), amongst others. Each of the variants aims to make improvements over the simplest GAN known as the vanilla GAN.

2.6 Chapter Summary

A comprehensive overview and survey of the most important techniques used to develop computer vision applications have been provided, which are the basic foundation of OCR: feature selection, image preprocessing, and an overview of ML and DL techniques. ML techniques have been successful for many years for computer vision applications. However, in recent years, with more computing power and the complexity of certain grid-like data, such as images, deep learning techniques were developed to solve these challenges by improving feature selection and representation learning. This section will be relevant to better understanding the contributions made in this research. In particular, feature extraction and convolutional layers are used extensively in Chapters 5 and 6. Dense layers and deep generative models are used in Chapter 7.

3.0 Literature Review

Application of traditional OCR systems in burned-in text data recognition in MIM still has spaces for improvement because of various limiting factors such as low resolution, small font size and background interference. A review of previous studies on the recognition of burned-in text data on MIM is provided in this chapter with reference to the state-of-the-art to show the existing gaps this research aims to fill. In the case of MIM, the burned-in textual information is small in font size and low resolution, making it difficult for traditional OCRs to accurately recognise these text data as these OCRs (For example, Tesseract) are trained to recognise text with a minimum resolution of 300dpi and of 12 pt font size, this research focuses on recognising burned-in textual data at about 96 DPI.

Florea et al. (2005) began one of the earliest works for the automatic indexing of medical images for image retrieval purposes inside a large online health database; they planned to achieve such a task by recognising and extracting textual annotation present in these medical images using image processing and OCR. The authors worked on these modalities: angiography, ultrasonography, magnetic resonance imaging, standard radiography, computer tomography (CT) and scintigraphy. After visual inspection and reviews, the authors concluded that the characteristics of medical images are alike irrespective of the type of modalities. Hence, a solution for textual annotation extraction from a particular modality image's pixel structure would also be usable in other modalities. The authors also noted that the visual content of medical imaging modalities is highly affected by its means of acquisition. Hence, the entire pixel structure containing this burned-in textual annotation is directly affected by the overall image resolution after its acquisition. The authors proposed a technique to recognise burned-in textual information at the character level, using a prior knowledge of the colour and thickness of each character, as well as applying standard morphological means before applying a commercial OCR software known as Abbyy FineReader 7.0 to identify the characters. They employed a manual image thresholding approach known as the TopHat filter set on each character thickness. This method can isolate objects lighter than the neighbourhood and smaller than a structural element conveniently chosen; this method was decided after a critical evaluation showed text regions of all modalities have relatively resembling characteristics (Florea et al., 2015) in font, colour and thickness, but distortion and sizes of these texts were largely dependent on the resolution of the image, which are most times close to the image borders. Though they achieved a recall rate of 60% in CT images, their method failed to recognise some burned-in text in various modalities, such as the angiography category, as this method failed to extract maximum features at the recognition stage and the image's low resolution which resulted to the small font size of the burned-in text data, hence poor performance of the commercial OCR. The OCR used was designed to recognise printed characters and not burned-in textual information.

Alter & Werner (2007) went further to extract the zoom factor of the small font size of characters in ultrasound having a low resolution in order to analyse the character thickness and hence identify the burned-in text. Similarly to the previous work by Florea et al., 2015, these authors also made use of an open-source OCR, which was not designed to deal with the low resolution of varied medical image modalities. The proposed method by Alter & Werner (2007) was not consistent when applied to varied modalities and required multi0ple image pre-processing steps before exporting to the open-source OCR.

Reul et al. (2016) applied region of interest detection, binarisation and segmentation to extract burned-in text-containing areas into lines and pass them into an OCR engine. Due to the large error from the recognition by the open-source OCR used, the authors proposed an OCR correction technique that involves an assisted user revision using Excel structured tables, where the errors can be checked and manually corrected. The authors used prior knowledge of the burned-in text data, resulting in a highly user-assisted approach, which is not feasible with the modern-day variety and quantity of medical image modalities. The authors concluded that the recognition errors of the samples used for evaluation were difficult to avoid due to the low quality of the burned-in text region in the image modalities caused by its low resolution (Ruel et al., 2016). Though they achieved a very low error rate of 0.6%, this approach only allows almost optimal recognition rates by making use of very specific constraints.

Monteiro et al. (2017) proposed a model using CNN to recognise burned-in text using a simple 6-layered network design, and a good accuracy rate was obtained. However, the evaluation was done using only a single type of medical image modality (Ultrasound) from a single institution with existing knowledge of the detected text data. The authors did not evaluate the proposed model to other modalities outside their institution to properly address the problem of recognition of burned-in text data regarding the problem of background interference, small font sizes and low resolution. Hence, their proposed system was limited and cannot be generalised.

Between 2005 and 2023, most studies focused on improving image pre-processing techniques and feeding them into an open-source or commercial OCR, which was not explicitly designed to meet the challenge of small font sizes and low resolution. A further literature review by this study on the quality assessment of these MIMs showed the small font size of the burned-in text; its low-resolution results from the overall low quality during storage (Chow et al., 2016).

The next sub-sections briefly explain the nature and acquisition of MIM, existing ML-based practices, related works on the domain of MICR, challenges and open issues, suitable evaluation metrics, and lastly, the research objectives based on the identified literature gaps.

3.1 Nature and Acquisition of Medical Image Modalities (MIM).

Medical imaging includes different modalities and processes to visualise the interior of the human body parts for varied clinical purposes (Firoz et al., 2017). However, the most common problems with these types of images are poor contrast quality, low resolution, and background interference (noise). The several objects coupled with the degradation stated above, including proximity of adjacent pixel values, which may lead to overlapping one object on another object in the same image, make the application of several traditional OCRs in identifying burned-in text data a difficult task.

During the image acquisition process, the imaging part is automatically combined with patient text data, and both are merged into the same pixel structure, resulting in the text data appearing as burned-in text data and not printed on the image. These burned-in text may contain sensitive information and vital information for diagnostic purposes, which may be useful for various post-processing needs. Hence, an efficient way of recognising these burned-in text data is required in such a constrained situation.

3.2 Character-level Recognition - A Review and Justification

In burned-in textual data recognition, the recognition choice is always at the character or word levels. The character-level approach involves recognising and classifying the extracted image patch according to a predetermined target class of characters. In contrast, the word-level approach recognises the extracted image patch as a word unit rather than a single character.

The word-level approach has the advantage of avoiding the problem of character segmentation and overcoming local errors in the character-level approach (Erlandson et al., 1996). The pipeline process for this approach begins with computing a vector of a query’s input image-morphological features and matching this vector against a predetermined database of vectors from a lexicon of computed wordlist. The vectors with the highest match score are returned as possible outcomes for the unknown input image containing a word.

The character-level approach aims to detect, segment, and recognise an input image patch into individual characters. The detection can be done using the popular OpenCV python library, and to effectively segment the individual characters or words for recognition, the situation of a uniform gap being maintained between each character or word on the input image has to be considered significantly to establish a threshold.

The main factor to be considered in the choice of approach in the level of recognition of an OCR model for a problem lies in the size of the lexicon of that problem domain. This research conducted a thorough physical inspection of varied medical image modalities to understand the lexicon in use. The burned-in textual data included sensitive data such as patient’s names, patient IDs, clinical parameters, and other unique clinical examination details (Tsui & Chan, 2012), depending on the patient and the diagnostic process in place. Unlike word-level recognition that require a predefined vocabulary, character-level recognition relies only on a combination of characters (Chung et al., 2019).

Therefore, it would be right to conclude that there is no fixed lexicon in the domain of the burned-in text data that could be built into medical image modalities. Using this as a guide, this research proposes recognising these burned-in text data at the character level.

This research proposes deep learning techniques designed based on CNN that achieved a high accuracy of approximately 98-99% on handwritten digit recognition (Kayed et al., 2020; Tabik et al., 2017; Buda et al., 2018). The proposed DL techniques are geared towards the intended task in MIM, which is to design highly specialised DL models to tackle the challenges in MICR.

3.3 Machine Learning-based Techniques for Burnt-in Textual Data Recognition

Designing and implementing an ML model that accurately recognises burned-in textual data on MIM has been challenging over the years. Several works have proposed different combinations of image pre-processing techniques and ML models to recognise burned-in textual data on low-resolution MIM either at the word level with a pre-determined wordlist or the character level. These varied MIMs are usually unstandardised images, making conventional OCR methods unreliable because of MIM variety in low contrast, distortion, low resolution, and background interference. The main problems in recognising burned-in textual data are tackling the unstandardised image problems with critical emphasis on low resolution and background interference. This is because the low contrast, distorted text lines, skew, and background noise result in poor OCR outcomes. In addition, the low resolution makes characters more zigged and merge with the image's background, leading to erroneous character recognition. This area needs to be investigated when considering high recognition rates from open-source and commercial OCRs on the standard printed text and unconstrained text. Different past authors have applied ML algorithms that have been successful in other classification tasks in medical image processing, such as computerised tomography (CT) image classification to detect lesion classification, X-ray classification to diagnose pneumonia and classifying Magnetic resonance (MR) brain images of patients for mild cognitive impairment (MCI). Mostly widely used algorithms for these general classification tasks in MIM include random forest, gradient boosting classifiers (Rabiei, 2022), support vector machine (SVM), support vector regression (SVR), naive Bayes, k-nearest neighbour algorithm (K-NN), decision tree (DT) algorithms (Amethiya et al., 2022). These ML algorithms classify which parts of the human body, presented by the medical image, are infected by the disease using various feature extraction and selection techniques. The poor outcomes of these ML algorithms in burned-in text recognition are due to the problems outlined above, requiring further in-depth research in this area.

A detailed analysis by Newhauser et al. (2014) showed that the most popular OCR (Tesseract) was specially designed for recognising text on office documents with character size scanned at a resolution of 300-400 dpi. These documents have a pixel dimension of 1700 X 2000 pixels. Therefore, a text with an 8pt font size under this resolution of 300-400dpi will be about 22 pixels and can be easily recognised using popular OCRs. A text of 22 pixels means that each character takes up 22 pixels on the image from the top of the character to its bottom. In contrast, a regular computerised tomography (CT) image has a pixel dimension of 512 X 512 pixels, and a burned-in text of 8pt font will be approximately 9 pixels (Newhauser et al.,2014). The resolution in MIM is significantly less than what the popular OCR can recognise. The accuracy of these popular OCRs drops rapidly for a text with an 8pt font size, resolution less than 300dpi, and fewer pixels than 22 pixels per character (Newhauser et al.,2014).

A study by Menasalvas & Gonzalo-Martin (2016) on the analysis of non-structured text on MIM also indicated that there would be a need to develop new algorithms and methodologies that can take full advantage of the burned-in textual data contained in these MIM. Menasalvas & Gonzalo-Martin (2016) stated that a significant problem in this area lies in the variety of these MIM in background content and low resolution, hindering the applicability of conventional OCR solutions. Hence, designing an ML-based OCR with a high recognition accuracy for any input MIM from any organisation, country, and lexicon is a technically challenging task. In the general domain of OCR under difficult conditions (such as text in natural scenes and degraded hand-held camera-captured document images), several authors have proposed various solutions, such as structure extraction by graph spectral decomposition and component selection criterion (Kawano et al.,2010), multiple commercial OCRs with majority logic (Miyao et al., 2004), reinforcement learning, and multiple recognisers (Park et al., 2020). Good results were achieved by these works, such as an 82.3% recognition rate for decorated characters (Kawano et al.,2010), a 98.83% recognition rate for printed Japanese characters (Miyao et al., 2004), and a 90.1% recognition rate for Chinese characters with unique character shapes (Park et al.,2020). However, these solutions are unsuitable for MIM because of their lower resolution and background interference.

Regarding the ML approach to recognising burned-in textual data, different authors have proposed combining various image-filtering algorithms with traditional ML models (Vcelak et al., 2019). More recently, convolutional neural networks (CNN) have been employed to solve multiple difficulties experienced in using open-source OCR (Mohsenzadegan et al.,2020). The challenge in the ML approach is the need for a high-performance classifier that can distinguish similar characters with low resolution in MIM, even in the presence of background interference.

3.4 Overview of previous works in the use of ML in recognising burned-in text in MIM.

In this subsection, notable works that have proposed ML-based solutions to recognise burned-in textual data will be reviewed, and the gaps presented to motivate this research will be presented. Table 3.1 below provides an overview of previous works on the use of ML techniques for the modelling of different solutions to recognise these burned-textual data and presents some main reference papers in the recent literature along with the authors' names, methods used, the dataset sources, the applied evaluation metric (see section 3.7), and the outcomes of the works. The ML approach usually requires a dataset collected from a public or private source (Segal & Hansen, 2021). The collected dataset can be divided into training and validation datasets. The training dataset is used to train the ML model. The validation dataset estimates the trained model's performance while tuning the model's hyperparameters.

Table 3. 1 : Overview of previous works in the use of ML in recognising burned-in text in MIM (Papers arranged in ascending order by publication year).

Works	**Image Pre-Processing Technique**	**Modelling Technique**	**Dataset**	**Evaluation Metrics**	**Outcomes**
Wang, (2002)	Daubechies wavelet's image transformation.	Open-source OCR.	100 medical images were collected from a public source and Stanford medical centre.	Character Recognition Rate (CRR)	The results cannot be generalised due to the small validation dataset used and the problem of low resolution.
Antunes et al. ,(2011)	Template matching from pre-existing MIM metadata	Open-source OCR	Several hundreds of ultrasound images. (Exact quantity not mentioned)	CRR	Poor performance in complex backgrounds with overlapping text data.
Tsui & Chan (2012)	Regional thresholding and morphology for character segmentation.	Tesseract OCR with weighted similarity.	189 ultrasound images from 6 volunteers. Simulated images were produced from 660 previously	Character Recognition Rate (CRR)	CRR of 99.5% but relied on a pre-determined dictionary and a human-assisted revision for error correction.

			anonymised medical images.		
Newhauser et al. (2014)	Threshold-redaction algorithm	Tesseract OCR	NIH-funded studies from 13 patients for cancer treatment.	Character Error Rate (CER)	50% CER achieved, but poor results on Xray Images.
Monteiro et al. (2015)	Total-variation denoising, adaptive bilateral filtering, and binary thresholding.	Restricted Boltzmann machine, and Random Forest Classifier	Training data was from a public Character MEDPIXnd validated on a 60 ultrasound image collected from a Portuguese medical centre.	False positive rate, false negative rate, F1-score, precision, and recall	Could not recognise certain small font sizes and types in low-resolution MIM.
Ma & Wang, (2015)	Using local features such as edge density.	Adaboost Classifier	100 medical images with text-ultrasound, MR, CT, X-ray. The size was between 300 X 600 pixels to 800 X 1200 pixels.	Computational cost, precision, and recall.	The precision was 74%, and the recall was 77%—difficulty in recognising varied fonts in low-resolution images.
Reul et al., (2016)	An expectation-driven method by using prior knowledge of	Open-source OCR	22,500 ultrasound images were collected from	Character Error Rate (CER),	A user-assisted revision method with a low error rate of 0.06%.

	the position and appearance of the textual data in the image.		an investigation of 26 peripheral nerves, 225 measurements are performed on at least 100 subjects.	Word Error Rate (WER)	poor generalisation with complex processes.
Monteiro et al (2017)	Total-variation de-noising, Adaptive bilateral filtering, and binary threshold	CNN	Training data was from a public character MEDPIXnd validated on privately collected 500 ultrasound images.	CNN model's Precision, recall and F1-score.	It depended on complex processes and could not be applied to varied MIM with low resolution.
Silva et al., (2018)	Adaptive bilateral filtering and total-variation de-noising.	CNN	400 high-resolution varied medical images were collected from a private clinic facility.	Character Error Rate (CER), Word Error Rate (WER)	The model could not recognise certain font types and similar characters. .
Vcelak et al. (2019)	Binarisation for image transformation.	Tesseract OCR	15,334 images for training and 70,191 for validation were collected from the University	Weighted average recall and inverse recall, Cohen's kappa	FPR of 1.81%-4.00% requires a pre-determined dictionary.

			hospital in the Czech Republic.	coefficient, False positive rate (FPR)	
Xu et al., (2021)	Image blending	CRNN	2500 images from the Medpix cardiac atlas (MRI) database.	CRNN model’s Precision, recall and F1-measure.	Could not recognise similar characters in low-resolution MIM.

Table 3.1 shows that these past studies encountered similar challenges in recognising burned-in textual data on MIM, which are (a) the problem of background interference and (b) the problem of low resolution. The problem of background interference in MIM occurs mainly due to a grey background, fuzzy font, overlapping text, and inconsistent image quality (too-bright or too-dark image). The problem of low resolution occurs at a resolution of 72-150 DPI for varied modalities such as ultrasound, CT, and others.

Additionally, in terms of the dataset key aspects used in these past works ; the problem of accessibility to MIM is seen, as some of these papers such as Xu et al. (2021), Monteiro et al. (2015) included synthetic character images in their study, even though the MIM used were mostly small dataset except for Vcelak et al. (2019). These accessibility challenge remains a significant issue in the research in medical imaging domain. The MIM used by these papers shared similar characteristics which are mainly poor contrast, background noise, low resolution, and distortion. These characteristics did not appear in a single form, but in composites and the extent of these characteristics are determined by the acquisition machine used, lightening condition and selected variables during image acquisition. However, these papers did not explicitly mention the magnitude of the image resolution in DPI which they worked on and this further limits the generability of their findings.

The studies by (Wang, 2002; Newhauser et al.,2014; Tsui & Chan, 2012; Reul et al., 2016; Vcelak et al., 2019) followed similar methodologies to recognise the burned-in textual data while focusing on the problem of background interference. The authors of these studies mainly proposed image transformation techniques to improve the background contrast and fed the enhanced image to an open-source or commercial OCR. Wang (2002) applied the Daubechies wavelet image processing, and Tsui & Chan (2012) performed morphological operations. Several studies (Newhauser et al.,2014; Reul et al., 2016; Vcelak et al., 2019) implemented multiple thresholding techniques to improve the background. These studies focused on increasing the image's local contrast, that is, the contrast between burned-in textual data and background pixels, to make it easier to recognise the characters.

Several studies (Monteiro et al., 2017; Xu et al., 2021; Badano et al., 2015) focused on the low-resolution problem in MIM and proposed various solutions to recognise burned-in textual data on these MIM. The work by Mário et al. (2011), after proposing a character template solution to recognise burned-in textual data focusing on the MIM's low resolution of 352 dpi, concluded that the quality (resolution) of the generated character dataset (before recognition) is a principal factor that determines the character recognition accuracy. Monteiro et al. (2017) performed recognition on ultrasound images with a 6-layer CNN, achieving a recognition rate of 89.2% on 500 processed images. The low-resolution problem was suggested as the reason why the CNN-based solution could not recognise certain font types and characters. Xu et al. (2021) went further to propose a solution for the low-resolution problem by using a Convolutional Recurrent Neural Network (CRNN) combining scale variant features during training. Though Xu et al. (2021) achieved a recall of 65%, a precision of 70% and an F1-measure of 67% in cardiac magnetic resonance imaging (MRI), the authors (Xu et al., 2021) concluded that the solution was not transferrable to other types of MIM. The system could not distinguish similar characters in low-resolution MIM. The study (Xu et al., 2021) did not evaluate the character recognition rate MIM but provided only the model's performance metrics.

This sub-section shows that there is a need to explore further research in ML and DL to improve state-of-the-art recognition accuracy of burned-in textual data. Such research focuses on designing a specialised ML or DL model to recognise these characters as accurately as possible under these problematic conditions of low resolution, background interference and noise corruption (Kociołek et al., 2020). The background interference occurs on varied MIM because of the lightning and image acquisition process (Maier-Hein et al., 2018). The low resolution is standard on varied MIM because of the limited storage of the acquisition machines, leading to reduced image quality (Aljabrin et al., 2022). Hence, this research suggests that a prompt understanding of MIM's background interference and low resolution is required to design an optimal classifier effectively. This understanding will enable research into the design of creating a specific classifier for each modality with a critical focus on the problem of low resolution and background interference.

3.5 Open issues and challenges in Burned-in Textual Data Recognition

Various works have been done using traditional image pre-processing techniques and open-source OCR, and more recently, using ML algorithms to recognise this burned-in text. Nevertheless, some challenges and issues still need to be solved. These challenges and issues exist because of the constraints of acquiring MIM in different difficult conditions, resulting in low resolution and background interference problems due to hardware limits (Li et al., 2021). The open issues discussed in this section include (a) No consensus to measure image enhancement and (b) small medical imaging dataset. The significant challenge identified from the different existing ML approaches from an extensive literature review is (a) the discrimination of visually similar characters in low-resolution MIM with background interference.

There is yet to be a consensus on measuring the image enhancement on these MIM to get a high recognition performance from the OCR due to the unique nature of the images. This is partly because there is yet to be a commonly accepted metric to measure the level of image enhancement, though some researchers have proposed the Peak-Signal-To-Noise-Ratio (PSNR). Michalak et al. (2019) used the PSNR to measure the success of a proposed image pre-processing methodology using local image entropy for an OCR in text recognition on illuminated document images. Their study suggested a more helpful approach would be the application of metrics calculated for recognised characters based on the Levenshtein distance, known as the Character Error Rate (CER), instead of individual pixels. Bieniecki et al. (2017) used the CER to evaluate the image’s pre-processing methods for text recognition in distorted document images using open-source ABBYY FineReader. Another study (Nomura et al., 2009) showed that the resulting models' CER is commonly used to measure the image pre-processing success rate. The study (Nomura et al., 2009) concluded on the average CER metric from a quantitative evaluation on a test dataset of 1194 degraded word images to show the essentiality and effectiveness of their proposed image pre-processing method to increase the character recognition rate. Nomura et al. (2009) applied a modified Otsu global thresholding technique, which reduced computational requirements and improved the CER in degraded digital word images on an open-source OCR system. Past researchers have applied these generalised metrics in evaluating their overall pipeline by mainly using the CER to measure the level of image enhancement without directly measuring the image enhancement stage. There is a need for a benchmark metric to determine the success of the image enhancement used as different image filtering algorithms from past works have been applied. This current research due to time constaints, will not focus on this issue, and that can be part of future works.

The issue of small datasets is of serious concern in applying ML to medical imaging, including the OCR for burned-in textual data recognition, because ML requires a lot of training data to enable optimal tuning of parameters by the learning algorithm. DL algorithms for image classification require large datasets to produce good results, and they perform poorly with small datasets (Davila et al., 2021). Privacy protection requirements and accessibility greatly hinder the availability of MIM. The resultant effect has led to small dataset of MIM available for the implementation and validation of pipelines, leading to slow progress in the field. Health centres housing MIM usually follow a regional regulatory framework such as the Health Insurance Portability and Accountability Act of 1996 in the United States (HIPAA), as a mandate on the privacy protection of patient's medical records, which ensures that a guarantee is given on the confidentiality of data during storage and transmission via any secured or unsecured means(Li et al.,2005). Due to the adapted regulatory framework, collecting a sufficiently large-scale, balanced MIM dataset is difficult (Qin et al., 2019). Several medical image classification competitions have been organised in recent years, motivated by the need to provide more datasets to the ML community to investigate novel ML algorithms on medical images. A notable competition was the Grand

Challenge for Biomedical Imaging, organised by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) in 2007 (Maier-Hein et al., 2018). This medical imaging competition uses annotated datasets to ensure a uniform validation protocol is available for all participants (Aljabrin et al., 2022). Such annotated datasets cannot be used for textual data recognition in MIM research.

With the existing problems of low resolution and background interference, discriminating visually similar characters (VSC) is a significant challenge in various ML approaches to recognising burned-in textual data in MIM. A poorly defined character due to low image resolution, background interference, and small fuzzy font sizes can often distort the geometric shape of the character (Pal et al., 2021). Human vision sometimes misinterprets VSC, especially when these characters stand alone. This challenge has guided research in improving the recognition rate of similar characters such as “0” and “O”, “5” and “S”. A study by Inkeaw et al. (2019) similarly identified this challenge in OCRs and suggested a classifier-based approach may be the solution to improve the recognition rate of these VSC. There is a need to develop a complex classifier using ML and DL techniques to adequately learn discriminative features of VSC existing on MIM. Monteiro et al. (2017) and Vcelak et al. (2019) identified and attempted to solve this significant challenge using DL techniques by applying a 6-layer CNN but could not recognise certain VSC.

3.6 Literature Review Conclusion

In this section, I have concentrated on related works regarding recognising burned-in textual data, which is low resolution and background interference. All these conditions have made it challenging for conventional OCRs, image pre-processing and ML approaches to recognise these textual data. As a result, an accurate medical image character recognition system will be considered a significant milestone in medical informatics. This will improve healthcare delivery by accessing and using these identified textual data in decision-making systems. This will also apply to scene text recognition for low-resolution images with noisy backgrounds.

Furthermore, the recognised burned-in textual data will be relevant for post-OCR purposes such as anonymisation, sensitive data obfuscation, automatic integration into EHR systems, and developing a controlled search mechanism for MIM in a large medical database using text-based queries (Safaei, 1995). As regards a way forward in solving these challenges and open issues discussed in the previous section, I suggest efforts in the areas of (a) Collaboration between medical imaging centres and the ML research community and (b) an advanced DL approach.

A possible approach to managing the issue of small datasets and the lack of image enhancement measurement consensus is, first of all, to encourage collaboration between medical imaging centres and the ML research community. This collaboration will enable seamless MIM dataset collection and sharing to allow the researchers to carry out a more massive pattern recognition. There is a need to share MIM with privacy considerations to researchers and scientists using acceptable ethical standards (Pal et al., 2021). This would come with similar efforts in finding optimal image pre-processing standards for each MIM to provide guidelines on steps that can be applied to MIM before feeding onto an ML algorithm. For instance, most of the ML and DL algorithms for OCR cannot be applied directly to the original image to avoid poor performance because a strong and significant representation of the pixel content of these images is highly relevant for the overall success of the algorithm (Inkeaw et al., 2019). With more datasets to conduct more research, ML communities can reach a consensus for acceptable standard metrics.

In recent years, the advancement of ML has led to the outstanding development of DL models that possess multiple optimisation strategies and deep-layer architectures. These advancements can solve the limitations posed by using single classifiers in past works. Each DL layer can capture patterns and deeper representation and abstraction, especially in image classification tasks (Menasalvas & Gonzalo-Martin, 2016). Though these are not new ideas in the area of medical informatics, a more advanced approach to DL techniques, such as the multi-column deep neural networks, weighted majority voting ensemble, stack ensembles, and DL-Tree Classifier ensembles, have not been researched in literature, to see how accurate these advanced techniques can recognise these burned-in textual data. These methods have been applied in other areas, such as handwriting benchmark recognition with averaged predictions on the benchmark MNIST dataset (Cireşan et al., 2012) and CNN to classify into 1000 class images in the ImageNet dataset (Krizhevsky et al., 2017). However, these advanced techniques are yet to be exploited to recognise burned-in textual data primarily because of small training data. A detailed analysis of classifiers’ voting techniques in the domain of pattern recognition by Lam & Suen (1997) showed that using a combination of classifiers resulted in an outstanding improvement in overall recognition results in the OCR domain, and this was not depending on the nature of the classifiers (Shlens, 2014). A detailed study by Kovács-V (1995) on the voting combination strategy on the NIST Special Database for hand-printed characters gave an error rate of 2.59% using three classifiers operated in parallel with a final supervisor classifier.

In summary, this literature review has shown there is still much research to be done regarding the recognition of burned-in textual data in MIM and revealed that their low-resolution mode of acquisition, complex background and small fonts remain significant constraints in this domain. Therefore, there is a need for further research in this area to create an automatic and highly accurate burned-in textual data recognition model. For a more comprehensive and in-depth review, each technical chapter from 5 to 7 includes related works: burned-in recognition in Chapter 5, Siamese neural network and attention module in Chapter 6, and the CVAE in Chapter 7.

3.7 Evaluation Metrics

The research would use the common error measurement in the OCR solution to evaluate the proposed pipeline, which is the Character Error Rate (CER). The CER is the percentage of erroneous characters identified in the model's output, and it is considered the most common metric in OCR-related tasks (Drobac & Linden, 2020). To derive the CER is shown in (3.1) below:

$$CER = \frac{S + D + I}{N} \qquad (3.1)$$

Where:

- **S** = No. of **S**ubstitutions
- **D** = No. of **D**eletions
- **I** = No. of **I**nsertions
- **N** = No. of characters in the ground truth

The S, D and I are all calculated based on the Levenshtein distance, where the minimum number of character-level modifications needed to transform the OCR output into the ground truth text is used to calculate the CER. The result of the equation (3.1) gives the percentage of erroneous characters in the OCR output. The lower the CER percentage, the better the performance of the OCR model. Notable OCR comparison studies by multiple authors (Bazzi et al., 1999; Vijayarani & Sakila, 2015), on the analysis of state-of-the-art OCR tools and a comparison of their performance, the authors concluded the comparison could be made only using two factors, model's accuracy, and CER. Natarajan et al. (2009) and Carrasco (2014) designed an opensource tool which computes the statistics of the difference between a provided ground truth and the output of an OCR model, and the result of this computation is used to analyse the performance of the model, this computation was referred to as the CER.

According to the literature review findings of the common use of the CER metric to evaluate the overall system performance of proposed OCR solutions in different studies, the research explored possible evaluation techniques for this present research from the literature. The findings show that the proposed DL techniques can be evaluated in two aspects; this is shown in Figure 3.1 below:

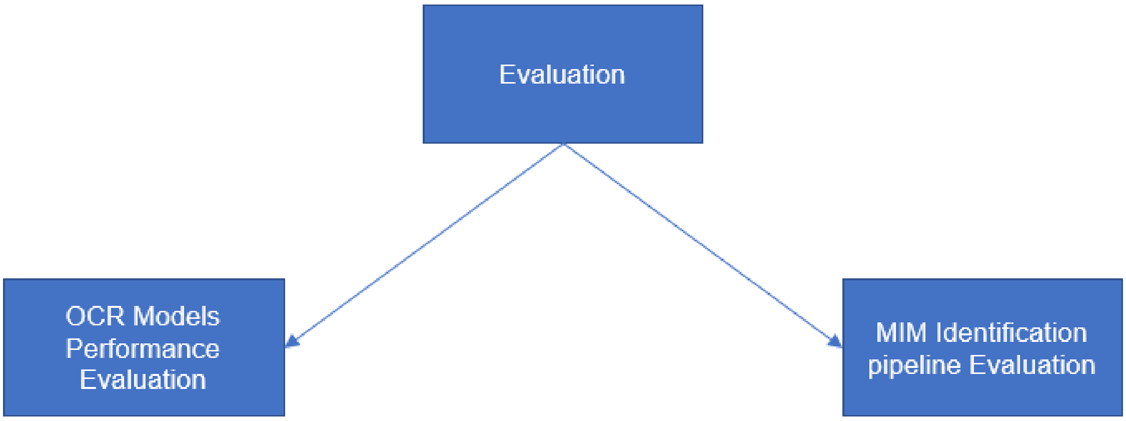

Evaluation
OCR Models Performance Evaluation
MIM Identification pipeline Evaluation

Figure 3. 1: Study's Evaluation Methods

3.7.1 OCR Models Performance Evaluation

The proposed models would be experimented, reported, and evaluated using the accuracy, precision, recall and F-1 score metrics. Hence, this research will consider different configurations of models and hyperparameter tunings, such as batch normalisation and data augmentation. The experimental results would be compared using the validation accuracy, precision, recall and F-1 score metrics; a similar direction was taken by Monteiro et al. (2017), Anand et al. (2020) and Shibly et al. (2021), where CNNs models were used for character-level identification.

Precision quantifies the fraction of predicted labels that is correct and corresponds to the target class. This can be expressed below:

$$Precision = \frac{TP}{TP+FP} \tag{3.2}$$

Recall indicates the fraction of the target class which is correctly identified among all of the samples, and this can be expressed below:

$$Recall = \frac{TP}{TP+FN} \tag{3.3}$$

The F-1 measure is a metric to determine the overall performance of the classifying model, and it can be expressed below:

$$F - 1\, Measure = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (3.4)$$

Where TP= True Positive, FP = False Positive

These formulas in 3.2, 3.3, and 3.4 are mainly for binary classification. However, to calculate for the mult-class problem, the microaverage technique following the one-vs-approach will be used to calculate the evaluation metrics separately for each character class with class predicted true and class predicted false, without regard to the wrong character predicted. This means that micro-averaging gives equal weight to each instance. The individual metrics will then be pooled to get the value of the final metrics for the model.

3.7.2 MIM Recognition Pipeline Evaluation

This evaluation involves visually inspecting if the low-resolution burned-in text data in small font sizes have been correctly recognised at the character level. The CER can be calculated from the results obtained in the model performance evaluation stage of different model configurations. The overall performance will be measured based on the accuracy obtained by the proposed models in predicting the correct characters, and the number of errors will be taken as the average CER over these medical image datasets during testing and evaluation.

3.8 Research Objectives and Research Questions

3.8.1 Research Objectives

1. Provide a critical analysis of the state-of-the-art techniques for recognising burned-in data in medical imaging modalities.

2. Carry out a data collection study in an on-site location to collect an original dataset vital for the proper evaluation of this study. This will follow an ethical approval process from the University ethics committee.

3. Propose deep learning techniques to recognise burned-in textual data in low-resolution medical imaging modalities with background interference.

4. Further investigations on recognising visually similar characters and generative modelling to tackle the issue of small dataset size for model training.

5. Evaluation of the proposed techniques and recommendations based on experimental results.

3.8.2 Research Questions

The research questions below are based on the gaps identified from the comprehensive literature review provided in this section, and this PhD thesis seeks to answer them.

1. Can a deep learning-based solution be designed to recognise burned-in text data with small font sizes, low resolution, and background interference in varied medical image modalities?
2. Can a deep learning-based solution based on few-shot metric learning be designed to recognise visually similar character images with a small dataset sample size in varied medical image modalities?

3. Can generative modelling be proposed to improve burned-in text data recognition by generating synthetic data samples for each character?

4.0 Research Methodology

The previous chapter reviewed existing methods for medical image character recognition. This chapter will address the chosen methodology within which this thesis will investigate the identified research gaps and propose, implement and validate solutions with a detailed description of the datasets used to evaluate the proposed techniques.

4.1 Overview of the experimental pipeline

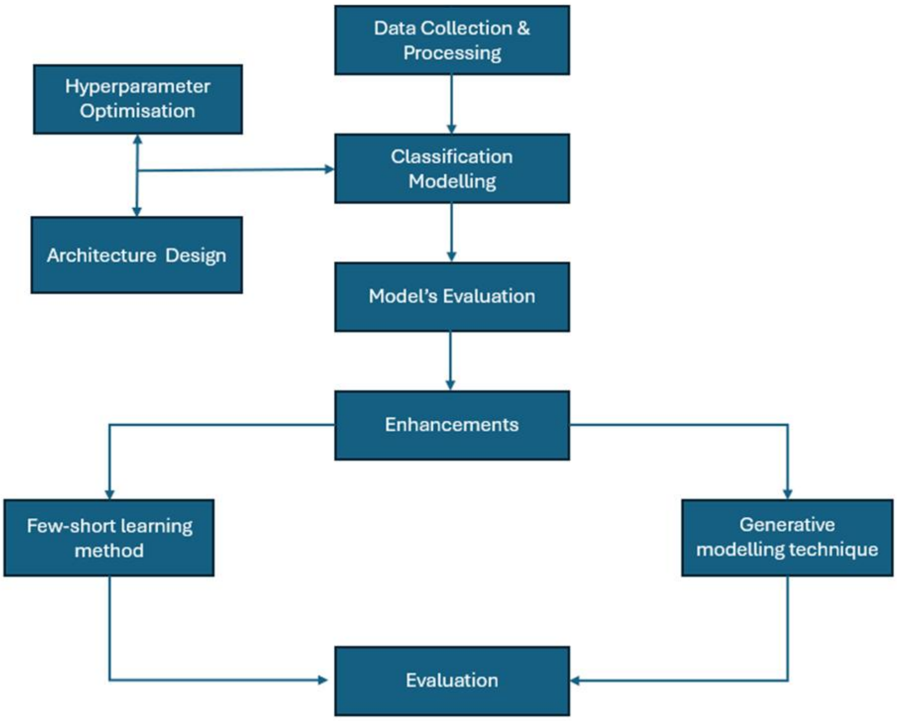

Data Collection & Processing
Hyperparameter Optimisation
Classification Modelling
Architecture Design
Model's Evaluation
Enhancements
Few-short learning method
Generative modelling technique
Evaluation

Figure 4. 1: Experimental Pipeline

The experimental pipeline will begin by collecting data from both public and private sources, carrying out data preprocessing, modelling, and evaluation, and improving performance by including techniques to deal with the limitations of small dataset sample sizes and visually similar characters.

4.2 Experimental Research Design

This thesis employed a quantitative method involving rigorous experimentations and investigation using various deep learning techniques, as supported by previous works on MICR. Quantitative research is chosen because it is more objective, focused, reliable, measurable, and suited to identify cause-and-effect relationships and produce results that can be replicated. The thesis provides justified explanations of the evaluation metrics, reproducibility, and validity of the depth of the proposed models' performance. This ensures that a third party can replicate the experiment to validate the findings.

4.3 Dataset Description

The medical image datasets used to test the proposed MICR models are open-source and originally collected (University Ethics Committee approved). These datasets were used for all the experiments in this thesis.

MEDPIX: Medpix medical image dataset is open source and contains 60,613 image collections of ultrasounds, X-rays, MRI and CT. I manually curated 3050 character image patches from the collection to form a character dataset. This means I carefully and manually selected character image patches using a simple image software (Microsoft Paint) instead of any specialised computer program, and this was to ensure the highest level of accuracy in data labelling, as automated character segmentation comes with many errors as supported by Sagar & Dixit (2019).

In addition, the dataset contains burned-in textual data representing various medical interpretations of the images. The dataset comprises 62 classes (A-Z, a-z, 0-9), averaging 50 samples per class with a dimension of (28,28,3). Checking the resolution of the datasets using the Python Image Pillow library gives a tuple of (96, 96), which is 96 dpi. The two values indicate dpi values across each image’s dimension, meaning each character image patch has 96 dots in 1 inch across the height and width dimension. The average sample size for each class was 50. However, character “R” had a class size of 145, and the lowest class size of 5 was for characters "v", "j", "q", and "y".

A tabular description of MEDPIX showing the frequency of each character is shown below in Table 4.3.

Table 4. 1: Character frequency of MEDPIX

CH.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
FQ.	116	123	101	60	60	67	40	43	51	58	129	48	89	82	161	42	54	78	73

CH.	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**	**a**	**b**
FQ.	17	16	120	85	82	80	92	**4**	145	111	140	37	44	38	9	29	16	47	9

CH.	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**p**	**q**	**r**	**s**	**t**	**u**
FQ.	12	28	64	5	20	9	33	5	7	19	35	29	25	18	5	28	35	31	11

CH.	**v**	**w**	**x**	**y**	**z**
FQ.	5	6	12	5	7

CH. = Character

FQ. = Frequency

PRIVATEDT: A private and original image dataset was collected after getting University ethics approval (Protocol number: SPECS/PGR/UH/05141). The data was collected on-site in Nigeria during the third year of the PhD. The image contains 3,000 image collections of ultrasound images from three medical laboratories with varied imaging acquisition techniques based on the approved guidelines. Appendix A-E provides information about the data collection requests and approval documents. Similarly, the dataset contains burned-in textual data representing various medical interpretations of the images of different internal human parts. This study manually curated 2076 character image patches from the collection to form a character dataset consisting of 62 classes (A-Z, a-z, 0-9), having an average of 34 samples per class with a dimension of (28,28,3). The resolution of the character images is 96 dpi.

