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

Efosa Osagie
Department of Computer Science
University of Hertfordshire,
College Ln, Hatfield AL10 9AB,
Hatfield, Hertfordshire, United Kingdom
e.osagie @ herts.ac.uk

Wei Ji
Department of Computer Science
University of Hertfordshire,
College Ln, Hatfield AL10 9AB,
Hatfield, Hertfordshire, United Kingdom
w.1.ji @ herts.ac.uk

Na Helian
Department of Computer Science
University of Hertfordshire,
College Ln, Hatfield AL10 9AB,
Hatfield, Hertfordshire, United Kingdom
n.helian @ herts.ac.uk

Abstract — Generally, Medical Imaging Modalities (MIM) have 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, an enhanced CNN model for medical image character recognition (MICR) is proposed in this paper. The Lenet-5 architecture inspires this proposed Model. To further enhance this new technique to recognise visually similar characters, this paper proposes an ensemble classifier of CNN base learners. Intensive experiments are done using an open-source medical imaging dataset. 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. Finally, our proposed models and the Lenet-5 are compared using the Medpix medical image collection. Further investigation shows that our proposed model shows a 10% increase in accuracy compared with the base model and other past algorithms in recognising burned-in textual data on medical imaging modalities.

Keywords— Medical Image Character Recognition, Ensemble, Burned-in Textual data recognition, Lenet-5, Optimization

I. 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. Medical Imaging Modalities (MIM) comprise imaging of human body parts scanned using specialised acquisition devices during a medical examination. These 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 Medical Image Character Recognition (MICR) research. However, these MIM possess complicated features: 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 Fig. 1 below.

Fig. 1: X-ray sample image

(The outlined patch on the X-ray image on the right is magnified and shown on the left, containing textual burned-in data)

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 [1] and can only recognise textual data on printed and scanned document images [2]. 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) [3]. 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 [4], while MIM are 150 dpi – 72 dpi. To solve this problem, earlier proposals used varying image pre-processing techniques to enhance these MIM and feed them to these conventional OCRs. After that, machine learning algorithms (ML) (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, deep learning (DL) models,

979-8-3503-1017-7/23/$31.00 ©2023 IEEE

especially convolutional neural networks (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 [5]. High classification accuracy of CNN has been recorded compared to Multiple Layer Perceptron and Probabilistic networks [6]. 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 [5]. The Lenet-5 uses an efficient combination of convolutional layers to extract important 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 [7].

Still, within the aspect of MICR, there is an associated challenge in recognising visually similar characters (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 [8]. 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 visually similar characters, reducing the model's confidence. In this study, we propose to use a consensus of enhanced models trained on different subsets of the datasets, where each model represented 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 to contain 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 been applied recently in recognising 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 study is to propose an enhanced CNN model inspired by the Lenet-5 architecture to recognise these burned-in textual in MIM at a character level. A majority voting algorithm is employed to improve the recognition of visually similar characters. This study will propose an enhanced CNN model suited for MICR. Finally, a comparison of the proposed enhanced CNN model and the state-of-the-art will be made to show the improvement achieved. The paper is organised in the following section. Section II reviews the related works in MICR. Section III provides the contributions of this study. Section IV discusses the proposed CNN-based ensemble model inspired by the Lenet-5 and includes justifications for the modifications done. Section V describes the experimental setup. Section VI presents the experimental result. Finally, section VII presents the conclusion of this study.

II. 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 [9] 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%. Ref [10] 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 [11] 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 [12]. The authors [12] included a weighted similarity in combination with the user-assisted revision. Ref [13] 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. The authors in [14] applied local feature extraction and the 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 [14]. Ref [15] 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 [15] 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. With further advancement in pattern recognition using ML algorithms, CNN was developed. CNN has received much attention in different image recognition tasks because of their advantage in learning the representation of images and its high classification accuracy. The learning layer of the CNN, known as the convolutional layer, can learn weights by sliding across an input tensor using pre-determined kernel sizes. Recent authors proposed a CNN-based recognition model for burned-in textual data on MIM [16]. The CNN model [16] proved better than previous ML algorithms, with an accuracy of 87.5%. The design in [16] 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 understand the spatial relationships between different parts of the image[16]. Ref [16] 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. [17] used the same CNN model as [16] 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 [16]. More recently, authors [18] 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 [18] 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 [18]. These past works [14, 15, 16, 18] concluded that their models were further limited by the challenge of recognising similar characters such as "U" and "V". Therefore, in recognising characters in MIM, consideration has also to be given to the similar characters to improve the model's confidence.

