Machine Learning 2
Group Project

Cover Page

ID	Contribution
ID1	100%
ID2	100%
ID3	100%

- Do all the members agree with the above contributions? Yes

- We have attached on link that contains both our trained model and the processed dataset (folder ‘emnist letters’).
  https://drive.google.com/drive/folders/1D1tBg7PMuALKkIesyAymA-vN-JBMTA8f?usp=sharing
  The link also contains a folder of some test images that are used and a requirements.txt file that contains all its dependencies if it was to be run in a Jupyter Environment.

- The dataset was obtained in binary format from:
  https://www.nist.gov/itl/products-and-services/emnist-dataset

Machine Learning 2
Group Project

Real Life Problem and Motivation

The recent outbreak of the COVID-19 pandemic was an unprecedented event that shook the world and forced many industrial sectors to change how their working systems work. One of the sectors that was greatly affected and caught off-guard was the educational sector. Schools all around the world were forced to shut down after the World Health Organisation issued a statement regarding the human-to-human transmission related to COVID-19, leaving over 1.2 billion children out of their classroom. This has consequently resulted in the emergence of a new hybrid teaching system where students would receive their teaching through online classes and courses (World Economic Forum, 2020). Before the pandemic students had to physically attend their final exams during their examination periods, but new plans had to be established by all schools and universities which would organise a new structure of online examinations for the students. An organisation that yearly has been marking over 8 million examination papers each year would be the Cambridge Assessment Group (OCR, 2021). They are the organisation responsible for the IGCSE and GCSE papers and have adapted the Optical Character Recognition (OCR) system which is utilised to scan handwritten text and process it to a digital document. This technology is processing over 1 million documents each year, giving the ability to examiners to mark the papers remotely from any place of the world (Marshall, 2020). The main advantage of this technology is that examiners are able to look at a digital document and not a handwritten paper, which is making it easier to be read and as a result this makes them more efficient and unbiased for all the papers. With the demand of such systems now rising more than ever, this research project aims to evaluate the accuracy and robustness of such OCR technologies as this new hybrid educational system may become a permanent phenomenon, even after the pass of the pandemic. Even though the concept of distant learning is associated with several advantaged and disadvantages, the driving motivation for such research is to enforce the educational sector with the necessary tools to create an unbiased system that is fair to all students, by assessing the challenges related with OCR systems. A replica OCR system will be created which will be used to evaluate its accuracy when trained to classifying different characters and reveal if there any limitations arise when there is added noise. This system will furthermore be used to process handwritten text to digital documents.

Dataset Used

The importance of selecting the appropriate dataset is often understated, as it will act as the benchmark for providing a solution that will fairly compare, analyse, and evaluate the different approaches and techniques used, while also offering the adequate insights about the performance of an algorithm (Cohen, Afshar, Tapson and van Schaik, 2017). The dataset that has been chosen to train the OCR system was the EMNIST dataset. The EMNIST dataset was created as an extended version of the MNIST dataset (Lecun, Bottou, Bengio and Haffner, 1998), with the purpose of ensuring the usefulness and longevity of the dataset in existing and new classification systems. It contains a set of handwritten characters, which were processed and converted to an 28x28 image in pixel format. The dataset is available to download at (The EMNIST Dataset, 2017), where there are several variations of character splits provided. For this project, the 'EMNIST ByLetter' dataset has been selected which contains 145,600 characters of 26 balanced classes. These classes include the 26 different classes of lowercase characters (a – z) and the 26 different classes of uppercase characters (A – Z) merged together. This dataset was chosen as it is big enough to represent a real-life situation of the OCR classifying text from many different handwriting styles, allowing the performance of the system to be established as trustworthy. When the model achieves the expected results, the

robustness of the classifier will be put on test by setting it to predict classes from images beyond the EMNIST dataset. This dataset will consist of images that you can draw on your computer using your mouse on screen, you can draw different numbers or letters. This will test the object detection, which allow the program to detect how many numbers or letters have you draw, and the classification which it will tell you what the object is from the 'EMNIST' dataset. An example of how the different classes in the dataset are represented by the 28x28 images is represented in figure 1.