A tabular description of **PRIVATEDT** showing the frequency of each character is shown below in Table 4.3.

Table 4. 2: Character frequency of PRIVATEDT

CH.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
FQ.	27	60	45	38	21	35	17	21	19	16	84	44	41	49	86	32	29	28	60

CH.	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**	**a**	**b**
FQ.	10	19	61	33	43	32	27	**5**	70	71	72	26	14	9	9	22	7	59	29

CH.	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**p**	**q**	**r**	**s**	**t**	**u**
FQ.	21	31	90	10	25	18	48	9	12	42	27	44	40	24	12	54	56	48	24

CH.	**v**	**w**	**x**	**y**	**z**
FQ.	13	12	8	30	8

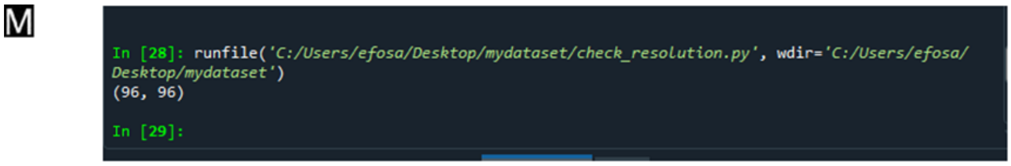
CH. = Character, FQ. = Frequency

For clarity, manual curation of the datasets involved manually segmenting words into characters using simple image software (Microsoft Paint). Data labelling was carried out by grouping each of the segmenting character images into 62 classes (A-Z, a-z, 0-9). The manual curation was usually in four steps to ensure accuracy:

- Segmentation of the burned-in on the MIMs text into lines
- Segmentation of the lines from the previous step into words,
- Segmentation of the words from the previous step in characters
- Grouping each character in a folder representing classes.

4.4 Resolution of Character Image Patches

The target resolution for the character recognition for this report is 96 dpi. The reason for the focus on the 96 DPI resolution is that the research carried out an extensive assessment of varied samples of MIM from various medical images online databases, including anonymised samples from a physical location where this PhD research collected original medical image dataset, after approval from the University Ethics committee. The images' resolution range was 96 DPI. Hence, this study was adapted to work with the 96 DPI. The resolution of the images collected was checked and confirmed using the PIL Python library. An image of the result using the PL Python library for random images from the dataset collected is shown in Figure 4.2 below:



```
In [28]: runfile('C:/Users/efosa/Desktop/mydataset/check_resolution.py', wdir='C:/Users/efosa/
Desktop/mydataset')
(96, 96)

In [29]:
```

Figure 4. 2: Resolution of a sample character “M” using the PIL Python library. “mydataset” here represents the MEDPIX dataset.

The PIL Python library image processing utility for Python is designed for fast access to data stored in basic image pixel formats. From Figure 4.2, the tuple (96, 96) indicates dpi values across each dimension (height and width), meaning the character “M” image has 96 dots per inch in both dimensions of the image’s height and width. The left side of Figure 4.2 shows the character patch, and the opposite shows the resulting resolution using the PIL Python library.

4.5 Challenges, Technical Factors and Ethical Considerations

This research utilised various technical platforms, including UHHPC, Google Computer Engine, and ML / DL library packages, to propose, implement, and validate solutions for the challenges in MICR identified in Chapter 3. This report will provide configurations of these usages as appropriate. However, some challenges were encountered, such as computational resources, hardware limitations, data quantity, and time constraints. This research addresses these challenges to develop effective and reliable DL models by implementing these strategies: online data augmentation to artificially increase data quantity, leveraging the Google cloud computing platform (Colab) to increase computational resources available, optimising algorithms and using advanced specialised hardware such as GPUs, TPUs, to reduce training time, and ensuring efficient data labelling to improve final model’s accuracy and generalisation.

This research will meticulously adhere to several ethical guidelines to safeguard the data collected according to the ethical approval by the university’s ethics committee, with protocol number SPECS/PGR/UH/05141. Data privacy and confidentiality will be strictly maintained.

4.6 Chapter Summary

This chapter provides an overview of the research methodology and experimental overview for this PhD thesis. It allows an understanding of the methods, steps, and techniques used for this research to allow reproducibility.

5.0 Ensemble Learning for Medical Image Character Recognition based on Enhanced Lenet-5

This chapter presents an enhanced CNN model for MICR and an ensemble classifier of CNN-based learners to enhance this new technique of recognising VSC. Intensive experiments are done using open-source and privately collected medical imaging datasets. Generally, MIM has a distinctive nature of low contrast, complex background, and low resolution, containing burned-in textual data of patients. The conventional OCRs hardly recognise these burned-in textual data under these conditions, as they are designed for mainly bi-level text with a minimum resolution of 300 dpi. With a focus on solving these challenges, this chapter proposes these models to aid a more accurate character recognition of burned-in textual data. The classical Lenet-5 architecture inspires this chapter. The problem of low resolution at 96dpi and background interference is targeted by using small 3 X 3 CNN filters to extract local features and changing the pooling layer to a learning layer by replacing it with 5 X 5 filters with a stride of 2 and training on a low-resolution character dataset. The final prediction is based on a majority voting algorithm. The consensus of the base learners improves the model's stability in recognising visually similar characters.

The work presented in this chapter was done during the second year of this PhD project (2021) and was published in an international conference proceedings in 2023 (Osagie et al., 2023). The content of this chapter has been adapted from Osagie et al., 2023, with some modifications and additional experimental results to suit the style better and ensure a logical presentation of the study.

5.1 Introduction

There has been a recent demand for the application and integration of artificial intelligence in medical imaging to understand their embedded patterns and use this extracted information to improve healthcare delivery and medical research. The imaging acquisition processes occurred under varying lighting conditions and distortions, resulting in low contrast with background interference. These MIM formats usually contain patients’ demographic and clinical examination data, but they exist as burned-in textual data. These specialised acquisition devices typically have low storage capacity. Hence, MIM has low resolution, resulting in burned-in textual data having a small font size. The need to recognise and extract this burned-in text for various post-identification purposes led to MICR research. However, these MIMs possess complicated features, such as commonly complex background interference and low resolution. These problematic conditions have resulted in poor performance accuracy when conventional optical character recognition (OCR) systems are applied. A common MIM showing these conditions included with the burned-in text magnified is shown in Figure 5.1 below.

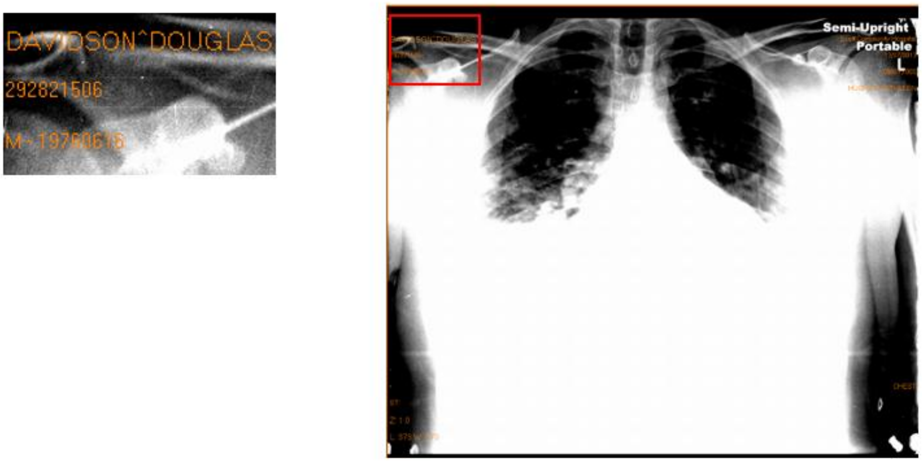


Figure 5. 1 : X-ray image (The Cancer Imaging Archive (TCIA) Public Access)

The low resolution makes the burned-in textual data appear in tiny font sizes, further increasing the complexity of using conventional OCR solutions to recognise them. Tesseract, Kraken, Calamari, Ocropy and Abby Reader are the most widely used OCRs (Drobac and Lindén, 2020) and can only recognise textual data on printed and scanned document images (Ramdhani et al., 2021). These conventional OCR solutions usually operate in two steps. (a) Divide the input image and determine the region of interest with the textual data and (b) Segment the character and do the recognition individually. However, these steps are inefficient in MIM, where the text is unstructured (Text may not appear in a straight horizontal line) (Istephan and Siadat, 2016). As a result, the background may overlap the text, with a resolution much lower than what these Conventional OCRs were designed for. Conventional OCR solutions require a minimum resolution of 300 dpi for good accuracy (Oni and Asahiah, 2020), while MIM are 150 dpi – 72 dpi. To solve this problem, earlier proposals used varying image pre-processing techniques to enhance these MIMs and feed them to these conventional OCRs. After that, ML algorithms (such as Random Forest, AdaBoost and Boltzmann Restricted Machine classifiers) were used to design specialised classifiers, but the performance was limited by the inability of these algorithms to learn optimally in the presence of noise. Recently, DL models, especially CNN-based models, have greatly succeeded in image classification tasks, as the convolutional layer can extract local features from input training samples using linear and nonlinear operations. A CNN variant explicitly designed for handwritten and machine-printed characters on document images, which achieved high success on the MNIST handwritten character dataset, is known as Lenet-5 (Lecun et al., 1998). The accuracy of CNN classification has been recorded as high compared to MLP and probabilistic networks (Wei et al., 2019). The Lenet-5 architecture consists of 5 learnable layers, with three sets of convolutional layers and average pooling layers, followed by two dense layers and a SoftMax classifier at the posterior (Lecun et al., 1998). The Lenet-5 uses an efficient combination of convolutional layers to extract essential features from input training samples while reducing training time through its simple yet efficient architecture. The Lenet-5 uses the gradient descent method for the global convergence of the algorithm (Zhang et al., 2019).

Still, within the aspect of MICR, there is an associated challenge in recognising VSC (such as “0” and “O”) in low-resolution MIM due to the poor quality, resolution, and dimension adjustment of textual data in the complex backgrounds of MIM (Pal et al., 2021). The low resolution often results in distortion in the shape of these characters, making it difficult for even a trained classifier to recognise the target class correctly. As a result, even a well-trained classifier may misclassify these VSC, reducing the model’s confidence. This chapter proposes a consensus of enhanced models trained on different subsets of the datasets, where each model represents learned significant discriminative features of characters from the training samples. Due to background interference, shade gradients, overlapping text and low resolution, developing a large all-inclusive dataset of characters in this domain is quite challenging. Recent OCR engines may have auto-correct functionalities based on language dictionaries. Still, in the MIM domain, it is difficult to have such a dictionary that contains all alphanumeric medical text and labels. Hence, there is a need to recognise these individual characters accurately and independently. It is relevant to employ ensemble techniques to improve the character recognition accuracy of classifiers by leveraging the advantages of a majority voting algorithm.

CNN has recently been applied to recognise these burned-in textual data on MIM. Still, these solutions are limited in low-resolution MIM of 96dpi and need further enhancement to distinguish visually similar characters. The primary focus of this chapter is to propose an enhanced CNN model inspired by the Lenet-5 architecture to recognise these burned-in textual data in MIM at a character level. A majority voting algorithm is employed to improve the recognition of visually similar characters. A comparison of the enhanced CNN models and the state-of-the-art will be made to show the improvement achieved.

This chapter investigates the applicability of specially designed and enhanced deep learning models to the recognition of burned-in textual data in MIM under the constraints of low resolution, background interference and tiny text.

The remainder of this chapter is organised in the following section. Section 5.2 reviews the related works in MICR. Section 5.3 provides the specific contributions of this Chapter. Section 5.4 discusses the proposed CNN-based ensemble model inspired by the Lenet-5 and includes justifications for the modifications done. Section 5.5 describes the experimental setup. Section 5.6 presents the results and evaluation. Finally, section 5.7 presents the conclusion of this chapter.

5.2 Related Works

Past works have proposed different solutions to recognise burned-in textual data on MIM by leveraging the pattern recognition ability of both ML and DL techniques, with recent methods being CNN-based. These works attempted to solve the challenge under the problematic conditions explained in the introduction by combining different image pre-processing techniques and ML or DL models. One early approach proposed for MICR was presented by authors in (Florea et al., 2005) using prior knowledge of the intended character, applied morphological transformations (TopHat filter) to thicken the edges, and finally fed to ABBYY FineReader opensource OCR. Still, the approach could not identify text in the angiography category and other textual annotations in varied MIM with a recognition rate of 58.8% and recall of 60.0%. Wang (2002) applied a wavelet-based medical image-filtering algorithm to recognise burned-in text containing areas into lines and passed into an OCR engine. The solution depended on the images' quality or sharpness and only recognised characters on the corners of grayscale medical images. A similar approach using open-source OCR and zoom factor extraction technique by (Alter and Werner, 2007) performed poorly in recognising burned-in textual data overlapping on the complex background due to high background interference. The zoom factor extraction largely depended on the quality of the images. Hence low-resolution images reduced the performance of this approach. Some authors saw the need for a pre-determined dictionary and a user-assisted revision stage. The user-assisted revision reduces errors based on a specified lexicon but cannot be automated (Tsui, and Chan, 2012). Tsui, and Chan, (2012) included a weighted similarity in combination with the user-assisted revision. Vcelak et al. (2019) applied binarisation with Tesseract OCR on ultrasound images. These proposed methods all suffered similar unreliable results, especially in low-resolution MIM containing overlapping textual data with background interference. Even though the various image pre-processing methods reduced the background noise, the OCRs were explicitly designed for printed and scanned document text. Due to the inadequacies of the conventional OCRs, more specialised solutions were designed.

Yu and Yuanyuan (2015) applied local feature extraction and Adaboost to recognise burned-in textual data. However, unreliable results were seen in low-resolution MIM with poor contrast and lightning. The background noise affected the learning ability of

Adaboost (Yu and Yuanyuan, 2015). Monteiro et al. (2015) used a random forest classifier and restricted Boltzmann machine. However, they could not recognise varied font styles and small font sizes on low-resolution MIM. The limitation in the random forest classifier in (Monteiro et al , 2015) is due to the model's poor performance in dealing with higher-order convolutional structures (images), as it is more accurate in learning features from tabular data. Recent authors proposed a CNN-based recognition model for burned-in textual data on MIM (Monteiro et al , 2017). The CNN model (Monteiro et al., 2017) proved better than previous ML algorithms, with an accuracy of 87.5%. The design of (Monteiro et al., 2017) was a shallow network with two convolution layers, two max-pooling layers and two dense layers. It was limited by its representational capacity to learn complex features and poor ability to learn spatial representations, which are essential to understanding the spatial relationships between different parts of the image (Monteiro et al., 2017). Monteiro et al. (2017) could not generalise the solution to varied MIM with different font styles and small sizes. They trained using a non-medical image character dataset and evaluated only on ultrasound imaging. They suggested that the background interference in the low-resolution MIM reduce the model's accuracy and reliability. Silva et al. (2018) used the same CNN model as Monteiro et al. (2017) and included complex user-assisted revision stages. The system had problems finding patterns for similar characters in dark backgrounds and relied on too many complex processes, such as multiple software integration (Silva et al., 2018). More recently, Xu et al. (2021) proposed a modified Convolutional recurrent neural network (CRNN) with a multiscale architecture learning scale variant feature. The result was a recall of 65.0%, a precision of 67% and an F-measure of 70%. Their proposed model (Xu et al., 2021) was poorly learned due to the large network width, the small dataset of 1500 images used and the background. The model could not recognise burned-in text reliably on varied MIM with low resolution and hence could not be generalised (Xu et al., 2021). These past works (Yu et al., 2015; Monteiro et al., 2015; Monteiro et al., 2017; Xu et al., 2021) concluded that their models were further limited by the challenge of recognising VSC such as “U” and “V”. Therefore, in recognising characters in MIM, consideration has also to be given to the VSC to improve the model’s confidence.

In the general OCR domain, several studies (Caruana, 1997; Hou et al., 2017; Chen et al., 2017) used ensemble learning to improve the recognition of handwritten characters while considering visually similar characters. These authors (Caruana, 1997; Hou et al., 2017; Chen et al., 2017) did not explicitly specify the resolution of their work but agreed on the problem of background interference. However, not much work has been done to recognise burned-in textual data in MIM using ensemble enhancement. Ensemble learning is an intensive pattern recognition technique that combines base models to improve the final model's generalisation ability. The MICR task can be enhanced with a higher performing accuracy by combining a group of base classifiers as an ensemble. This consensus prediction is advantageous, especially in recognising visually similar characters. Creating multiple classifiers and manipulating the training data in an organised or random way, as well as changing the hyper-parameters, will give rise to different hypotheses by each classifier as they converge individually on a different space. Combining these classification rules learned from different convergence and applying a majority voting method, this study achieved diversity in each CNN member by training on different subsets of the data while carrying out online augmentation. The Lenet-5 is recognised as a pioneer model from which other advanced models were developed (Emmert-Streib et al., 2020). A notable improvement of the Lenet-5 is the AlexNet, which won the ImageNet Large

Scale Visual Recognition Challenge (ILSVRC) in 2012 with a top-5 error rate of 15.3% (Emmert-Streib et al., 2020). Most recent studies using CNN for MICR have designed only a single classifier (Monteiro et al., 2017; Silva et al., 2018; Xu et al., 2021). However, a single CNN classifier may show poor accuracy due to a limited set of possible approximations the model can create for a target function and its representational capacity or have been stuck on a local minimum due to a stalled weight update. Furthermore, the single outcome cannot be appropriately aligned with the desired outcome when considering the difficulty in recognising characters. These limitations motivated this chapter's work to propose an enhanced CNN model using Bayesian reasoning for the MICR task. A majority ensemble is employed to tackle the problem of distinguishing VSC using a consensus algorithm. Based on an extensive search of notable article databases for the last 10 years, no past works have employed the optimisation of the Lenet-5 architecture and implemented the advantage of ensemble learning to recognise burned-in textual data on MIM. This chapter aims to investigate and contribute to this aspect to tackle these challenges in burned-in text recognition in MIM.

5.3 Contributions

The main contributions of this chapter are :

- This study proposes an enhanced CNN model motivated by the classical Lenet-5 model. The enhanced model is optimised using Bayesian reasoning. The Lenet-5 uses a filter size of 5x5 in its first convolutional layer, followed by average pooling. In this study, these are replaced by a 3x3 filter size and a 5x5 filter size with a stride of 2, respectively. This enhancement ensures that the proposed CNN model can learn more local features, which are essential in designing a MICR solution for low-resolution MIM with background interference.

- Performing MICR on burned-in textual data at a low resolution of 96 dpi with background interference. The proposed models are evaluated using open-source and privately collected datasets. An outstanding accuracy score was achieved, and MICR at such low resolution has not been previously reported in the literature.
- A majority voting ensemble algorithm is proposed to enhance the model's performance. The research uses the bootstrapping method to create 3 subsets of character datasets. A classifier is fitted to each of these subsets and evaluated. An ensemble is designed using the trained classifiers of the training subsets, and a final classification outcome is based on a majority voting algorithm. This improves the model's performance in distinguishing VSC.

5.4 Proposed Model

The proposed model is an enhancement of the classic Lenet-5 model suited for the task of MICR. The model will be used to form an ensemble model based on a majority voting algorithm. The enhancement is done by optimising the base model using a combination of optimisation techniques presented in the network hyperparameter optimisation sub-section. The following sub-sections discuss the network design and optimisation techniques.

5.4.1 Network Hyperparameter Optimisation

This study used the Bayesian Optimisation (BO) algorithm to decide the optimal architecture of the hyper-parameters and efficiently modify the base model for the task. The BO is a sequential design strategy for the global optimisation of objective functions that may be expensive to evaluate (Zhang et al., 2021), such as the hyperparameters in neural networks. It can efficiently reduce the computational cost of fine-tuning hyperparameters compared to brute-force methods (Gridsearch and Randomsearch) by reducing the number of search iterations by choosing the input values based on the past outcome of a previous configuration. The BO uses the informed learning method based on the Gaussian process by using a surrogate function to model the black box function and then uses an acquisition function to find the next point of evaluation. The goal is to get very close to the optimum values with very few iterations of the black box functions. BO can fit the observed values of the black-box function and interpolate between observed data points, with increasing statistical uncertainty the farther you move away from the observed data. These properties are essential for this study, as I know the function values taken from the Lenet-5 as the base model, but I am not certain of the impact of increasing or decreasing these functional values. BO can achieve the global minima with the smallest loss function value (Gao et al., 2019). Compared with the popular Genetic Algorithm (GA), the GA must move from one generation to the next, so it trains the same configuration on multiple hyperparameters. In contrast, BO can train a single configuration and update the posterior information based on learned history, hence reducing computational costs. However, the BO has some instability limitations, particularly in dealing with a large hyperparameter search space because of the curse of dimensionality (Eriksson and Jankowiak, 2021). Several recent empirical studies (Moriconi et al., 2019; Frazier, 2018; Awal et al., 2021) have shown that BO is practically limited to optimising less than 20 parameters. Although the parameters are less than 20 in this study, I desired a reduction of the iterations needed for BO. This study combined a search space pruning mechanism known as the Successive Halving Algorithm (SHA) to reduce the computational cost for BO iterations. SHA is an advanced early-stopping method that determines the most useful hyperparameter search values that may lead to good results by allocating minimum resources (such as the number of epochs) to each configuration and terminating unpromising trials by monitoring each trial learning curve. Basically, SHA determines the useful search space with very soon promising configurations and the BO uses its reasoning properties to find the optimal configuration. The SHA can be run in parallel and simultaneously with BO to reduce the search space, overcoming a major shortcoming of BO. In this study, I focus only on optimising the base model for the MICR by leveraging a combination of the techniques of SHA and BO to determine the optimal hyperparameters. Hence, no detailed derivation of the optimisation algorithms will be provided. SHA determines how many configurations to evaluate with which budget, but the BO replaces the default random search. Once the desired number of configurations is reached, the SHA reduces the number of configurations using a reduction factor. The SHA-BO combination is implemented using the Optuna Python library, which allows input of various parameters that can affect the optimisation and create trials known as a study (Akiba et al., 2019). In the Optuna library, I used the Gaussian process-based algorithm for the BO. The Gaussian process-based algorithm can build a model by applying Bayesian reasoning to balance the exploration versus exploitation trade-off. Several recent studies (Watanabe and Hutter, 2022; Bergstra et al., 2011; Rong et al., 2021; Ozaki et al., 2020) have agreed that it is a notable BO estimator to optimise hyperparameters to ensure the strong performance of DL models. I can pass a function for the optimisation, specify the number of iterations, and visualise the importance of the hyperparameters. As a first study, I ran hyperparameter tuning for about 100 trials and then checked which hyperparameters were the most important. Next, I omitted the less important hyperparameters for the subsequent studies up to 1500 trials. The flow chart of the optimisation process is shown below in Figure 5.2.

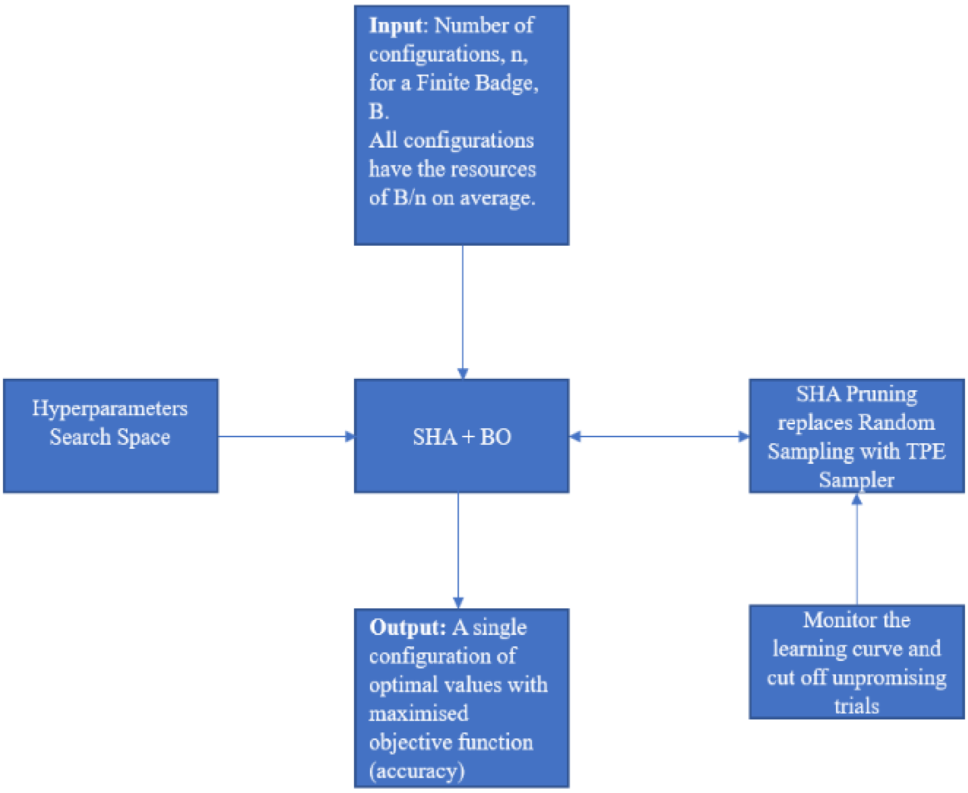

Input: Number of configurations, n, for a Finite Badge, B.
All configurations have the resources of B/n on average.
Hyperparameters Search Space
SHA + BO
SHA Pruning replaces Random Sampling with TPE Sampler
Output: A single configuration of optimal values with maximised objective function (accuracy)
Monitor the learning curve and cut off unpromising trials

Figure 5. 2: Flow chart of the BO hyperparameter optimisation process

The hyperparameters setting was carefully selected after careful observations of notable models, datasets and key values affecting the optimised objective function. This ensured that no computational cost was spent on running iterations on already known, likely not promising settings. Hyperparameters search space included activation, learning rate, optimisation, kernel size, strides, number of convolutional layers, number of filters, number of layers, number of dense units and drop-out rate.

For clarification, the configuration of the hyperparameter optimisation process is provided below:

- The Optuna library[9] with the GPSampler is used for the hyperparameter optimisation process.
- Objective (a trial) was defined according to this search space; filters (32, 64, 128), kernel size (3,5,7), strides (1,2), activation (ReLU, sigmoid, Leaky ReLU), dropout rate (0.2, 0.3, 0.4, 0.5), Con2D (1,2,4,5,6), learning_rate (1e-5, 1e-1) and Dense (64, 128, 256, 512).
- Pruner was set to ‘successivehalvingpruner’, and the direction of the trial was set to ‘maximise’ accuracy.
- The number of trials was initially set to 100, and the trials ran 30 times to establish the importance of hyperparameters. This helped to refine the search space, since the most important hyperparameters are known after this and are being focused on.
- Next, the number of trials was set to 1500, which ran 30 times to determine the best hyperparameters for the proposed CNN architecture.The results from all the runs were similar; hence, only one result was taken.
- On each trial, this sampler fits a Gaussian process (GP) to the objective function and optimises the acquisition function to suggest the next parameters.
- The GP configuration used was Matern kernel with nu=2.5, Automatic relevance determination (ARD) for the length scale of each parameter, Log Expected Improvement (logEI) as the acquisition function, and Quasi-Monte Carlo (QMC) sampling to optimise the acquisition function. These other nu

---

[9] https://optuna.readthedocs.io/en/stable/

values for the Matern kernel were all evaluated [0.5, 1.5, 2.5] and compared based on faster convergence for hyperparameter searches. Other values above 2.5 incurred more expensive computational costs during the optimisation process and reached up to 10 times more resources when running trials. Hence, they were not used. The nu value of 2.5 was kept constant through the optimisation process after prior evaluation of the other values.

- After the first 100 trials, the optional has a module 'importance' that provides functionality to evaluate hyperparameter importance from completed trials.
- After evaluating the importance of the hyperparameters from the initial completed trials of 100, the search space was adjusted.
- The choice of the matern kernel was based on its performance, as seen in past works (Gao et al., 2017) regarding its robustness in predicting uncertainties in hyperparameter optimisation (Wood et al., 2022).

5.4.2 Designed Model

The detailed layerwise summary of the MICR model is shown in Table 5.1. The enhanced MICR model consists of multiple convolutional layers (Conv2D) and dense layers at the end. A 2D convolution is done in each convolutional layer, followed by Relu activation. I applied a 3x3 filter initially to learn most local features across all channels while keeping padding at zero. This is followed closely by the 5x5 filter across the Conv2D. The 5 X 5 Conv2D with a stride of 2 replaces the pooling layer on Lenet-5 to allow more representation learning of local features while downsampling the image simultaneously. This was discovered after running over 300 iterations during the model optimisation step. Recent studies (Springenberg et al., 2014; Muresan &

Oltean, 2017) agreed that this replacement improves the model's expressiveness ability. I applied 128 neurons for the dense layers. In addition, the visual representation of the model is shown in Fig 5.3. to show the network architecture. The enhanced CNN model is a relatively simple yet efficient model for the desired task of recognising burned-in textual data on MIM. Experimental results showed that accuracy reduced drastically as the network became deeper. This was due to the problem of information loss, vanishing gradient, and the small dataset. It is agreed that the deeper the architecture, the more information loss can occur during the downsampling process, as the dataset has a small amount of data (Tomasini et al., 2022; Alzubaidi et al., 2021). Dropout was added to avoid over-fitting (Goodfellow et al., 2016), which is important when dealing with a small dataset. The dense layer converted the 2-dimensional feature maps into 1-D vectors. All neurons are fully connected to the neurons in the adjacent and subsequent layers. The output layer uses a Softmax function to predict the final classification outcome.

Compared with the past works (Monteiro et al., 2017; Silva et al., 2018; Xu et al., 2021), the design solved the poor learning ability due to the large network width and information loss by using an optimal configuration of 3x3 filters which is able to reduce information loss. The architecture design solved the limitations of Monteiro et al. (2017) to recognise certain font sizes and styles by replacing pooling layers with learnable downsampling layers, which is targeted at the problem of recognising characters in low-resolution MIM as key features are small and local. This approach is inspired by Springenberg et al. (2014), and it increases the model’s expressiveness ability.

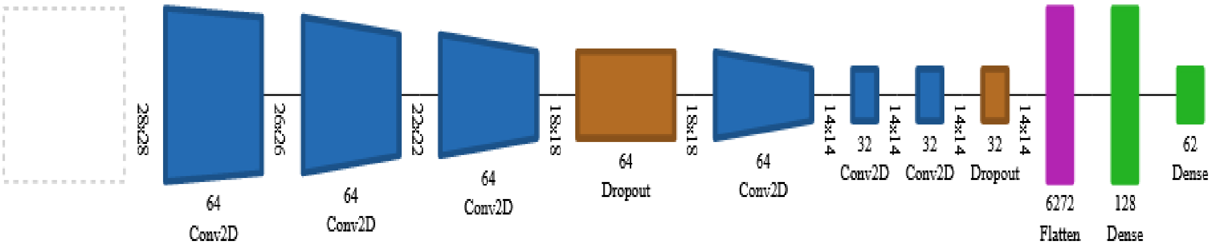


Figure 5. 3: Proposed Enhanced MICR model

Table 5. 1: Layerswise Summary of the proposed CNN model.

Layer (Type)	**Output Shape**	**Learnable Parameters**	**Filter**	**Stride**
conv2d_22 (Conv2D)	26, 26, 64	1792	3x3	
conv2d_23 (Conv2D)	11, 11, 64	102464	5x5	2
conv2d_24 (Conv2D)	7, 7, 64	102464	5x5	
dropout_5 (Dropout) (0.4)	7, 7, 64	0		
conv2d_25 (Conv2D)	2, 2, 64	102464	5x5	2
conv2d_26 (Conv2D)	2, 2, 32	18464	3x3	
conv2d_27 (Conv2D)	2, 2, 32	25632	5x5	
dropout_6 (Dropout) (0.4)	2, 2, 32	0		
flatten_1 (Flatten)	128	0		
dense_2 (Dense)	128	16512		
dense_3 (Dense)	62	7998		
Trainable parameters: 377,790				

The justification for the modification achieved on the proposed MICR model is further explained in Table 5.2 in the following sub-section.

5.4.3 Modifications on Lenet-5

Table 5. 2: Modifications carried out on the Lenet-5 base model

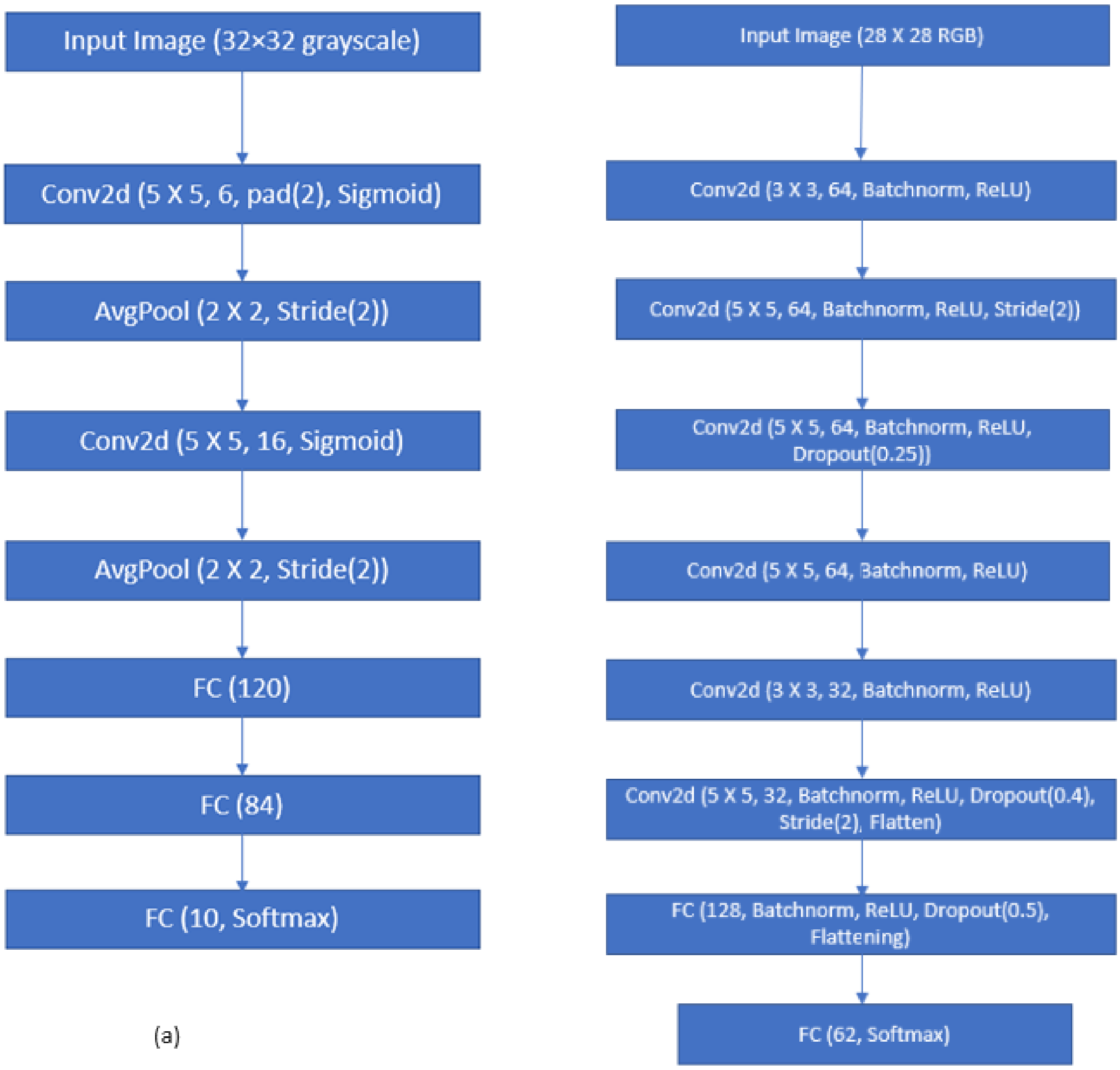
Lenet-5	**EfeNet22**	**Justification**
Input layer size of 32 X 32 and grayscale	Input layer size of 28 X 28 and RGB	The input images in this research are 28 X 28 and RGB, hence the modification of the input layer's size.
Two Conv2D layer with 5x5 filter	Two stacked Conv2D layers with 3X3 filters to replace each of the default 5 X 5	The choice of filters, especially in replacing the 5X5 filter with two stacked 3X3 filters, is drawn from an understanding of the automatic feature extraction ability of CNN at each layer, determined by the kernel size. Most of the useful features in an image are local, and to effectively learn these features, it is better to apply small convolutions to take a few pixels at a time. Therefore, choosing a 3X3 to replace the original 5X5 filter reduces the computational requirements of learning features

		and the weight sharing when dealing with these noisy medical images. Also, the 3X3 ensures simplicity in implementation and extracting features from localised pixels of interest and their neighbours from all sides; this can efficiently learn useful local features such as vertical edges, which is essential in this case. From the notable inception model done by Szegedy et al (2014), the authors agreed with the fact that replacing the 5X5 filter in the Lenet-5 with the two stacked 3X3 filters results in a (9 + 9)/25x reduction in computation requirement because using two filters means 2(3*3 + 3*3) individual weights compared to (5*5) individual weights in a 5X5 single filter; hence fewer parameters reduce the computational resources needed.
Sub-sampling layers	Conv2D layer with stride 2	This research replaced the subsampling layers in the Lenet-5 with learnable convolutional layers with strides of 2 because pooling is a fixed operation while convolution can be learned,

		even though it is a more expensive operation in terms of computational requirements, this is required as the dataset contain some amount of noise. Therefore it is appropriate to learn as much representations as possible. The learnable layers also increase the model's expressiveness ability, generalisation, and overall accuracy. A notable study by Springenberg et al. (2015) demonstrates that this action improves the model's overall accuracy with the same depth and width, leading to increased model stability. Also, other widely referenced studies on image classification tasks implemented learnable Conv2D layers instead of Maxpooling or other sub-sampling methods (Mureşan and Oltean, 2017). ResNet, a popular CNN, has also agreed and embraced this finding by using convolutions with strides rather than sub-sampling to reduce spatial dimensions in between residual modules and result in higher accuracy.

No batch normalisation	Batch normalisation is added	Batch normalisation improves performance, speed, and stability during the training of DL models (Ioffe et al., 2015). It helps reach convergence faster (Bjorck et al., 2018) and makes the optimisation landscape significantly smoother (Santurkar et al., 2018)
The Sigmoid activation function	ReLU activation function	ReLU is simple, fast and solves the problem of vanishing gradient due to slow convergence by having a derivative of 0 or 1 during weight multiplication, compared to sigmoid having a derivative between 0-1; this was reported from multiple experiments by Nwankpa et al., 2015 and Szandała, 2021.
No Drop-out regularisation	Drop-out regularisation is added	Drop-out was added to avoid over-fitting (Srivastava et al., 2014), which is important when dealing with a small dataset.
7-Layers with three fully connected layers.	8-Layers with two fully connected layers.	Reduction of the fully connected layers from three to two, to reduce network complexity and training time (Ma et al, 2018).