In the general OCR domain, several studies [19, 20, 21] used ensemble learning to improve the recognition of handwritten characters with the consideration of visually similar characters. These authors [19, 20, 21] did not explicitly specify the resolution of their work but agreed on the problem of background interference. But there has not been much work in recognising 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, 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.

It is seen that the problem of designing a reliable MICR solution remains an unsolved problem considering the challenges of low-resolution and background interference in MIM. Therefore, there is a need for a model which can learn enough representation from the low-resolution MIM dataset. In the general domain of OCR, the Lenet-5 is recognised as a pioneer model from which other advanced models were developed [22]. 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% [22]. Most recent studies using CNN for MICR have designed only a single classifier [16, 17, 18]. 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 visually similar characters. These limitations motivated this paper's work to propose an enhanced CNN model enhanced using Bayesian reasoning for the MICR task. A majority ensemble is employed to tackle the problem of distinguishing visually similar characters using a consensus algorithm. The proposed CNN model is motivated by the Lenet-5 architecture. To the best of our knowledge, no literature has employed the optimisation of the Lenet-5 architecture and implemented the advantage of ensemble learning to recognise burned-in textual data on MIM.

III. CONTRIBUTION

In this study, we have taken a novel approach to the problem of recognising burned-in textual data on MIM with background interference and a low resolution of 96 dpi. This study builds on specific existing ideas. Our main contributions 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 and in our study, these are replaced by a 3x3 filter size and a 5x5 filter size with a stride of 2, respectively. This enhancement ensures that our 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 Medpix medical image collection is used to train and evaluate our enhanced CNN model. We manually annotated a low-resolution character dataset from the Medpix medical image collection. We achieved an outstanding accuracy score. MICR at such low resolution has not been previously reported in the literature.

• To analyse the impact of dataset size on the enhanced model's performance. We have used the bootstrapping method to create three (3) training sample subsets of the low-resolution character dataset. 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 our model's performance in distinguishing visually similar characters.

IV. PROPOSED MODEL

Our 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 design sub-section. The following sub-sections discuss the network design and optimisation techniques.

A. NETWORK HYPERPARAMETER OPTIMIZATION

Numerous CNN architectures and modifications have been done, including existing ones from past works in the medical imaging domain, such as segmentation and classification. Although many of these network designs have

been successful in various tasks, no particular network design can suit all existing problems. Hence, there is a need to design a model to suit a specific task. 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 [23], 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 we know the function values taken from the Lenet-5 as our base model, but we are not certain of the impact of increasing or decreasing these functional values. BO can achieve the global minima with the smallest loss function value [24]. 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 less computational cost. However, the BO has some instability limitations, particularly in dealing with a large hyperparameter search space because of the curse of dimensionality [25]. Several recent experimental studies 26, 27, 28] have shown that BO is practically limited to optimising less than 20 parameters. In this study, though our parameters are less than 20, we desired a reduction of the iterations needed for BO. To reduce the computational cost for BO iterations, this study combined a search space pruning mechanism known as the Successive Halving Algorithm (SHA). SHA is an advanced early-stopping method to determine 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, we focus only on optimising our 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 will be provided about the optimisation algorithms. 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 study [29]. The Optuna library uses the Tree-structured Parzen Estimator (TPE) for the BO. The TPE can build a model by applying Bayesian reasoning to balance the exploration versus exploitation trade-off. Several recent studies [30, 31, 32, 33] have agreed that the TPE is a notable BO estimator to optimise hyperparameters to ensure the strong performance of deep learning models. We can pass a function for the optimisation, specify the number of iterations and visualise the hyperparameter importance. As a first study, we ran hyperparameter tuning for about 100 trials and then checked which hyperparameters were the most important. Next, we omitted the less important hyperparameters for the subsequent studies up to 1500 trials. The flow chart of the optimisation process is shown below in Fig. 2.