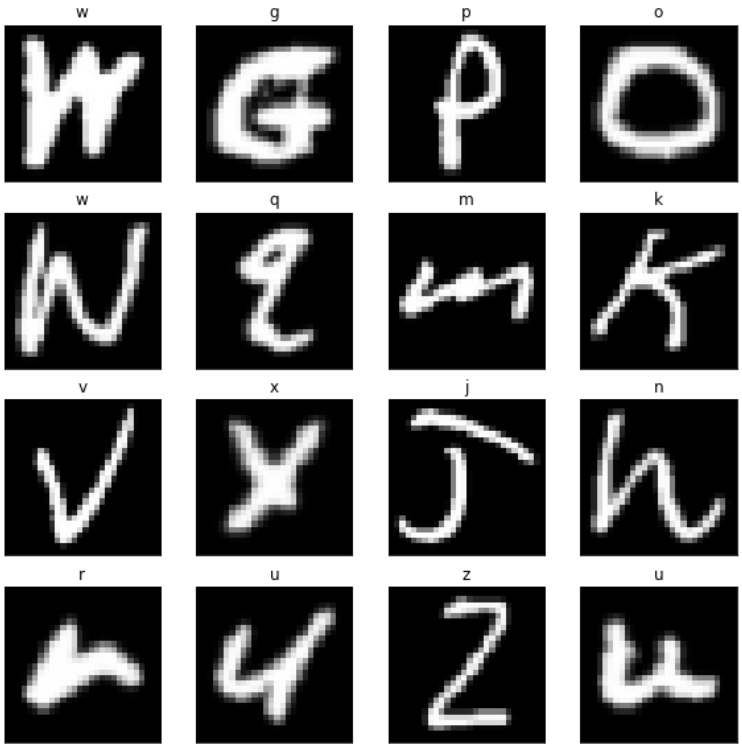

Figure 1: Dataset used with classes and labels

Deep learning models used

Deep learning has networks where each layer is able of gradually extracting information from the input data (What is Deep Learning?, 2020). The model used consists of a sequential model since the dataset consist of a single input which is an image of 28x28 pixels and a single output which is the number of classes that we are classifying. The model at the first layer flattens features extracted from the 28x28 pixels image to a vector. The vector is then passed to the fully connected layers which apply an activation function called Rectified Linear Unit (ReLu) twice. The 'ReLu' activation function which is used at the second layer consist of dense 512 neurons. Moreover, the next layer consists of the 'ReLu' activation function but utilises 128 neurons. Basically, using the fully connected layers we have that all the inputs from the previous layer are connected to every activation unit of the next layer. The model for output layer of activation function uses the 'SoftMax' activation function which is used for multi-class classification problems. It is used as the activation function of the output layer since it will be used to normalise the different probabilities of each input being classified to a specific class (Elijah Koech, 2020). Lastly, the model uses the Adam optimizer to train the model using stochastic gradient descent. Adam optimizer was chosen since it is well suited for noisy data and in general it is applicable on problems that have a large set of datasets (Kingma and Ba, 2017).

Train and Evaluation

To have a good representation of the actual performance of the classifier that will be used to predict the 26 different classes of characters, the EMNIST dataset will be split into an 80% train dataset and 20% test dataset. On many occasions, in order to validate the accuracy of the training set the K-fold cross validation method is used which splits the training dataset into a predefined number of '*K'* folds and then make predictions and evaluate them on each fold by using a model trained on the rest of the remaining folds. This will not be used as the evaluation method for this classifier though, as the classes of the dataset are unbalanced, and it will not result in reliable accuracy results. Instead of that a confusion matrix will be created which will illustrate the strengths and weaknesses of the algorithm when it comes to predicting classes, by creating the rows representing the actual classes and the columns representing the predicted classes. Furthermore, the confusion matrix can be used to calculate the 'precision' which illustrates the accuracy of positive predictions and 'recall' which illustrates the ratio of positive instances that have been correctly detected. After calculating the prediction probability of classes, a variable threshold value can be set which will set the Precision/Recall trade-off of the algorithm.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

After setting the desirable threshold, the robustness of the classifier will be put on test by setting it to predict classes from images beyond the EMNIST dataset. This will allow an observation of how well the OCR system will perform in both the object detection and classification tasks on different image structures.

Results found

The results that have been collected after training the algorithm were very promising as a 95.27% training accuracy and a 91.97% testing accuracy has been found. Figure 2 illustrates how the training and validation loss and accuracy have altered during each epoch, representing a fall in the loss and a rise in the accuracy. The model was to run for a total of 50 epochs, but an early stopping function was used which was set to stop training if the past 10 epochs did not show any significant changes, therefore avoiding any potential overfitting of the model, and this occurred at epoch number 26.