5.4.4 Visual description of the models' architecture


Input Image (32×32 grayscale)
Conv2d (5 X 5, 6, pad(2), Sigmoid)
AvgPool (2 X 2, Stride(2))
Conv2d (5 X 5, 16, Sigmoid)
AvgPool (2 X 2, Stride(2))
FC (120)
FC (84)
FC (10, Softmax)
Input Image (28 X 28 RGB)
Conv2d (3 X 3, 64, Batchnorm, ReLU)
Conv2d (5 X 5, 64, Batchnorm, ReLU, Stride(2))
Conv2d (5 X 5, 64, Batchnorm, ReLU, Dropout(0.25))
Conv2d (5 X 5, 64, Batchnorm, ReLU)
Conv2d (3 X 3, 32, Batchnorm, ReLU)
Conv2d (5 X 5, 32, Batchnorm, ReLU, Dropout(0.4), Stride(2), Flatten)
FC (128, Batchnorm, ReLU, Dropout(0.5), Flattening)
FC (62, Softmax)

(a)

(b)

Figure 5. 4: Visual description of model's enhancement (a) Lenet-5 architecture: Input is grayscale handwritten digits, output is 10 possible outcomes, FC = Fully connected layers, AvgPool = Average Pooling layer. (b) Proposed CNN Member architecture.

The Lenet-5 is the basis of numerous models due to its high performance in OCR applications and foundations for notable architectures like AlexNet and VGG. The Lenet-5 is made up of 7 layers, which are 3 convolutional layers, 2 subsampling layers and 2 fully connected layers; the input layer is not included in the total number of layers, as no learning occurs at this layer but only takes in 32 X 32 images which are passed to the next layer. In the original paper by Lecun et al. (1998), the Lenet-5 experimented on grayscale images with normalised pixel values of -0.1 to 1.175 to ensure the batch of images had a mean of 0 and a standard deviation of 1; this resulted in a reduction in the overall training time. As seen in Figure 5.4(a) above, the Lenet-5 is built on two significant layers: the subsampling and the convolutional layer.

The first Convolutional layer produces as output 6 feature maps with a kernel size of 5 X 5, a sigmoid activation function, and the dimensions of the 6 feature maps are 28 X 28. The second convolutional layer also uses a 5 X 5 kernel, a sigmoid activation function, and outputs 16 feature maps. Each 2 X 2 subsampling layer reduces the dimension by a factor of 4 via spatial downsampling and outputs the corresponding feature maps received by the previous layer. The inputs are flattened after the last subsampling layer, from a 4-dimensional input into the 2-dimensional input expected by the three fully connected layers with 120, 84 and 10 outputs accordingly, where the 10 outputs correspond to the possible classes for the image classification task. On the other hand, as seen in Figure 5.4(b), this thesis proposes an enhancement of the Lenet-5 architecture by modification suited for the task at hand: the recognition of the burned-in text data on MIM with small font size and low resolution. Table 5.2 shows detailed enhancements and justification accordingly.

5.4.5 Majority Voting Algorithm

The majority voting ensemble algorithm used to improve the recognition accuracy of visually similar characters consists of two steps: (a) train the enhanced model on three (3) subsets of the training sample based on the bootstrapping method and (b) combine each prediction of the ensemble members, to get a consensus classification outcome. The bootstrap method involves iteratively randomly resampling the dataset with replacement and determining the expected size of the subsets and the number of subsets required. The experimental results show that the ensemble is better in accuracy because different models will usually not make the same error across the testing set (Goodfellow et al., 2016). There are different ways to vary the members of the ensemble. They include (a) choice of data, (b) choice of models' architecture, and (c) choice of outcome consensus technique. In this study, I use the varying data approach by splitting it into three subsets and estimating the generalisation error of the enhanced MICR model configuration. The resulting three models are represented by $MICR_7$, $MICR_8$ and $MICR_9$, with the subscript stating the percentage of the training samples subset used for the model. This approach was supported by the statistical studies by Gareth et al. (2013), who said having access to multiple training sets is not always practical. Instead, bootstrapping can be done by taking repeated samples from the training sets. This reduces the variance of each member of the ensemble (Gareth et al., 2013). The bootstrapping method used is the replication method, and the number of repeats was 500. The number of repeats is determined iteratively to allow significant variability of the fitted models trained on each bootstrap subset (Walters & Campbell, 2004). Three bootstrap subsets were created from 70%, 80%, and 90% splits of the original dataset, while the remaining 30%, 20% and 10% were kept for testing each model's performance on the respective subset. Each bootstrap subset was fitted to a given model, and results averaged for 30 runs. This means that due to the replication method of sampling, some data samples from the original dataset will not appear in the bootstrap subset, and some will be repeated.

Figure 5.5 below shows the framework of the majority voting algorithm used in this study.

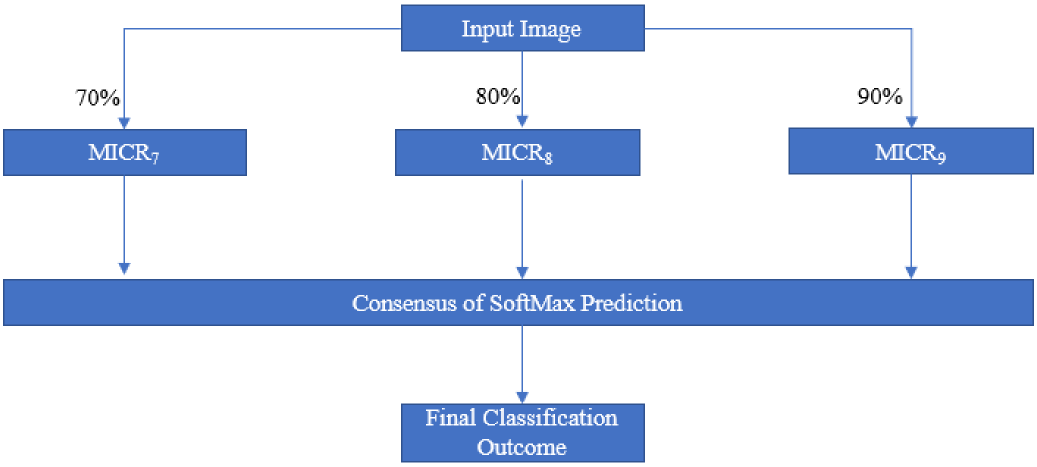


Figure 5. 5: Majority voting ensemble approach used

As mentioned earlier, the subscript on the MICR model in Figure 5.5 above represents the percentage of the training subset. The split of 70%, 80%, and 90% was chosen to use a significant amount of the dataset for training so that the model can learn effectively while keeping a sufficient amount for testing. This ensures that the base learners’ performance can be assessed reliably on test data.

The bootstrapping used in the majority voting ensemble reduces the final prediction model's variance, reduces overfitting, and balances the bias-variance trade-off. After creating three subsets from the training dataset, each classifier is fitted to a classifier and trained using data augmentation techniques. The softmax prediction shown in the ensemble set up above provides a distinct probability distribution for each character class, enabling a more confident measure of the model's prediction, and it is faster to compute when compared to exponential functions. Combining these softmax predictions in an odd-numbered format will improve the overall model performance by reducing the bias of a particular base model to a particular character class. This has been used in similar instances for classification problems in X-ray images (Chandra et al., 2021), breast cancer images (Naji et al., 2021), and handwritten text recognition (Hamida et al., 2023), amongst others.

5.5 Experimental Set-Up

5.5.1 Data Preparation and Training Strategy

The datasets were split into different subsets to create diversity in the models for the ensemble, as previously explained in the majority voting sub-section. Online data augmentation was used to improve the model’s generalisation by varying the data and minimising data overfitting (Shorten, and Khoshgoftaar, 2019). The SHA and BO combination, as explained in Network Design Section 5.4.1, was used to optimise the hyperparameters of the proposed model. For the random translation, the image is randomly shifted horizontally and vertically up to 10% of its size. For the random rotation, each image was rotated up to 20 degrees, either clockwise or anticlockwise. The random translation and random rotation increase the diversity of the training set, thereby improving the generalisation of the final model, and they were implemented using the Keras library's ImageDataGenerator preprocessing layers module. All the original images were transformed (rotation and translation) during every epoch and then used to train the model. Therefore, the total number of data samples per class did not change and remains equal to the number of original images per class. Using the online data augmentation based on the Keras library does not mean increasing the total number of totally distinct training data samples; it simply creates different variations of existing training data samples used for the model's training. This improves the robustness of the final model. For example, for 10 epochs, it simply means 10 variations of the image from each class have been used instead of just using the same single original image in the whole training.

Checkpoints were initialised during the training to determine the best epoch for the training. The validation accuracy (testing accuracy) was monitored here, and only the best weights were saved. The Adam optimiser was used to ensure faster convergence. A batch size 28 was used during the training, determined after several training iterations. To control the training and test ratio of each class, no distinct undersampling or oversampling was done, but a careful sampling of the dataset and manually checking each folder were done to ensure each class was represented in significant proportion in the test and train subsets. The main disadvantage of undersampling is that potentially useful data samples' critical information will be removed (Gnip et al., 2021), while oversampling may lead to limited information gain and overfitting of the model since we are making replicated copies of the data samples (Hassanat et al., 2022).

5.6 Results and Evaluation

The experimental environment's configuration, including the optimisation, is as follows: Python 3 Google Compute Engine backend (GPU) of 83 GB RAM A100 GPU.

The experimental results, including the time taken to train the proposed MICR model and the Lenet-5, are shown in Table 5.5 below. All results presented are from experiments carried out on MEDPIX and PIRVATEDT, and the improvement is shown.

The proposed model is simpler with fewer parameters, yet more efficient compared to the Lenet-5 model in recognising burned-in textual data on MIM. The proposed ensemble is represented as MICR (n). n is an odd number, as each ensemble member is entitled to a single vote. The proposed MICR model is compared with Lenet-5 on the bootstrapped subsets of the training samples. I will represent the trained models for the ensemble as explained in the majority voting sub-section 5.4.5. The results are shown below, averaged on 30 runs at 100 epochs with checkpoints to save the best weights. The train-test split ratio is 8:2 for MICR (1) and Lenet-5. At the same time, random data reshuffling, bootstrapping subsets, and augmentation are performed on the ensemble members. After an extensive literature review on recognising burnt-in textual data on MIM, the experiments aim to support the hypothesis that the enhanced MICR model is more accurate than the base OCR model (Lenet-5) and other existing algorithms. Furthermore, it shows that a majority voting ensemble algorithm can have improved results than a single model.

5.6.1 Models’ Results and Optimal Number (n) of Ensemble Members

Table 5. 3: Results for different ensemble members' configuration

Model	**Accuracy (%)**	**Time Taken in Seconds (TTS)**	**Precision (%)**	**Recall (%)**	**F1-measure (%)**
MEDPIX					
Lenet-5	70.14 ±0.04	445.92 ±0.05	68.07 ±0.09	69.30 ±0.03	67.00 ±0.09
MICR(1)	91.54 ±0.06	243.75 ±0.07	92.62 ±0.02	92.52 ±0.05	92.12 ±0.06
MICR(3)	94.05 ±0.08	781.25 ±0.06	94.12 ±0.08	94.78 ±0.05	93.35 ±0.04

$MICR_7$ +$MICR_8$ + $MICR_9$					
MICR(5) **$MICR_7$ +$MICR_8$ + $MICR_9$ + $MICR_7$ +$MICR_8$**	93.78 ±0.10	1303.75 ±0.08	93.70 ±0.09	93.94 ±0.08	93.55 ±0.11
MICR(7) **$MICR_7$ +$MICR_8$ + $MICR_9$ + $MICR_7$ +$MICR_8$+ $MICR_9$ + $MICR_7$**	93.71 ±0.14	1853.25 ±0.04	93.36 ±0.17	93.66 ±0.09	93.12 ±0.11
MICR(9) **$MICR_7$ +$MICR_8$ + $MICR_9$ + $MICR_7$ +$MICR_8$+ $MICR_9$ + $MICR_7$ +$MICR_8$ + $MICR_9$**	93.45 ±0.16	2119.25 ±0.09	93.42 ±0.08	93.45 ±0.13	93.09 ±0.12
PRIVATEDT					
Lenet-5	71.53 ±0.18	122.93 ±0.04	72.48 ±0.13	71.53 ±0.10	68.86 ±0.16
MICR(1)	92.02 ±0.04	165.85 ±0.03	92.37 ±0.02	92.03 ±0.04	91.19 ±0.07
MICR(3) **$MICR_7$ +$MICR_8$ + $MICR_9$**	94.31 ±0.06	517.55 ±0.01	94.79 ±0.08	94.30 ±0.03	94.05 ±0.09
MICR(5)	93.84 ±0.08	889.25 ±0.08	94.52 ±0.07	93.84 ±0.11	93.51 ±0.09

MICR$_7$ +MICR$_8$ + MICR$_9$ + MICR$_7$ +MICR$_8$					
MICR(7) **MICR$_7$ +MICR$_8$ + MICR$_9$ + MICR$_7$ +MICR$_8$+ MICR$_9$ + MICR$_7$**	93.84 ±0.11	1230.95 ±0.03	94.52 ±0.14	93.84 ±0.10	93.51 ±0.13
MICR(9) **MICR$_7$ +MICR$_8$ + MICR$_9$ + MICR$_7$ +MICR$_8$+ MICR$_9$ + MICR$_7$ +MICR$_8$ + MICR$_9$**	92.48 ±0.13	1577.65 ±0.09	93.26 ±0.10	92.48 ±0.14	92.23 ±0.10

For Table 5. 3, the bootstrap subset was used to train only the ensemble members (MICR(3), MICR(5), MICR(7), and MICR(9), whose CNN architectural hyperparameters were optimised using BO. A train-test split ratio of 8:2 is used for MICR (1) and Lenet-5 without bootstrapping.

As earlier mentioned, the odd number of the ensemble member is used to allow a definitive final prediction of the queried input image. The results in Table 5.3, with different odd number configurations, allow a decision on the optimal number of members based on their accuracy on the testing set. The best model of the model as MICR(3) is justified given the imbalance class, based on multiple evaluation metrics, most notably the F1 measure, which is the harmonic mean of the precision and recall, and the f1-measure only increases if the prediction quality improves. Table 5.3 shows that the F1 measure at the MICR(3) is higher than every other model evaluated at 93.35% and 94.05% for MEDPIX and PRIVATEDT, respectively.

A plot of Table 5.3 is shown below in Figure 5.6

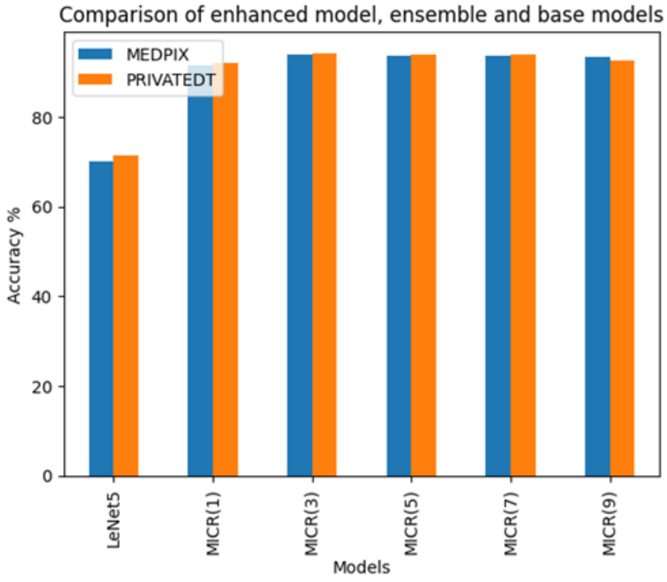


Figure 5. 6: Plot for the optimal number of ensemble members.

The Chart in Figure 5.6 shows the accuracy of LeNet-5 and proposed models with the two datasets. The Chart shows that the number of ensemble members is between 1 and 9, with the peak experienced at No. 3.

Figure 5.6 confirms that increasing the number of ensemble members further after three does not significantly increase the model accuracy. After using the bootstrap sampling method to create the training subsets, the MICR$_7$, MICR$_8$, and MICR$_9$ test datasets were 30%, 20%, and 10% of the overall dataset, respectively. The charts in Figure 5.6 above show that the MICR model, either as a single classifier or ensemble model, has higher precision, recall and F1-score evaluation metrics than the Lenet-5 model. For MEDPIX, Figure 5.6 shows a 21.40% increase in accuracy, 24.55% increase in precision, a 23.22 % increase in recall, and a 25.12% increase in F1-score for the single MICR model, while a 23.91% increase in accuracy, 26.05 % increase in precision, a 25.48 % increase in recall and a 26.35 % increase in F1-score for the MICR(3) model when compared with the Lenet-5 model. Similarly, for PRIVATEDT, Figure 5.6 shows a 20.49% increase in accuracy, 19.89% increase in precision, a 20.50% increase in recall, and a 22.33% increase in F1-score for the single MICR model, while a 22.78% increase in accuracy, 22.31% increase in precision, a 22.77% increase in recall and a 25.19% increase in F1-score for the MICR(3) model.

It is noted that with the bootstrapping sampling technique, where multiple samples are taken from the dataset with replacement to form subsets, no improvement was seen after three ensemble members (MICR(3) - MICR_7 +MICR_8 + MICR_9)). That is because the basis of the ensembling technique is to reduce variance in a model and improve the final prediction accuracy. However, the ensemble model's performance will decline when there is no further diversity in the ensemble members due to the small and insufficient training data. In order to get the optimal number of ensemble members, this study evaluated different number combinations and considered the training subsets before concluding that the three-member ensemble model had the highest accuracy.

Figure 5.7 below shows the learning curves for the Lenet-5 model on MEDPIX, showing the accuracy and loss over time for 100 epochs. The plot's ‘validation’ legend represents the testing accuracy and loss.

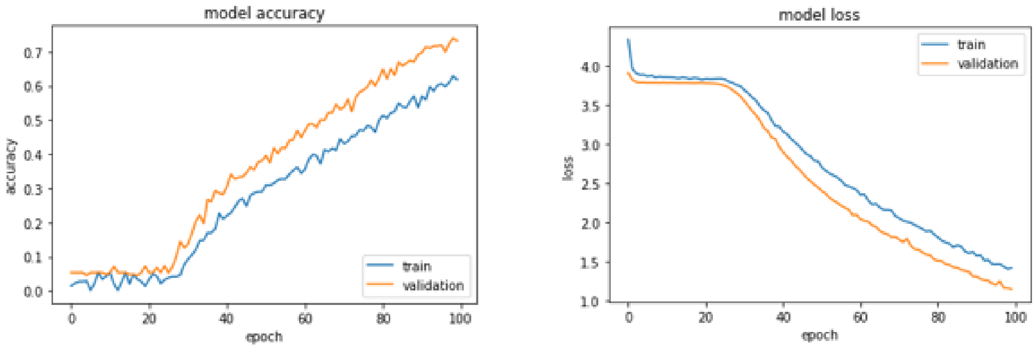


Figure 5. 7: Lenet-5 learning curves on MEDPIX

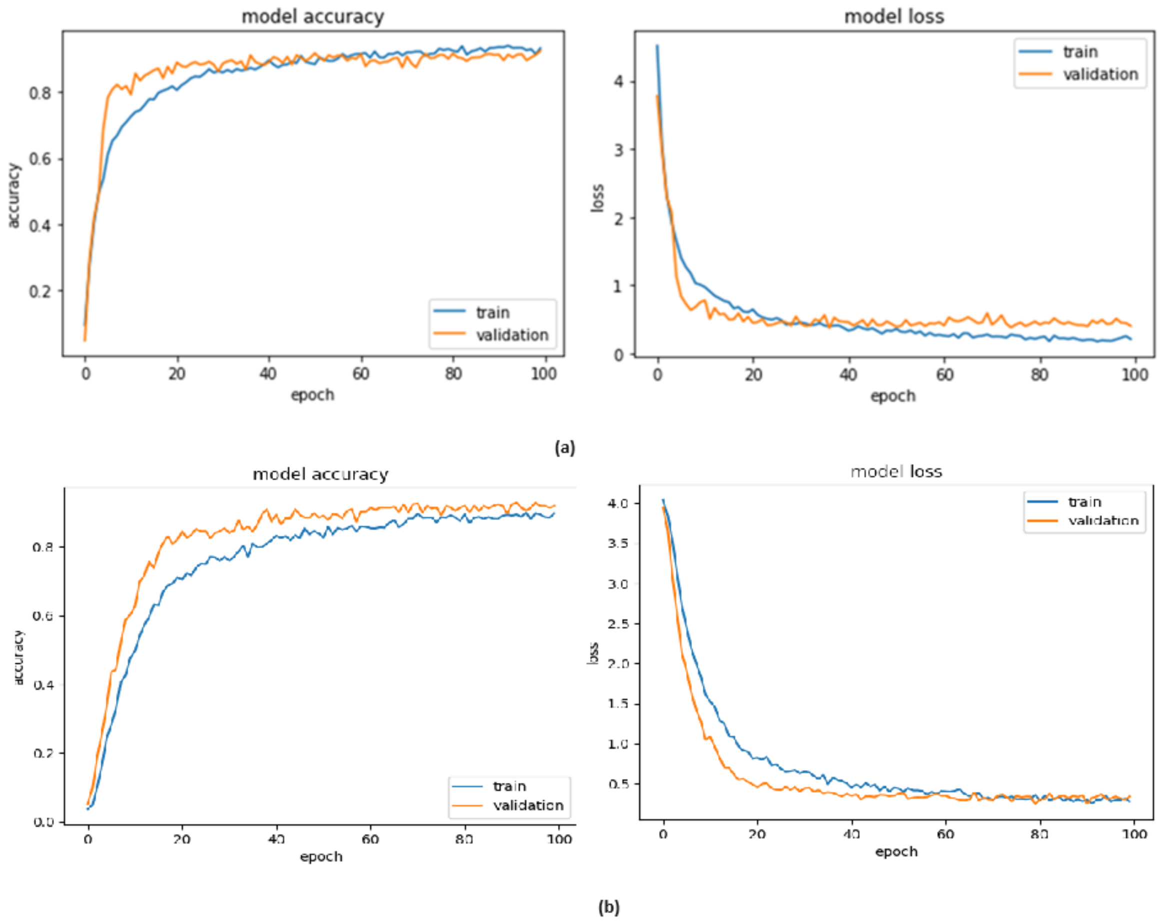


Figure 5. 8: Learning curves for the MICR model (a) MEDPIX (b) PRIVATEDT

The learning curves in Figure 5.8 show that the MICR(1) model converges faster than the Lenet-5. This is due to the architectural improvements in the MICR model and activation functions used. It can also be inferred that the model understands the training dataset.

5.6.2 Analysis of hyperparameters

The choice of hyperparameters is always based on the problem and the context. The search space was initially defined according to section 5.4.1, guided by literature works such as using a small kernel size for small input images (Hashemi, 2019; Tang et al., 2023). The hyperparameters' importance is checked after the 100 trials set using the Optuna importance module. The module compares the hyperparameters with their effect on the object function being minimised during each trial and returns values for their importance represented by non-negative floating numbers, where higher values mean the hyperparameter is more important than others with lower values. The correlation between the hyperparameters and the objective function defined mainly defines it. That is a high correlation means that when the hyperparameter has a higher value, the objective function also has a higher value. The module can evaluate all the defined hyperparameters and provide a report. The report is visualised below in Figure 5.9.

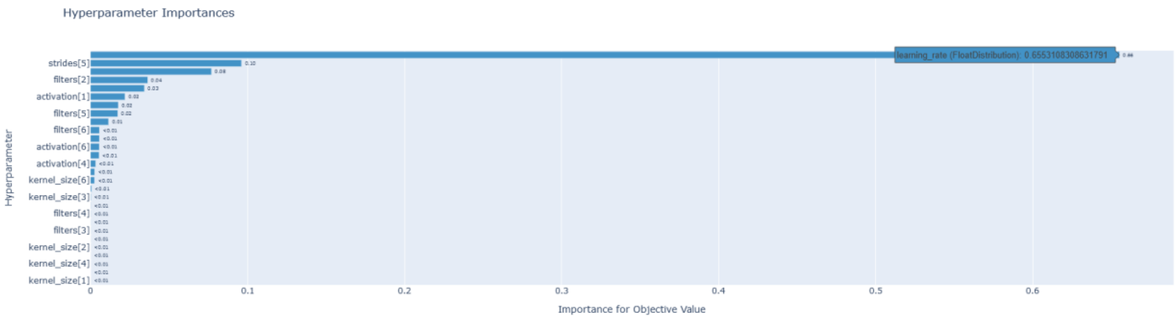


Figure 5.9: Hyperparemeters' importance

Figure 5.9 shows that the learning rate, the second stride, the filters in the second layer, activation in the first layer, and the activation and kernel size in the sixth layer had more effect on the objective value in terms of importance. On the other hand, the kernel size of the first and fourth layers and filters of the third and fourth layers showed less importance. This allows an optimal adjustment of the search space around these more impacting values to get the model's architecture presented in Table 5.1 with training hyperparameters of 0.000443917297 as learning rate, ReLU for the activation, batch size of 28 and drop out of 0.4 each at the third and sixth convolutional layers.

5.6.3 Classification report for all character classes

It is essential to evaluate the classwise evaluation metrics to enable a better understanding of the strengths and weaknesses of the MICR model. Please see Subsection 4.3 for sample sizes.

The classwise classification report for MEDPIX using the MICR(1) model is shown in Table 5.4 below.

Table 5. 4: Classification report on all classes in MEDPIX

Class	**Precision**	**Recall**	**f1-score**	**Class**	**Precision**	**Recall**	**F1-measure**
0	0.77	0.83	0.80	**a**	0.82	0.90	0.86
1	1.00	0.96	0.98	**b**	1.00	0.50	0.67
2	0.83	0.95	0.89	**c**	0.75	1.00	0.86
3	0.85	0.92	0.88	**d**	1.00	1.00	1.00

4	0.92	1.00	0.96
5	1.00	1.00	1.00
6	1.00	1.00	1.00
7	1.00	1.00	1.00
8	1.00	0.64	0.78
9	1.00	0.83	0.91
A	0.96	1.00	0.98
B	1.00	0.90	0.95
C	1.00	0.94	0.97
D	0.85	1.00	0.92
E	0.94	1.00	0.97
F	0.90	1.00	0.95
G	1.00	1.00	1.00
H	0.94	1.00	0.97
I	0.79	0.73	0.76
J	0.80	1.00	0.89
K	1.00	0.75	0.86
L	1.00	1.00	1.00
M	0.89	1.00	0.94
N	1.00	1.00	1.00
O	0.69	0.69	0.69
P	1.00	1.00	1.00
Q	1.00	1.00	1.00
R	0.97	0.97	0.97

e	0.86	0.92	0.89
f	1.00	1.00	1.00
g	0.80	1.00	0.89
h	0.00	0.00	0.00
i	0.86	0.86	0.86
j	0.00	0.00	0.00
k	1.00	1.00	1.00
l	0.40	0.50	0.44
m	1.00	1.00	1.00
n	0.67	0.67	0.67
o	0.50	0.20	0.29
p	1.00	0.75	0.86
q	1.00	1.00	1.00
r	1.00	1.00	1.00
s	0.67	0.57	0.62
t	1.00	0.86	0.92
u	0.67	0.67	0.67
v	0.00	0.00	0.00
w	0.00	0.00	0.00
x	0.50	0.33	0.40
y	1.00	1.00	1.00
z	0.00	0.00	0.00

S	0.92	1.00	0.96	
T	1.00	1.00	1.00	
U	0.86	0.75	0.80	
V	0.89	0.89	0.89	
W	0.78	0.88	0.82	
X	0.50	1.00	0.67	
Y	1.00	0.83	0.91	
Z	1.00	1.00	1.00	

For PRIVATEDT using the MICR(1) model, the classification report for all classes is shown below in Table 5.5.

Table 5. 5: Classification report on all classes in PRIVATEDT

Class	**Precision**	**Recall**	**f1-score**		**Class**	**Precision**	**Recall**	**F1-measure**
0	1.00	0.50	0.67		**a**	1.00	1.00	1.00
1	1.00	1.00	1.00		**b**	1.00	1.00	1.00
2	1.00	1.00	1.00		**c**	1.00	0.60	0.75
3	1.00	1.00	1.00		**d**	1.00	1.00	1.00
4	1.00	1.00	1.00		**e**	1.00	1.00	1.00
5	1.00	1.00	1.00		**f**	1.00	0.50	0.67
6	1.00	1.00	1.00		**g**	1.00	1.00	1.00
7	1.00	1.00	1.00		**h**	1.00	1.00	1.00

8	1.00	1.00	1.00
9	1.00	1.00	1.00
A	1.00	0.94	0.97
B	1.00	1.00	1.00
C	0.82	1.00	0.90
D	0.83	1.00	0.91
E	1.00	1.00	1.00
F	1.00	1.00	1.00
G	1.00	1.00	1.00
H	1.00	1.00	1.00
I	0.83	0.83	0.83
J	1.00	1.00	1.00
K	1.00	1.00	1.00
L	1.00	1.00	1.00
M	1.00	1.00	1.00
N	1.00	1.00	1.00
O	0.83	0.71	0.77
P	1.00	0.83	0.91
Q	1.00	1.00	1.00
R	1.00	1.00	1.00
S	0.68	1.00	0.81
T	0.94	1.00	0.97
U	0.75	1.00	0.86
V	0.00	0.00	0.00

i	0.90	0.90	0.90
j	0.67	1.00	0.80
k	1.00	1.00	1.00
l	0.88	0.78	0.82
m	1.00	1.00	1.00
n	0.90	1.00	0.95
o	0.60	0.75	0.67
p	0.83	1.00	0.91
q	1.00	1.00	1.00
r	0.92	1.00	0.96
s	1.00	0.42	0.59
t	1.00	0.90	0.95
u	1.00	0.60	0.75
v	0.50	1.00	0.67
w	1.00	0.67	0.80
x	0.00	0.00	0.00
y	1.00	1.00	1.00
z	0.50	1.00	0.67

W	0.67	1.00	0.80	
X	0.50	1.00	0.67	
Y	1.00	1.00	1.00	
Z	0.00	0.00	0.00	

As seen in the classification results for each character, the models’ poor results were only seen in visually similar characters such as “2” and “Z”, where “Z” had a small sample size compared to “2”. This affected the models' ability to learn effective and discriminative representations between these characters. The models performed better in other classes, with larger sample sizes, as seen in the evaluation metrics presented in Tables 5.4 and 5.5 above. Classwise comparison in PRIVATEDT reveals that small class sizes such as “c” had 1.00, 0.60, and 0.75 for precision, recall, and F1-measure, respectively, for the MICR model compared to 0.67, 0.67, and 0.67 for Lenet-5 on similar evaluation metrics. Larger class sizes such as “S” had 0.55, 0.75, and 0.63 for precision, recall, and F1-measure, respectively, for Lenet-5, while MICR had 0.68, 1.00 and 0.81 on similar evaluation metrics. Similar results were shown in MEDPIX; Hence, the proposed models improve the recognition accuracy of large and small class-size character classes compared to Lenet-5.

5.6.4 Result Comparison with past works

This study's results outperformed most existing works in the domain of burned-in textual recognition at the character level. After an extensive literature review, Table 5.6 compares this study with other works that designed classifiers. This study compared the proposed models with existing algorithms in the MIM domain of burned-in textual recognition, as shown in Table 5.6 below. Online data augmentation was used to train all the models using random translation and rotation. For the random translation, the image is randomly shifted horizontally and vertically up to 10% of its size. For the random rotation, each image was rotated up to 20 degrees, either clockwise or anticlockwise. Bootstrapping was used only for the MICR(3), as it is the only ensemble model on the table, while the train-test split ratio for the other models is 70:30. These related works evaluated their models on MEDPIX; hence, this was used only to compare performance.

Table 5. 6: Comparison with Related Works on MEDPIX

Method	**Recall (%)**	**Precision (%)**	**F1-measure (%)**
Modified CRNN (Xu et al., 2021)	65.00	67.00	70.00
CNN (Monteiro et al., 2017)	78.95	83.05	79.73
MICR	89.89	88.61	90.56
MICR(3)	94.46	94.49	94.49

The result proved that the work has outstanding results in better performance in classifying characters in low-resolution MIM with background interference. Moreover, this study performed better than a more complex model designed by authors in Xu et al. (2021), which was a multiscale CRNN. Xu et al. (2021) is the most recent work and used the same MEDPIXs as ours (Medpix dataset). Xu et al. (2021) reported an F-1 score of 70.00%, while the proposed MICR model outperformed with 90.56%. The majority voting ensemble also performed better than the CRNN model from the work of Xu et al. (2021). The proposed MICR model also outperformed notable works by authors (Monteiro et al., 2017; Silva et al., 2018) whose CNN model obtained an F1-measure of 79.73% on MEDPIX. The proposed MICR model was evaluated on 62 classes of characters, which were manually annotated by this study. Based on the bootstrapping data varying method, the majority voting ensemble had more accuracy than most existing works in the literature on the MICR domain. The practical application results of the proposed ensemble model in recognising burned-in textual data in MIM are shown in Figure 5.10 below.

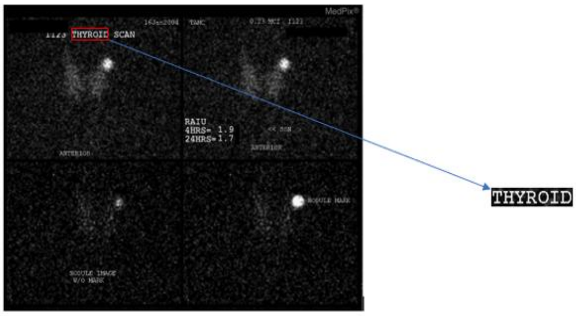


Figure 5. 10: Recognition of low-resolution MIM sample with background interference

As seen in Figure 5.10, the proposed MICR model has a good accuracy rate in dealing with low-resolution MIM with background interference. The word "**THYROID**" was recognised according to individual characters. The proposed MICR model can also recognise fuzzy words in MIM irrespective of font type and style.

5.7 Chapter Summary

This chapter introduces an enhanced CNN model inspired by the Lenet-5 classical OCR model for the task of medical image character recognition. The enhanced CNN model is optimised using Bayesian reasoning to determine the optimal combination of hyperparameters. Experimental results demonstrated that replacing the initial 5x5 filters and average pooling layers in Lenet-5 with 3x3 filters and 5x5 with a stride of 2, respectively, increased the accuracy of the enhanced CNN model. Several iterations were performed during optimisation to decide the optimal depth of the model to achieve a good performance. An ensemble model was introduced based on a majority voting algorithm to enhance the recognition of visually similar characters. Training subsets were created based on a bootstrapping method. Each classifier was trained on each subset and evaluated on the remaining test data in each training iteration. The enhanced CNN and ensemble models achieved an outstanding accuracy score in MICR at a previously unreported low resolution of 96 dpi compared to the state-of-the-art. The empirical observations generally indicated that a simple CNN architecture with initial small filter sizes and learnable downsampling layers could achieve better performance in MICR in low-resolution MIM with background interference. In future work, a more specialised CNN architecture and a more advanced ensemble will be considered to boost the performance relating to visually similar characters with a small sample size and low accuracy. In addition, I may include an attention mechanism to selectively give more relevance to some areas of the input image than others. The attention mechanism will increase the representation power of interests, as supported by past works (Li et al., 2022; Guo et al., 2022).

6.0 Medical Image Character Recognition using Attention-based Siamese Networks for Visually Similar Characters with Low Resolution.

This chapter proposes a channel attention-based Siamese Network to accurately recognise VSC in burned-in textual data in MIM by efficiently learning the semantic similarities between the extracted embeddings from the input character images. Intensive experiments are done using open-source and privately collected medical imaging datasets. The learned similarities and attention-focused feature extraction layer enable the proposed model to discriminate between different character classes efficiently, with only small samples available. Bayesian optimisation is used to determine optimal network parameters. I aim to set a benchmark for the performance of the Siamese network in OCR in MICR in terms of parameter size and accuracy at a determined sample size per class.

The work presented in this chapter was done during the third year of this PhD project (2022), was presented in international conference proceedings in 2024 (Osagie et al., 2024b) and was published by the Springer book series Lecture Notes in Networks and Systems (Osagie et al., 2024b). The content of this chapter has been adapted from Osagie et al. (2024b), with some modifications and additional experimental results to suit the style better and ensure a logical presentation of the study.

6.1 Introduction

OCR is an important computer vision application that converts text into images in easily accessible forms. It is widely used in numerous applications, such as industrial, medical, and educational institutions, mainly in automating data entry and other database-driven processes. However, numerous documents and images, such as MIM, may have certain constraints, such as low resolution, character distortion, text overlapping and background interference. These can be caused either by the mode of acquisition or storage. The textual data are usually burned in on the MIM. Due to distortion, poor image quality, background noise and low resolution, certain characters may appear visually similar in their structure and appearance (Röhrbein et al., 2015). These can be termed visually similar characters (VSC). Recognising these characters may become more challenging due to the nature of the images by conventional OCRs. Even with the rapid growth in the application of deep learning techniques in the field of OCR, the problem of recognising VSC remains unsolved, resulting in various research to find a solution (Inkeaw et al., 2019). This is because conventional DL techniques rely on a large and equally distributed dataset to achieve good performance. However, collecting a large dataset in certain domains, such as MIM, requires a lot of resources, such as privacy permissions (Padmapriya and Parthasarathy, 2024). Hence, developing a MICR solution that can learn highly discriminative features from low-resolution images with background interference becomes important to classify VSC with only a small sample size available. This will enable further adoption of OCR in low-resource domains where data accessibility is highly limited. This chapter proposes a few-shot learning method based on the Siamese neural network (SNN) and channel attention mechanism to deal with these issues. The SNN is a major component of few-shot learning methods (He et al., 2023; Müller et al., 2022; Dey et al.,2017; Cao et al., 2013).

The SNN can learn semantic similarities between classes of images by minimising the metric distance between the same class and maximising the metric distance between different classes. However, using the concept to define a fine-tuned classification decision boundary for VSCs on these complex images is a major challenge when the issue of tiny text, low resolution, and background interference must be considered. This is because the complex nature of the character images may affect the extracted feature embeddings to be compared. Hence, combining a channel-wise attention mechanism will enable the SNN to focus on the image's critical discriminative region by exploiting the features' inter-channel relationship. Since each channel of a feature map can be considered a feature detector, an SNN with a channel-wise attention mechanism focuses on the meaningful aspect of an input image that sets it apart for effective representation learning. An accurate MICR solution can improve health data analytics by allowing a more accessible and more accurate extraction of data from medical images, which can assist in analysing image data to identify patterns that are not easily visible to the naked eye. Hence improving patient care and diagnosis.

This chapter investigates SNN and channel attention modules' recognition of VSCs in low-resolution images under the limited sample size constraint. Section 6.2 presents related work. Section 6.3 provides the contributions. Section 6.4 presents the proposed methods. Section 6.5 shows the experimental setup. Section 6.6 shows the results and analysis. Section 6.7 provides the conclusion regarding this chapter.