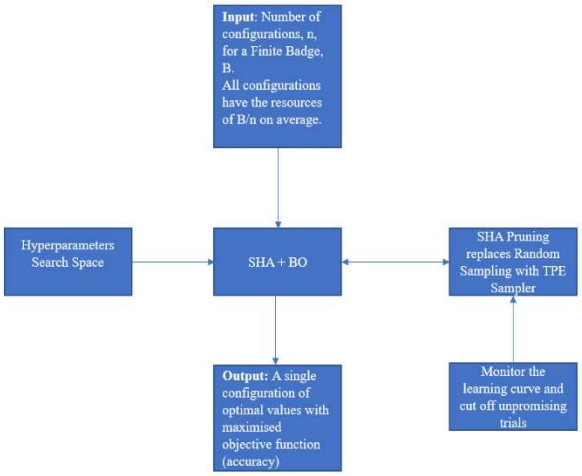

Fig 2. Flow chart of the optimisation process

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, optimiser, kernel size, strides, Number of convolutional layers, Number of filters, Number of layers, Number of dense units and drop-out rate.

B. DESIGNED MODEL

The detailed layerwise summary of our MICR model is shown in Table 1. Our 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. We 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 [34, 35] agreed that this replacement improves the model's expressiveness ability. We applied 128 neurons for the dense layers. In addition, the visual representation of the model is shown in Fig 3. to show the network architecture. The enhanced CNN model is a relatively simple yet efficient model for the desired task to recognise 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 will be 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 [36, 37]. Dropout was added to avoid over-fitting [38], 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 in [16, 17, 18], our 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. Our design solved the limitations of [17] to recognise certain font sizes and styles by replacing pooling layers with learnable downsampling layers, which is targeted at our problem of recognising characters in low-resolution MIM as key features are small and local. This approach is an inspiration from [34], and it increases our model's expressiveness ability.

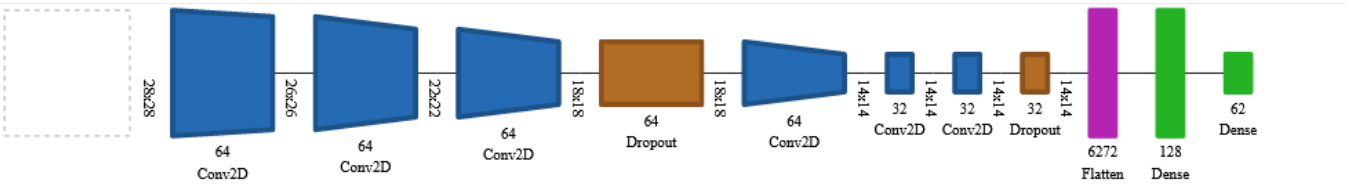

Fig. 3: Proposed Enhanced MICR model

TABLE 1: LAYERWISE SUMMARY OF THE MICR 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)	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)	2, 2, 32	0		
flatten_1 (Flatten)	128	0		
dense_2 (Dense)	128	16512		
dense_3 (Dense)	62	7998		
Trainable parameters: 377,790				

C. *MAJORITY VOTING ENSEMBLE*

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 bootstrapping method and (b) combine each prediction of the ensemble members, to get a consensus classification outcome. 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 [38]. 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, we 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. [39] that 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 [39]. Fig. 4 below shows the framework of the majority voting algorithm used in this study.

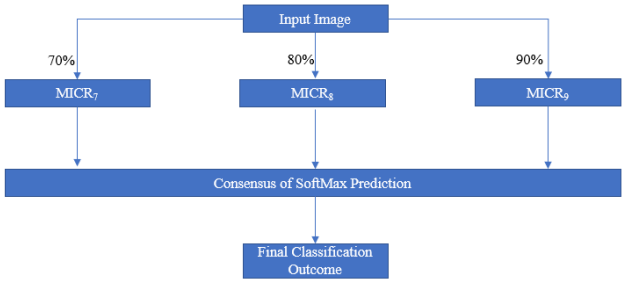

Fig. 4: Majority voting ensemble approach used

The bootstrapping used in the majority voting ensemble reduces the variance of the prediction model, reduces the overfitting and balances the bias-variance trade-off. These advantages are useful when recognising visually similar characters. After creating three subsets from the training dataset, a classifier is fitted on each classifier and trained along with data augmentation techniques.