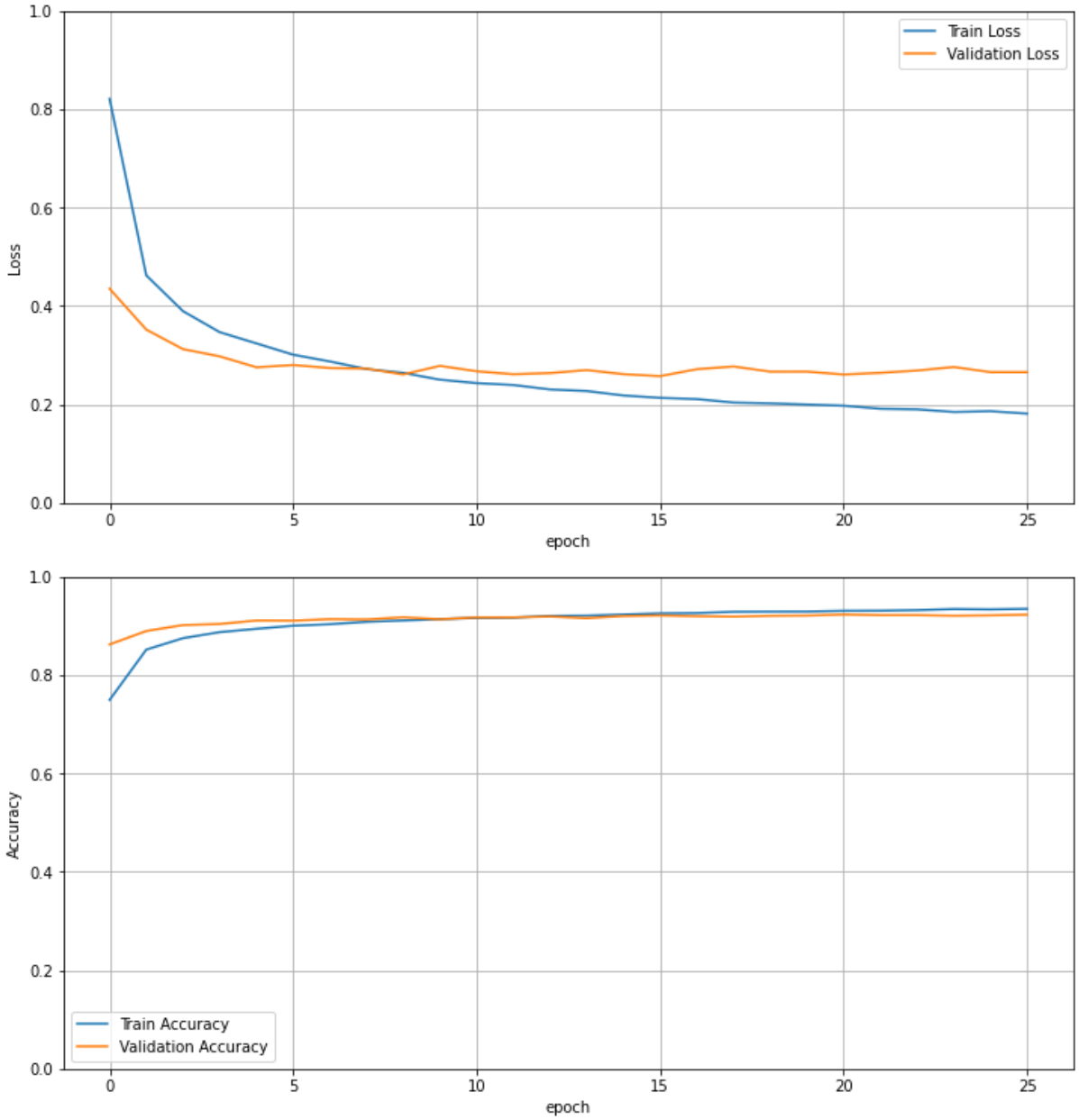

Figure 2: Accuracy and Loss during each epoch

To get a better insight about the performance of the algorithm, a confusion matrix has been formed which illustrated the predictions and actual labels of each of the 26 characters and is illustrated in figure 3 below.