6.2 Related Works

In this section, the study will review related works on applying SNN in the general field of OCR because extensive reviews have shown that SNN has not been applied in the MICR.

SNN and K-Nearest Neighbour classification algorithms were used to classify similar text by Hosseini-Asl and Guha (2015). An evaluation was done on machine-printed and handwritten text, and they reported an accuracy of 99.5%. Hosseini-Asl and Guha (2015) used a large dataset containing over 188,526-character images. A combined loss function was used, which caused difficulty during training, and the dataset was of high quality. Good accuracy of 97%, 79% and 89% were reported on three datasets, but this method will not be efficient in situations where the dataset is much more limited in sample size. With more focus on leveraging the advantages of the feature extraction capabilities on the radical-level composition of characters, Wang et al. (2019) proposed a radical aggregation network for few-shot recognition of handwritten character recognition. Their network used a convolutional block, ResNet, and an attention module. It performed an efficient radical feature selection using a radical mapping encoder to map the input into a radical representation sequence, where each representation is a high-dimensional feature vector. A distance metric is calculated between these radical representations and radical prototypes, and a character analysis decoder does transcription to a character. A 96.97% accuracy on the CASIA-HWDB character dataset was obtained using only 6,391 training samples. Although the accuracy was good, the representation mapping of distorted characters and VSCs was poor based on comparison with human performance, meaning that the radical representation learned by the network is still ineffective. It was complex and highly resource-demanding, more than twice the baseline CNN-based classifier model used in the study. Another study that leverages the use of a prototype was done by Snell et al. (2017) to compute an N-dimensional representation of each class through an embedding function with learnable parameters. Each prototype is the mean vector of the embedded support points belonging to its class. Their proposed method obtained an accuracy of 49.42% on only 1623 samples of handwritten characters with 50 classes and 68.20% on training with 5 samples per class. However, the study did not propose any defined architecture; there was no consideration for low resolution and background interference in these characters. According to Snell et al. (2017), episodic training was done to simplify the training. However, this depends is based on the idea that there exists an embedding in which data points cluster around a single prototype representation for each class and that a model can learn a non-linear mapping of the input into an embedding and take a class's prototype to be the mean of its support set in the embedding space. Classification is then done for an embedded query point by finding the nearest class prototype. This technique becomes inefficient when characters are blurred, distorted, or degraded.

6.3 Contributions

The main contributions of this chapter are.

- Proposes a Siamese neural network to learn semantic similarities between extracted embeddings of image pairs in metric space in the presence of a limited dataset, low image resolution, and background interference. The resulting model can discriminate between visually similar characters by learning a fine-tuned decision boundary.

- Propose a channel attention mechanism combined with a Siamese neural network to learn meaningful parts of an input image that discriminate when compared to a visually similar image.
- Provide a benchmark for using similarity learning in medical image character recognition (MICR). After an extensive literature review in the past 10 years, no previous work has been done regarding MICR and SNN with and without channel attention mechanisms.

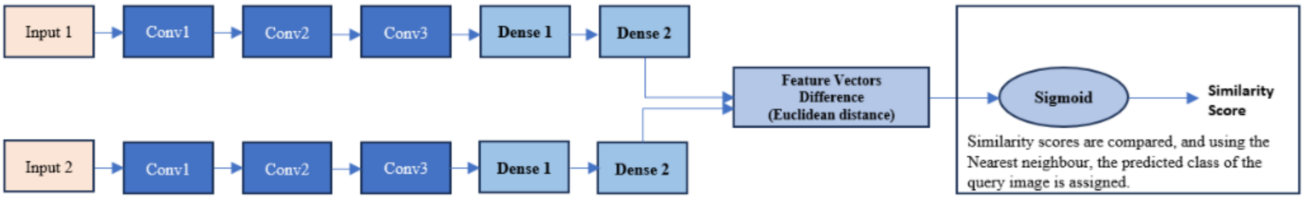
  This study's extensive reviews, which searched notable databases such as Elsevier, IEEE, Nature, and Science, did not reveal any existing work for MICR that used the Siamese network. There are no reviews that discuss or propose Siamese-based methods for MICR. The popular methods included enhanced Tesseract or other open-source OCR, RNN, and CNN-based methods. I aim to support this by comparing the architecture’s performance with related past works on OCR with the Siamese network based on a medical image character dataset, sharing the constraint of background interference.

6.4 Proposed Method

6.4.1 Model architecture

This chapter proposes a SNN to learn semantic similarity between small samples of VSC, with the constraints of low resolution with background interference. The SNN is two CNNs that are joined at the end. Before being joined, each CNN has 5 layers (3 convolutional and 2 dense layers). Then, a Euclidean distance layer merges both CNNs with a single output. The weights are shared between the two CNNs, and the goal is to compute similarity functions between input images to identify whether an image pair is similar or not. For clarity, each single CNN (with the same configuration) outputs an embedding of the input image, and the Euclidean distance layer calculates the Euclidean distance between the two feature embeddings from each of the CNN outputs and scores the similarity between the two feature embeddings (Koch et al., 2015). Weight sharing between each single CNN is achieved when both networks are backpropagated with the same loss function since they are joined at the end. If each CNN is represented by CNN_1 and CNN_2, when you compute forward the gradient for CNN_1 and then also for CNN2, then a concatenation of both gradients is done at the lamba layer where both CNNs are joined. According to (Koch et al., 2015), during updating with the averaged gradients, both CNNs are updated simultaneously. This architecture of using a twin CNN network is the standard for SNN, as supported by similar studies in medical imaging with SNN (Chung& Weng, 2017; Deepak & Ameer, 2021).

This study used Bayesian optimisation (BO) based on a tree-structured Parzen estimator to find the CNNs' optimal hyperparameters. The BO is a sequential design strategy for the global optimisation of objective functions that may be expensive to evaluate, such as the hyperparameters in neural networks (Osagie et al., 2023). The BO uses the informed learning method based on the Gaussian process by using a surrogate function to model the black box function and then uses an acquisition function to find the next point of evaluation (Osagie et al., 2023). The goal is to get very close to the optimum values with very few iterations of the black box functions. BO can fit the observed values of the black-box function and interpolate between observed data points, with increasing statistical uncertainty the farther you move away from the observed data. This study will not focus on the optimisation technique as it is not part of the aims. The BO's overall goal was to find the maximum value of the objective function, which is the similarity score between a query image and a set of support images. I ran hyperparameter tuning for about 200 trials during the first study using the Optuna library and checked the most important hyperparameters. Next, I omitted the less important hyperparameters for the subsequent studies up to 2000 trials to find the optimal CNN's hyperparameters. The final layerwise summary is three Convolutional layers with filters of 16, 32 and 64, each with a kernel size of 3x3 and a stride of 2. Two dense layers follow with 128 and 254 units, respectively. ReLU is used as an activation function in the hidden layers. Fig 6.1 below shows a visual representation of SNN. For ease of reference, this model will be referred to as SIAM-MICR.



Similarity scores are compared, and using the Nearest neighbour, the predicted class of the query image is assigned.

Figure 6. 1: SIAM-MICR

The CNNs were designed as pairs, and training was achieved using the two parallel CNNs with shared weights, trained on matched and unmatched character image pairs. Each image is fed through one branch of the CNN, generating a d-dimensional embedding for the image. The loss function optimised is based on contrastive representation learning, which aims to learn such an embedding space in which similar image pairs are close to each other while dissimilar image pairs are far from each other (Chopra et al., 2005). Contrastive loss aims to predict relative distances between model inputs when projected onto a hyperspace. The embeddings between the pairs are used to calculate the Euclidean distance to measure similarity. In the SNN architecture, the Lambda layer computes the Euclidean distances between the outputs of the two parallel CNNs.

6.4.2 Model + Attention Mechanism

This study proposed using a channel attention mechanism as motivated by notable works by Wang et al. (2019) and Shen et al. (2018) to improve the previously designed SIAM-MICR model. The channel attention mechanism in each CNN generates channel-wise responses by using global average pooling to aggregate spatial information (Hu et al., 2017). Given the aggregated features obtained from the global average pooling (GAP), a fast 1D convolution of kernel size, k, is performed to generate the output channel weights. k is the kernel size of the 1D convolutional layer. It represents the coverage of local cross-channel interactions, that is, the number of pixel neighbours taking part in the output of one channel map. Using a 1D convolution avoids dimension reduction and allows efficient learning across the channel for significant and discriminating features of the input images for the SNN. Much investigation via experiments was carried out to determine the optimal position for the attention module on the SNN architecture, and these are presented in the result section. The optimal position for the attention module was investigated by alternating the insertion position and comparing it with the average accuracy achieved at that position. This setup improved the network’s ability to focus on learning weights for more primitive and discriminative features, such as curves, lines, and edges, which may appear similar across the character classes.

The channel attention mechanism used in the SIAM-MICR is motivated by a notable work by Wang et al. (2019). The input tensor to the module is the output of a convolutional layer and has a 4-D shape of B, C, H, and W, where B is the batch size, C is the number of channels, and H and W are the dimensions of each feature map. The output of the attention module is also a 4-D tensor of the same shape. Figure 6.2 shows the SIAM-MICR with the attention module after layer 1.

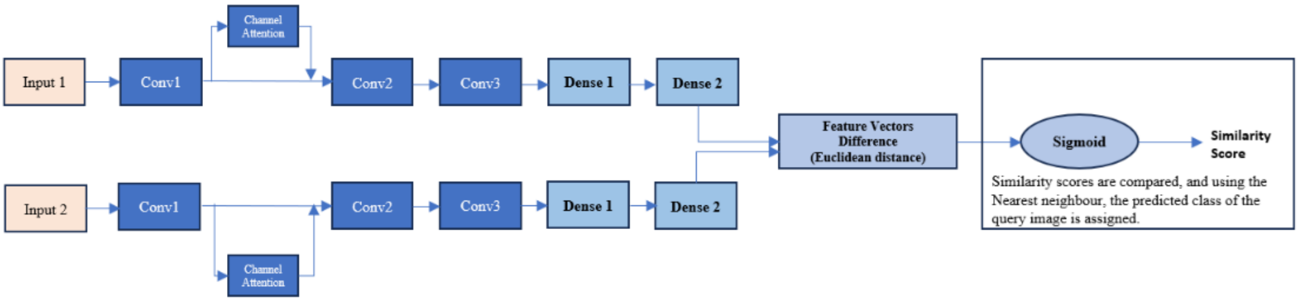

Channel Attention
Input 1
Conv1
Conv2
Conv3
Dense 1
Dense 2
Feature Vectors Difference (Euclidean distance)
Sigmoid
Similarity Score
Similarity scores are compared, and using the Nearest neighbour, the predicted class of the query image is assigned.
Input 2
Conv1
Conv2
Conv3
Dense 1
Dense 2
Channel Attention

Figure 6. 2: SIAM-MICR + Attention

6.5 Experimental Set-Up

6.5.1 Dataset Description

The datasets used are the same as described in Chapter 4 (Research Methodology), which were used for this chapter.

6.5.2 Training Strategy

To train and evaluate the Siamese network, the 62-class dataset is changed into a binary classification problem by creating a new dataset of pairs, where matched images are labelled 0.0 and unmatched images are labelled 1.0. RMSprop optimiser was used because of its advantage in fast convergence speed over a few iterations (Kandel et al., 2020; Hassan et al., 2023; Lee et al., 2022). The training pairs were formed randomly and were balanced across classes. The sample size taken means the total data points per a single class. The pairing is done via the following procedure;

- A list of indexes for each class label is built using a for loop, and there are 62 class labels. This outputs each data point's current class label, total sample size, and indexes.
- The algorithm efficiently generates the positive and negative pairs based on the data point indexes. It selects a particular image and then randomly picks an image that belongs to the same class (positive pair). With both images, the current and the same class image, the new dataset of pairs list is created with a 2-tuple of the selected image and the same class image, and the target label of this new dataset is updated with a value of 0 to indicate a positive pair.
- To generate the negative pair, the algorithm selects all indexes of class labels not equal to the selected image and randomly selects one of these indexes as the negative image. Similarly, the target label of this new dataset is updated with a value of 1, indicating a negative pair.
- Finally, the newly formed dataset of pair images and pair labels is returned. A train-test split ratio of 70:30 is set and used for the model's training.

For balanced pairing, classes with smaller class sizes were repeatedly paired for the training. Figure 6.3 shows samples of the training data after pairing for “W”..

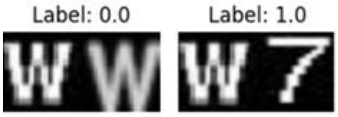


Figure 6. 3: Pairing of images for training

Based on visual inspection and taking note of the highest and lowest sample sizes, 15, 20, and 25 samples per class were used for this pairing and training. The characters that are to be paired are 0~9, A~Z, and a~z, a total of 62 characters. In my dataset, there are 50 images for each character, e.g, 50 different “W” images in “W” sub-image set. So there are total of 62 sub-image sets for the 62 characters. Figure 6.3 shows two ”W” images in the “W” sub-image set, one is clear and the other much fused. When pairing, e.g., to pair “W”, firstly randomly pick up one “W” from the “W” sub- image set. Secondly, randomly pick up another “W” from the same “W” sub-image set, this constitutes a positive match as they are similar. This pair is assigned “0” class label. Thirdly, randomly pick up an image from “W” sub-image set and another image from any other sub-image set, this constitutes a negative match because they are dissimilar then a “1” class label is assigned to this pair. Figure 6.3 shows a positive match and a negative match labelled with 0 and 1, respectively. This process carries on until all images are paired using non-replacement sampling. Below Table 6.1 is the summary of the pairing for the MEDPIX dataset.

Table 6.1. Summary of pairing results for MEDPIX dataset

Characters	No. of images in each sub-image set	No. of positive pairs for each character	No of negative pairs for each character	Total number of pairs for each character
0~9	50	25	25	50
A~Z	50	25	25	50
A~z	50	25	25	50
	Total No of images in the dataset	Total No of positive pairs that are labelled as 0	Total No of negative pairs that are labelled as 1	Total No of pairs in the dataset
Total	3100	775	775	1550

The same process of pairing is implemented for the PRIVATEDT dataset, and the summary is in Table 6.2.

Table 6.2. Summary of pairing results for PRIVATEDT dataset

Characters	No of images in each sub-image set	No of positive pairs for each character	No of negative pairs for each character	Total number of pairs for each character
0~9	50	25	25	50
A~Z	50	25	25	50
A~z	50	25	25	50
	Total No of images in the dataset	Total No of positive pairs that are labelled as 0	Total No of negative pairs that are labelled as 1	Total No of pairs in the dataset
Total	3100	775	775	1550

As seen in Tables 6.1 and 6.2, the ratio between similar and dissimilar images is 1:1, and the pairing is randomly done without replacement. During training, a train-test split ratio of 70:30 was used; bootstrapping and cross-validation were not used in the experiments. Online data augmentation was used during training to improve the model's generalisation (Shorten and Khoshgoftaar, 2019). This was done by random rotation; each image was rotated up to 20 degrees, either clockwise or anticlockwise. This encodes rotational invariance in the SNN, increasing each CNN representational power and its classification accuracy (Quiroga et al., 2019). However, this leads to increased training time. The main goal is to learn a twin network on a small dataset and do classification based on the similarity between pair images (Li et al., 2022).

6.6 Results and Analysis

This study investigated the accuracy of the SNN model with and without channel attention on sample sizes of 15, 20, and 25. Due to the overall small dataset size, the study preferred to use a maximum of 25 sample sizes to consider the classes with small sample sizes. The experimental results are presented in Table 6.1.

Table 6. 3. Comparison of Model's Accuracy with/without channel attention at 100 epochs (Attention module inserted after 3rd CNN layer)

Sample size	**Accuracy- SIAM-MICR (%)**	**Accuracy - SIAM-MICR + Attention (%)**
MEDPIX		
25 samples per class	87.73±0.92	90.77±0.80
20 samples per class	85.73±0.61	87.58±0.45
15 samples per class	82.35±0.56	85.67±0.78
PRIVATEDT		
25 samples per class	95.45 ±0.13	97.66 ±0.22
20 samples per class	95.79 ±0.11	97.58 ±0.16
15 samples per class	93.64 ±0.24	94.72 ±0.43

For MEDPIX and PRIVATEDT, the standard deviation is represented as ±SD in Table 6.3 to show the average dispersion of the results relative to the mean. The results from Table 6.3 show that adding the channel attention mechanism on the base SNN improved the accuracy by approximately 3.0%. The results further reveal that on the privately collected dataset, PRIVATEDT, the SIAM-MICR + Attention model remains high-performing and stable even with a reduction of 20% of its sample size from 25 to 20. This supports the practical application and generalisation of the proposed channel attention-based model. However, without the attention module, 20 samples per class perform slightly better than 25 samples per class because a small increase in sample size does not always lead to increased performance for all models (Bailly et al., 2022). The increase does not equally mean increased data quality available to the model (Alwosheel et al., 2018).

The results also show that PRIVATEDT has a smaller standard deviation than MEDPIX; this indicates less variability in the image dataset and shows that the images in PRIVATEDT are more consistent and uniform than those in MEDPIX, which is open source. This is easily understood because this research collected PRIVATEDT from the designated location using similar acquisition machines, whereas MEDPIX is an open-source collection with diverse contributions from different sources.

The study investigated the optimal layer for the attention module insertion on the SNN, and the results showed an accuracy of 96.29% ±0.16, 93.14% ±0.27, and 96.93% ±0.23, at CNN layers 1, 2 and 3, respectively, averaged at 30 runs on the PRIVATEDT. Similarly, on MEDPIX, results on the optimal layer showed an accuracy of 88.64% ±0.23, 89.81% ±0.27, and 89.89% ±0.33, at CNN layers 1, 2 and 3, respectively, averaged at 30 runs. During the investigation the training and test sets are split in a 70:30 ratio, with a sample size of 25 per class and no online data augmentation. It is seen that the absence of online data augmentation in this investigation led to a reduction in the attention-based model's accuracy for both datasets when compared to Table 6. 3; for instance, at the same 3rd CNN layer of insertion, MEDPIX reduced slightly from 90.77% ±0.80 to 89.89% ±0.33 while PRIVATEDT reduced slightly from 97.66 ±0.22 to 96.93% ±0.23.

Note that only three convolutional layers are present in the proposed model's architecture, hence the range of investigation. This is presented in the plot in Figure 6.4 below.

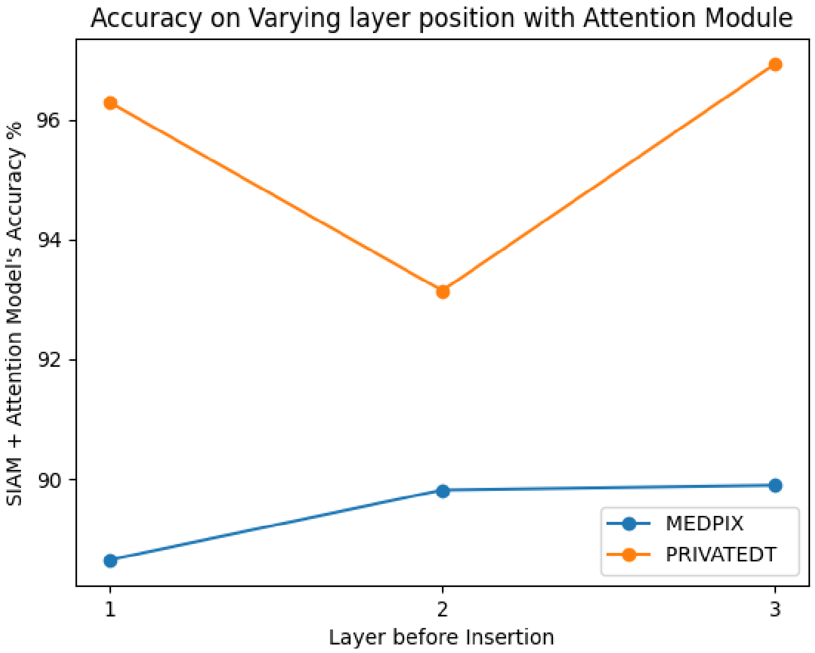
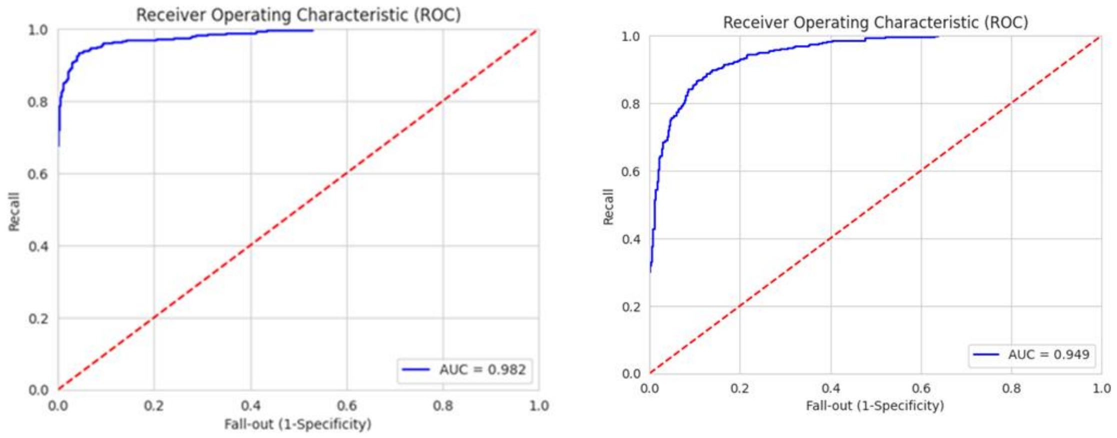


Figure 6.4: Optimal layer of insertion of attention module

Figure 6.4 shows that the attention-based model's performance remains highest at layers 1 and 3 for the PRIVATEDT and layers 2 and 3 for the MEDPIX. These positions signify the best layers that learn the attention weights for each discriminating part of the input image by exploring the optimal insertion positions of the channel attention module. Figure 6.4 also shows that the change in the layer number has a bigger impact on PRIVATEDT than on MEDPIX, with the latter having a linear line on the plot. The reason is that the attention module selectively highlights salient features from the images and concatenates them with original input to improve the model's overall performance, and this will be more impactful in the PRIVATEDT, which has less noisy, irrelevant features, highly consistent and uniform images compared to the MEDPIX. Hence, changing the layers allows the attention module to simultaneously learn weights for features at different positions, leading to a bigger impact.

6.6.1 Performance Analysis on AUC - ROC Curve

This is a performance measurement for classification that tells how much the models can distinguish the classes. The higher the AUC value, the better the model can distinguish whether the actual Euclidean distance is 0 or 1. This study’s experiments on 25 sample sizes on MEDPIX, as shown in Figure 6.5, show a 98.2% AUC value for the SIAM-MICR + Attention and a 94.9% AUC value for the SIAM-MICR. Similarly, for PRIVATEDT, the AUC values are closer to 1, as shown in Figure 6.5. From these results, it is agreeable that the model has a good measure of separability since the AUC values are closer to 1 than 0.



(a)

(b)

Figure 6. 5: ROC AUC on MEDPIX (a) SIAM-MICR + Attention (b) SIAM-MICR

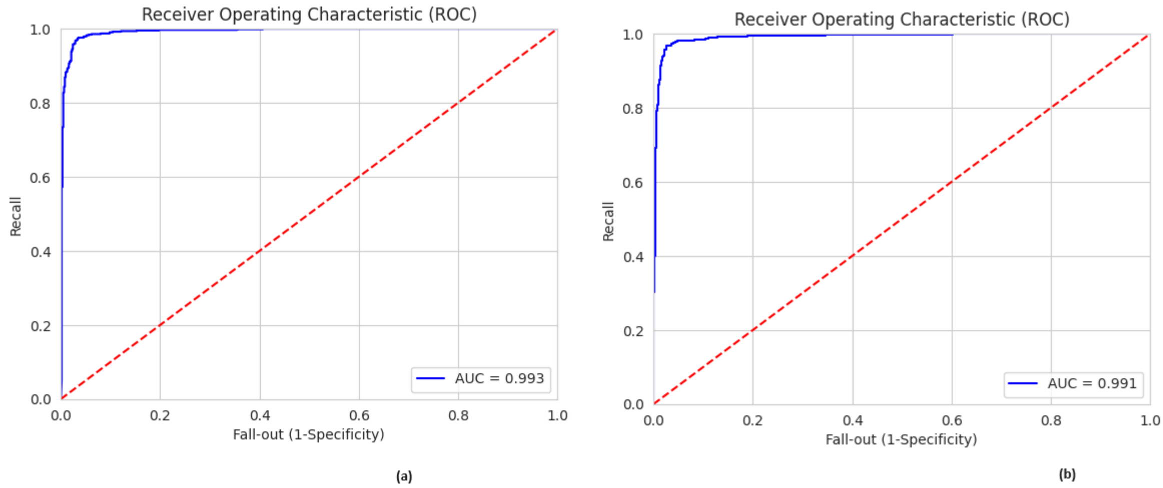


Figure 6. 6: ROC AUC on PRIVATEDT (a) SIAM-MICR + Attention (b) SIAM-MICR

However, only a slight increase in the AUC value of 0.002% is seen when the attention module is included for the PRIVATEDT; this shows that both models are high-performing at distinguishing between the matched and unmatched image pairs in PRIVATEDT compared to MEDPIX. An increase in ROC AUC usually indicates better performance in distinguishing positive and negative classes, and a 0.002% increase, as seen in Figure 6. 6 for the PRIVATEDT, shows the discriminative ability of the SIAM-MICR model is not significantly affected by the inclusion of the attention module. This may be due to noise in the dataset because the attention mechanism can be sensitive to noise, and the PRIVATEDT images contain noise due to their acquisition mode. Also, attention mechanisms may have limited representation learning ability, especially when dealing with small sample sizes, as focus may be made on specific patterns or pixels in the images, leading to ignoring other potentially relevant information (Ou, 2023) that may be useful for distinguishing positive and negative classes. Weng et al. (2023) examined similar attention and non-attention deep learning models and agreed that not every time an attention mechanism improved the model's performance, especially with small sample sizes.

6.6.2 Performance Analysis Using Feature Map Visualisation.

The intuition here is that the channel attention module inserted after the optimal CNN layer acts as a masking matrix that identifies and locates the prominent regions that contain significant morphological characteristics and passes these reinforced identified representations to the subsequent layers for better representation learning. This enables the network to focus only on a certain part of the feature maps that is more prominent and, therefore, more discriminating. This leads to a lower loss and a finer decision boundary between classes. This is demonstrated visually in Fig 6.8 below, where the feature map visualisation of the output of the 2nd CNN layer with the channel attention module shows that more prominent regions of the characters are densely populated with pixels when compared with the output of the 2nd CNN layer without the attention module in Fig 6.7, where these prominent regions' pixels are missing or very limited. Fig 6.7 and Fig 6.8 are shown below.

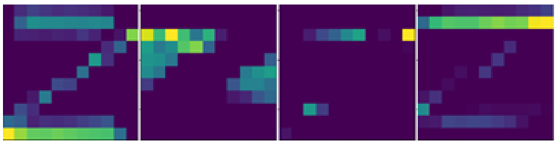


Figure 6. 7: SIAM-MICR's Second layer output for Character "Z."

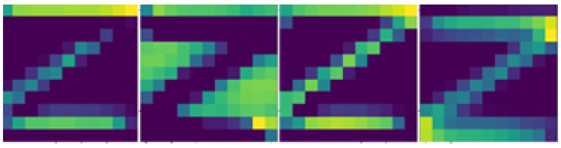


Figure 6. 8: SIAM-MICR + Attention's Second layer output after channel attention module insertion -Character "Z."

It has been shown that a prominent data representation improves performance compared to a poor data representation, as the DL algorithm is highly dependent on the integrity of the input-data representation (Alzubaidi et al., 2021).

6.6.3 Performance Analysis Using Confusion Matrix

The study computed the confusion matrix for all the classes in both datasets and the class-wise accuracy metrics. The test set for the confusion matrix evaluation has only 6 images per class based on the train-test split ratio. As mentioned earlier in the experiment design section, 30 runs are implemented. On analysing the confusion matrix for each of the 30 runs, the accuracies were the same up to two decimal places. The confusion matrix presented below shows one of the results. The 62 classes comprising “0-9,” “A-Z,” and “a-z” were used for the confusion matrix computation. The confusion matrices are presented in Figure 6.9 and Figure 6.10 for MEDPIX and PRIVATEDT, respectively.

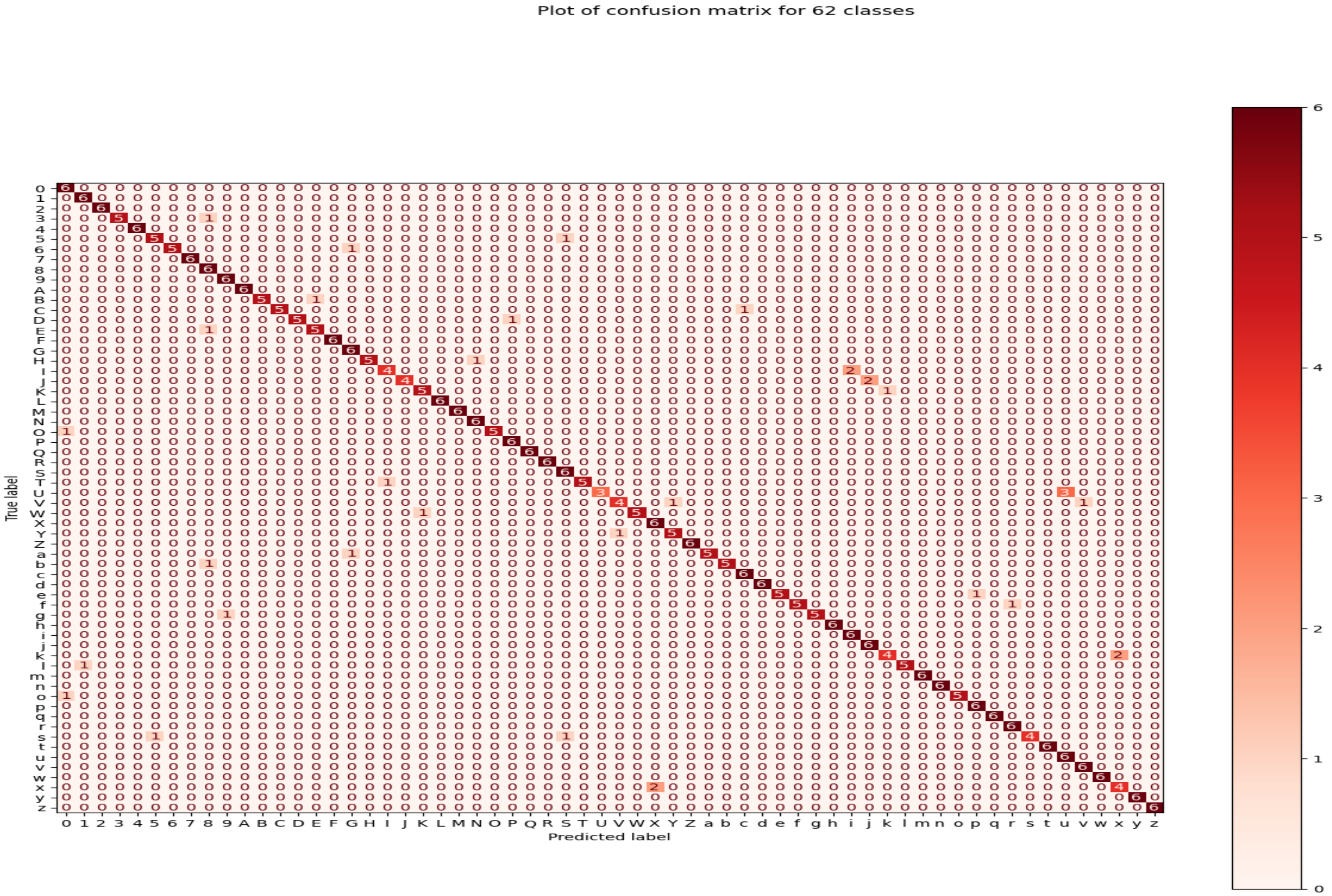


Figure 6. 9: Confusion Matrix for MEDPIX (SIAM-MICR + Attention)

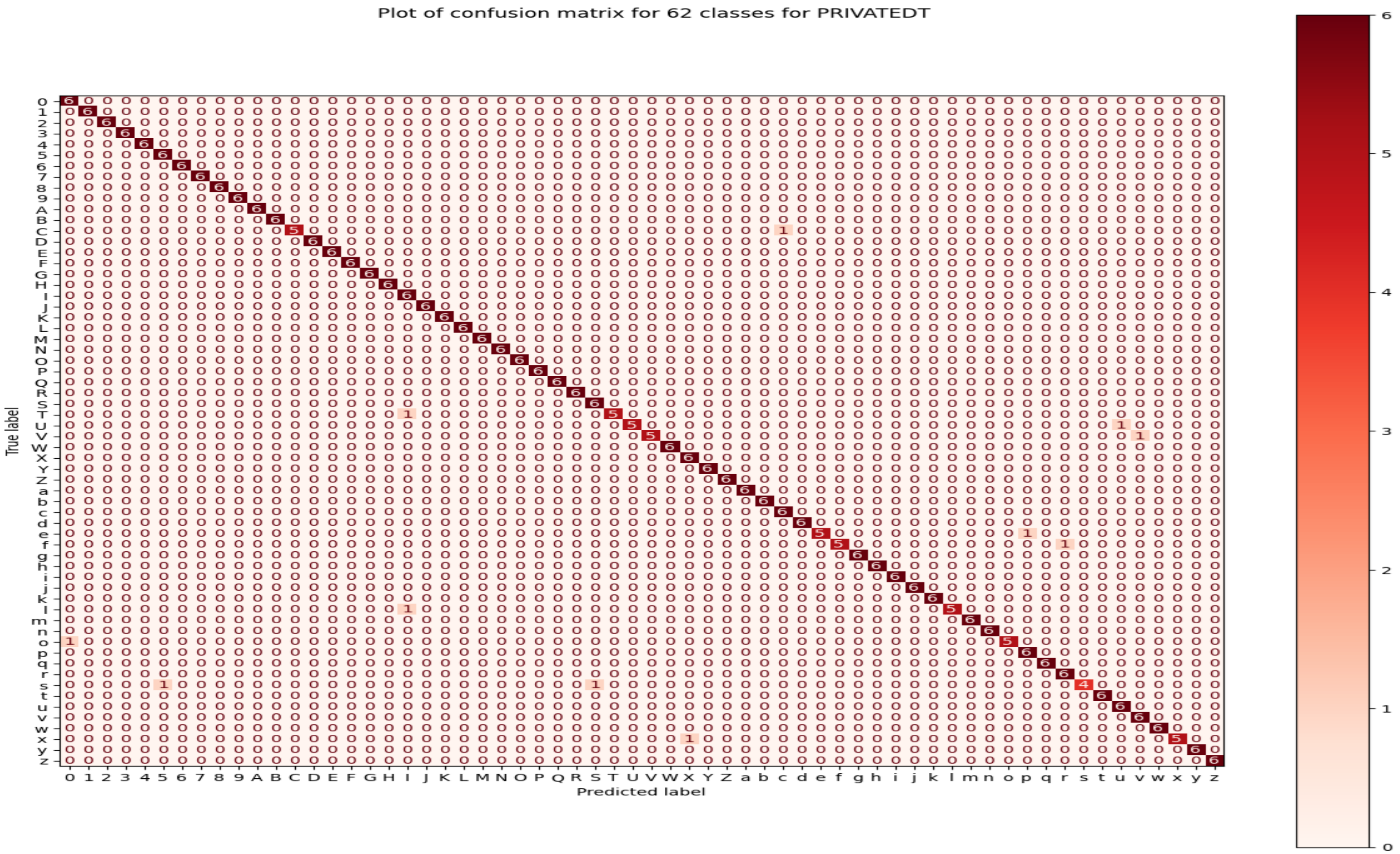


Figure 6. 10: Confusion Matrix for PRIVATEDT (SIAM-MICR + Attention)

6.6.4 Interpretation of confusion matrix and class-wise evaluation metrics

The confusion matrix in Figure 6.9 is obtained by training a classifier and evaluating the trained model on a test set for MEDPIX. Let that matrix be called "*M*," and each element in the matrix be denoted by "*M_i_j*," where "*i*" is the row number (predicted class), and "*j*" is the column number (expected class).

- As usual, the diagonal elements are the correctly predicted samples. Out of 372 samples of 6 test samples per each of the 62-character classes, the model accuracy is 90.59% by correctly predicting 337 samples.
- A lot of the elements in *M_i_j, are equals to 0, such as M_1_8* = 0 and *M_2_8* =0. T*his* implies that the model does not confuse samples originally belonging to different classes, i.e., the SNN learned the classification boundary well.
- Looking closer at the confusion matrix, the SNN confuses some upper-case characters with their lower-case counterparts, such as incorrectly predicting *I* as *i*, with an error rate of 33.3%, that is correctly predicting 4 samples as *I* and 2 samples incorrectly as *i*, in a total of 6 samples.
- For visually similar characters *"0"* and "O", the model got 100% accuracy in predicting all samples correctly, as shown in *M_0_0* = 6 and did not confuse it with O. This is encouraging considering the model was trained based on Euclidean distance, with only 17 samples per class.
- To improve the model's performance, the study can focus on the predictive results in the confusion matrix, which has the highest misclassification rate among all the classes*, such as M_30_58* = 0.

Similarly, the following observations were made for the PRIVATEDT confusion matrix in Figure 6.10.

- Out of 372 samples of 6 test samples per each of the 62-character classes, the model accuracy is 97.04% by correctly predicting 361 samples.
- The model does not confuse samples originally belonging to different classes; i.e., the SNN learned the classification boundary well, as seen in most of the cells on the confusion matrix, which are equal to 0.
- For visually similar characters "*5"* and *"S"* and "0" and "O", the model got 100% accuracy in predicting all samples correctly for "5" and "0", respectively.
- However, the model confuses some upper-case characters with their lower-case counterparts, such as "T" and "t" and "C" and "c."

When comparing Figure 6.9 and Figure 6.10 for the performance of the SIAM-MICR + Attention model for the MEDPIX and PRIVATEDT, the PRIVATEDT shows higher performance, with an increased accuracy of 6.42%. This difference in performance is because different datasets have different properties, no matter their similarities. Privately collected datasets are usually of higher quality, and a high-quality dataset can accurately represents real-world scenarios, have less noise and be free from biases (Gong et al.m, 2023). Hence, the quality of the PRIVATEDT had a significant impact on the accuracy and effectiveness of the SIAM-MICR + Attention model when compared to the MEDPIX, which is open source.

6.6.5 Classwise Evaluation Metrics on the SIAM-MICR + Attention

Table 6. 4: Classification report for all character classes for MEDPIX.