V. EXPERIMENTAL SET-UP

A. *Dataset Description*

The medical image dataset used to test our proposed MICR model is the Medpix medical image collection. Medpix medical image dataset is open source and contains 60,613 image collections of ultrasounds, X-rays, MRI and CT. We manually curated 3050 character image patches from the collection to form a character dataset. In addition, the dataset contains burned-in textual data representing various medical interpretations of the images. The dataset consists of 62 classes (A-Z, a-z, 0-9), having an average of 40 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.

B. 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 variability to the data and minimising data overfitting [40]. The SHA and BO combination, as explained in Network Design Section IV, 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 20 degrees, either clockwise or anticlockwise. Checkpoints were initialised during the training to determine the best epoch for the training. The validation accuracy was monitored here, and only the best weights were saved. The Adam optimiser was used to ensure faster convergence. A batch size of 28 was used during the training.

VI. RESULTS

The configuration of the experimental environment, including the optimisation, is as follows; Python 3 Google Compute Engine backend (GPU) of 15.0 GB RAM. The comparison of the proposed MICR model and the Lenet-5 is shown in Table 3 below. All results presented are from experiments carried out on the Medpix Medical image dataset, and the improvement is shown. Our proposed model is simpler with fewer parameters, yet more efficient compared to past works in recognising burned-in textual data on MIM. The proposed ensemble is represented as MICR (n). n is an odd number, as each member of the ensemble is entitled to a single vote. The proposed MICR model is compared with Lenet-5 on the bootstrapped subsets of the training samples. We will represent the trained models for the ensemble as explained in the majority voting sub-section in IV. The results are shown below, averaged on 30 runs at 100 epochs with checkpoints to save the best weights. The experiments support our hypothesis that our enhanced MICR model is more accurate than the base OCR model (Lenet-5) and other existing algorithms after an extensive literature review.

TABLE 2: COMPARISON OF THE LENET-5 AND MICR MODELS

Model	**Subset (%)**	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**F-1 Score (%)**	**Training Time (s)**
Lenet-5	70	78.36	78.36	76.32	75.89	287.79
	80	77.69	77.69	74.11	75.03	292.90
	90	69.18	69.18	64.84	65.23	311.50
MICR	70	89.82	89.82	88.72	88.68	238.72
	80	89.87	89.87	87.71	88.29	248.23
	90	90.33	90.33	89.43	89.24	273.70
MICR(3) $MICR_7$ +$MICR_8$ +$MICR_9$	70	94.49	94.46	94.49	94.08	713.53
	80	93.99	93.99	93.59	93.29	765.99
	90	93.05	93.05	91.39	91.82	836.10

The test dataset for $MICR_7$, $MICR_8$ and $MICR_9$ was 30%, 20% and 10% of the overall dataset, respectively, after using the bootstrap sampling method to create the training subsets. The training subsets of the overall dataset are shown in Table 2. The learning curves for the MICR model showing the accuracy and loss over time for 100 epoch is shown in Fig 5 below. The testing accuracy and loss are represented by the "validation" legend on the plots.