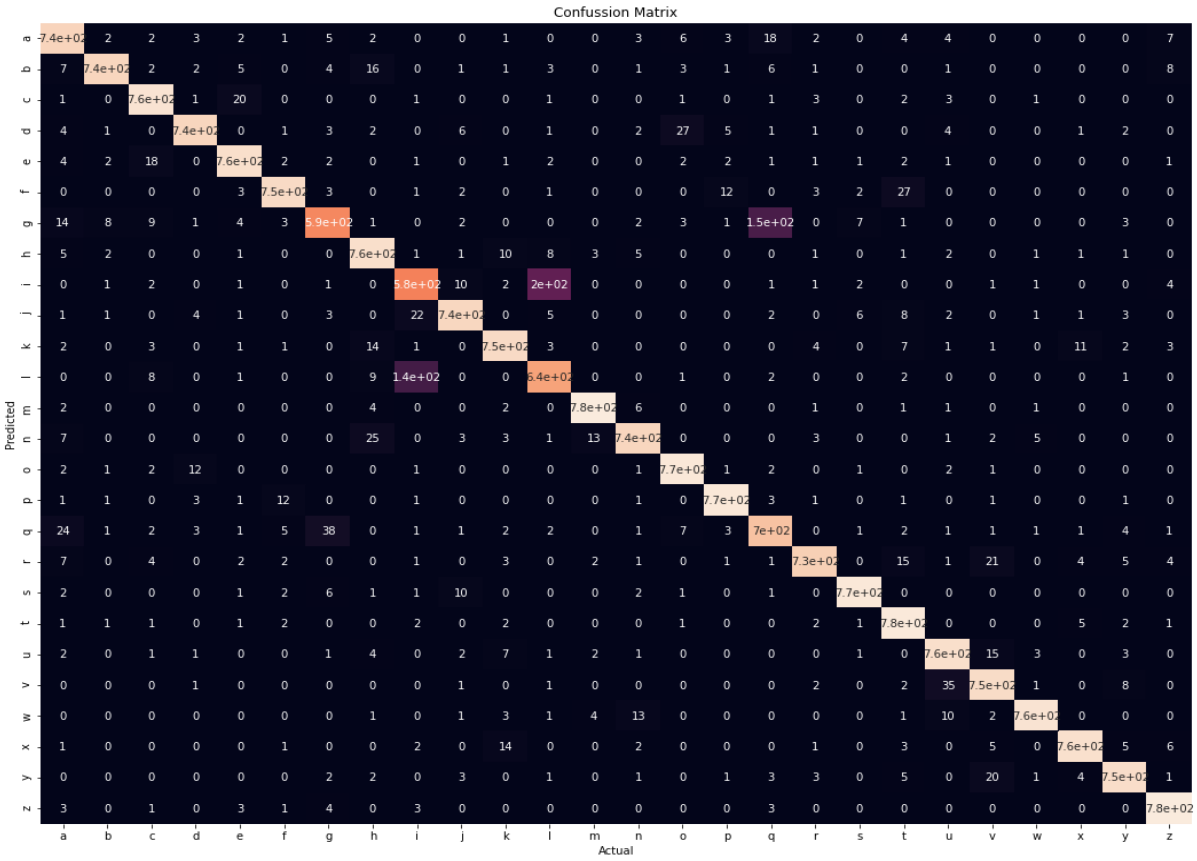

Figure 3: Confusion Matrix of classifier

The diagonally coloured line on the confusion matrix is evidence that most of the predictions made by the algorithm were the actual classes of each character, while the purple boxes indicate the weaknesses that the classifier has. As expected, the classifier's most common mistakes were confusing an 'l' with an 'i', and a 'g' with a 'q', since the two letters are very identically shaped which on some occasions did not allow the classifier to distinguish between them.

A demo program has been created which was further used to test the classifier and object detector with our own handwritings. Some examples showing the successful algorithms performance of the algorithms are shown in figure 4.