Class	Precision	Recall	f1-score
0	1.00	0.75	0.86
1	1.00	0.86	0.92
2	1.00	1.00	1.00
3	0.83	1.00	0.91
4	1.00	1.00	1.00
5	0.83	0.83	0.83
6	0.83	1.00	0.91
7	1.00	1.00	1.00
8	1.00	0.67	0.80
9	1.00	0.86	0.92
A	1.00	1.00	1.00
B	0.83	1.00	0.91
C	0.83	1.00	0.91
D	0.83	1.00	0.91
E	0.83	0.83	0.83
F	1.00	1.00	1.00
G	1.00	0.75	0.86
H	0.83	1.00	0.91
I	0.67	0.80	0.73

Class	Precision	Recall	f1-score
a	0.83	1.00	0.91
b	0.83	1.00	0.91
c	1.00	0.86	0.92
d	1.00	1.00	1.00
e	0.83	1.00	0.91
f	0.83	1.00	0.91
g	0.83	1.00	0.91
h	1.00	1.00	1.00
i	1.00	0.75	0.86
j	1.00	0.75	0.86
k	0.67	0.80	0.73
l	0.83	1.00	0.91
m	1.00	1.00	1.00
n	1.00	1.00	1.00
o	0.83	1.00	0.91
p	1.00	0.86	0.92
q	1.00	1.00	1.00
r	1.00	0.86	0.92
s	0.67	1.00	0.80

J	0.67	1.00	0.80
K	0.83	0.83	0.83
L	1.00	1.00	1.00
M	1.00	1.00	1.00
N	1.00	0.86	0.92
O	0.83	1.00	0.91
P	1.00	0.86	0.92
Q	1.00	1.00	1.00
R	1.00	1.00	1.00
S	1.00	0.75	0.86
T	0.83	1.00	0.91
U	0.50	1.00	0.67
V	0.67	0.80	0.73
W	0.83	1.00	0.91
X	1.00	0.75	0.86
Y	0.83	0.83	0.83
Z	1.00	1.00	1.00
t	1.00	1.00	1.00
u	1.00	0.67	0.80
v	1.00	0.86	0.92
w	1.00	1.00	1.00
x	0.67	0.67	0.67
y	1.00	1.00	1.00
z	1.00	1.00	1.00

Precision measures the accuracy of positive predictions, while recall measures the completeness of positive predictions. These are highly relevant evaluation metrics for data science models in medical applications (Hicks et al., 2022).

The class-wise classification reports show improvement in recognition of visually similar character images compared to the MICR CNN classifier reports in section 6.6.3. Instances of these are presented in Table 6.2 below

Table 6. 5. Improvement in VSC recognition compared to the multi-class MICR model (MEDPIX)

Multi-class MICR model				**SIAM-MICR + Attention Model**			
Class	**Precision**	**Recall**	**F1-score**	**Class**	**Precision**	**Recall**	**F1-score**
v	0.00	0.00	0.00	**v**	1.00	0.86	0.92
z	0.00	0.00	0.00	z	1.00	1.00	1.00
j	0.00	0.00	0.00	j	1.00	0.75	0.86

Table 6.5 shows that the SIAM-MICR + Attention Model improves recognition accuracy in certain characters with visually similar pairs and small sample sizes. Hence, the improvement is noted using the metric learning technique proposed in this chapter.

Table 6. 6: Classification report for all character classes for PRIVATEDT.

Class	**Precision**	**Recall**	**f1-score**
0	1.00	0.86	0.92
1	1.00	1.00	1.00
2	1.00	1.00	1.00
3	1.00	1.00	1.00
4	1.00	1.00	1.00
5	1.00	0.86	0.92
6	1.00	1.00	1.00

Class	**Precision**	**Recall**	**f1-score**
a	1.00	1.00	1.00
b	1.00	1.00	1.00
c	1.00	0.86	0.92
d	1.00	1.00	1.00
e	0.83	1.00	0.91
f	0.83	1.00	0.91
g	1.00	1.00	1.00

7	1.00	1.00	1.00
8	1.00	1.00	1.00
9	1.00	1.00	1.00
A	1.00	1.00	1.00
B	1.00	1.00	1.00
C	0.83	1.00	0.91
D	1.00	1.00	1.00
E	1.00	1.00	1.00
F	1.00	1.00	1.00
G	1.00	1.00	1.00
H	1.00	1.00	1.00
I	1.00	0.75	0.86
J	1.00	1.00	1.00
K	1.00	1.00	1.00
L	1.00	1.00	1.00
M	1.00	1.00	1.00
N	1.00	1.00	1.00
O	1.00	1.00	1.00
P	1.00	1.00	1.00
Q	1.00	1.00	1.00
R	1.00	1.00	1.00
S	1.00	0.86	0.92
T	0.83	1.00	0.91
U	0.83	1.00	0.91

h	1.00	1.00	1.00
i	1.00	1.00	1.00
j	1.00	1.00	1.00
k	1.00	1.00	1.00
l	0.83	1.00	0.91
m	1.00	1.00	1.00
n	0.83	1.00	0.91
o	0.83	1.00	0.91
p	1.00	0.86	0.92
q	1.00	1.00	1.00
r	1.00	0.86	0.92
s	0.67	1.00	0.80
t	1.00	1.00	1.00
u	1.00	0.86	0.92
v	1.00	0.86	0.92
w	1.00	1.00	1.00
x	0.83	1.00	0.91
y	1.00	1.00	1.00
z	1.00	1.00	1.00

V	0.83	1.00	0.91
W	1.00	1.00	1.00
X	1.00	0.86	0.92
Y	1.00	1.00	1.00
Z	1.00	1.00	1.00

Table 6. 7. Improvement in VSC recognition compared to the multi-class MICR model (PRIVATEDT)

Multi-class MICR model				**SIAM-MICR + Attention Model**			
Class	**Precision**	**Recall**	**F1-score**	**Class**	**Precision**	**Recall**	**F1-score**
v	0.00	0.00	0.00	**v**	1.00	0.86	0.92
z	0.00	0.00	0.00	z	1.00	1.00	1.00
j	0.00	0.00	0.00	j	1.00	1.00	1.00

Similarly, Table 6.6 and Table 6.7 show for the PRIVATEDT that the SIAM-MICR + Attention Model improved recognition accuracy in certain characters with visually similar pairs and small sample sizes. However, the performance declined considering certain characters, such as G, in the MEDPIX, in Table 5.4 and the classwise results in Table 6.4. The difference in dataset size explains this: for instance, for letter G, the result in Table 5.4 was obtained by training on a larger sample size of 43, while only a sample size of 17 was used for training in Table 6.4. This is supported by Mehmood et al. (2020) on how an increased number of samples would improve a model's performance.

6.6.6 Quantitative Analysis with Related Works with Background Interference

To further support setting a benchmark on SNN for MICR, the study investigated notable past works on OCR using an SNN on a small dataset, which attempted to recognise VSC having complex backgrounds. MEDPIX, with 25 samples per class, was used for the quantitative analysis with the network architectures in Wang and Lu (2017) and Koch et al. (2015). This experimental set-up's train: test ratio was 70:30, online augmentation by random rotation; each image was rotated up to 20 degrees, either clockwise or anticlockwise, and the attention module was inserted after the 3rd CNN layer. BO was not used for this experiment, as already existing models' architectures were used for the training and comparison. 30 runs were conducted, and the results were averaged. The results are presented in Table 6.6 below.

Table 6. 8**:** Comparison with related works on MEDPIX on 62-way 25-shot learning averaged over 30 runs.

Works	**Trainable Parameters**	**Accuracy (%)**
Wang and Lu, (2017)	50,184,000	83.12 ±0.34
Koch et al., (2015)	10,234,502	80.24 ±0.28
SIAM-MICR	187,406	87.36 ±0.64
SIAM-MICR + Attention	187,409	90.58 ±0.71

The results shown in Table 6.8 show that the proposed models require fewer parameters than existing SNNs from past works. Therefore, it can be agreed that the proposed models are more efficient than Wang and Lu (2017) and Koch et al. (2015) in the MICR task. Hence, the models are more memory efficient, and less computational power is an advantage, setting a benchmark for SNN on VSCs for MICR with small sample sizes. Table 6.8 is presented visually in Figure 6.11 below;

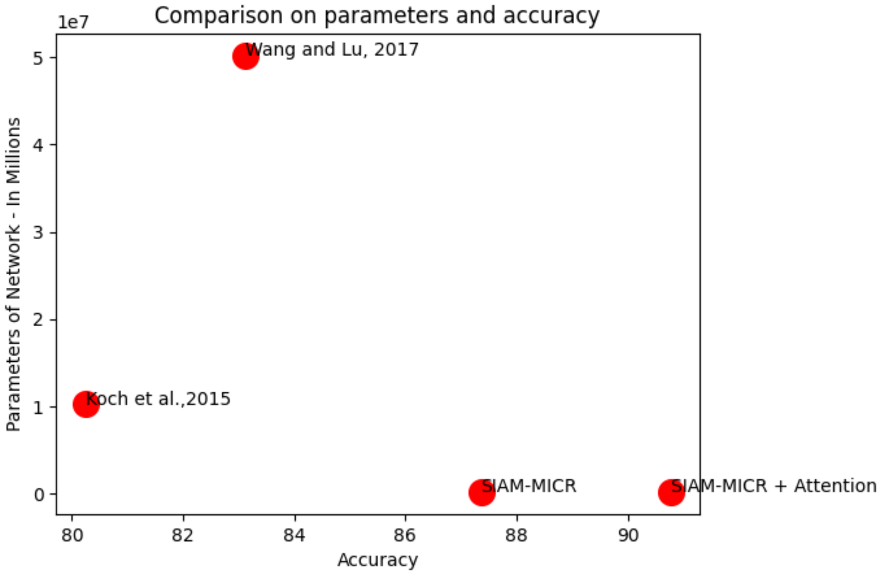


Figure 6.11: Comparison with related works in accuracy and parameter size

6.7 Chapter Summary

This chapter proposed a channel attention-based SNN suited for metric learning to adequately learn a discriminative pattern of individual classes for MICR, with the existing problems of low resolution and background interference. The experiments revealed that the channel attention module, inserted after the third convolutional layer, can perform better than the SNN without attention. There is an overall increase in accuracy, especially with a small sample size for a class with visually similar character images and reduced training parameters compared to related past works. Hence, the proposed models achieved good character recognition with less computational resources. Furthermore, the architecture of the proposed model is generic. It can be applied for any few-shot learning task, where there are cases of small sample size per class, visually similar images, and problems of low resolution with background interference.

In future work, I will consider generative modelling techniques. They may help increase the sample size per class so that a deterministic model can learn more features from low-resolution images at different image scales, as seen in Mishra et al. (2022) and Yuan et al. (2018). I will also consider transfer learning for metric learning, which leverages feature representations from a pre-trained model.

7.0 Generating Synthetic Training Data to Improve Character Recognition Accuracy using a Conditional variational autoencoder

7.1 Introduction

This chapter proposes a variant of the Variational Autoencoder (VAE) generative model, known as the Conditional variational autoencoder (CVAE), focusing on finding a solution to the small dataset problem in the medical image character dataset. This study proposes that increasing the dataset size will increase deterministic deep learning model diversity, model generalisation and better performance. Hence, this study aims to use supporting experimental evidence to investigate this proposal.

In the domain of OCR for burnt-in textual data in MIM, the problem of large dataset availability to train DL algorithms is a pressing issue. Unfortunately, DL classification models may perform worse when trained with small datasets because small datasets typically contain fewer details. Hence, the classification model cannot generalise patterns in training data. In addition, over-fitting becomes much harder to avoid as it sometimes goes beyond training data to affect the validation set (Rahman et al., 2017). Obtaining a large dataset is a major challenge due to privacy concerns in accessing medical images with patients' interpretations in burnt-in text and the significant cost associated with data acquisition and labelling for research. The available medical datasets used for training these OCR models are relatively small, significantly affecting models’ generalisation and performance.

This study aims to solve the problem of small dataset size by proposing a specialised generative model, Conditional Variational Autoencoder (CVAE), as an effective and practical data augmentation approach to synthesise data character images to improve the character recognition rate of deterministic models. In the proposed approach, the condition represents the label of the images, and the CVAE learns the probability distribution of image data conditioned on optimally determined latent variables and the corresponding labels. Bayesian optimisation determines the CVAE's architecture for the problem being investigated optimally. The trained CVAE model can be implemented as a data augmentation solution to synthesise low-resolution new images. Experimental results will demonstrate the approach's effectiveness on two independent medical image datasets consisting of open-source and privately collected images with textual interpretations.

7.2 Background

A generative modelling approach deals with models of distribution $P(X)$, defined over data points X in some potentially high-dimensional space (Doersch, 2016). An image can be termed a data point, and these data points describe the image (Cromey, 2012). The task of a generative modelling approach is to capture the dependencies between these pixels on how they are organised and the pattern they appear morphologically. The generative modelling approach allows numerical computation of the distribution $P(X)$ and, when trained successfully, can create new samples from the underlying distribution (Ruthotto and Haber, 2021). This approach can be utilised in the medical image character recognition (MICR) task to solve the problem of small dataset size and improve discriminative models' performance.

The overall task of this chapter of the study will be to collect the datasets (MEDPIX and PRIVATDT), which have samples X distributed according to an unknown distribution $P_{un}\ (X)$ and propose a model that is model P, that can learn and produce synthetic samples such that P is as similar as possible to get $P_{un}\ (X)$. The generated samples will augment the training dataset to improve the performance of deterministic models. This will improve the reliability of DL solutions in automated burnt-in textual data extraction, allowing better insights into MIM and pattern analysis and thereby easing data entry. Overall, this will improve healthcare service delivery and treatment plans.

Training generative models have major setbacks, such as strong assumptions of data structures (Sabuncu et al., 2010), making suboptimal approximations leading to substantially increased uncertainty (Beck et al., 2012) and relying on heavy computational inference methods (Bond-Taylor et al., 2022). These challenges prompt major advancements in using neural networks as powerful approximators due to their stability of numerical approximation (Tang and Yang, 2021; DeVore et al., 2020). The variational autoencoder (VAE) is a major approach due to its weak assumptions and reasonably fast training (Kingma and Welling, 2019).

VAE is a generative model whose training can be regularised to avoid overfitting and ensure that the latent space has good properties to enable the generative process. In contrast to the autoencoder training, where the input is encoded as a single point, the VAE is fed with the input as a distribution over the latent space. During training, a data point from the latent space is sampled, then the sampled data point is decoded, and the reconstruction error is calculated and backpropagated through the network. It assumes that the data is generated by some random process involving an unobserved continuous random variable z and that z is generated from some prior distribution

$P_\theta(z)$ and the data, X, is generated from some condition distribution $P_\theta(X|Z)$. z can be referred to as the hidden representation of data X (latent space). These encoded distributions are Gaussian distributions to enable the training to return the mean and covariance matrix (Liu et al., 2023). To form the Gaussian distribution in the latent space, the encoder neural network layers compute the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ from **x**.

For further clarification, below is a step-by-step explanation of how VAEs work (Doersch, 2016; Bond-Taylor et al., 2022) with a visual representation in Fig 7.1.

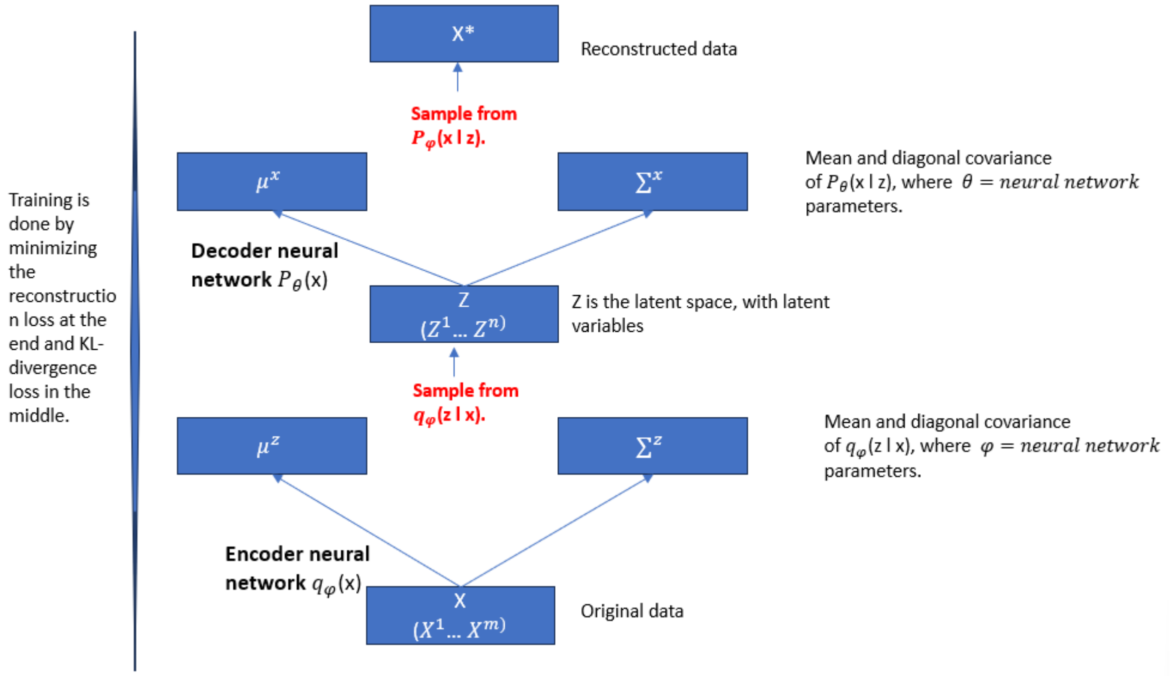

X*
Reconstructed data
Sample from P_φ(x l z).
μ^x
Σ^x
Mean and diagonal covariance of P_θ(x l z), where $\theta = neural\ network$ parameters.
Training is done by minimizing the reconstruction loss at the end and KL-divergence loss in the middle.
Decoder neural network P_θ(x)
Z $(Z^1 \ldots Z^n)$
Z is the latent space, with latent variables
Sample from q_φ(z l x).
μ^z
Σ^z
Mean and diagonal covariance of q_φ(z l x), where $\varphi = neural\ network$ parameters.
Encoder neural network q_φ(x)
X $(X^1 \ldots X^m)$
Original data

Figure 7. 1: VAE generative modelling process (source: author). Latent variables in the latent space are transformations of the data points into continuous lower-dimensional space and $Z^1 \ldots Z^n$ is less than $X^1 \ldots X^m$. In exploring the latent variable concept in terms of images, it is known that neighbouring pixels in an image are highly dependent on each other, as this determines the image's colour, size and layout. In this case, the latent variable is the underlying hidden features that determine the pixels and interactions in the original input image. These latent variables are not explicitly known. For more clarification, a 28*28 input image will have observed variables of $X^1 \dots X^{784}$ and latent variables of $Z^1 \dots Z^n$, and n is less than 784.

Step 1: Encoder Network maps the input data, denoted as $\boldsymbol{x}$, to parameters of latent space distribution. The output of the encoder network, denoted as $\boldsymbol{h_{en}}$, is computed as:

$$h_{en} = f_{en}(x)$$

where $\boldsymbol{f_{en}}$ represents the encoder network's transformation.

The encoder neural network outputs parameters that define a probability distribution for each dimension of the latent space (standard normal distribution). For each data point, the encoder neural network outputs a mean vector $\boldsymbol{\mu}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}$ (diagonal covariance simplifies the computation) for each dimension of latent space.

Step 2: Latent Space Gaussian Distribution: The latent space follows a multivariate Gaussian distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$**.** The neural network estimates these parameters. The Gaussian distribution is chosen due to its many advantages, such as faster analytical evaluation of the KL divergence in the variational loss and the ability to use the reparameterisation trick for more effective gradient computation. The reparameterisation trick allows the backpropagation during the training by approximating Z using the decoder parameters and another parameter, ϵ, where ϵ is a random noise to enable the stochasticity of Z. Mathematically, reparameterisation is shown below.

$$Z = \mu + \sigma . \epsilon$$

σ is the standard deviation of a Gaussian distribution derived from the decoder output.

The Gaussian distribution allows easy sampling of the latent space to generate new samples. $\boldsymbol{\mu^z}$ and $\boldsymbol{log}(\sum^{\boldsymbol{z}})$, is computed mathematically as:

$$\mu^z = \int_{\mu^z} (h_{en})$$

$$log(\sum^z) = \int_{\sum^z} (h_{en})$$

Step 3: Sampling: A random vector is sampled on the latent space to generate a sample from the latent space distribution.

Step 4: Decoder Network takes the sampled latent vector $\boldsymbol{z}$ and maps it back to the data space to reconstruct the original input. The output of the decoder network, denoted as $\boldsymbol{h_{de}}$, is computed as:

$$h_{de} = f_{de}(z)$$

where $\boldsymbol{f_{de}}$ is the decoder neural network's transformation.

Step 5: Reconstructed output, denoted as ***x*,*** is obtained by applying a suitable activation function $\boldsymbol{g_{de}}$ to the decoder output $\boldsymbol{h_{de}}$:

$$X^* = g_{de}(h_{de})$$

Step 6: VAE Loss Function: In VAE, the loss function combines reconstruction loss and Kullback-Leibler (KL) divergence loss. The reconstruction loss measures the difference between reconstructed output ***x**** and original data ***x***. It can be defined using a suitable distance metric such as mean squared error (MSE). The KL divergence loss quantifies the difference between the learned latent distribution and the assumed prior distribution.

The model is trained to minimise reconstruction error and Kullback-Leibler divergence (regularisation). The reconstruction error measures how much the decoder leans to reconstruct the samples from the latent distribution, and a higher error indicates the decoder's poor performance in the reconstruction of the data. The Kullback-Leibler divergence measures how much information is lost while encoding the data points into the latent space. It shows the difference between two probability distributions and quantifies how much extra information is needed to approximate the true distribution using an estimated distribution. Once the VAE is trained, it can generate new samples by sampling from the prior distribution and passing them through the decoder network.

However**,** suppose the latent space is too small or restrictive based on the chosen latent variables. In that case, the generated data may be limited and not similar in structure to the original data. On the other hand, if latent space is too large or too unconstrained, the generated data may be unrealistic or difficult to interpret. Choosing the optimal value of the latent variables is essential for the performance of VAE. A common approach is to use a search over a range of latent variable values and evaluate the model's performance on a validation set based on the loss functions. This study experimentally evaluated the optimal latent variables for modelling the latent space of the medical image character datasets, and the experimental results are presented in the result section of this chapter.

VAE is appealing as it is built upon standard neural networks and can be trained using stochastic gradient descent (Pu et al., 2017). An extensive literature search has shown its successful application in areas with optical character recognition (OCR), which motivated this study. Some of these relevant works with citations of over 3000+ each, Handwritten digit recognition (Rezende et al., 2014) and House number data recognition (Kingma et al., 2014)

Motivated by these works in digit and character recognition (Rezende et al., 2014; Kingma et al., 2014), this study aims to explore the VAE approach for data augmentation in solving the small dataset size problem in MICR, which affects the ability of classifiers to learn effective discriminative representation. However, the research on CVAE in low-resolution images and medical image character recognition is still very limited, as seen by extensive literature searches. This research problem can be better answered by asking these questions.

- How well can CVAE generate a synthetic image from a low-resolution medical image character sample of 96 dpi?
- How much accuracy improvement does the discriminative model show when trained and evaluated with the CVAE augmented dataset compared to the original dataset?

A literature search up to 10 years ago shows that the VAE approach has not been explored in MICR for burnt-in textual data. Hence, this study will provide a comprehensive guide for research into this area.

7.3 Conditional Variational Autoencoders (CVAE) and Limitations of VAE

The problem with the regular VAE is that there is no control over what kind of data is generated since generation is done by sampling grid points on the latent space. The decoder attempts to generate new data based on these points. The latent space is the hidden layer that contains the latent variables used to generate the outputs. To further understand this, if VAE is trained with a digit dataset, and I attempt to feed a label into the decoder to generate, the output will be randomly generated digits. Even if the training is good and the reconstruction loss is minimal, the output will remain randomly generated. Hence, there is a need to propose a conditional variant of the VAE, known as CVAE, as a more specialised generative model for the research problem. The CVAE process can be summarised as:

- If given an input label **L** and the expected output from the generative model will be **X**, the image.

So, the generative model process will be modified as follows:

Given label, L, z is drawn from the prior distribution $P_\theta(z|L)$ and the output X^* is generated from the $P_\theta(x|L,z)$, which is in contrast to the regular VAE, where the prior is $P_\theta(z)$ and the output is generated by $P_\theta(x|z)$. So, data X's encoding in the latent space is conditioned by L, and the generated data's decoding is conditioned by L. Fig 7.2 provides a simple visual explanation of the CVAE process.

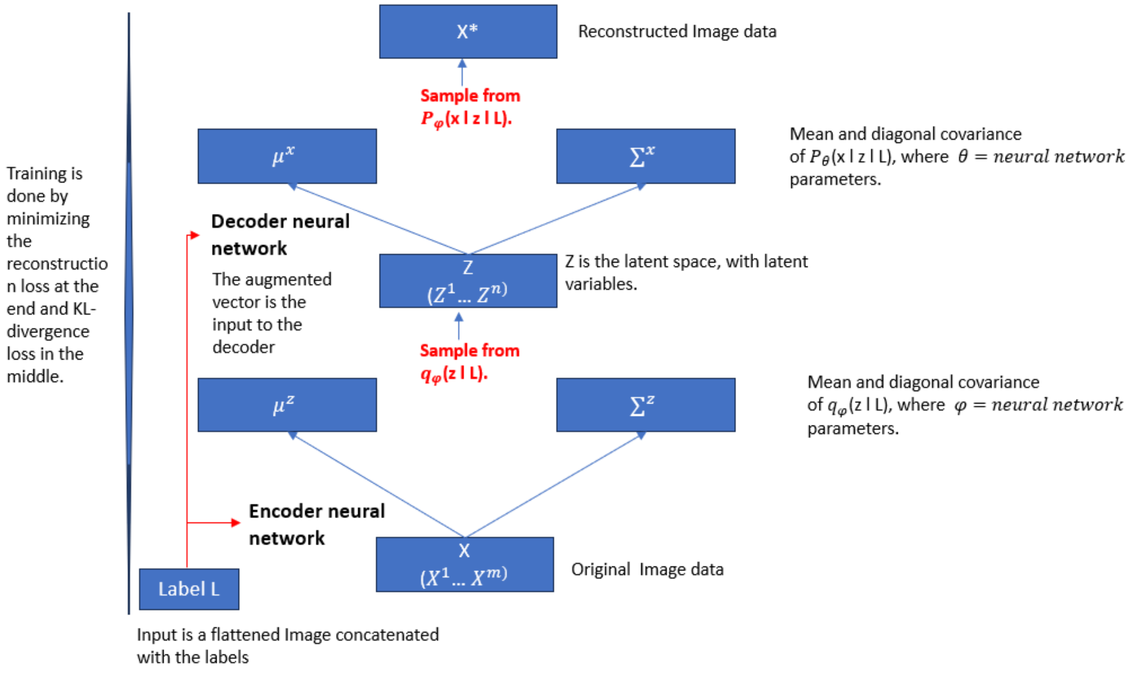

X*
Reconstructed Image data
Sample from P_φ(x I z I L).
μ^x
Σ^x
Mean and diagonal covariance of P_θ(x I z I L), where $\theta = neural\ network$ parameters.
Training is done by minimizing the reconstruction loss at the end and KL-divergence loss in the middle.
Decoder neural network
The augmented vector is the input to the decoder
Z
$(Z^1 \dots Z^n)$
Z is the latent space, with latent variables.
Sample from q_φ(z I L).
μ^z
Σ^z
Mean and diagonal covariance of q_φ(z I L), where $\varphi = neural\ network$ parameters.
Encoder neural network
X
$(X^1 \dots X^m)$
Original Image data
Label L
Input is a flattened Image concatenated with the labels

Figure 7. 2: CVAE generative modelling process (source: author).

Fig 7.2 shows training the CVAE model to learn the representation of the original image data by feeding the concatenation of the flattened image and the label as a one-hot encoding to the encoder. Using one-hot encoding increases the dimensionality of the data set, as a separate column is added for each category in the vector. The flattening converts the pixel values into a single continuous vector, which retains the spatial information of the original image but in an organised linear form. The label **L** is the condition which gives this variant the name “conditional”, and images can be generated by feeding the label into the decoder, and the model returns the specific data required. The input to the decoder is the concatenation of the normal distribution sampling corresponding to the latent variable Z and the label information. This allows the resampling range to be constrained in the specified label space rather than the entire normal distribution.

The CVAE model depends not only on the latent space for the encoding but also on the label information to encode other information, such as character stroke width, curves, and angles of these characters. To allow an easier comparison, Fig 7.1 shows the VAE, which takes input as image data only and produces probability distributions in the latent space, and the decoder takes sampled vectors in the latent space and returns generated image data. There is no control of the data generation process in the VAE, which is problematic when there is a need to generate specific data (Lavda et al., 2019). For example, if there is a need to generate a digit based on a query, rather than randomly sampling the latent space using a random vector, which may generate varying data. Suppose the query character is “5”; how do the VAE generate the images that are only character “5”? Random points have to be sampled and can give varied results. This is a major limitation of the VAE. The CVAE solves this problem, where labels are added to enforce the latent space to learn independent features per class (Pesteie et al., 2019).

This chapter aims to propose a generic generative model based on a CVAE for data augmentation for MICR, which can be modified and extended into other medical image data generation, especially in generating low-resolution image modalities. The architecture configuration was optimised using the Bayesian optimisation algorithm, where the objective function to minimise is the reconstruction error.

7.4 Justification of CVAE instead of GAN for this research problem

The choice of a model depends heavily on the problem and the data available. This study did an extensive literature review on generative models that can learn a low-dimensional representation of the medical image character patch and generate new data similar to the original data. In this case, the original textual data has the constraints of low resolution, blurry when enlarged, complex background, and tiny text. The closest option was the Generative Adversarial Networks (GANs); the justification for choosing CVAE is below.

- Regarding architecture and the ability to train with a small dataset, the GANs are difficult to train, as the discriminator quickly overfits the training data due to its deep layers. With the very deep layers, this overfitting causes a limited flow of feedback to the generator, which results in a training collapse due to the inability of the generator to learn salient feature representation. Hence, GANs require a large amount of training data for good performance. During training, the convolutional layer’s kernels are applied to the image's pixel according to a defined stride and padding, using element-wise multiplication, and the result is summed. This means that the output for a convolution of a pixel is a value that takes information from the pixel's neighbourhood. With the deep layers in GANs, applying more convolution operations as I move deeper into the GAN architecture may lead to a loss of information and overfitting when there is a small MEDPIXnd low-resolution images, leading to a collapse. For instance, the Vanilla GAN (Goodfellow et al., 2014), the simplest of all GANs, has 4 layers in the generator and 4 layers in the discriminator, making a total of 8 layers with a dropout regularisation. A more popular and efficient GAN architecture known as CycleGAN (Zhu et al., 2017) is composed of 2 GANs, making it a total of 2 generators and 2 discriminators, having with each generator have 7 layers, with 9 consecutive residual blocks, making a total of 14 layers, and 18 residual blocks, each of the discriminators has 6 layers, making a total of 12 layers. In total, the CycleGAN has a total of 26 layers and 18 consecutive residual blocks. These GANs are deep and pose a problem as this study's medical image character dataset is small; hence, training these GANs using this small dataset typically leads to discriminator overfitting, causing training to diverge. Performing MEDPIXugmentations (such as rotation) causes a case of augmentation leakage, where the GAN generates the augmented images that have been rotated.

- GANs have been used on medical images, magnetic resonance images (MRIs), computed tomography (CT), X-ray, and positron emission tomography (PET) and demonstrated promising results. However, these studies involved a quite large training set, averaging above 1000 training samples per class (Chadebec et al., 2022), whereas, in the present task of MICR, it remains very challenging to gather such large cohorts of labelled medical imaging modalities with burned-in textual representation on it due to privacy reasons and acquisition cost. Compared to GANs, CVAE can learn an efficient latent space to generate new data samples from a small dataset (Clément and Stéphanie, 2021), which further justifies the choice of CVAE for this study.

- CVAE generates mainly low-resolution images from low-resolution inputs, which are similar to the original dataset of this study (He, 2023; Chen and Guo, 2023), while studies have shown GAN generate high-resolution images from low-resolution inputs (Wang et al., 2023; Aggarwal et al., 2021). In deep learning, it is important that the training and testing datasets come from the same distribution, and it is also ideal that the inputs, once the model is deployed for practical use, come from the same distribution. Otherwise, the model's predictions and quality will be highly inaccurate. This means that models are useless for inputs far from the training data in terms of distribution. An empirical study by Alkhalifah et al. (2023) supports this viewpoint that a different distribution of the synthetic data may lack many realistic features embedded in the original data, which results in poor performance of the trained neural network model during inference. Therefore, ensuring that the synthetic images generated by the chosen generative model are as close as possible to the original data and have a similar distribution is highly necessary. A low-resolution image has smaller pixels with less than 300 pixels per inch, which is the opposite of a high-resolution image with more than 300 pixels per inch. The original dataset in this study is low-resolution images of 96dpi, as confirmed using the pillow library. CVAE produces these same low-resolution images and is a more appropriate choice for this study.

This study argues that CVAE can reliably be used for data augmentation in MICR to improve the deterministic models' performance during inference by optimally modelling the architecture and latent space and amending how the data is generated through conditioned sampling.

7.5 Models' Architecture

The baseline model is a classification model used to evaluate the effect of the proposed augmentation method for MICR. Fig 7.3 shows the architecture of the base model, which is the MICR classification model from the previous chapter. The networks are initialised with random weights and trained based on a train-test split ratio of 70:30.

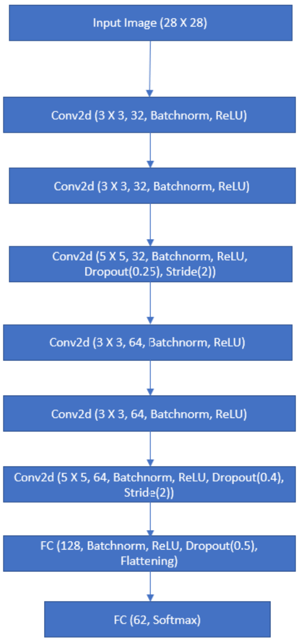

Input Image (28 X 28)
Conv2d (3 X 3, 32, Batchnorm, ReLU)
Conv2d (3 X 3, 32, Batchnorm, ReLU)
Conv2d (5 X 5, 32, Batchnorm, ReLU, Dropout(0.25), Stride(2))
Conv2d (3 X 3, 64, Batchnorm, ReLU)
Conv2d (3 X 3, 64, Batchnorm, ReLU)
Conv2d (5 X 5, 64, Batchnorm, ReLU, Dropout(0.4), Stride(2))
FC (128, Batchnorm, ReLU, Dropout(0.5), Flattening)
FC (62, Softmax)

Figure 7. 3: MICR Model
(from Chapter 5, source: author)

To show the effect of the augmentation data generated by the proposed CVAE model, the training data will be mixed with certain synthetic samples and used to train the MICR model. The model's accuracy will be presented to determine how much improvement can be achieved using CVAE-augmented character image samples for training models for MICR.

The proposed CVAE model is motivated by the notable study proposing a Gaussian Stochastic CVAE by Sohnet et al. (2015). Their CVAE model was deep with 21 layers, and the aim was to generate an output y, conditioned on z. The authors experimented with corrupted input data and attempted to reconstruct a clean copy of the data, and the results were convincing. Hence, this study made significant architectural changes to design a CVAE that is suited to the problem of small dataset size in MICR.

This chapter proposed a model shown in Figure 6.4, comprising two dense layers in the encoder and three in the decoder. The non-linear activation for each layer is ReLU, and the output is activated via the sigmoid function. The dimension of the latent space is set to 2 after investigations to determine the optimal dimension by checking the minimum reconstruction loss for each value of latent variables, averaged after 50 iterations each of 1000 epochs. The value with the lowest reconstruction loss is the optimal latent variable size for the CVAE model. The Bayesian optimisation (BO) technique was applied to determine the optimal architecture's hyperparameters by taking the objective function as the reconstruction loss and balancing the exploration-exploitation trade-off of hyperparameter search space. BO has the advantage of reducing computational cost by strategically choosing configurations to evaluate based on an informed approach rather than a random, expensive tunning process. In this work, the proposed CVAE architecture can learn independent features per class for small dataset sizes, increasing the model's ability to learn from limited information from low-resolution samples by introducing a dense-only connected layer in both the encoder and decoder parts. This yields improved approximation to allow concatenation of the label encoding, allowing the model to capture richer discriminative features from the input images.

To ensure consistency, this architecture is maintained through the experiments. See Figure 7.4 below.

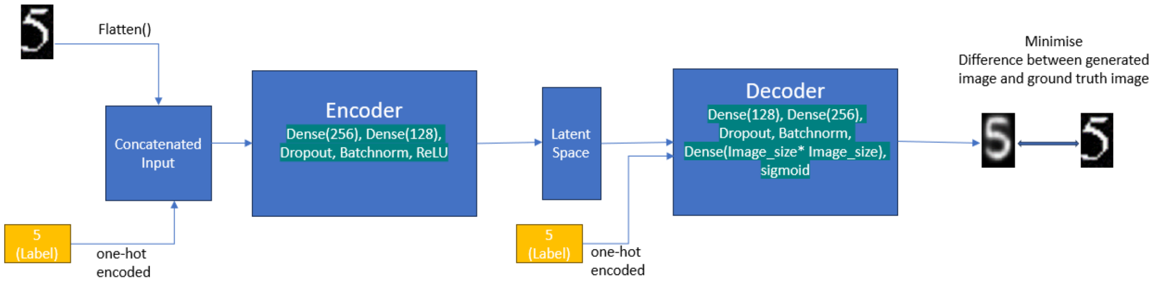

Flatten()
Concatenated Input
5 (Label)
one-hot encoded
Encoder
Dense(256), Dense(128), Dropout, Batchnorm, ReLU
Latent Space
5 (Label)
one-hot encoded
Decoder
Dense(128), Dense(256), Dropout, Batchnorm, Dense(Image_size* Image_size), sigmoid
Minimise
Difference between generated image and ground truth image

Figure 7. 4: Proposed CVAE Architecture after Bayesian Optimisation with a Dropout of 0.3. *Please see Figure 7.2 for the training and generative process.*

Figure 7.4 provides a visual representation of the proposed CVAE architecture. The input, consisting of the original image and a label encoded as a one-hot vector, is fed from the left, and the generated images are extracted on the right. The encoder processes the input image and the conditional variable (labelled as a one-hot encoded vector). The latent space, represented by the central box, captures the relevant features of the input data given the condition (label). The decoder then takes the latent representation and the same conditional variable to generate an output that matches the data and the condition. This unique capability allows us to generate data that meet a specific condition, which is the label, and it enables the model to learn additional information about the input images based on their classes. Including labels in the learning process allows the model to learn conditional relationships, thereby enhancing the control over the data generation process. This practical implication of the proposed architecture underscores its potential in various real-world applications.

7.6 Results and Evaluation

This chapter's experiments were conducted using a Python 3 Google Compute Engine backend, 12.7 GB system RAM, and TensorFlow and Keras libraries. Datasets described in Section 4.3 were used for this chapter.

7.6.1 Latent variable investigation

This study conducted several experiments to find the optimal latent variables for the latent space that gives the lowest reconstruction loss of the input image based on the following procedures.