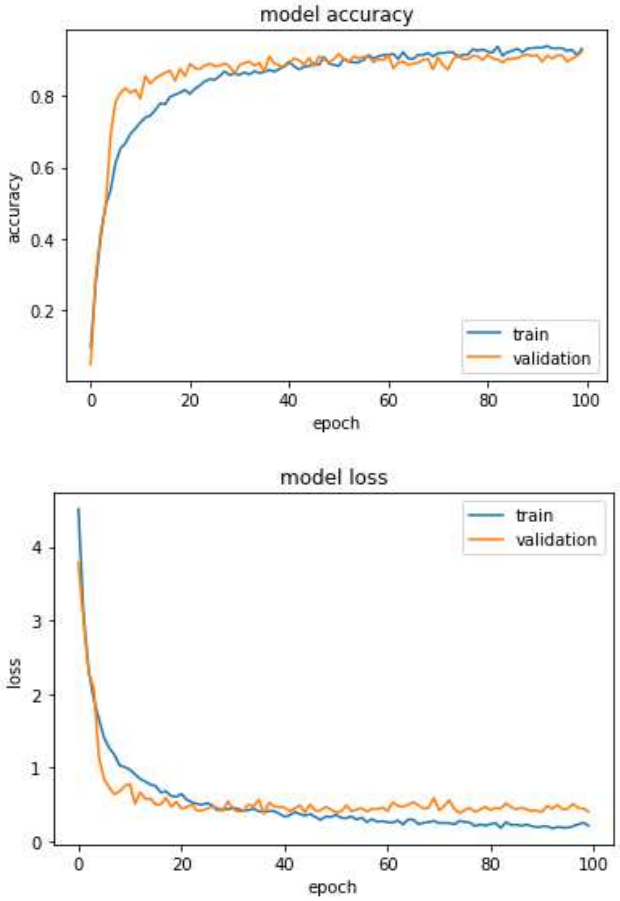

Fig 5: learning curves for MICR model

We further compare our work with existing algorithms in the domain of burned-in textual recognition in medical imaging modalities, which is shown in Table 3 below. The train and test datasets for the proposed ensemble are 70% and 30%, respectively.

TABLE 3: COMPARISON OF PROPOSED MODELS AND RECENT WORKS ON THE MEDPIX DATASET

Method	**Recall (%)**	**Precision (%)**	**F-1 Score (%)**
Modified CRNN [18]	65.00	67.00	70.00
CNN [16]	78.95	83.05	79.73
Proposed MICR	**89.89**	**88.61**	**90.56**
Proposed Ensemble	**94.46**	**94.49**	**94.49**

A. RESULT COMPARISON

The result of this study outperformed most of the existing works in the domain of burned-in textual recognition at the character level. In Table 3, a comparison of this study with other works which designed a classifier was presented after an extensive literature review. The result proved that our work

has outstanding results in terms of 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 [18], which was a multiscale CRNN. The work [18] is the most recent work and used the same dataset as ours (Medpix dataset). The authors in [18] reported an F-1 score of 70.00%, while our proposed CNN model outperformed with 90.56%. Our majority voting ensemble also performed better than the CRNN model from work in [18]. Our proposed model also outperformed notable works by authors [16,17], whose CNN model obtained an F-1 score of 79.73% on the same dataset as ours. Our proposed MICR model and evaluated on 62 classes of characters, which were manually annotated by this study. The majority voting ensemble based on the bootstrapping data varying method had more accuracy than most existing works from the literature on the MICR domain.

The application results of our proposed ensemble model in recognising burned-in textual data in MIM are shown in Fig 6 below;

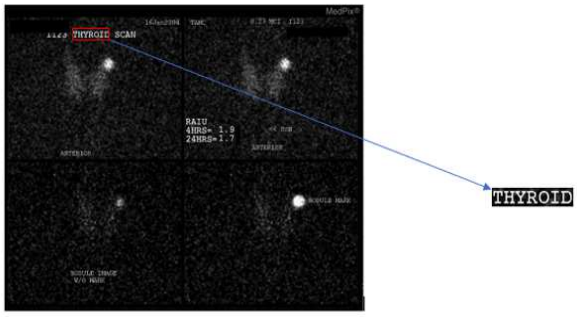

Fig 6. : Recognition results of low-resolution MIM with background interference

As seen in Fig. 6, 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.

VII. CONCLUSION

This study 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 our 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. Three (3) 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. Our 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 multi-scale CNN architecture and a more advanced ensemble will be considered to boost the performance. In addition, we will include an attention mechanism to selectively give more relevance to some areas of the input image compared to others. The attention mechanism will increase the representation power of interests, as supported by authors in [41, 42].