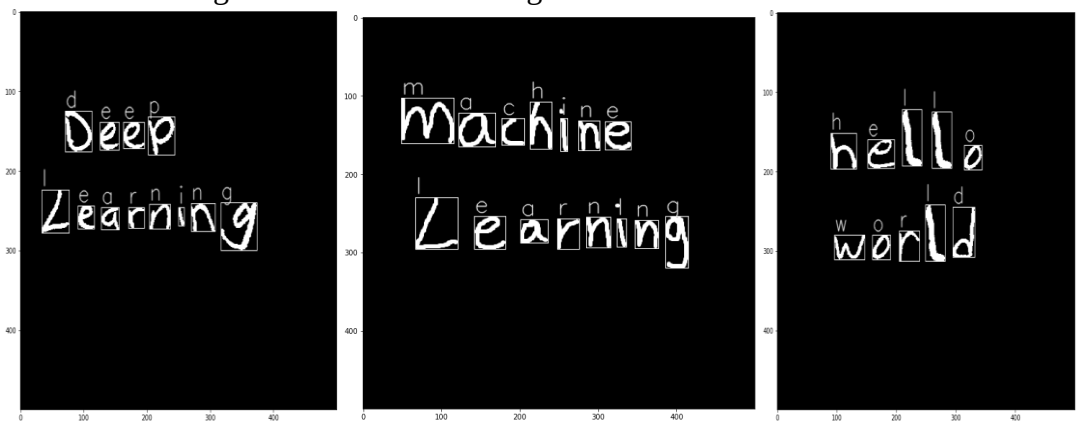

Figure 4: Classification and Object detection examples of the OCR system created

Alternative methods

An alternative method that would have been a much more suitable method for an OCR approach is to use an out of the box object detection method. This alternative would have consisted of the use of feature extractions within the deep convolutional neural networks and proper classification on all of the classes that have chosen to be used. Object detection usually uses features that are handcraft and architects that use a lot of training (Zhao, Zheng, Xu and Wu, 2019). For this project, a picture of low-level quality has been used which underestimates the results of the object detection model. In contrast deep learning-based object detection frameworks use only Convolutional Neural Networks (CNN) which consist of image input and extraction of region proposals which is the main characteristic which has not been utilised, and it could have made the model better. Using the extraction of region proposals, the program can separate the objects in the picture into different pictures and use the OCR method to predict the letter or number that each picture has and combine them to conclude to what are the predictions of the letters that are input (Zhao, Zheng, Xu and Wu, 2019).

From other papers which have been studied, it can be observed though that the classification approach which has been utilised in this project has performed much better. When comparing other classification approaches on the same dataset it has been found that:

- (Ahlawat and Choudhary, 2020) has achieved an 88.56% accuracy using a KNN classifier and an 89.51% using a SVM classifier.
- (Cavalin, Sabourin, Suen and Britto Jr., 2009) applied a hidden Markov model, achieving an accuracy of 90.00%
- (Coerih and Calva, 2005) applied a multi-layer perceptron achieving an accuracy of 87.79%
- (Dufourq and Bassett, 2017) used evolutionary deep networks for classification achieving an accuracy of 88.30%

References

1. A. L. Koerich and P. R. Kalva, 2005, "Unconstrained handwritten character recognition using metaclasses of characters," *IEEE International Conference on Image Processing 2005*, pp. II-542, doi: 10.1109/ICIP.2005.1530112.
2. Ahlawat, S. and Choudhary, A., 2020. Hybrid CNN-SVM Classifier for Handwritten Digit Recognition. Procedia Computer Science, 167, pp.2554-2560.
3. Cavalin, P., Sabourin, R., Suen, C. and Britto Jr., A., 2009. Evaluation of incremental learning algorithms for HMM in the recognition of alphanumeric characters. Pattern Recognition, 42(12), pp.3241-3253.
4. Cohen, G., Afshar, S., Tapson, J. and van Schaik, A., 2017. EMNIST: an extension of MNIST to handwritten letters.
5. Dufourq, E.; Bassett, B.A., 2017, EDEN: Evolutionary Deep Networks for Efficient Machine Learning., arXiv:1709.09161.
6. Elijah Koech, K., 2020. Softmax Activation Function—How It Actually Works.
7. Ibm.com. 2020. What is Deep Learning?.
8. Kingma, D. and Ba, J., 2017. Adam: A Method for Stochastic Optimization.
9. 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.
10. Marshall, C., 2020. How is OCR used in the real world?. [online] Medium.
11. NIST. 2017. The EMNIST Dataset.
12. Ocr.org.uk. 2021. Who we are - OCR. [online]
13. World Economic Forum. 2020. The COVID-19 pandemic has changed education forever. This is how. [online]
14. Zhao, Z., Zheng, P., Xu, S. and Wu, X., 2019. Object Detection with Deep Learning: A Review.