1. Prepare the datasets, resizing the image size to 28x28 and manually checking that each class is represented in the train-test ratio of 80:20.
2. Normalise the dataset by dividing each pixel value by 255 to scale the values to the range [0, 1]
3. Initiate the CVAE model based on the optimised architecture configuration and a starting point of 2 as the latent variables.
4. Train the CVAE for the generated output image and calculate the overall minimum reconstruction loss.
5. Choose a different value for the latent variables for the CVAE model. Redo (4) to find the optimal value of latent variables based on the minimum reconstruction loss comparison.
6. Online data augmentation was not applied for this experiment to maintain a high-quality dataset and ensure the CVAE model learned representative latent variables that can generate outputs as close as possible to the original image.

The results of this procedure are shown in Table 7.1 below.

Table 7. 1: Latent variables and Reconstruction loss for the CVAE model

Latent variables	**Minimum Reconstruction Loss (MEDPIX)**	**Minimum Reconstruction Loss (PRIVATEDT)**
2	27.03 ±0.02	13.23 ±0.11
3	28.22 ±0.03	14.58 ±0.04

4	28.78 ±0.04	14.69 ±0.02
5	28.65 ±0.02	14.65 ±0.04
6	28.72 ±0.03	14.87 ±0.04
7	28.86 ±0.14	14.77 ±0.10
8	29.62 ±0.39	15.29 ±0.07

The ± represents the standard deviation, which ranges between 0.02 and 0.39. This low standard deviation indicates that data are clustered tightly around the mean, less dispersed, and therefore more precise. This is further presented visually in Figure 7.5 below.

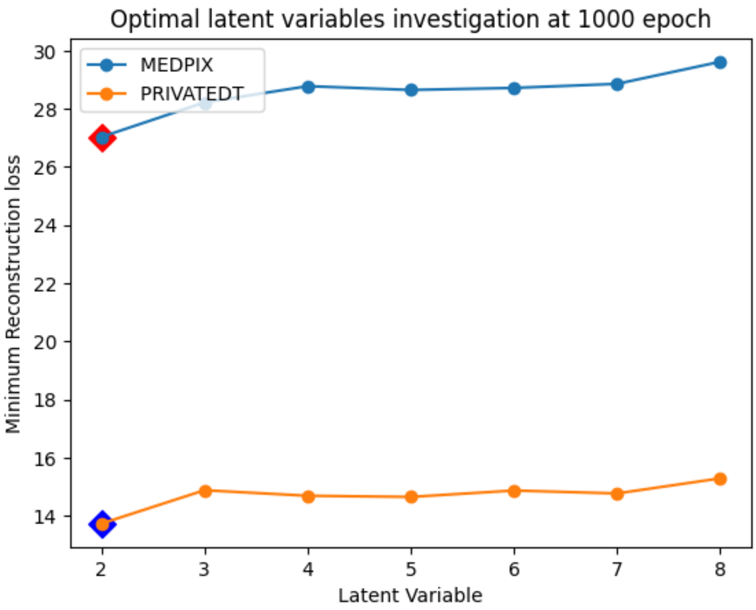


Figure 7. 5: Optimal Latent Variables Investigation. Two (2) latent variables are shown in **red and blue** as the best choice for the CVAE model latent space configuration on the two datasets. The figure shows that the minimum reconstruction loss increases as the latent variables increase.

This study did not include 1 latent variable size, as it may be insufficient to effectively encode the image’s pixel relationship and the one-hot labels' one-hot encodings.

When the latent space size is limited to 1, it becomes clear that the space is too small to encode the dataset effectively. The CVAE model finds filling the concatenated data information of the input image and label one-hot encoding into the 1 latent variable data space challenging, leading to an inconsistency between the encoded and latent space distribution. This may result in a large reconstruction error for the decoded data points and a failure to cover the entire data distribution with the samples from the latent space. Increasing to large latent variable sizes may result in collapsed dimensions due to the encoder predicting a mean 0 and unit variance for the Gaussian, and this increases the reconstruction error, leading to an unnecessarily large model, which can be improved by finding the optimal variable value for the latent space with fewer parameters and less computational requirement. Relevant past studies show that going above 8 may result in using large latent variables, which greatly complicates training and convergence modelling (Ji and Lu, 2021). The latent space is visualised in Figure 7.6 below.

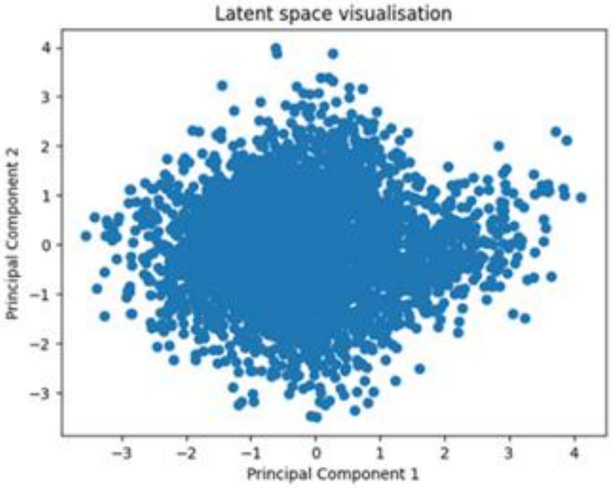
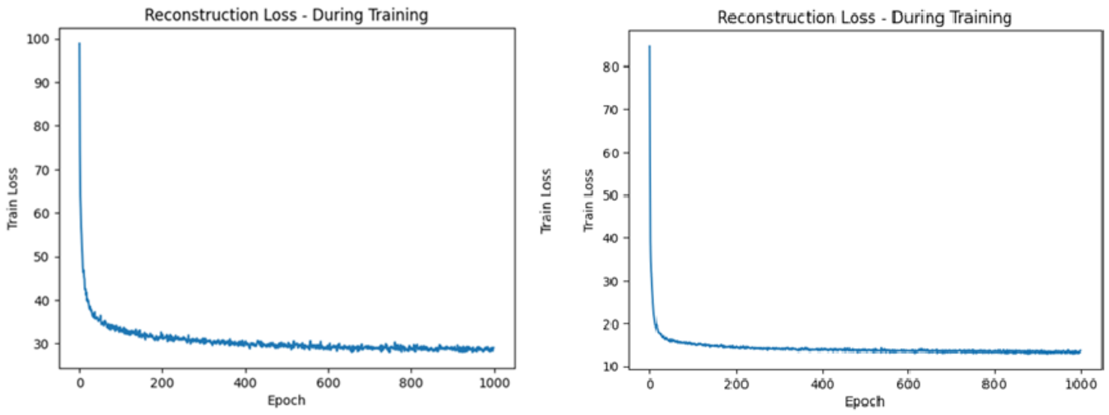


Figure 7. 6: Scatter plot of latent space. Two latent variables only to do principal component analysis, and the variance captured by the two PCs in percentage - 58.34% and 41.66%.

As seen in Figure 7.6, the ideal goal is for encodings to be as close as possible to each other while being distinctly separated. This allows smooth interpolation and enables the generation of new image samples. The KL divergence loss function optimises the probability distribution to be close to that of the target distribution, thereby reducing the reconstruction error. This loss function enables the distribution of all encodings of the classes for the characters evenly around the centre of the latent space, as seen in the latent space visualisation in Figure 7.6. Thus, it follows a Gaussian distribution with most values concentrated and clustered around the centre region of the curve. Addition of the class label as a condition enforced the latent space to learn independent features per class. This is particularly noteworthy, as the training was done using a small dataset, yet the latent space was able to learn independent features per class effectively. The latent space maintains the similarity of nearby encoding by clustering the data points locally while globally packed at the latent space origin. This arrangement allows for a smooth mix of features when interpolating to generate new image samples. In Fig 7.6, the two latent variables were used to do a principal component analysis, and summing up the variance values in this array for each of the principal components [0.5834254, 0.4165746]. it is equal to the explained variance ratio of 1.0000, which measures the relative variance amount explained by each of the principal components, thus, indicating that the two principal components together explain 100% of the variance of the data.

The training curve based on two latent variables is shown in Fig 7.7 below.



(a) (b)

Figure 7. 7: Minimum reconstruction loss curve (a) MEDPIX (b) PRIVATEDT

Figures 7.5 and 7.6 show that the MEDPIX has a higher reconstruction error than the open-source dataset, PRIVATEDT. This may be due to noise associated with background interference, poor background contrast, and irrelevant distribution areas at the pixel level, such as white noise images (Pividori et al., 2019) from the publicly accessible images compared to the private onsite data collection done by this research.

7.6.2 Generated Character Images – Samples

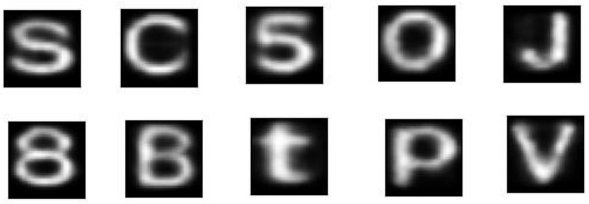


Figure 7. 8: Generated character images

The samples of generated character images shown in Fig 7.8 were chosen randomly by querying the generative process with the required character. The repeated images show the stability of the CVAE model in generating image data by simply entering the condition for the generative action, that is, the label. The generated image has a low resolution of 96 dpi, the same as the original image data. This was confirmed using the pillow image library.

7.6.3 Qualitative Analysis - Augmenting Datasets with Synthetic Images

This section provides a qualitative analysis of the synthetic images generated by the proposed CVAE and investigates the feasibility of augmenting medical image character image datasets with synthetic images. The augmented datasets are used to train the MICR models (discriminative models) to determine whether they achieve higher accuracy than models trained with original character images alone. The default sample size for each class is shown below in section 4.3 in the data description. The augmentation process involves adding **N**, the number of synthetic images in each class, to increase the total size of the dataset. Table 7.2 reports the following result, which investigates the impact of the generated samples on the accuracy of the MICR classification models based on a Train-Test ratio of 70:30. This experiment aims to reveal the impact of different augmentation sizes on the deterministic model's accuracy in MICR.

Table 7. 2: Accuracy of MICR(1) model on augmented datasets—averaged on 20 runs.

	Number of synthetic images per class (N)			
	0	**50**	**100**	**150**
MEDPIX	87.13 ±0.18%	90.33 ±0.12%	90.63 ±0.10%	88.92 ±0.02%
PRIVATEDT	91.42+0.14%	93.83 ±0.02	98.27 ±0.06	93.02+0.06%

Table 7.2 is shown visually in the chart below in Figure 7.9 below.

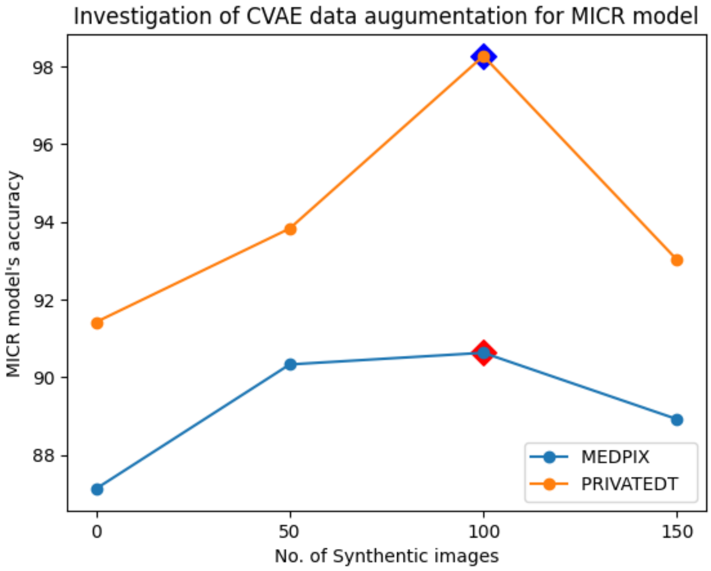


Figure 7. 9: Accuracy vs Sample Size Augmentation for both datasets

In the training set, multiple synthetic images were generated by sampling different vectors based on the label's condition by querying the required character label. These generated images were used to augment the training set, according to Table 7.2. On MEDPIX and PRIVATEDT, the accuracy of the models trained with augmented datasets increases with respect to those trained without synthetic images. Compared to Table 5.3, the results are different, as Table 7.2 uses a train-test split ratio of 70:30, compared to 80:20 in Table 5.3 for the MICR(1) model. I decided to increase the testing ratio in Table 7.2 so that more original data samples can be used to evaluate the models trained. The test set consists of original data samples only to evaluate the model’s effectiveness in a real-world case. For the training set, the synthetic images are generated and selected based on visual quality. The number of synthetic images, according to Table 7.2 (50, 100 and 150), is added to the training set. For instance, N = 50 means the training set consists of 70% original images with 50 extra synthetic images per class.

Maximum accuracy improvements of +3.2%, +3.5% and +1.79% were obtained when 50, 100, and 150 synthetic images per class were added to the Medpix dataset, respectively. For PRIVATEDT, an increase of approximately +2.41%, +6.85%, and +1.60, when 50, 100, and 150 synthetic images per class were added, respectively. This shows that the privately collected dataset benefitted more from this data augmentation approach.

These results are consistent with the intuition that adding synthetic images to smaller datasets should result in more significant improvement than adding them to larger datasets (Anderson et al., 2022). The results in Table 7.2 also suggest an optimal balance between the number of original and synthetic images per class in the dataset. Adding 150 synthetic images per class to the dataset resulted in a lower accuracy than adding 100 synthetic images per class since the proportion of original images per class becomes smaller. Furthermore, these results show that adding synthetic images to smaller datasets improves the predictive accuracy of deterministic models.

7.6.4 Evaluation of individual classes' Improvements

Given that the classes are very imbalanced in both datasets, it would be insightful to investigate the impact of the synthetic data augmentation on individual classes and determine how well the performances are on small and oversized classes, respectively. The augmentation was done by adding 100 synthetic images per class, and the classification report showing the precision, recall, and F1-score evaluation metrics was reported together with each class’s sample size. The synthetic data was added only to the training data, and evaluation was done on the original data. The results for MEDPIX are summarised in Table 7.3 below.

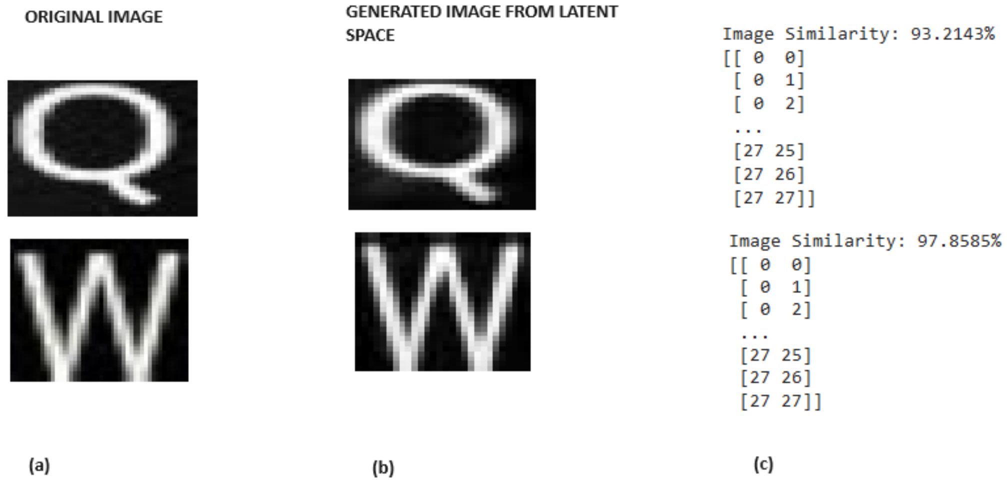
Table 7.3. Investigation of Individual Classes' Improvements

MICR model					**MICR + Synthetic Data Augmentation**				
Class	**Sample Size**	**Precision**	**Recall**	**F1-score**	**Class**	**Sample Size**	**Precision**	**Recall**	**F1-score**
v	5	0.00	0.00	0.00	**v**	105	1.00	1.00	1.00
w	6	0.00	0.00	0.00	w	106	0.80	1.00	0.89
j	7	0.00	0.00	0.00	j	107	1.00	1.00	1.00
S	111	0.92	1.00	0.96	S	211	1.00	1.00	1.00
T	140	1.00	1.00	1.00	T	240	1.00	0.93	0.97
2	101	0.83	0.95	0.89	2	201	1.00	1.00	1.00

Smaller classes (“v,” “k,” and “j”) and big classes (“S,” “T,” and “2”) were taken as sample instances, and the improvement was seen irrespective of previous class sizes. However, the smaller classes benefitted more from the augmentation, as shown in Table 7.3.

7.6.5 Investigation of Impact classes with small sample size on latent space

This study opines that investigating the pixel difference and similarity between the original input image and the generated image from the latent space would reveal how much information the CVAE model learns in encoding these small sample-size classes. This is done by loading the images and checking their structural similarity. A higher percentage of similarity shows that the CVAE model can learn from these small-size classes and generate highly similar images with minimum pixel difference. The results of these classes with small size, being compared with the generated output from the latent space, are shown in Figure 6.10 below.



```
Image Similarity: 93.2143%
[[ 0  0]
 [ 0  1]
 [ 0  2]
 ...
 [27 25]
 [27 26]
 [27 27]]
```

```
Image Similarity: 97.8585%
[[ 0  0]
 [ 0  1]
 [ 0  2]
 ...
 [27 25]
 [27 26]
 [27 27]]
```

Figure 7.10: Similarity comparison of images from small classes.

Figure 7.10 shows that there is a high structural similarity between the (a) original characters “Q” and “W”, at 93.21% and 97.86% approximately and (b) the generated image from the latent space. The arrays in (c) below the similarity score show the different pixels between the two images. This structural similarity technique of comparing local pixel difference and intensity is supported by the notable work of Wang et al. (2014). It can be implemented using the sci-kit-image Python library.

These results comprehensively highlight the benefits of data augmentation in improving deterministic model performance with small-size datasets using a specially designed CVAE approach.

7.7 Chapter Summary

This chapter has explored the feasibility of augmenting medical image character datasets with realistic synthetic character images for medical image character recognition under a small dataset problem with low-resolution images. Specifically, this chapter proposed a CVAE model that can generate realistic synthetic character images from latent variables encoding attributes and decoding via fully connected layers. The output of the model has the same distribution in terms of low resolution as the original medical image modalities of 96 dpi. This model’s architecture experimentally has resulted in faster convergence during training, reducing training time compared with a deeper generative counter of the GANs family. Several CNN-based MICR models with different combinations of real and synthetic images were trained to demonstrate the benefit of augmenting small datasets with the proposed method. Results show that the discriminative models trained with the augmented datasets outperformed those trained with original images alone. Compared to other character image synthesis methods explicitly designed to generate character images, the proposed generative method is more generic and can generate low-resolution images when the available dataset is small. The proposed method is useful for generating low-resolution MIM, as seen in the performance analysis of the two datasets MEDPIX and PRIVATEDT, which represent the open-source and originally collected images.

8.0 Conclusion

8.1 Overview

This PhD thesis aimed to explore and develop efficient DL-based solutions to recognise burned-in textual data in medical imaging modalities under the constraints of low resolution, background interference, tiny text, and small datasets. This chapter concludes this research by reviewing the findings, generalisation, significant contributions, limitations, and future works. The current works on recognising burned-in textual data are still very limited, and many of these reasons are presented comprehensively in Chapter 3.

8.2 Research Findings

The primary research goal of this thesis was to develop DL techniques for MICR, with the constraints explained in the previous section. A thorough review of the existing literature highlighted significant research gaps, which served as a primary source of motivation for this research. These gaps were presented comprehensively in Chapter 3, where the open issues and challenges in MICR were discussed; the content of this section was published in Osagie et al. (2023) as a review article. Experiments were conducted for each primary research question, and the research findings answered these questions through critical interpretations and investigation of the experimental results. The summary of the research findings are summarised in Table 7.1 below.

Table 8. 1: Research questions and findings from experimental investigations.

Research Questions	**Research Findings**

1. Can a deep learning-based solution be designed to recognise burned-in text data with small font sizes, low resolution, and background interference in varied medical image modalities?	An enhanced and Bayesian-optimised CNN model based on Lenet-5 architecture was designed to recognise burned-in text data in varied medical imaging modalities, with the constraints of low resolution and background interference. The proposed model can recognise burned-in textual data in low-quality medical images with a low resolution of 96 dpi and a small font size. A majority voting ensemble model was designed based on the proposed model through a series of investigations to determine the best number of members. This can reliably recognise burned-in textual data in varied medical image modalities. The evaluation used publicly and privately collected datasets with varied imaging: ultrasounds, X-rays, MRI, and CT. Results showed improvement in the proposed model compared to existing works and the Lenet-5 classical OCR model. The investigation showed further improvement in the ensemble's performance due to reduced variance between members.
2. Can a deep learning-based solution based on	A Siamese neural network based on a twin CNN with a channel attention module was designed to

few-shot metric learning be designed to recognise visually similar character images with a small dataset sample size in varied medical image modalities?	employ metric learning with contrastive loss to recognise visually similar characters with a small data sample size. The evaluation used publicly and privately collected datasets with varied sample sizes. Results show that the proposed Siamese neural network can classify visually similar characters without difficulty based on their metric distance, even with the small sample sizes.
3. Can generative modelling be proposed to improve burned-in text data recognition by generating synthetic data samples for each character?	A generative model based on the conditional variant of the variational autoencoder was designed to improve MICR accuracy by generating synthetic data samples based on a Bayesian optimised architecture and best latent variables based on experimental investigations. The evaluation used publicly and privately collected datasets with varied sizes to add the synthetic data samples. Results reveal a high similarity between the original and generated images from the latent space, even for small sample size classes. The result also showed improved performance in deterministic models when trained with generated data samples from the proposed generative model.

8.3 Research Limitations

The limitations of this research are :

- Firstly, the hyperparameter optimisation techniques were chosen based on the justification of their effectiveness, speed, and flexibility in implementation in modelling complex relationships in data. This was used across all the modelling done in the three technical chapters of this thesis. However, experimentally, it is unclear if different hyperparameter optimisation techniques would be more appropriate for the different modelling methods. There was time constraint in experimentally testing several methods of optimisation, as they are a lot of varieties.
- DL-based solutions have revolutionised image analysis in various domains, but implementation is computationally and resource-intensive. This research addressed this limitation by implementing a commercial Google Colab platform using TPUs and GPUs. However, accessibility to a high-powered computing platform would have allowed for more training, optimisation and evaluation experimentation. This research made optimum use of the available resources to achieve the results presented in this thesis.
- The limitations of the dataset, as previously discussed, are that the DL-based solution requires a large amount of training samples per class to prevent over-fitting during model training. The most successful results in DL are thousands to millions of samples per class, which allows for efficient representation learning. However, using a small dataset in this research is, at best, difficult and, frequently, challenging.

8.4 Research Contributions to Knowledge

The main contributions of this thesis are :

- This study introduces an enhanced CNN model motivated by the classical Lenet-5 model. The enhanced model is optimised using Bayesian reasoning. The Lenet-5 uses a filter size of 5x5 in its first convolutional layer, followed by average pooling. In this study, these are replaced by a 3x3 filter size and a 5x5 filter size with a stride of 2, respectively. This enhancement ensures that the proposed CNN model can learn more local features, essential in designing a MICR solution for low-resolution MIM with background interference.
- Performing MICR on burned-in textual data at a low resolution of 96 dpi with background interference. An outstanding accuracy score was achieved, and MICR at such low resolution has not been previously reported in the literature.
- To analyse the impact of the ensemble technique on the enhanced model's performance. The bootstrapping method was used to create three (3) subsets of the dataset. A classifier is fitted to each of these subsets and evaluated. An ensemble is designed using the trained classifiers of these subsets, and a final classification outcome is based on a majority voting algorithm. This improves the model's performance in distinguishing visually similar characters.
- This study proposes a Siamese neural network to learn semantic similarities between extracted embeddings of image pairs in metric space in the presence of small sample size, low image resolution, and background interference. The resulting model can discriminate between visually similar characters by learning a fine-tuned decision boundary.

- This study proposes a channel attention mechanism combined with a Siamese neural network to learn meaningful parts of an input image that discriminate when compared to a visually similar image.
- This study provides a benchmark for using similarity learning in medical image character recognition (MICR). After an extensive literature review in the past 10 years, no previous work has been done regarding MICR and SNN with and without channel attention mechanisms. Extensive reviews conducted by this study by searching notable databases such as Elsevier, IEEE, Nature, and Science did not reveal any existing work for medical image character recognition that used the Siamese network. There are no reviews that discuss or propose Siamese-based methods for MICR.
- This study proposes a generic generative model based on a conditional variational autoencoder for data augmentation for MICR. This model can be modified and extended to generate other medical image data, especially low-resolution image modalities. The architecture configuration was optimised using the Bayesian optimisation algorithm, where the objective function to minimise is the reconstruction error. The best latent variable dimension was determined via experimental investigations, which considered generating the same distribution of data samples as the training dataset.

8.5 Research Significance

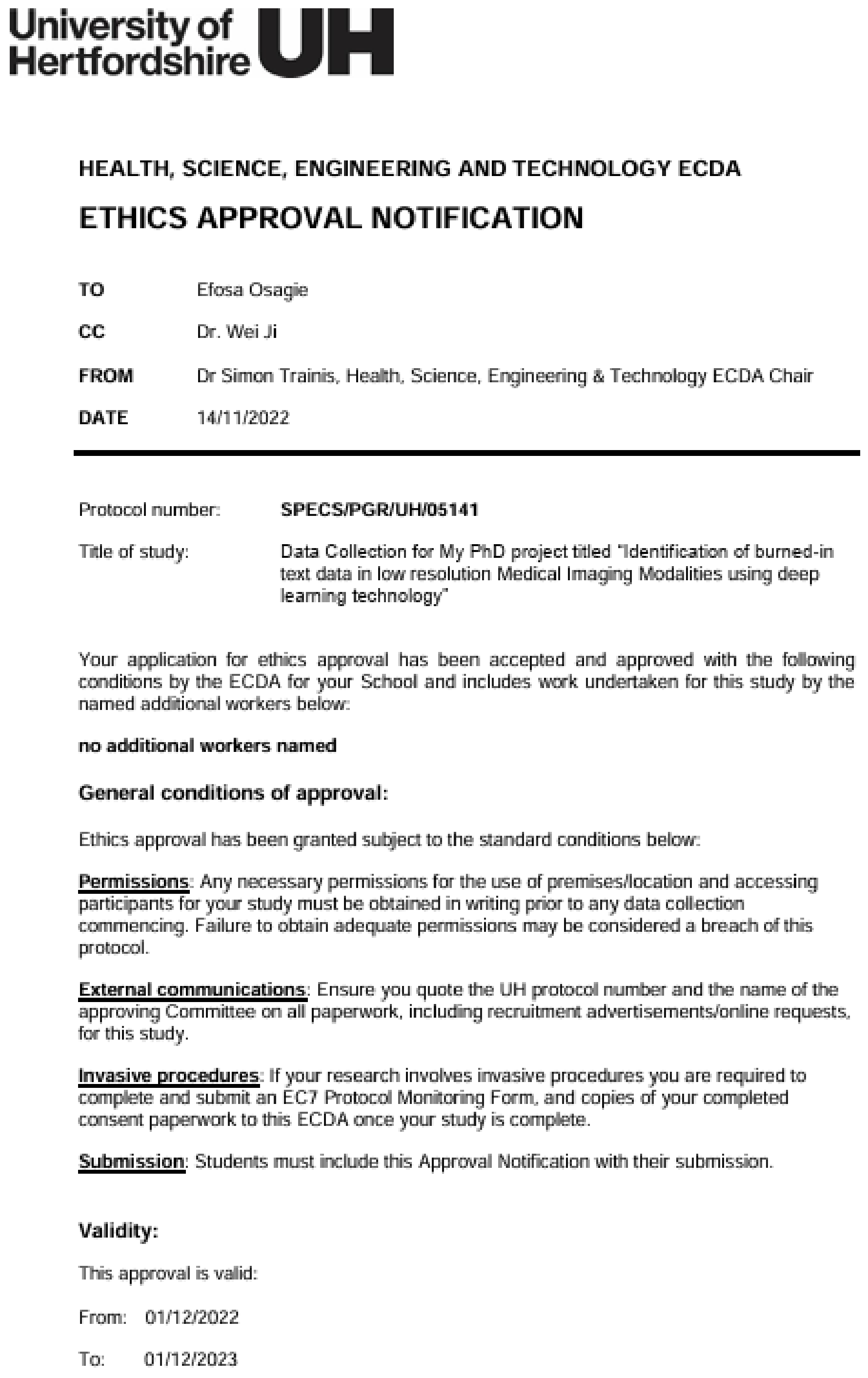
A reliable and automated DL-based solution for recognising burned-in textual data from MIM will improve medical informatics by significantly improving the speed, accuracy, and management of medical data entry systems. It will also allow easy integration of heterogeneous data from multiple sources, including burned-in textual data on MIM, in making critical decisions in diagnosis, prognosis, and patient treatment plans.

Regarding the generalisability and transferability of these research findings, the modelling approach could be applied to image retrieval systems, image anonymisation, and other cases where text recognition is required in low-resolution images such as historical text documents and degraded text documents, amongst others. This research demonstrated the importance of applying specialised network architecture and small kernel size in CNN networks for character-wise recognition in low-resolution images and background interference. The results from the enhanced Lenet-5 CNN model corroborated this. Having domain knowledge of the images will contribute more meaningfully to the network architecture and hyperparameter configuration design and improve character recognition performance. Another fact that this research highlighted was the application of few-shot learning techniques, such as the Siamese neural networks, in providing an algorithm that can learn a similarity function from a dataset having small sample sizes and yet produce more meaningful results compared to conventional multi-class classification algorithms trained on the same dataset. The evaluation results highlight that few-shot learning and Siamese neural networks are a solution to low-resource domains, where gathering large sample sizes for class in a dataset is challenging, such as satellite imaging, electronic health records, and degraded and historical documents. Following the application of CVAE in the augmentation of the medical imaging dataset, this research highlights the importance of using a training set augmented with synthetic images up to a determined peak to improve the classification performance of deterministic models trained on augmented training sets. This thesis has shown that these techniques; specially designed network architecture, few-shot learning, and CVAE, are possible solutions to the problems of low-resolution images with background interference and small sample sizes per class, and results from this thesis provide an early investigation into these techniques as regards imaging analysis

8.6 Future Outlook

Even with research limitations such as data acquisition, time constraints, and limited past works in this domain, this thesis provides an early investigation into medical image character recognition based on DL techniques. In future work, a multi-scale CNN architecture and a more advanced ensemble will be considered. This research will also consider multi-scale modelling techniques in metric learning with Siamese neural networks; they may help learn more features at different scales from low-resolution images. Transfer learning for metric learning to leverage feature representations from a pre-trained model will also be considered.

Appendix A – Data Collection Approval (UH Ethics Committee)



University of Hertfordshire UH

HEALTH, SCIENCE, ENGINEERING AND TECHNOLOGY ECDA

ETHICS APPROVAL NOTIFICATION

TO	Efosa Osagie
CC	Dr. Wei Ji
FROM	Dr Simon Trainis, Health, Science, Engineering & Technology ECDA Chair
DATE	14/11/2022

Protocol number:	**SPECS/PGR/UH/05141**
Title of study:	Data Collection for My PhD project titled "Identification of burned-in text data in low resolution Medical Imaging Modalities using deep learning technology"

Your application for ethics approval has been accepted and approved with the following conditions by the ECDA for your School and includes work undertaken for this study by the named additional workers below:

no additional workers named

General conditions of approval:

Ethics approval has been granted subject to the standard conditions below:

Permissions: Any necessary permissions for the use of premises/location and accessing participants for your study must be obtained in writing prior to any data collection commencing. Failure to obtain adequate permissions may be considered a breach of this protocol.

External communications: Ensure you quote the UH protocol number and the name of the approving Committee on all paperwork, including recruitment advertisements/online requests, for this study.

Invasive procedures: If your research involves invasive procedures you are required to complete and submit an EC7 Protocol Monitoring Form, and copies of your completed consent paperwork to this ECDA once your study is complete.

Submission: Students must include this Approval Notification with their submission.

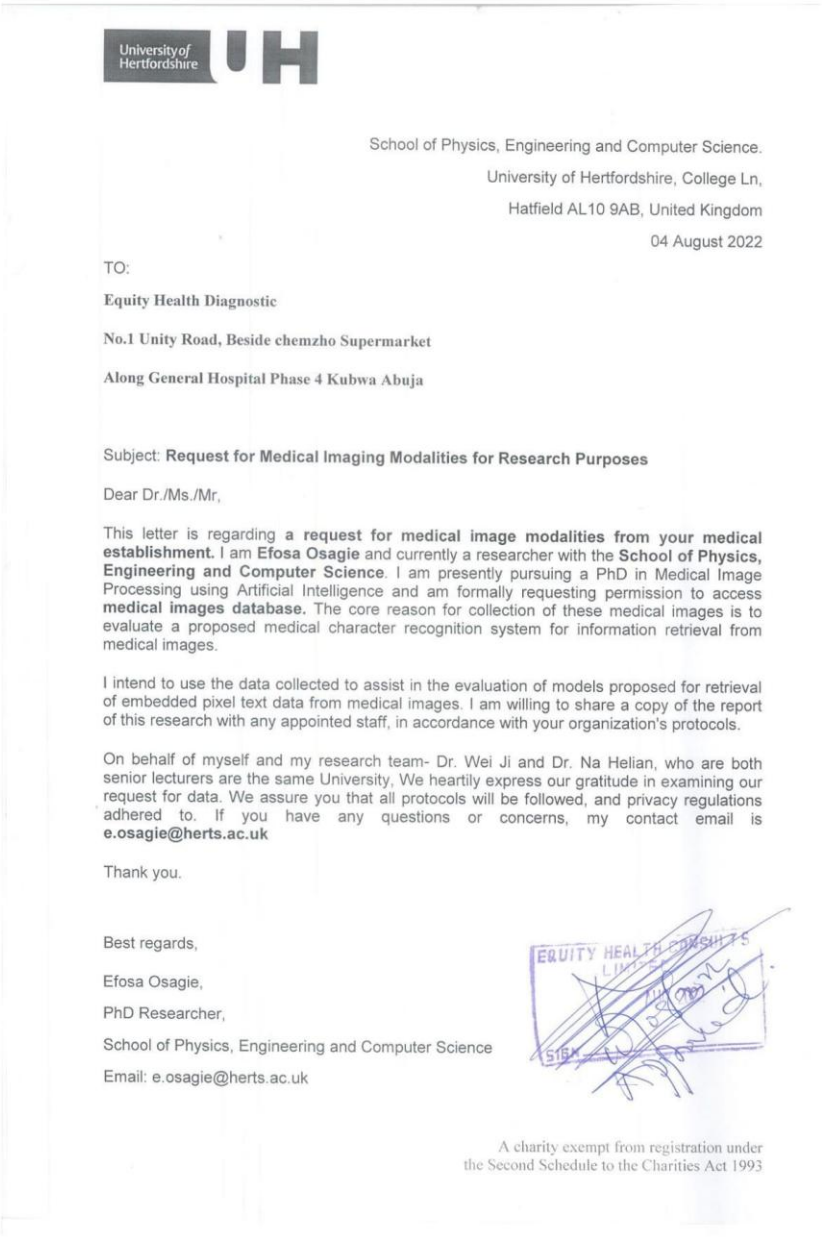
Validity:

This approval is valid:

From: 01/12/2022

To: 01/12/2023

Appendix B – Data Request to Data Collection Location 1



University of Hertfordshire UH

School of Physics, Engineering and Computer Science.
University of Hertfordshire, College Ln,
Hatfield AL10 9AB, United Kingdom

04 August 2022

TO:

Equity Health Diagnostic

No.1 Unity Road, Beside chemzho Supermarket

Along General Hospital Phase 4 Kubwa Abuja

Subject: **Request for Medical Imaging Modalities for Research Purposes**

Dear Dr./Ms./Mr,

This letter is regarding **a request for medical image modalities from your medical establishment.** I am **Efosa Osagie** and currently a researcher with the **School of Physics, Engineering and Computer Science**. I am presently pursuing a PhD in Medical Image Processing using Artificial Intelligence and am formally requesting permission to access **medical images database.** The core reason for collection of these medical images is to evaluate a proposed medical character recognition system for information retrieval from medical images.

I intend to use the data collected to assist in the evaluation of models proposed for retrieval of embedded pixel text data from medical images. I am willing to share a copy of the report of this research with any appointed staff, in accordance with your organization's protocols.

On behalf of myself and my research team- Dr. Wei Ji and Dr. Na Helian, who are both senior lecturers are the same University, We heartily express our gratitude in examining our request for data. We assure you that all protocols will be followed, and privacy regulations adhered to. If you have any questions or concerns, my contact email is **e.osagie@herts.ac.uk**

Thank you.

Best regards,

Efosa Osagie,

PhD Researcher,

School of Physics, Engineering and Computer Science

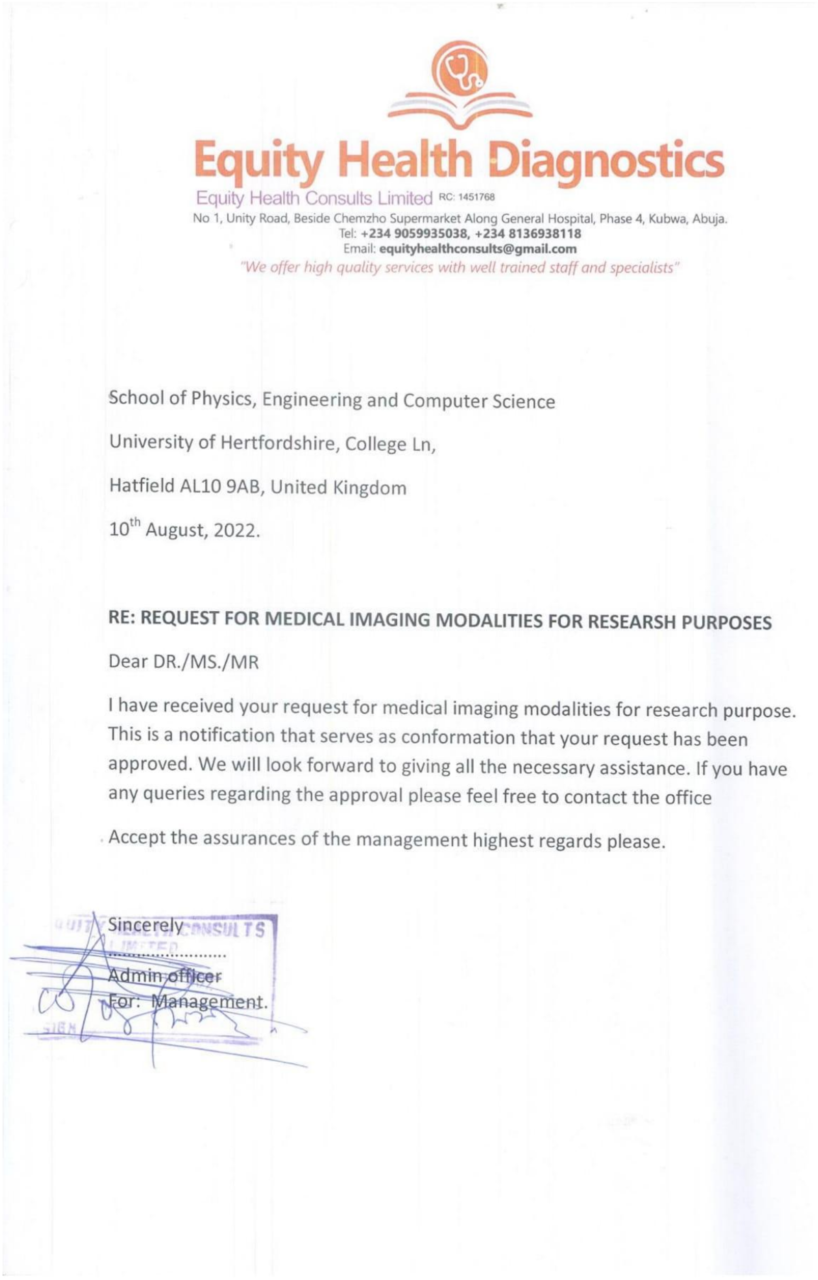
Email: e.osagie@herts.ac.uk

EQUITY HEALTH CONSULTS
LIMITED
SIGN

A charity exempt from registration under
the Second Schedule to the Charities Act 1993

Appendix C – Data Approval to Data Collection

Location 2



Equity Health Diagnostics
Equity Health Consults Limited RC: 1451768
No 1, Unity Road, Beside Chemzho Supermarket Along General Hospital, Phase 4, Kubwa, Abuja.
Tel: **+234 9059935038, +234 8136938118**
Email: **equityhealthconsults@gmail.com**
"We offer high quality services with well trained staff and specialists"

School of Physics, Engineering and Computer Science
University of Hertfordshire, College Ln,
Hatfield AL10 9AB, United Kingdom
10th August, 2022.