REFERENCES

[1] S. Drobac and K. Lindén, "Optical character recognition with neural networks and post-correction with finite state methods," IJDAR, vol. 23, no. 4, pp. 279–295, Dec. 2020, doi: 10.1007/s10032-020-00359-9.

[2] T. W. Ramdhani, I. Budi, and B. Purwandari, 'Optical Character Recognition Engines Performance Comparison in Information Extraction', IJACSA, vol. 12, no. 8, 2021, doi: 10.14569/IJACSA.2021.0120814.

[3] S. Istephan and M.-R. Siadat, 'Unstructured medical image query using big data – An epilepsy case study', Journal of Biomedical Informatics, vol. 59, pp. 218–226, Feb. 2016, doi: 10.1016/j.jbi.2015.12.005.

[4] O. J. Oni and F. O. Asahiah, 'Computational modelling of an optical character recognition system for Yorùbá printed text images', Scientific African, vol. 9, p. e00415, Sep. 2020, doi: 10.1016/j.sciaf.2020.e00415.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[6] G. Wei, G. Li, J. Zhao, and A. He, 'Development of a LeNet-5 Gas Identification CNN Structure for Electronic Noses', Sensors, vol. 19, no. 1, p. 217, Jan. 2019, doi: 10.3390/s19010217.

[7] Z.-H. Zhang, Z. Yang, Y. Sun, Y.-F. Wu, and Y.-D. Xing, '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, Dec. 2019, pp. 196–199. doi: 10.1109/ICCWAMTIP47768.2019.9067661.

[8] D. Pal, A. Alladi, Y. Pothireddy, and G. Koilpillai, '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, Jun. 2021, pp. 1–6. doi: 10.1109/EUVIP50544.2021.9483980.

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

[10] Wang, J. 'Security filtering of medical images using OCR. School of Information Sciences and Technology'. Pennsylvania State University, 2002

[11] Alter D, and Werner, A. 'Automatische texterkennung (ocr) in ultraschall'. Konferenz der SAS-Anwender in Forschung und Entwicklung,2007

[12] G. K. Tsui and T. Chan, 'Automatic Selective Removal of Embedded Patient Information From Image Content of DICOM Files', American Journal of Roentgenology, vol. 198, no. 4, pp. 769–772, Apr. 2012, doi: 10.2214/AJR.10.6352.

[13] P. Vcelak, M. Kryl, M. Kratochvil, and J. Kleckova, 'Identification and classification of DICOM files with burned-in text content', International Journal of Medical Informatics, vol. 126, pp. 128–137, Jun. 2019, doi: 10.1016/j.ijmedinf.2019.02.011.

[14] Yu Ma and Yuanyuan Wang, '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, Aug. 2015, pp. 953–958. doi: 10.1109/FSKD.2015.7382072.

[15] E. Monteiro, C. Costa, and J. L. Oliveira, '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, Aug. 2015, pp. 1381–1384. doi: 10.1109/EMBC.2015.7318626.

[16] E. Monteiro, C. Costa, and J. L. Oliveira, 'A De-Identification Pipeline for Ultrasound Medical Images in DICOM Format', J Med Syst, vol. 41, no. 5, p. 89, May 2017, doi: 10.1007/s10916-017-0736-1.

[17] J. M. Silva, E. Pinho, E. Monteiro, J. F. Silva, and C. Costa, 'Controlled searching in reversibly de-identified medical imaging archives', Journal of Biomedical Informatics, vol. 77, pp. 81–90, Jan. 2018, doi: 10.1016/j.jbi.2017.12.002.

[18] X. Xu, W. Wang, and Q. Liu, '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, Jun. 2021, pp. 408-412. doi: 10.1109/SPAC53836.2021.9539999.

[19] R. Caruana, " Multitask Learning " Machine Learning, vol. 28, no. 1, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.

[20] J. Hou, H. Zeng, L. Cai, J. Zhu, J. Cao, and J. Hou, 'Handwritten numeral recognition using multi-task learning', in 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Xiamen, China: IEEE, Nov. 2017, pp. 155–158. doi: 10.1109/ISPACS.2017.8266464.

[21] Z. Chen, Y. Wu, F. Yin, and C.-L. Liu, "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, Nov. 2017, pp. 525–530. doi: 10.1109/ICDAR.2017.92.

[22] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, 'An Introductory Review of Deep Learning for Prediction Models With Big Data', Front. Artif. Intell., vol. 3, p. 4, Feb. 2020, doi: 10.3389/frai.2020.00004.

[23] M. Zhang, A. Parnell, D. Brabazon, and A. Benavoli, 'Bayesian Optimisation for Sequential Experimental Design with Applications in Additive Manufacturing', 2021, doi: 10.48550/ARXIV.2107.12809.

[24] Y. Gao, T. Yu, and J. Li, "Bayesian optimization with local search," 2019, doi: 10.48550/ARXIV.1911.09159.

[25] D. Eriksson and M. Jankowiak, 'High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces', 2021, doi: 10.48550/ARXIV.2103.00349.

[26] R. Moriconi, M. P. Deisenroth, and K. S. S. Kumar, "High-dimensional Bayesian optimization using low-dimensional feature spaces," 2019, doi: 10.48550/ARXIV.1902.10675.

[27] P. I. Frazier, 'A Tutorial on Bayesian Optimization', 2018, doi: 10.48550/ARXIV.1807.02811.

[28] Md. A. Awal, M. Masud, Md. S. Hossain, A. A.-M. Bulbul, S. M. H. Mahmud, and A. K. Bairagi, "A novel bayesian optimization-based machine learning framework for covid-19 detection from inpatient facility data," IEEE Access, vol. 9, pp. 10263–10281, 2021, doi: 10.1109/ACCESS.2021.3050852.

[29] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, 'Optuna: A Next-generation Hyperparameter Optimization Framework', 2019, doi: 10.48550/ARXIV.1907.10902.

[30] S. Watanabe and F. Hutter, "C-tpe: generalizing tree-structured parzen estimator with inequality constraints for continuous and categorical hyperparameter optimization," 2022, doi: 10.48550/ARXIV.2211.14411.

[31] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "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.

[32] G. Rong et al., 'Comparison of Tree-Structured Parzen Estimator Optimization in Three Typical Neural Network Models for Landslide Susceptibility Assessment', Remote Sensing, vol. 13, no. 22, p. 4694, Nov. 2021, doi: 10.3390/rs13224694.

[33] Y. Ozaki, Y. Tanigaki, S. Watanabe, and M. Onishi, 'Multiobjective tree-structured parzen estimator for computationally expensive optimization problems', in Proceedings of the 2020 Genetic and Evolutionary Computation Conference, Cancún Mexico: ACM, Jun. 2020, pp. 533–541. doi: 10.1145/3377930.3389817.

[34] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, 'Striving for Simplicity: The All Convolutional Net', 2014, doi: 10.48550/ARXIV.1412.6806.

[35] H. Mureşan and M. Oltean, "Fruit recognition from images using deep learning," 2017, doi: 10.48550/ARXIV.1712.00580.

[36] U. M. Tomasini, L. Petrini, F. Cagnetta, and M. Wyart, "How deep convolutional neural networks lose spatial information with training," 2022, doi: 10.48550/ARXIV.2210.01506.

[37] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J Big Data, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[38] Goodfellow, I., Bengio, Y., & Courville, A. " Deep Learning" . MIT Press, 2016.

[39] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. " An introduction to statistical learning : with applications in R ". New York :Springer, 2013

[40] C. Shorten and T. M. Khoshgoftaar, 'A survey on Image Data Augmentation for Deep Learning', J Big Data, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[41] G. Li, Q. Fang, L. Zha, X. Gao, and N. Zheng, 'HAM: Hybrid attention module in deep convolutional neural networks for image classification', Pattern Recognition, vol. 129, p. 108785, Sep. 2022, doi: 10.1016/j.patcog.2022.108785.

[42] M.-H. Guo et al., 'Attention mechanisms in computer vision: A survey', Comp. Visual Media, vol. 8, no. 3, pp. 331–368, Sep. 2022, doi: 10.1007/s41095-022-0271-y.