RE: REQUEST FOR MEDICAL IMAGING MODALITIES FOR RESEARSH PURPOSES

Dear DR./MS./MR

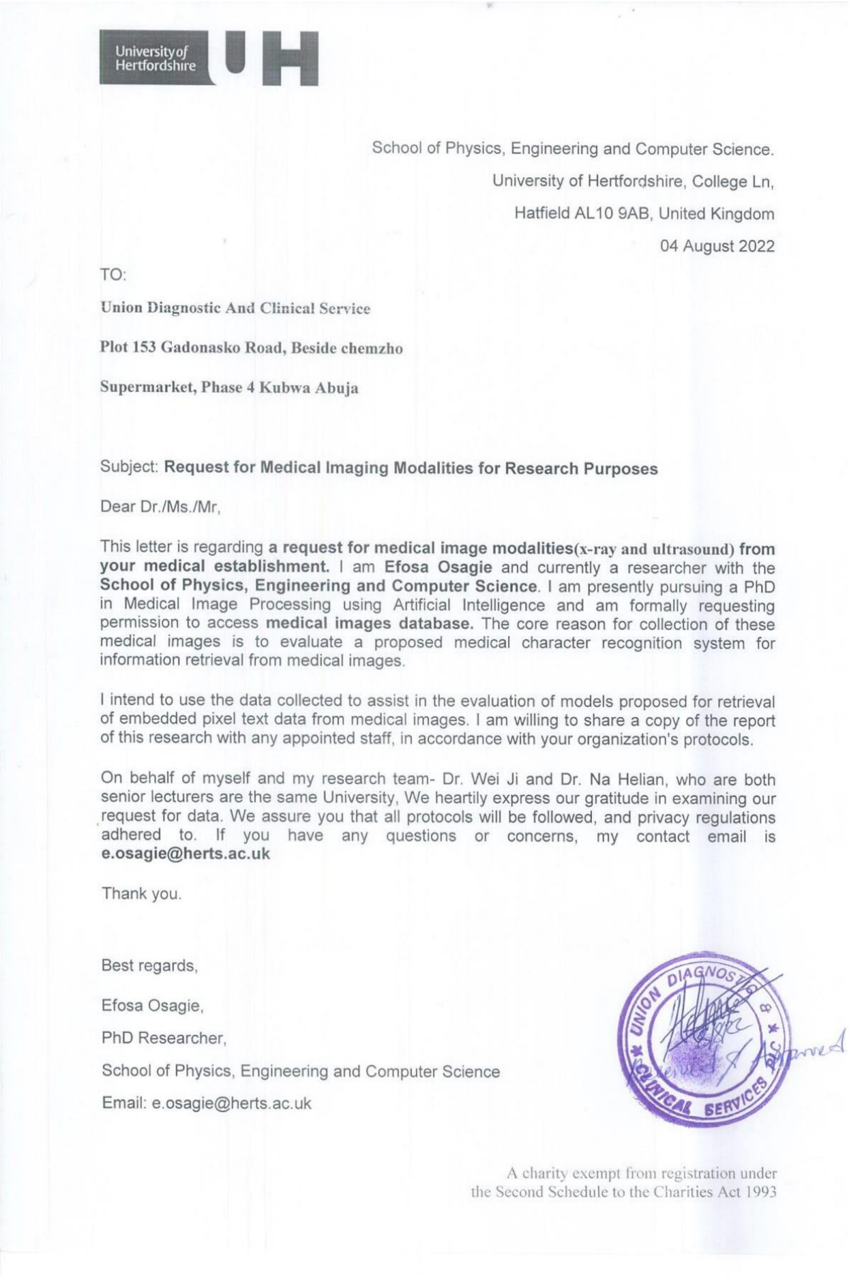
I have received your request for medical imaging modalities for research purpose. This is a notification that serves as conformation that your request has been approved. We will look forward to giving all the necessary assistance. If you have any queries regarding the approval please feel free to contact the office

Accept the assurances of the management highest regards please.

Sincerely
……………………
Admin officer
For: Management.

CONSULTS LIMITED
SIGN

Appendix D – Data Request and Approval to Data Collection Location 2



University of Hertfordshire UH

School of Physics, Engineering and Computer Science.
University of Hertfordshire, College Ln,
Hatfield AL10 9AB, United Kingdom

04 August 2022

TO:

Union Diagnostic And Clinical Service

Plot 153 Gadonasko Road, Beside chemzho

Supermarket, Phase 4 Kubwa Abuja

Subject: **Request for Medical Imaging Modalities for Research Purposes**

Dear Dr./Ms./Mr,

This letter is regarding **a request for medical image modalities(x-ray and ultrasound) from your medical establishment.** I am **Efosa Osagie** and currently a researcher with the **School of Physics, Engineering and Computer Science**. I am presently pursuing a PhD in Medical Image Processing using Artificial Intelligence and am formally requesting permission to access **medical images database.** The core reason for collection of these medical images is to evaluate a proposed medical character recognition system for information retrieval from medical images.

I intend to use the data collected to assist in the evaluation of models proposed for retrieval of embedded pixel text data from medical images. I am willing to share a copy of the report of this research with any appointed staff, in accordance with your organization's protocols.

On behalf of myself and my research team- Dr. Wei Ji and Dr. Na Helian, who are both senior lecturers are the same University, We heartily express our gratitude in examining our request for data. We assure you that all protocols will be followed, and privacy regulations adhered to. If you have any questions or concerns, my contact email is **e.osagie@herts.ac.uk**

Thank you.

Best regards,

Efosa Osagie,

PhD Researcher,

School of Physics, Engineering and Computer Science

Email: e.osagie@herts.ac.uk

UNION DIAGNOSTIC & CLINICAL SERVICES PLC

Reviewed & Approved

A charity exempt from registration under
the Second Schedule to the Charities Act 1993

Appendix E - Risk Assessment for Data Collection

Risk Assessment involving the consideration of physical and psychological risks and data privacy protection during the data collection stage have been adequately considered. The study has developed procedures that reduce and minimise risks to human participants, which is vital in the original data collection. These are itemised below:

1. Documentation of informed consent would be collected using the Ethics' EC3 Consent form.
2. Only approved medical laboratories with an in-house licensed Technologist were used for the data collection.
3. The EC3 Consent form required only minimal data (Name and signature only) to ensure an acceptable level of anonymity for the participants' personal information.
4. The licensed technologist, an assisting nurse, and the participant were present during the capture, as explained in the ethics application form.
5. The MIM with textual data already burned in after image acquisition was collected on a secured USB drive and then transferred along with an electronically scanned copy of the EC3 Consent form to the University's UH OneDrive Storage.
6. The USB and the paper copies of the EC3 Consent form were destroyed after this.

Appendix F- Results from Commercial OCRs

Note: The original image is on the left, and the OCRs' result is on the right.

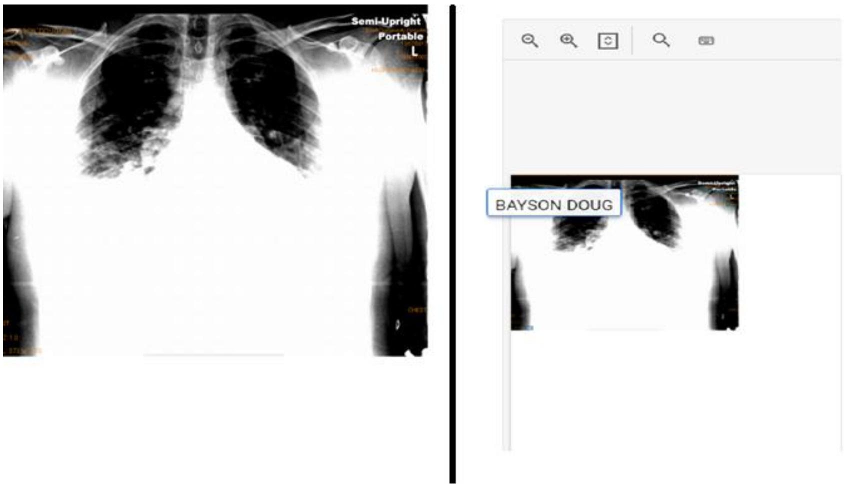


Figure F1: Google Document AI results on sample medical image

As seen in Figure F1, The Google Document AI could not recognise "DAVIDSON DOUGLAS accurately." Moreover, it did not provide any result for "HUGHES FATHLEEN" and other text on the image's upper right side.

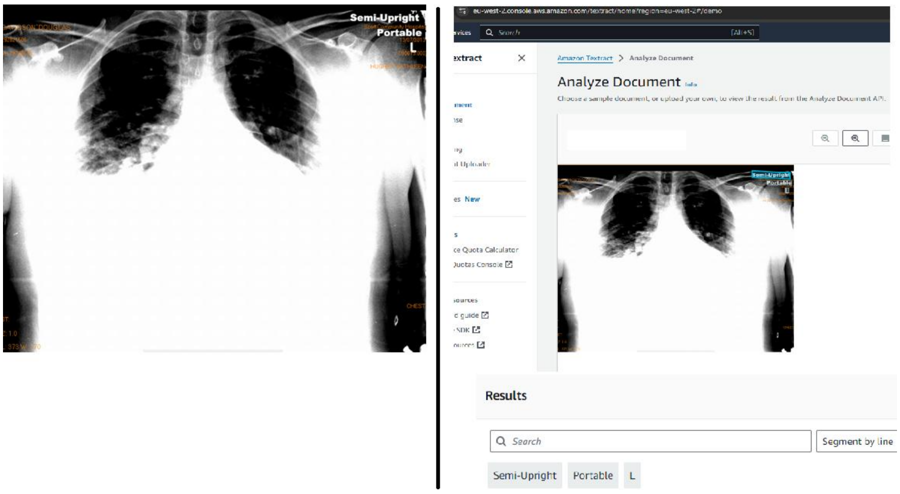

Semi-Upright
Portable
L
eu-west-2.console.aws.amazon.com/textract/home?region=eu-west-2#/demo
Search
[Alt+S]
Amazon Textract
Analyze Document
Analyze Document
Choose a sample document, or upload your own, to view the result from the Analyze Document API.
Results
Search
Segment by line
Semi-Upright
Portable
L

Figure F2: Amazon AWS Textract results on sample medical image

Similarly in Figure F2, Amazon AWS Textract could not recognise the patient's name or clinical information on the left side of the image.

Bibliography

1) A. Mumuni and F. Mumuni. (2022) Data augmentation: A comprehensive survey of modern approaches, Array, vol. 16, p. 100258, Dec. 2022, doi: 10.1016/j.array.2022.100258.

2) Ackland, P., Resnikoff, S., & Bourne, R. (2017). World blindness and visual impairment: despite many successes, the problem is growing. Community eye health, 30(100), 71–73.

3) Adnan, K. and Akbar, R. (2019) Limitations of information extraction methods and techniques for heterogeneous unstructured big data, *International Journal of Engineering Business Management*, 11, pp. 184797901989077. DOI:10.1177/1847979019890771.

4) Agazzi, O. E. and Kuo, S. (1993) Hidden Markov model-based optical character recognition in the presence of deterministic transformations, *Pattern Recognition*, 26 (12), pp. 1813–1826. DOI:10.1016/0031-3203(93)90178-Y.

5) Aggarwal, A., Mittal, M. and Battineni, G. (2021) Generative adversarial network: An overview of theory and applications, *International Journal of Information Management Data Insights*, 1 (1), pp. 100004. DOI:10.1016/j.jjimei.2020.100004.

6) Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019, July 25) Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv. Available from: http://arxiv.org/abs/1907.10902 [Accessed 10 April 2024].

7) Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M. and Collado-Mesa, F. (2022) Towards a better understanding of annotation tools for medical imaging:

a survey, *Multimedia Tools and Applications*, 81 (18), pp. 25877–25911. DOI:10.1007/s11042-022-12100-1.

8) Alkhalifah, T., Wang, H. and Ovcharenko, O. (2022) MLReal: Bridging the gap between training on synthetic data and real data applications in machine learning, Artificial Intelligence in Geosciences, 3, pp. 101–114. DOI:10.1016/j.aiig.2022.09.002.
9) Alter D, and Werner, A. (2007) Automatische texterkennung (ocr) in ultraschall . Konferenz der SAS-Anwender in Forschung und Entwicklung,2007
10) Alter D, and Werner, A. (2007). Automatische texterkennung (ocr) in ultraschall. Konferenz der SAS-Anwender in Forschung und Entwicklung.
11) Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size
12) Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of Big Data, 8 (1), pp. 53. DOI:10.1186/s40537-021-00444-8.
13) Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, Journal of Big Data, 8 (1), pp. 53. DOI:10.1186/s40537-021-00444-8.
14) Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In Journal of Big Data (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-021-00444-8

15) Amethiya, Y., Pipariya, P., Patel, S. and Shah, M. (2022) Comparative analysis of breast cancer detection using machine learning and biosensors, *Intelligent Medicine*, 2 (2), pp. 69–81. DOI:10.1016/j.imed.2021.08.004.

16) Anand, R., Shanthi, T., Sabeenian, R. S., & Veni, S. (2020). Real time noisy dataset implementation of optical character identification using CNN. International Journal of Intelligent Enterprise, 7(1/2/3), 67. https://doi.org/10.1504/IJIE.2020.104646

17) Anderson, J. W., Ziolkowski, M., Kennedy, K. and Apon, A. W. (2022) Synthetic Image Data for Deep Learning. DOI:10.48550/ARXIV.2212.06232.

18) Anil, R., Manjusha, K., Kumar, S. S., & Soman, K. P. (2015). Convolutional neural networks for the recognition of malayalam characters. In S. C. Satapathy, B. N. Biswal, S. K. Udgata, & J. K. Mandal (Eds.), Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 (Vol. 328, pp. 493–500). Springer International Publishing. https://doi.org/10.1007/978-3-319-12012-6_54

19) Antunes, Mário & Machado, Ricardo & Silva, Augusto. (2011). Anonymization of burned-in annotations in ultrasound imaging. Electrónica e Telecomunicações. 5. 360-364.

20) Awal, Md. A., Masud, M., Hossain, Md. S., Bulbul, A. A.-M., Mahmud, S. M. H. and Bairagi, A. K. (2021) A Novel Bayesian Optimization-Based Machine Learning Framework for COVID-19 Detection From Inpatient Facility Data, IEEE Access, 9, pp. 10263–10281. DOI:10.1109/ACCESS.2021.3050852.

21) Badano, A., Revie, C., Casertano, A., Cheng, W.-C., Green, P., Kimpe, T., *et al.* (2015) Consistency and Standardization of Color in Medical Imaging: a

Consensus Report, *Journal of Digital Imaging*, 28 (1), pp. 41–52. DOI:10.1007/s10278-014-9721-0.

22) Badla, S. (2014). IMPROVING THE EFFICIENCY OF TESSERACT OCR ENGINE. San Jose State University Library. https://doi.org/10.31979/etd.5avd-kf2g

23) Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate (Version 7). arXiv. https://doi.org/10.48550/ARXIV.1409.0473

24) Bai, J.-W., Qiu, S.-Q., & Zhang, G.-J. (2023). Molecular and functional imaging in cancer-targeted therapy: current applications and future directions. In Signal Transduction and Targeted Therapy (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41392-023-01366-y

25) Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., & Roy, P. (2022). Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. In Computer Methods and Programs in Biomedicine (Vol. 213, p. 106504). Elsevier BV. https://doi.org/10.1016/j.cmpb.2021.106504

26) Baldominos, A., Saez, Y., & Isasi, P. (2019). A survey of handwritten character recognition with mnist and emnist. Applied Sciences, 9(15), 3169. https://doi.org/10.3390/app9153169

27) Banerjee, I., Ling, Y., Chen, M. C., Hasan, S. A., Langlotz, C. P., Moradzadeh, N., Chapman, B., Amrhein, T., Mong, D., Rubin, D. L., Farri, O., & Lungren, M. P. (2019). Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report

classification. In Artificial Intelligence in Medicine (Vol. 97, pp. 79–88). Elsevier BV. https://doi.org/10.1016/j.artmed.2018.11.004

28) Baucum, M., Khojandi, A. and Papamarkou, T. (2020) Hidden Markov models as recurrent neural networks: an application to Alzheimer’s disease. DOI:10.48550/ARXIV.2006.03151.

29) Bergeron, B. (2005) Clinical data capture: OMR and OCR and your flatbed scanner, *MedGenMed: Medscape General Medicine*, 7 (2), pp. 66.

30) Bieniecki, W., Grabowski, S. ,and Rozenberg, W. (2007) . Image Preprocessing for Improving OCR Accuracy . *2007 International Conference on Perspective Technologies and Methods in MEMS Design*, 2007, pp. 75-80, doi: 10.1109/MEMSTECH.2007.4283429.

31) Bieniecki, W., Grabowski, S., & Rozenberg, W. (2007). Image Preprocessing for Improving OCR Accuracy. In *2007 International Conference on Perspective Technologies and Methods in MEMS Design* (pp. 75-80).

32) Bjorck, J., Gomes, C.P., & Selman, B. (2018). Understanding Batch Normalization. NeurIPS.

33) Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2022). Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 44, Issue 11, pp. 7327–7347). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tpami.2021.3116668

34) Bredell, G., Flouris, K., Chaitanya, K., Erdil, E. and Konukoglu, E. (2023) Explicitly Minimizing the Blur Error of Variational Autoencoders. DOI:10.48550/ARXIV.2304.05939.

35) Briechle, K., & Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation&lt;/title&gt; In D. P. Casasent & T.-H. Chao (Eds.), SPIE Proceedings. Aerospace/Defense Sensing, Simulation, and Controls. SPIE. https://doi.org/10.1117/12.421129

36) Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

37) C. Chadebec and S. Allassonnière, ‘Data Augmentation with Variational Autoencoders and Manifold Sampling’, 2021, doi: 10.48550/ARXIV.2103.13751.

38) C. Doersch, ‘Tutorial on Variational Autoencoders’, 2016, doi: 10.48550/ARXIV.1606.05908.

39) Cai, A., Hu, W., & Zheng, J. (2020). Few-Shot Learning for Medical Image Classification. In Lecture Notes in Computer Science (pp. 441–452). Springer International Publishing. https://doi.org/10.1007/978-3-030-61609-0_35

40) Campos T. E, Babu B. R. , and Varma, M. (2009) Character recognition in natural images. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, February 2009.

41) Cao, Q., Ying, Y., and Li., P. (2013) Similarity metric learning for face recognition. In: Proc. IEEE Int. Conf. on Computer Vision. 1–8 December, Darling Harbour, Sydney, pp. 2408–2415. https://doi.org/10.1109/ICCV.2013.299

42) Caruana, R. (1997) Multitask Learning, Machine Learning, 28 (1), pp. 41–75. DOI:10.1023/A:1007379606734.

43) Chadebec, C., Thibeau-Sutre, E., Burgos, N. and Allassonniere, S. (2022) Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-

Based Variational Autoencoder, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–18. DOI:10.1109/TPAMI.2022.3185773.

44) Chandra, T. B., Verma, K., Singh, B. K., Jain, D., and Netam, S. S. (2021). Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble. Elsevier BV. https://doi.org/10.1016/j.eswa.2020.113909

45) Chen, S. and Guo, W. (2023) Auto-Encoders in Deep Learning—A Review with New Perspectives, Mathematics, 11 (8), pp. 1777. DOI:10.3390/math11081777.

46) Chen, Z., Wu, Y., Yin, F. and Liu, C.-L. (2017) Simultaneous Script Identification and Handwriting Recognition via Multi-Task Learning of Recurrent Neural Networks, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, pp. 525–530.

47) Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., & Lähdesmäki, H. (2019). An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. In Nature Communications (Vol. 10, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-019-09785-8

48) Cho, W., Kwon, J., Kwon, S., & Yoo, J. (2019). A comparative study on ocr using super-resolution for small fonts. International Journal of Advanced Smart Convergence, 8(3), 95–101. https://doi.org/10.7236/IJASC.2019.8.3.95

49) Cho, W., Kwon, J., Kwon, S., & Yoo, J. (2019). A comparative study on ocr using super-resolution for small fonts. International Journal of Advanced Smart Convergence, 8(3), 95–101. https://doi.org/10.7236/IJASC.2019.8.3.95

50) Chopra, S., Hadsell, R. and LeCun, Y. (2005) Learning a Similarity Metric Discriminatively, with Application to Face Verification, in: 2005 IEEE Computer

Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA: IEEE,1, pp. 539–546.

51) Chow, L. S., & Paramesran, R. (2016). Review of medical image quality assessment. Biomedical Signal Processing and Control, 27, 145–154. https://doi.org/10.1016/j.bspc.2016.02.006

52) Chung, T., Xu, B., Liu, Y., Ouyang, C., Li, S., & Luo, L. (2019). Empirical study on character level neural network classifier for Chinese text. In Engineering Applications of Artificial Intelligence (Vol. 80, pp. 1–7). Elsevier BV. https://doi.org/10.1016/j.engappai.2019.01.009

53) Chung, Y.-A., & Weng, W.-H. (2017). Learning Deep Representations of Medical Images using Siamese CNNs with Application to Content-Based Image Retrieval (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1711.08490

54) Cireşan, D., Meier, U. and Schmidhuber, J. (2012) Multi-column Deep Neural Networks for Image Classification. DOI:10.48550/ARXIV.1202.2745.

55) Clémen,C. and Stéphanie, A. Data Augmentation with Variational Autoencoders and Manifold Sampling. DALI 2021 : 1st MICCAI Workshop on Data Augmentation, Labeling, and Imperfections, Oct 2021, Strasbourg, France.

56) Collin, C. B., Gebhardt, T., Golebiewski, M., Karaderi, T., Hillemanns, M., Khan, F. M., *et al.* (2022) Computational Models for Clinical Applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation, *Journal of Personalized Medicine*, 12 (2), pp. 166. DOI:10.3390/jpm12020166.

57) Condorcet, M. (1785). Essay on the Application of Analysis to the Probability of Majority Decisions. https://www.britannica.com/topic/Essay-on-the-Application-of-Analysis-to-the-Probability-of-Majority-Decisions

58) Côté, M., Lecolinet, E., Cheriet, M., & Suen, C. Y. (1998). Automatic reading of cursive scripts using a reading model and perceptual concepts. International Journal on Document Analysis and Recognition, 1(1), 3–17. https://doi.org/10.1007/s100320050002

59) Cunningham, P., Cord, M., & Delany, S. J. (n.d.). Supervised Learning. In Cognitive Technologies (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

60) D. J. Rezende, S. Mohamed, and D. Wierstra, 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models', 2014, doi: 10.48550/ARXIV.1401.4082.

61) D. P. Kingma and M. Welling, 'An Introduction to Variational Autoencoders', FNT in Machine Learning, vol. 12, no. 4, pp. 307–392, 2019, doi: 10.1561/2200000056.

62) D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, 'Semi-Supervised Learning with Deep Generative Models', 2014, doi: 10.48550/ARXIV.1406.5298.

63) D. W. Cromey, 'Digital Images Are Data: And Should Be Treated as Such', in Cell Imaging Techniques, vol. 931, D. J. Taatjes and J. Roth, Eds., Totowa, NJ: Humana Press, 2012, pp. 1–27. doi: 10.1007/978-1-62703-056-4_1.

64) Dash, S., Shakyawar, S. K., Sharma, M. and Kaushik, S. (2019) Big data in healthcare:management, analysis and future prospects, Journal of Big Data. Springer Science and Business Media LLC. DOI:10.1186/s40537-019-0217-0.

65) Davila Delgado, J. M. and Oyedele, L. (2021) Deep learning with small datasets: using autoencoders to address limited datasets in construction management, *Applied Soft Computing*, 112, pp. 107836. DOI:10.1016/j.asoc.2021.107836.

66) Deepak, S., & Ameer, P. M. (2021). Brain tumour classification using siamese neural network and neighbourhood analysis in embedded feature space. In International Journal of Imaging Systems and Technology (Vol. 31, Issue 3, pp. 1655–1669). Wiley. https://doi.org/10.1002/ima.22543

67) Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

68) Dey, S., Dutta, A., Toledo, J. I., Ghosh, S. K., Llados, J. and Pal, U. (2017) SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification. DOI:10.48550/ARXIV.1707.02131.

69) Diessner, M., O'Connor, J., Wynn, A., Laizet, S., Guan, Y., Wilson, K., & Whalley, R. D. (2022). Investigating Bayesian optimization for expensive-to-evaluate black box functions: Application in fluid dynamics. In Frontiers in Applied Mathematics and Statistics (Vol. 8). Frontiers Media SA. https://doi.org/10.3389/fams.2022.1076296

70) Doersch, C. (2016). Tutorial on Variational Autoencoders (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1606.05908

71) Drobac, S. and Lindén, K. (2020) Optical character recognition with neural networks and post-correction with finite state methods, International Journal on Document Analysis and Recognition (IJDAR), 23 (4), pp. 279–295. DOI:10.1007/s10032-020-00359-9.

72) Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. International Journal on

Document Analysis and Recognition (IJDAR), 23(4), 279–295. https://doi.org/10.1007/s10032-020-00359-9

73) Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (IJDAR), 23(4), 279–295. https://doi.org/10.1007/s10032-020-00359-9

74) Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. In International Journal on Document Analysis and Recognition (IJDAR) (Vol. 23, Issue 4, pp. 279–295). Springer Science and Business Media LLC. https://doi.org/10.1007/s10032-020-00359-9

75) Due Trier, Ø., Jain, A. K., & Taxt, T. (1996). Feature extraction methods for character recognition-A survey. Pattern Recognition, 29(4), 641–662. https://doi.org/10.1016/0031-3203(95)00118-2

76) Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S. and Dehmer, M. (2020) An Introductory Review of Deep Learning for Prediction Models With Big Data, Frontiers in Artificial Intelligence, 3, pp. 4. DOI:10.3389/frai.2020.00004.

77) Eriksson, D. and Jankowiak, M. (2021) High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces. DOI:10.48550/ARXIV.2103.00349.

78) Erlandson, E. J., Trenkle, J. M., & Vogt III, R. C. (1996). Word-level recognition of multifont Arabic text using a feature vector matching approach (L. M. Vincent & J. J. Hull, Eds.; pp. 63–70). https://doi.org/10.1117/12.234725

79) Erlandson, E. J., Trenkle, J. M., & Vogt III, R. C. (1996). Word-level recognition of multifont Arabic text using a feature vector matching approach (L. M. Vincent & J. J. Hull, Eds.; pp. 63–70). https://doi.org/10.1117/12.234725

80) Feng, Y., Wang, Y., Li, H., Qu, M., & Yang, J. (2023). Learning what and where to segment: A new perspective on medical image few-shot segmentation. In Medical Image Analysis (Vol. 87, p. 102834). Elsevier BV. https://doi.org/10.1016/j.media.2023.102834

81) Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In The Springer Series on Challenges in Machine Learning (pp. 3–33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

82) Florea F, Rogozan A, Bensrhair A, Dacher JN, Darmoni S. (2005) Modality categorisation by textual annotations interpretation in medical imaging. Medical Informatics Europe (MIE 2005). 2005 Oct:1270-5.

83) Floreaa, F., Rogozana, A., Bensrhaira, A., Dacherc, J. and Darmonia, S. (2005). Modality Categorization by Textual Annotations Interpretation in Medical Imaging- Connecting Medical Informatics and Bio-Informatics R. Engelbrecht et al. (Eds.) ENMI, 2005

84) Frazier, P. I. (2018) A Tutorial on Bayesian Optimization. DOI:10.48550/ARXIV.1807.02811.

85) Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2021). Ensemble deep learning: A review. https://doi.org/10.48550/ARXIV.2104.02395

86) Gao, W., Karbasi, M., Hasanipanah, M., Zhang, X., & Guo, J. (2017). Developing GPR model for forecasting the rock fragmentation in surface mines. In Engineering with Computers (Vol. 34, Issue 2, pp. 339–345). Springer Science and Business Media LLC. https://doi.org/10.1007/s00366-017-0544-8

87) Gao, Y., Yu, T. and Li, J. (2019) Bayesian optimization with local search. DOI:10.48550/ARXIV.1911.09159.

88) Gareth, J., Daniela W., Trevor H., and Robert T. (2013) An introduction to statistical learning : with applications in R . New York: Springer, 2013

89) Ghahramani, Z. (2004). Unsupervised Learning. In Lecture Notes in Computer Science (pp. 72–112). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_5

90) Gilbey, J. D., & Schönlieb, C.-B. (2021). An end-to-end Optical Character Recognition approach for ultra-low-resolution printed text images. https://doi.org/10.48550/ARXIV.2105.04515

91) Gnip, P., Vokorokos, L., & Drotár, P. (2021). Selective oversampling approach for strongly imbalanced data. In PeerJ Computer Science (Vol. 7, p. e604). PeerJ. https://doi.org/10.7717/peerj-cs.604

92) Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. In Information and Software Technology (Vol. 162, p. 107268). Elsevier BV. https://doi.org/10.1016/j.infsof.2023.107268

93) Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Networks. DOI:10.48550/ARXIV.1406.2661.

94) Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Networks. DOI:10.48550/ARXIV.1406.2661.

95) Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

96) Goodfellow, I., Bengio, Y., & Courville, A. (2016) Deep Learning . MIT Press, 2016

97) Guibas, J. T., Virdi, T. S., & Li, P. S. (2017). Synthetic medical images from dual generative adversarial networks. https://doi.org/10.48550/ARXIV.1709.01872

98) Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022) Attention mechanisms in computer vision: A survey, Computational Visual Media, 8 (3), pp. 331–368. DOI:10.1007/s41095-022-0271-y.

99) Hamida, S., Cherradi, B., El Gannour, O., Raihani, A., and Ouajji, H. (2023). Cursive Arabic handwritten word recognition system using majority voting and k-NN for feature descriptor selection. Springer Science and Business Media LLC. https://doi.org/10.1007/s11042-023-15167-6

100) Hamoudi, Y., Amimeur, H., Aouzellag, D., Abdolrasol, M. G. M., & Ustun, T. S. (2023). Hyperparameter Bayesian Optimization of Gaussian Process Regression Applied in Speed-Sensorless Predictive Torque Control of an Autonomous Wind Energy Conversion System. In Energies (Vol. 16, Issue 12, p. 4738). MDPI AG. https://doi.org/10.3390/en16124738

101) Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. In Journal of Big Data (Vol. 6, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40537-019-0263-7

102) Hashemi, N. S., Aghdam, R. B., Ghiasi, A. S. B., & Fatemi, P. (2016). Template Matching Advances and Applications in Image Analysis (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1610.07231

103) Hassan, E., Shams, M. Y., Hikal, N. A. and Elmougy, S. (2023) The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study, Multimedia Tools and Applications, 82 (11), pp. 16591–16633. DOI:10.1007/s11042-022-13820-0.

104) Hassanat, A. B., Tarawneh, A. S., & Altarawneh, G. A. (2022). Stop Oversampling for Class Imbalance Learning: A Critical Review. Research Square Platform LLC. https://doi.org/10.21203/rs.3.rs-1336037/v1

105) He, K., Pu, N., Lao, M. and Lew, M. S. (2023) Few-shot and meta-learning methods for image understanding: a survey, International Journal of Multimedia Information Retrieval, 12 (2), pp. 14. DOI:10.1007/s13735-023-00279-4.

106) He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. https://doi.org/10.48550/ARXIV.1512.03385

107) He, L. (2023) Comparison of improved variational autoencoder models for human face generation, *Journal of Physics: Conference Series*, 2634 (1), pp. 012042. DOI:10.1088/1742-6596/2634/1/012042.

108) Hegghammer, T. (2021). Ocr with tesseract, amazon textract, and google document ai: A benchmarking experiment. Journal of Computational Social Science. https://doi.org/10.1007/s42001-021-00149-1

109) Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

110) Hom, J., Nikowitz, J., Ottesen, R., & Niland, J. C. (2022). Facilitating clinical research through automation: Combining optical character recognition with natural language processing. In Clinical Trials (Vol. 19, Issue 5, pp. 504–511). SAGE Publications. https://doi.org/10.1177/17407745221093621

111) Hosseini-Asl, E. and Guha, A. (2015) Similarity-based Text Recognition by Deeply Supervised Siamese Network. DOI:10.48550/ARXIV.1511.04397.

112) Hou, J., Zeng, H., Cai, L., Zhu, J., Cao, J. and Hou, J. (2017) Handwritten numeral recognition using multi-task learning, in: 2017 International Symposium on

Intelligent Signal Processing and Communication Systems (ISPACS). Xiamen, China: IEEE, pp. 155–158.

113) Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2017) Squeeze-and-Excitation Networks. DOI:10.48550/ARXIV.1709.01507.

114) Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely connected convolutional networks. https://doi.org/10.48550/ARXIV.1608.06993

115) Inkeaw, P., Bootkrajang, J., Marukatat, S., Gonçalves, T. and Chaijaruwanich, J. (2019) Recognition of similar characters using gradient features of discriminative regions, *Expert Systems with Applications*, 134, pp. 120–137. DOI:10.1016/j.eswa.2019.05.050.

116) Inkeaw, P., Bootkrajang, J., Marukatat, S., Gonçalves, T. and Chaijaruwanich, J. (2019) Recognition of similar characters using gradient features of discriminative regions, Expert Systems with Applications, 134, pp. 120–137. DOI:10.1016/j.eswa.2019.05.050.

117) Inunganbi, S., & Katariya, R. S. (2022). Transfer learning for handwritten character recognition. In A. K. Nagar, D. S. Jat, G. Marín-Raventós, & D. K. Mishra (Eds.), Intelligent Sustainable Systems (Vol. 334, pp. 691–699). Springer Singapore. https://doi.org/10.1007/978-981-16-6369-7_63

118) Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://doi.org/10.48550/ARXIV.1502.03167

119) Iqbal, T., & Ali, H. (2018). Generative adversarial network for medical images(Mi-gan). Journal of Medical Systems, 42(11), 231. https://doi.org/10.1007/s10916-018-1072-9

120) Isa, I. S., Sulaiman, S. N., Mustapha, M., & Darus, S. (2015). Evaluating denoising performances of fundamental filters for t2-weighted mri images. Procedia Computer Science, 60, 760–768. https://doi.org/10.1016/j.procs.2015.08.231

121) Islam, M. A., & Iacob, I. E. (2023). Manuscripts Character Recognition Using Machine Learning and Deep Learning. In Modelling (Vol. 4, Issue 2, pp. 168–188). MDPI AG. https://doi.org/10.3390/modelling4020010

122) Istephan, S. and Siadat, M.-R. (2016) Unstructured medical image query using big data – An epilepsy case study, Journal of Biomedical Informatics, 59, pp. 218–226. DOI:10.1016/j.jbi.2015.12.005.

123) J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. (2011) Algorithms for hyper-parameter optimization. in Proceedings of the 24th International Conference on Neural Information Processing Systems, in NIPS’11. Red Hook, NY, USA: Curran Associates Inc., Dec. 2011, pp. 2546–2554.

124) J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, and A. Pouget, ‘Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability’, Neuron, vol. 74, no. 1, pp. 30–39, Apr. 2012, doi: 10.1016/j.neuron.2012.03.016.

125) Ji, Y. and Lu, Z. (2021) The Theoretical Breakthrough of Self-Supervised Learning : Variational Autoencoders and Its Application In Big Data Analysis, Journal of Physics: Conference Series, 1955 (1), pp. 012062. DOI:10.1088/1742-6596/1955/1/012062.

126) K. Baskar, ‘A survey on feature selection techniques in medical image processing’, 2018. https://www.semanticscholar.org/paper/A-Survey-on-Feature-Selection-Techniques-in-Medical-

Baskar/280694439253fc179a5a4157af18f09177af105c (accessed Nov. 06, 2022).

127) Kandel, I., Castelli, M. and Popovič, A. (2020) Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images, Journal of Imaging, 6 (9), pp. 92. DOI:10.3390/jimaging6090092.

128) Kawano, H., Shimamura, A., Maeda, H., Orii, H., Ikoma, N., and Faculty of Engineering, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu 804-8550, Japan (2010) Structure Extraction from Decorated Characters by Graph Spectral Decomposition and Component Selection Criterion, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 14 (2), pp. 179–184. DOI:10.20965/jaciii.2010.p0179.

129) Kayed, M., Anter, A., & Mohamed, H. (2020). Classification of garments from fashion mnist dataset using cnn lenet-5 architecture. 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), 238–243. https://doi.org/10.1109/ITCE48509.2020.9047776

130) Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H. and Kim, N. (2019) Deep Learning in Medical Imaging, Neurospine, 16 (4), pp. 657–668. DOI:10.14245/ns.1938396.198.

131) Kingma, D. P. and Welling, M. (2013) Auto-Encoding Variational Bayes. DOI:10.48550/ARXIV.1312.6114.

132) Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes (Version 11). arXiv. https://doi.org/10.48550/ARXIV.1312.6114

133) Koch, G., Zemel, Richard, and Salakhutdinov, R., 2015. Siamese Neural Networks for One-shot Image Recognition

134) Kociołek, M., Strzelecki, M. and Obuchowicz, R. (2020) Does image normalization and intensity resolution impact texture classification?, *Computerized Medical Imaging and Graphics*, 81, pp. 101716. DOI:10.1016/j.compmedimag.2020.101716.

135) Koga, T., Nonaka, N., Sakuma, J., & Seita, J. (2018). General-to-detailed gan for infrequent class medical images. https://doi.org/10.48550/ARXIV.1812.01690

136) Kovács-V, Zs. M. (1995) A novel architecture for high quality hand-printed character recognition, *Pattern Recognition*, 28 (11), pp. 1685–1692. DOI:10.1016/0031-3203(95)00044-Z.

137) Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017) ImageNet classification with deep convolutional neural networks, *Communications of the ACM*, 60 (6), pp. 84–90. DOI:10.1145/3065386.

138) Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

139) L. Ruthotto and E. Haber, ‘An Introduction to Deep Generative Modeling’, 2021, doi: 10.48550/ARXIV.2103.05180.

140) Lam, L. and Suen, S. Y. (1997) Application of majority voting to pattern recognition: an analysis of its behavior and performance, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27 (5), pp. 553–568. DOI:10.1109/3468.618255.

141) Lam, L., & Suen, S. Y. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 27(5), 553–568. https://doi.org/10.1109/3468.618255

142) Lat, A., & Jawahar, C. V. (2018). Enhancing ocr accuracy with super resolution. 2018 24th International Conference on Pattern Recognition (ICPR), 3162–3167. https://doi.org/10.1109/ICPR.2018.8545609

143) Lavda, F., Gregorová, M. and Kalousis, A. (2019) Improving VAE generations of multimodal data through data-dependent conditional priors. DOI:10.48550/ARXIV.1911.10885.

144) Lavrenko, V., Rath, T. M., & Manmatha, R. (2004). Holistic word recognition for handwritten historical documents. First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings., 278–287. https://doi.org/10.1109/DIAL.2004.1263256

145) Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (11), pp. 2278–2324. DOI:10.1109/5.726791.

146) Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

147) Lee, H., Lee, J., Kwon, Y., Kwon, J., Park, S., Sohn, R. and Park, C. (2022) Multitask Siamese Network for Remote Photoplethysmography and Respiration Estimation, Sensors, 22 (14), pp. 5101. DOI:10.3390/s22145101.

148) Lee, S., Yun, J. S., & Yoo, S. B. (2022). Alternative collaborative learning for character recognition in low-resolution images. IEEE Access, 10, 22003–22017. https://doi.org/10.1109/ACCESS.2022.3153116

149) Li, G., Fang, Q., Zha, L., Gao, X. and Zheng, N. (2022) HAM: Hybrid attention module in deep convolutional neural networks for image classification, Pattern Recognition, 129, pp. 108785. DOI:10.1016/j.patcog.2022.108785.

150) Li, G., Fang, Q., Zha, L., Gao, X., & Zheng, N. (2022). HAM: Hybrid attention module in deep convolutional neural networks for image classification. In Pattern Recognition (Vol. 129, p. 108785). Elsevier BV. https://doi.org/10.1016/j.patcog.2022.108785

151) Li, M., Poovendran, R. and Narayanan, S. (2005) Protecting patient privacy against unauthorized release of medical images in a group communication environment, *Computerized Medical Imaging and Graphics*, 29 (5), pp. 367–383. DOI:10.1016/j.compmedimag.2005.02.003.

152) Li, R., Dai, G., Wang, Z., Yu, S., & Xie, Y. (2018). Using signal-to-noise ratio to connect the quality assessment of natural and medical images. In X. Jiang & J.-N. Hwang (Eds.), Tenth International Conference on Digital Image Processing (ICDIP 2018) (p. 211). SPIE. https://doi.org/10.1117/12.2503084

153) Li, X., Li, M., Yan, P., Li, G., Jiang, Y., Luo, H., & Yin, S. (2023). Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds. In International Journal of Network Dynamics and Intelligence (pp. 93–116). Australia Academic Press Pty Ltd. https://doi.org/10.53941/ijndi0201006

154) Li, Y., Chen, C. L. P., & Zhang, T. (2022). A Survey on Siamese Network: Methodologies, Applications, and Opportunities. In IEEE Transactions on Artificial Intelligence (Vol. 3, Issue 6, pp. 994–1014). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tai.2022.3207112

155) Li, Y., Sixou, B. and Peyrin, F. (2021) A Review of the Deep Learning Methods for Medical Images Super Resolution Problems, *IRBM*, 42 (2), pp. 120–133. DOI:10.1016/j.irbm.2020.08.004.

156) Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G. E., Smith, J. L., Mohtashamian, A., Olson, N., Peng, L. H., Hipp, J. D., & Stumpe, M. C. (2018). Artificial Intelligence–

Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. In Archives of Pathology &amp; Laboratory Medicine (Vol. 143, Issue 7, pp. 859–868). Archives of Pathology and Laboratory Medicine. https://doi.org/10.5858/arpa.2018-0147-oa

157) Liu, Y., Wang, Y., & Shi, H. (2023). A Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application. In Symmetry (Vol. 15, Issue 4, p. 849). MDPI AG. https://doi.org/10.3390/sym15040849

158) Liu, Y., Yang, Z., Yu, Z., Liu, Z., Liu, D., Lin, H., et al. (2023) Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, Journal of Materiomics, 9 (4), pp. 798–816. DOI:10.1016/j.jmat.2023.05.001.

159) M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, ‘A Generative Model for Image Segmentation Based on Label Fusion’, IEEE Trans. Med. Imaging, vol. 29, no. 10, pp. 1714–1729, Oct. 2010, doi: 10.1109/TMI.2010.2050897.

160) Ma, M., Gao, Z., Wu, J., Chen, Y., & Zheng, X. (2018). A smile detection method based on improved lenet-5 and support vector machine. 446–451. https://doi.org/10.1109/SmartWorld.2018.00104

161) Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., *et al.* (2018) Why rankings of biomedical image analysis competitions should be interpreted with care, *Nature Communications*, 9 (1), pp. 5217. DOI:10.1038/s41467-018-07619-7.

162) McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., Erickson, B. J. and Kallmes, D. F. (2015) The Effects of Changes in Utilization and Technological Advancements of Cross-Sectional

Imaging on Radiologist Workload, *Academic Radiology*, 22 (9), pp. 1191–1198. DOI:10.1016/j.acra.2015.05.007.

163) McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., Erickson, B. J., & Kallmes, D. F. (2015). The Effects of Changes in Utilization and Technological Advancements of Cross-Sectional Imaging on Radiologist Workload. In Academic Radiology (Vol. 22, Issue 9, pp. 1191–1198). Elsevier BV. https://doi.org/10.1016/j.acra.2015.05.007

164) Mehmood, A., Maqsood, M., Bashir, M., & Shuyuan, Y. (2020). A Deep Siamese Convolution Neural Network for Multi-Class Classification of Alzheimer Disease. In Brain Sciences (Vol. 10, Issue 2, p. 84). MDPI AG. https://doi.org/10.3390/brainsci10020084

165) Menasalvas, E. and Gonzalo-Martin, C. (2016) Challenges of Medical Text and Image Processing: Machine Learning Approaches, in: Holzinger, A. (ed.) *Machine Learning for Health Informatics*. Cham: Springer International Publishing,9605, pp. 221–242.

166) Michalak, H. and Okarma, K. (2019) Improvement of Image Binarization Methods Using Image Preprocessing with Local Entropy Filtering for Alphanumerical Character Recognition Purposes, *Entropy*, 21 (6), pp. 562. DOI:10.3390/e21060562.

167) Michalak, H., & Okarma, K. (2019). Improvement of image binarization methods using image preprocessing with local entropy filtering for alphanumerical character recognition purposes. Entropy, 21(6), 562. https://doi.org/10.3390/e21060562

168) Mishra, N. K., Dutta, M. and Singh, S. K. (2021) Multiscale parallel deep CNN (mpdCNN) architecture for the real low-resolution face recognition for

surveillance, Image and Vision Computing, 115, pp. 104290. DOI:10.1016/j.imavis.2021.104290.

169) Miyao, H., Nakano, Y., Tani, A., Tabaru, H., Hananoi, T., Shinshu University, 4-17-1, Wakasato, Nagano 380-8553, Japan, Kyushu Sangyo University, 2-3-1 Matsukadai, Higashi-ku, Fukuoka 813-8503, Japan, Hitachi Software Engineering Co., Ltd., 5030 Totsuka-cho, Totsuka-ku, Yokohama 244-8555, Japan, and Fuji Xerox Co., Ltd., 2-3-1 Matsukadai, Higashi-ku, Fukuoka 813-8503, Japan (2004) Printed Japanese Character Recognition Using Multiple Commercial OCRs, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 8 (2), pp. 200–207. DOI:10.20965/jaciii.2004.p0200.

170) Mohamed, R. Sh., & Yousif, G. A. (2010). Input resolution and its effect of the printed image quality on digital toner printing systems (case study – Sinai, Egypt). In The Egyptian Journal of Remote Sensing and Space Science (Vol. 13, Issue 1, pp. 75–80). Elsevier BV. https://doi.org/10.1016/j.ejrs.2010.07.009

171) Mohana, R. S., Kousalya, K., Sasipriyaa, N., Krishnakumar, B., & Gayathri, S. (2021). Investigation on deep learning for handwritten English character recognition. 140028. https://doi.org/10.1063/5.0068646

172) Mohsenzadegan, K., Tavakkoli, V. and Kyamakya, K. (2022) Deep Neural Network Concept for a Blind Enhancement of Document-Images in the Presence of Multiple Distortions, *Applied Sciences*, 12 (19), pp. 9601. DOI:10.3390/app12199601.

173) Monteiro, E., Costa, C. and Oliveira, J. L. (2015) A machine learning methodology for medical imaging anonymization, in: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Milan: IEEE, pp. 1381–1384.

174) Monteiro, E., Costa, C. and Oliveira, J. L. (2015) A machine learning methodology for medical imaging anonymization, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Milan: IEEE, pp. 1381–1384.

175) Monteiro, E., Costa, C. and Oliveira, J. L. (2017) A De-Identification Pipeline for Ultrasound Medical Images in DICOM Format, *Journal of Medical Systems*, 41 (5), pp. 89. DOI:10.1007/s10916-017-0736-1.

176) Monteiro, E., Costa, C. and Oliveira, J. L. (2017) A De-Identification Pipeline for Ultrasound Medical Images in DICOM Format, Journal of Medical Systems, 41 (5), pp. 89. DOI:10.1007/s10916-017-0736-1.

177) Moriconi, R., Deisenroth, M. P. and Kumar, K. S. S. (2019) High-dimensional Bayesian optimization using low-dimensional feature spaces. DOI:10.48550/ARXIV.1902.10675.

178) Moscoso, A., Silva-Rodríguez, J., Aldrey, J. M., Cortés, J., Fernández-Ferreiro, A., Gómez-Lado, N., Ruibal, Á., & Aguiar, P. (2019). Prediction of Alzheimer’s disease dementia with MRI beyond the short-term: Implications for the design of predictive models. In NeuroImage: Clinical (Vol. 23, p. 101837). Elsevier BV. https://doi.org/10.1016/j.nicl.2019.101837

179) Müller, T., Pérez-Torró, G. and Franco-Salvador, M. (2022) Few-Shot Learning with Siamese Networks and Label Tuning. DOI:10.48550/ARXIV.2203.14655.

180) Mureşan, H. and Oltean, M. (2017) Fruit recognition from images using deep learning. DOI:10.48550/ARXIV.1712.00580.

181) Mureşan, H., & Oltean, M. (2017). Fruit recognition from images using deep learning. https://doi.org/10.48550/ARXIV.1712.00580

182) Mustafa, S., Mohammed, B., & Abbosh, A. (2013). Novel preprocessing techniques for accurate microwave imaging of human brain. IEEE Antennas and Wireless Propagation Letters, 12, 460–463. https://doi.org/10.1109/LAWP.2013.2255095

183) Naji, M. A., Filali, S. E., Bouhlal, M., Benlahmar, E. H., Abdelouhahid, R. A., and Debauche, O. (2021). Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier. Elsevier BV. https://doi.org/10.1016/j.procs.2021.07.061

184) Newhauser, W., Jones, T., Swerdloff, S., Newhauser, W., Cilia, M., Carver, R., Halloran, A. and Zhang, R. (2014) Anonymization of DICOM electronic medical records for radiation therapy, *Computers in Biology and Medicine*, 53, pp. 134–140. DOI:10.1016/j.compbiomed.2014.07.010.

185) Nguyen, T. T. H., Jatowt, A., Coustaty, M. and Doucet, A. (2022) Survey of Post-OCR Processing Approaches, *ACM Computing Surveys*, 54 (6), pp. 1–37. DOI:10.1145/3453476.

186) Noack, M. M., Krishnan, H., Risser, M. D., & Reyes, K. G. (2023). Exact Gaussian processes for massive datasets via non-stationary sparsity-discovering kernels. In Scientific Reports (Vol. 13, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41598-023-30062-8

187) Nomura, S., Yamanaka, K., Shiose, T., Kawakami, H. and Katai, O. (2009) Morphological preprocessing method to thresholding degraded word images, *Pattern Recognition Letters*, 30 (8), pp. 729–744. DOI:10.1016/j.patrec.2009.03.008.

188) Nomura, S., Yamanaka, K., Shiose, T., Kawakami, H., & Katai, O. (2009). Morphological preprocessing method to thresholding degraded word images.

Pattern Recognition Letters, 30(8), 729–744. https://doi.org/10.1016/j.patrec.2009.03.008

189) Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. https://doi.org/10.48550/ARXIV.1811.03378

190) Oni, O. J. and Asahiah, F. O. (2020) Computational modelling of an optical character recognition system for Yorùbá printed text images, Scientific African, 9, pp. e00415. DOI:10.1016/j.sciaf.2020.e00415.

191) Osagie, E., Ji, W. and Helian, N. (2023) Ensemble Learning for Medical Image Character Recognition based on Enhanced Lenet-5, in: *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Eindhoven, Netherlands: IEEE, pp. 1–8.

192) Osagie, E., Ji, W. and Helian, N. (2024a) Burnt-in Text Recognition from Medical Imaging Modalities: Existing Machine Learning Practices, Journal of Advanced Computational Intelligence, and Intelligent Informatics, 28 (1), pp. 103–110. DOI:10.20965/jaciii.2024.p0103.

193) Osagie, E., Ji, W. and Helian, N. (2024b) Medical Image Character Recognition using Attention-based Siamese Networks for Visually Similar Characters with Low Resolution. In Lecture Notes in Networks and Systems (pp. 3–12). Springer Nature Switzerland.

194) Ou, L. (2023) Biological Image Processing Algorithm Based on Attention Mechanism and Convolutional Neural Network N. Govindan, ed. Advances in Multimedia. 2023, 1–7.

195) Ozaki, Y., Tanigaki, Y., Watanabe, S. and Onishi, M. (2020) Multiobjective tree-structured parzen estimator for computationally expensive optimization problems,

in: Proceedings of the 2020 Genetic and Evolutionary Computation Conference. Cancún Mexico: ACM, pp. 533–541.

196) Pachetti, E., & Colantonio, S. (2023). A Systematic Review of Few-Shot Learning in Medical Imaging (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2309.11433

197) Padmapriya, S. T. and Parthasarathy, S. (2024) Ethical Data Collection for Medical Image Analysis: a Structured Approach, Asian Bioethics Review, 16 (1), pp. 95–108. DOI:10.1007/s41649-023-00250-9.

198) Pal, D., Alladi, A., Pothireddy, Y. and Koilpillai, G. (2021) MSHSCNN: Multi-Scale Hybrid-Siamese Network to Differentiate Visually Similar Character Classes, in: *2021 9th European Workshop on Visual Information Processing (EUVIP)*. Paris, France: IEEE, pp. 1–6.

199) Pal, D., Alladi, A., Pothireddy, Y. and Koilpillai, G. (2021) MSHSCNN: Multi-Scale Hybrid-Siamese Network to Differentiate Visually Similar Character Classes, in: 2021 9th European Workshop on Visual Information Processing (EUVIP). Paris, France: IEEE, pp. 1–6.

200) Park, J., Lee, E., Kim, Y., Kang, I., Koo, H. I. and Cho, N. I. (2020) Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter, *IEEE Access*, 8, pp. 174437–174448. DOI:10.1109/ACCESS.2020.3025769.

201) Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source ocr tool tesseract: A case study. International Journal of Computer Applications, 55(10), 50–56. https://doi.org/10.5120/8794-2784

202) Paulsen, A., Overgaard, S. and Lauritsen, J. M. (2012) Quality of Data Entry Using Single Entry, Double Entry and Automated Forms Processing–An Example

Based on a Study of Patient-Reported Outcomes, *PLoS ONE*, 7 (4), pp. e35087. DOI:10.1371/journal.pone.0035087.

203) *Perumal, S.V., & Velmurugan, T. (2018). Preprocessing by Contrast Enhancement Techniques for Medical Images.*

204) Pesteie, M., Abolmaesumi, P. and Rohling, R. N. (2019) Adaptive Augmentation of Medical Data Using Independently Conditional Variational Auto-Encoders, *IEEE Transactions on Medical Imaging*, 38 (12), pp. 2807–2820. DOI:10.1109/TMI.2019.2914656.

205) Phung and Rhee (2019) A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets, *Applied Sciences*, 9 (21), pp. 4500. DOI:10.3390/app9214500.

206) Pinto-Coelho, L. (2023). How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. In Bioengineering (Vol. 10, Issue 12, p. 1435). MDPI AG. https://doi.org/10.3390/bioengineering10121435

207) Pividori, M., Grinblat, G. L., and Uzal, L. C. (2019). Exploiting GAN Internal Capacity for High-Quality Reconstruction of Natural Images (Version 1). Version 1. arXiv. https://doi.org/10.48550/ARXIV.1911.05630

208) Qaroush, A., Awad, A., Modallal, M. and Ziq, M. (2022) Segmentation-based, omnifont printed Arabic character recognition without font identification, *Journal of King Saud University - Computer and Information Sciences*, 34 (6), pp. 3025–3039. DOI:10.1016/j.jksuci.2020.10.001.

209) Qin, X., Bui, F. M. and Nguyen, H. H. (2019) Learning from an Imbalanced and Limited MEDPIXnd an Application to Medical Imaging, in: *2019 IEEE Pacific Rim*

Conference on Communications, Computers and Signal Processing (PACRIM). Victoria, BC, Canada: IEEE, pp. 1–6.

210) Quiroga, F., Ronchetti, F., Lanzarini, L., & Bariviera, A. F. (2019). Revisiting Data Augmentation for Rotational Invariance in Convolutional Neural Networks. In Advances in Intelligent Systems and Computing (pp. 127–141). Springer International Publishing. https://doi.org/10.1007/978-3-030-15413-4_10

211) R. DeVore, B. Hanin, and G. Petrova, 'Neural Network Approximation', 2020, doi: 10.48550/ARXIV.2012.14501.

212) Rabiei, R. (2022) Prediction of Breast Cancer using Machine Learning Approaches, Journal of Biomedical Physics and Engineering, 12 (3). DOI:10.31661/jbpe.v0i0.2109-1403.

213) Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77 (2), pp. 257–286. DOI:10.1109/5.18626.

214) Rahman, M. S. and Sultana, M. (2017) Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data, BMC Medical Research Methodology, 17 (1), pp. 33. DOI:10.1186/s12874-017-0313-9.

215) Rajeshwari, S. and Sharmila, T. S. (2013). Efficient quality analysis of MRI image using preprocessing techniques. *2013 IEEE Conference on Information & Communication Technologies*, 2013, pp. 391-396, doi: 10.1109/CICT.2013.6558127.

216) Rakhshan V. (2014). Image resolution in the digital era: notion and clinical implications. J Dent (Shiraz). Dec;15(4):153-5. PMID: 25469352; PMCID: PMC4247836.

217) Ramdhani, T. W., Budi, I. and Purwandari, B. (2021) Optical Character Recognition Engines Performance Comparison in Information Extraction, International Journal of Advanced Computer Science and Applications, 12 (8). DOI:10.14569/IJACSA.2021.0120814.

218) Rao, A., Park, J., Woo, S., Lee, J.-Y., & Aalami, O. (2021). Studying the Effects of Self-Attention for Medical Image Analysis. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE. https://doi.org/10.1109/iccvw54120.2021.00381

219) Rashid, S. F., Schambach, M.-P., Rottland, J., & von der Nüll, S. (2013). Low resolution Arabic recognition with multidimensional recurrent neural networks. Proceedings of the 4th International Workshop on Multilingual OCR - MOCR ’13, 1. https://doi.org/10.1145/2505377.2505385

220) Rashid, S. F., Schambach, M.-P., Rottland, J., & von der Nüll, S. (2013). Low resolution Arabic recognition with multidimensional recurrent neural networks. Proceedings of the 4th International Workshop on Multilingual OCR - MOCR ’13, 1. https://doi.org/10.1145/2505377.2505385

221) requirements when using artificial neural networks for discrete choice analysis. Journal of Choice Modelling,

222) Reul, C., Köberle, P., Üçeyler, N. and Puppe, F. (2016) Expectation-Driven Text Extraction from Medical Ultrasound Images, *Studies in Health Technology and Informatics*, 228, pp. 712–716.

223) Reul, C., Köberle, P., Üçeyler, N., & Puppe, F. (2016). Expectation-Driven Text Extraction from Medical Ultrasound Images. Studies in health technology and informatics, 228, 712–716. doi:10.3233/978-1-61499-678-1-712

224) Reul, C., Springmann, U., Wick, C., & Puppe, F. (2018). State of the art optical character recognition of 19th century fraktur scripts using open source engines. https://doi.org/10.48550/ARXIV.1810.03436

225) Rodemann, J., & Augustin, T. (2024). Imprecise Bayesian optimization. In Knowledge-Based Systems (Vol. 300, p. 112186). Elsevier BV. https://doi.org/10.1016/j.knosys.2024.112186

226) Röhrbein, F., Goddard, P., Schneider, M., James, G. and Guo, K. (2015) How does image noise affect actual and predicted human gaze allocation in assessing image quality?, Vision Research, 112, pp. 11–25. DOI:10.1016/j.visres.2015.03.029.

227) Rong, G., Li, K., Su, Y., Tong, Z., Liu, X., Zhang, J., Zhang, Y. and Li, T. (2021) Comparison of Tree-Structured Parzen Estimator Optimization in Three Typical Neural Network Models for Landslide Susceptibility Assessment, Remote Sensing, 13 (22), pp. 4694. DOI:10.3390/rs13224694.

228) S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, ‘Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models’, IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 7327–7347, Nov. 2022, doi: 10.1109/TPAMI.2021.3116668.

229) S. Tang and Y. Yang, ‘Why neural networks apply to scientific computing?’, Theoretical and Applied Mechanics Letters, vol. 11, no. 3, p. 100242, Mar. 2021, doi: 10.1016/j.taml.2021.100242.

230) S.J. Young and S. Young. (1994) The HTK Hidden Markov Model Toolkit: De sign and Philosophy. Entropic Cambridge Research Laboratory, Ltd., 2:2 44

231) Safaei, A. (2021) Text-based multi-dimensional medical images retrieval according to the features-usage correlation, Medical & Biological Engineering & Computing, 59 (10), pp. 1993–2017. DOI:10.1007/s11517-021-02392-0.

232) Sagar, S., & Dixit, S. (2019). A Comprehensive Study on Character Segmentation. In Lecture Notes in Computational Vision and Biomechanics (pp. 1509–1515). Springer International Publishing. https://doi.org/10.1007/978-3-030-00665-5_141

233) Sangiacomo, A., Hogenbirk, H., Tanasescu, R., Karaisl, Antonia, White, N. (2022) Reading in the mist: high-quality optical character recognition based on freely available early modern digitized books. Digital Scholarship in the Humanities. 37(4), 1197–1209.

234) Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2018). How does batch normalization help optimization? https://doi.org/10.48550/ARXIV.1805.11604

235) Sarvamangala, D. R., & Kulkarni, R. V. (2021). Convolutional neural networks in medical image understanding: a survey. In Evolutionary Intelligence (Vol. 15, Issue 1, pp. 1–22). Springer Science and Business Media LLC. https://doi.org/10.1007/s12065-020-00540-3

236) Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. Neural Computation, 4(2), 234–242. https://doi.org/10.1162/neco.1992.4.2.234

237) Segal, J. P. and Hansen, R. (2021) Medical images, social media and consent, *Nature Reviews Gastroenterology & Hepatology*, 18 (8), pp. 517–518. DOI:10.1038/s41575-021-00453-1.

238) Seo, J.-W., Jung, H.-G., & Lee, S.-W. (2021). Self-augmentation: Generalizing deep networks to unseen classes for few-shot learning. In Neural Networks (Vol. 138, pp. 140–149). Elsevier BV. https://doi.org/10.1016/j.neunet.2021.02.007

239) Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2018) Efficient Attention: Attention with Linear Complexities. arXiv, 2018.

240) Shi, P., Zhao, Z., Liu, K., & Li, F. (2022). Attention-based spatial–temporal neural network for accurate phase recognition in minimally invasive surgery: feasibility and efficiency verification. In Journal of Computational Design and Engineering (Vol. 9, Issue 2, pp. 406–416). Oxford University Press (OUP). https://doi.org/10.1093/jcde/qwac011

241) Shlens, J. (2014) A Tutorial on Principal Component Analysis. DOI:10.48550/ARXIV.1404.1100.

242) Shorten, C. and Khoshgoftaar, T. M. (2019) A survey on Image Data Augmentation for Deep Learning, Journal of Big Data, 6 (1), pp. 60. DOI:10.1186/s40537-019-0197-0.

243) Shteingart, H., Marom, E., Itkin, I., Shabat, G., Kolomenkin, M., Salhov, M., & Katzir, L. (2020). Majority voting and the condorcet’s jury theorem. https://doi.org/10.48550/ARXIV.2002.03153

244) Silva, J. M., Pinho, E., Monteiro, E., Silva, J. F. and Costa, C. (2018) Controlled searching in reversibly de-identified medical imaging archives, *Journal of Biomedical Informatics*, 77, pp. 81–90. DOI:10.1016/j.jbi.2017.12.002.

245) Silva, J. M., Pinho, E., Monteiro, E., Silva, J. F. and Costa, C. (2018) Controlled searching in reversibly de-identified medical imaging archives, Journal of Biomedical Informatics, 77, pp. 81–90. DOI:10.1016/j.jbi.2017.12.002.

246) Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/ARXIV.1409.1556

247) Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/ARXIV.1409.1556

248) Singh, M., Nagpal, S., Vatsa, M., & Singh, R. (2021). Enhancing fine-grained classification for low resolution images. https://doi.org/10.48550/ARXIV.2105.00241

249) Singh, P. and Budhiraja,S. (2011) Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey. International Journal of Engineering Research and Applications (IJERA).Vol. 1, Issue 4, pp. 1736-1739

250) Singh, R., Bharti, V., Purohit, V., Kumar, A., Singh, A. K., & Singh, S. K. (2021). MetaMed: Few-shot medical image classification using gradient-based meta-learning. In Pattern Recognition (Vol. 120, p. 108111). Elsevier BV. https://doi.org/10.1016/j.patcog.2021.108111

251) Smith, D. C. (2012). OCR enhancement through neighbor embedding and fast approximate nearest neighbors (A. G. Tescher, Ed.; p. 84991I). https://doi.org/10.1117/12.928865

252) Smith, R. (2007). An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September 2007 DOI: 10.1109/ICDAR.2007.4376991

253) Snell, J, Swersky, K, Zemel, Prototypical Networks for Few-Shot Learning. (2017) In Pro-ceedings of the 31st International Conference on Neural Information Processing Systems 2017 (pp. 4080–4090). Curran Associates Inc..

254) Sohn, K., Honglak L., and Xinchen Y. “Learning Structured Output Representation using Deep Conditional Generative Models.” Advances in Neural Information Processing Systems. 2015.

255) Sporici, D. Cusnir, E. and Boiangiu, C. (2020). Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. MDPI - Computer Science and Engineering Department, Faculty of Automatic Control and Computers, Politehnica

256) Sporici, D. Cusnir, E. and Boiangiu, C. (2020). Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. MDPI - Computer Science and Engineering Department, Faculty of Automatic Control and Computers, Politehnica

257) Springenberg, J. T., Dosovitskiy, A., Brox, T. and Riedmiller, M. (2014) Striving for Simplicity: The All Convolutional Net. DOI:10.48550/ARXIV.1412.6806.

258) Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. https://doi.org/10.48550/ARXIV.1412.6806

259) Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res., 15, 1929-1958.

260) Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. https://doi.org/10.48550/ARXIV.1507.06228

261) Stoddard, J. G., Birpoutsoukis, G., Schoukens, J., & Welsh, J. S. (2019). Gaussian process regression for the estimation of generalized frequency response functions. In Automatica (Vol. 106, pp. 161–167). Elsevier BV. https://doi.org/10.1016/j.automatica.2019.05.010

262) Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., & Paisley, J. (2022). Few-shot medical image segmentation using a global correlation network with discriminative embedding. In Computers in Biology and Medicine (Vol. 140, p. 105067). Elsevier BV. https://doi.org/10.1016/j.compbiomed.2021.105067

263) Szandała, T. (2021). Review and comparison of commonly used activation functions for deep neural networks (A. K. Bhoi, P. K. Mallick, C.-M. Liu, & V. E. Balas, Eds.; Vol. 903, pp. 203–224). Springer Singapore. https://doi.org/10.1007/978-981-15-5495-7_11

264) Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. https://doi.org/10.48550/ARXIV.1409.4842

265) Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. https://doi.org/10.48550/ARXIV.1512.00567

266) T. Davenport and R. Kalakota, ‘The potential for artificial intelligence in healthcare’, Future Healthc J, vol. 6, no. 2, pp. 94–98, Jun. 2019, doi: 10.7861/futurehosp.6-2-94.

267) Tabik, S., Peralta, D., Herrera-Poyatos, A., & Herrera, F. (2017). A snapshot of image pre-processing for convolutional neural networks: Case study of MNIST: International Journal of Computational Intelligence Systems, 10(1), 555. https://doi.org/10.2991/ijcis.2017.10.1.38

268) Tang, S., Jing, C., Jiang, Y., Yang, K., Huang, Z., Wu, H., Cui, C., Shi, S., Ye, X., Tian, H., Song, D., Xu, J., & Dong, F. (2023). The effect of image resolution on convolutional neural networks in breast ultrasound. In Heliyon (Vol. 9, Issue 8, p. e19253). Elsevier BV. https://doi.org/10.1016/j.heliyon.2023.e19253

269) Thambawita, V., Strümke, I., Hicks, S. A., Halvorsen, P., Parasa, S., & Riegler, M. A. (2021). Impact of image resolution on deep learning performance in endoscopy image classification: An experimental study using a large dataset of endoscopic images. Diagnostics, 11(12), 2183. https://doi.org/10.3390/diagnostics11122183

270) Tomasini, U. M., Petrini, L., Cagnetta, F. and Wyart, M. (2022) How deep convolutional neural networks lose spatial information with training. DOI:10.48550/ARXIV.2210.01506.

271) Tsai, D.-Y., Matsuyama, E., & Chen, H.-M. (2013). Improving image quality in medical images using a combined method of undecimated wavelet transform and wavelet coefficient mapping. International Journal of Biomedical Imaging, 2013, e797924. https://doi.org/10.1155/2013/797924

272) Tsui, G. K. and Chan, T. (2012) Automatic Selective Removal of Embedded Patient Information From Image Content of DICOM Files, *American Journal of Roentgenology*, 198 (4), pp. 769–772. DOI:10.2214/AJR.10.6352.

273) Tsui, G. K. and Chan, T. (2012) Automatic Selective Removal of Embedded Patient Information From Image Content of DICOM Files, American Journal of Roentgenology, 198 (4), pp. 769–772. DOI:10.2214/AJR.10.6352.

274) Tsui, G. K., & Chan, T. (2012). Automatic selective removal of embedded patient information from image content of dicom files. American Journal of Roentgenology, 198(4), 769–772. https://doi.org/10.2214/AJR.10.6352

275) Tsui, G. K., & Chan, T. (2012). Automatic selective removal of embedded patient information from image content of dicom files. American Journal of Roentgenology, 198(4), 769–772. https://doi.org/10.2214/AJR.10.6352

276) V. Ehrenstein, H. Kharrazi, H. Lehmann, and C. O. Taylor, Obtaining data from electronic health records. Agency for Healthcare Research and Quality (US), 2019. Accessed: Nov. 12, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK551878/

277) Vcelak, P., Kryl, M., Kratochvil, M. and Kleckova, J. (2019) Identification and classification of DICOM files with burned-in text content, *International Journal of Medical Informatics*, 126, pp. 128–137. DOI:10.1016/j.ijmedinf.2019.02.011.

278) Vcelak, P., Kryl, M., Kratochvil, M. and Kleckova, J. (2019) Identification and classification of DICOM files with burned-in text content, International Journal of Medical Informatics, 126, pp. 128–137. DOI:10.1016/j.ijmedinf.2019.02.011.

279) Volusonclub. (2017). Available at: www.Volusonclub.net (Accessed 3 February 2022).

280) W. James, 'Security filtering of medical images using OCR', 2002. School of Information Sciences and Technology. Pennsylvania State University

281) Wachenfeld, S., & Klein, H.-U. (01 2006). Recognition of Screen-Rendered Text. 1086–1089. doi:10.1109/ICPR.2006.974

282) Wachenfeld, S., Klein, H.-U., & Jiang, X. (2007). Annotated Databases for the Recognition of Screen-Rendered Text. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 1, 272–276. doi:10.1109/ICDAR.2007.4378718

283) Walters, S. J., & Campbell, M. J. (2004). In Health and Quality of Life Outcomes (Vol. 2, Issue 1, p. 70). Springer Science and Business Media LLC. https://doi.org/10.1186/1477-7525-2-70

284) Wang, J. (2002) Security filtering of medical images using OCR. School of Information Sciences and Technology. Pennsylvania State University, 2002

285) Wang, J. (2002). Security filtering of medical images using OCR. School of Information Sciences and Technology - Pennsylvania State University

286) Wang, Q. and Lu, Y. (2017) Similar Handwritten Chinese Character Recognition Using Hierarchical CNN Model, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, pp. 603–608.

287) Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q. (2019) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. DOI:10.48550/ARXIV.1910.03151.

288) Wang, T., Xie, Z., Li, Z., Jin, L. and Chen, X. (2019) Radical aggregation network for few-shot offline handwritten Chinese character recognition, Pattern Recognition Letters, 125, pp. 821–827. DOI:10.1016/j.patrec.2019.08.005.

289) Wang, X., Sun, L., Chehri, A. and Song, Y. (2023) A Review of GAN-Based Super-Resolution Reconstruction for Optical Remote Sensing Images, *Remote Sensing*, 15 (20), pp. 5062. DOI:10.3390/rs15205062.

290) Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. In IEEE Transactions on Image Processing (Vol. 13, Issue 4, pp. 600–612). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tip.2003.819861

291) Watanabe, S. and Hutter, F. (2022) c-TPE: Tree-structured Parzen Estimator with Inequality Constraints for Expensive Hyperparameter Optimization. DOI:10.48550/ARXIV.2211.14411.

292) Wei, G., Li, G., Zhao, J. and He, A. (2019) Development of a LeNet-5 Gas Identification CNN Structure for Electronic Noses, Sensors, 19 (1), pp. 217. DOI:10.3390/s19010217.

293) Weng, W., Zhu, X., Jing, L., Dong, M. (2023) Attention Mechanism Trained with Small Datasets for Biomedical Image Segmentation. Electronics. 12(3), 682.

294) Wikipedia contributors. (2022). Kernel (image processing) --- Wikipedia, The Free Encyclopedia. Ανακτήθηκε από https://en.wikipedia.org/w/index.php?title=Kernel_(image_processing)&oldid=1090475536

295) Wiley, V. and Lucas, T. (2018) Computer Vision and Image Processing: A Paper Review, International Journal of Artificial Intelligence Research, 2 (1), pp. 22. DOI:10.29099/ijair.v2i1.42.

296) Wood, K., Dunton, A. M., Muyskens, A., & Priest, B. W. (2022). Scalable Gaussian Process Hyperparameter Optimization via Coverage Regularization (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2209.11280

297) Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17, 26–40. doi:10.11989/JEST.1674-862X.80904120

298) Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Journal of Electronic Science and Technology, 17(1), 26–40. doi:10.11989/JEST.1674-862X.80904120

299) Xu, F., Chen, C., Shang, Z., Peng, Y., & Li, X. (2023). A CRNN-based method for Chinese ship license plate recognition. In IET Image Processing (Vol. 18, Issue 2, pp. 298–311). Institution of Engineering and Technology (IET). https://doi.org/10.1049/ipr2.12949

300) Xu, X., Wang, W. and Liu, Q. (2021) Medical Image Character Recognition Based on Multi-scale Neural Convolutional Network, in: *2021 International Conference on Security, Pattern Analysis, and Cybernetics（SPAC)*. Chengdu, China: IEEE, pp. 408–412.

301) Xu, X., Wang, W. and Liu, Q. (2021) Medical Image Character Recognition Based on Multi-scale Neural Convolutional Network, in: 2021 International Conference on Security, Pattern Analysis, and Cybernetics（SPAC). Chengdu, China: IEEE, pp. 408–412.

302) Y. Pu, Z. Gan, R. Henao, C. Li, S. Han, and L. Carin, ‘VAE Learning via Stein Variational Gradient Descent’, 2017, doi: 10.48550/ARXIV.1704.05155.

303) Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. In Neurocomputing (Vol. 415, pp. 295–316). Elsevier BV. https://doi.org/10.1016/j.neucom.2020.07.061

304) Yang, P., Yang, J., Zhou, B., & Zomaya, A. (2010). A Review of Ensemble Methods in Bioinformatics. Current Bioinformatics, 5. doi:10.2174/157489310794072508

305) Yasrab, R., Pound, M. P., French, A. P., & Pridmore, T. P. (2020). PhenomNet: Bridging Phenotype-Genotype Gap: A CNN-LSTM Based Automatic Plant Root Anatomization System. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2020.05.03.075184

306) Yu Ma and Yuanyuan Wang (2015) Text detection in medical images using local feature extraction and supervised learning, in: *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. Zhangjiajie, China: IEEE, pp. 953–958.

307) Yu Ma and Yuanyuan Wang (2015) Text detection in medical images using local feature extraction and supervised learning, in: 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Zhangjiajie, China: IEEE, pp. 953–958.

308) Yu, S., Dai, G., Wang, Z., Li, L., Wei, X., & Xie, Y. (2018). A consistency evaluation of signal-to-noise ratio in the quality assessment of human brain magnetic resonance images. BMC Medical Imaging, 18(1), 17. https://doi.org/10.1186/s12880-018-0256-6

309) Yuan, Q., Wei, Y., Meng, X., Shen, H. and Zhang, L. (2018) A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11 (3), pp. 978–989. DOI:10.1109/JSTARS.2018.2794888.

310) Yuan, Y., Li, H., & Wang, Q. (2019). Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network. In IEEE Access (Vol. 7, pp. 64676–64685). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/access.2019.2916989

311) Zhang, M., Parnell, A., Brabazon, D. and Benavoli, A. (2021) Bayesian Optimisation for Sequential Experimental Design with Applications in Additive Manufacturing. DOI:10.48550/ARXIV.2107.12809.

312) Zhang, X., Shams, S. P., Yu, H., Wang, Z., & Zhang, Q. (2023). A Similarity Measure-Based Approach Using RS-fMRI Data for Autism Spectrum Disorder Diagnosis. In Diagnostics (Vol. 13, Issue 2, p. 218). MDPI AG. https://doi.org/10.3390/diagnostics13020218

313) Zhang, Y.-D., Satapathy, S. C., Zhang, X., & Wang, S.-H. (2021). Covid-19 diagnosis via densenet and optimization of transfer learning setting. Cognitive Computation. https://doi.org/10.1007/s12559-020-09776-8

314) Zhang, Z.-H., Yang, Z., Sun, Y., Wu, Y.-F. and Xing, Y.-D. (2019) Lenet-5 Convolution Neural Network with Mish Activation Function and Fixed Memory Step Gradient Descent Method, in: 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing. Chengdu, China: IEEE, pp. 196–199.