

Configuration Manual

MSc Research Project Data Analytics

Paras Jain Student ID: 18182119

School of Computing National College of Ireland

Supervisor:

Mr. Hicham Rifai

National College of Ireland

MSc Project Submission Sheet



School of Computing

Student Name:	Paras Jain		
Student ID:	18182119		
Programme:	MSc DATA ANALYTICS	Year:	2019-2020
Module:	MSc RESEARCH PROJECT		
Lecturer:	Mr. Hicham Rifai		
Date:	17/08/2020		
Project Title:	Liver Disease Detection from CT scan image and Transfer Learning	es using	Deep Learning

Word Count:697Page Count:14

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

<u>ALL</u> internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Paras Jain 18182119

1 Introduction

This configuration manual details the system setup, software and hardware specifications and the step taken to carry out the Research Project: Liver Disease Detection from CT scan images using Deep Learning and Transfer Learning.

In this manual, Section 2 contains information on software and hardware specifications. Section 3 discusses the environment setup, data collection and preparation, importing of libraries and mounting of Google Drive. Section4 explains the various steps involved in image processing. Section 5 describes the design of the models and show their implementation.

2 System Configuration

This section of the configuration manual provides information about the system configuration utilised to implement the project.

2.1 Hardware Requirements

Operating System	Windows 10
RAM	12.72 GB (Google Colaboratory)
Disk Space	107.77 GB (Google Colaboratory)
Storage Space	15 GB (Google Drive)

Table 1 Hardware Configuration

2.2 Software Requirements

Programming Language Tools	Google Colaboratory (Cloud-based Jupyter notebook environment), Python 3.6.9
Web Browser	Google Chrome or Mozilla Firefox
Email Account	Gmail account for Google Drive and
	Colaboraory
Other software	Microsoft Word and Microsoft Excel

Table 2 Software Configuration

3 Project Development

This section of the configuration manual provides information about the setup of environment and collection and usage of data.

3.1 Google Colaboratory Environment Setup

Google Colaboratory¹ environment is used for the research implementation. It provides Python version 3.6.9 along with powerful RAM and GPU support and is hosted on Google Cloud. Gmail account is essential to use Google Colab as it requires authentication.



Figure 1 Google Colaboratory Homepage

3.2 Data Collection

The data is collected from $Medpix^2$ which provides an online collection of medical images for research purposes. It is publicly available and an ethical data source. Image data is gathered in the form of '.jpg' files for the two categories.



Figure 2 Data Source - Medpix

3.3 Data Preparation

This section describes the approach used for preparing data. The data is divided into 3 folders on Google Drive namely, Training, Validation, and Testing.

¹ https://colab.research.google.com/notebooks/

² https://medpix.nlm.nih.gov/

	Drive	Q Search in Drive			•
+	New	My Drive > Data -			
• 4	My Drive	Folders			
õ	Shared with me	Testing	Training	Validation	
()	Recent				
☆	Starred				
Ī	Bin				
:	Storage				
	1,019.8 MB of 15 GB used				
	Buy storage				

Figure 3 Google Drive Data

My Drive > Data > Training -
Folders
Abnormal Normal
My Drive > Data > Testing -
Folders
Abnormal Normal
My Drive > Data > Validation -
Folders
Abnormal Normal

Figure 4 Train, Valid and Test folders

3.4 Importing libraries

This section shows the libraries that are needed for the implementation of the research project. The libraries available in Colab are imported and unavailable libraries are installed using pip command. OpenCV³, Tensorflow⁴ and Keras⁵ are the primarily used libraries in the research.

³ https://pypi.org/project/opencv-python/

⁴ https://www.tensorflow.org/

⁵ https://keras.io/



Figure 5 Importing Libraries

3.5 Mounting Google Drive

Google Drive needs to be mounted in Google Colab to use the data. It requires authentication through the Gmail account used for Colab.



Figure 6 Mounting Google Drive on Colab Notebook

4 Image Processing

This section provides knowledge about different steps involved in the processing of images.

4.1 Image Pre-processing

There are various image processing techniques used in the research such as Resize, Histogram Equalisation, Noise removal, etc.



Figure 7 Image Pre-processing

```
[ ] #histogram equalisation
    img = cv2.imread(image,0)
    hist,bins = np.histogram(img.flatten(),256,[0,256])
    cdf = hist.cumsum()
    cdf_normalized = cdf * hist.max()/ cdf.max()
    plt.plot(cdf normalized, color = 'b')
    plt.hist(img.flatten(),256,[0,256], color = 'r')
    plt.xlim([0,256])
    plt.legend(('cdf', 'histogram'), loc = 'upper left')
    plt.show()
[ ] cdf_m = np.ma.masked_equal(cdf,0)
    cdf_m = (cdf_m - cdf_m.min())*255/(cdf_m.max()-cdf_m.min())
    cdf = np.ma.filled(cdf_m,0).astype('uint8')
[] img2 = cdf[img]
[ ] #histogram equalisation
    img = cv2.imread(image,0)
    equ = cv2.equalizeHist(img)
    res = np.hstack((img,equ)) #stacking images side-by-side
    img = cv2.imread(image,0)
    # create a CLAHE object (Arguments are optional).
    clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(8,8))
    cl1 = clahe.apply(img)
```

Figure 8 Image Pre-processing (contd.)

4.2 Image Augmentation

Image augmentation is done to increase the size of the dataset. Various techniques such as flip, rotate, gaussian blur, morphological gradient, etc are used.



Figure 9 Image Augmentation



Figure 10 Augmentation Results

4.3 Image Segmentation

Watershed Image segmentation is implemented to segment the required parts from the entire image.

```
#thresholding
    img = cv2.imread(image)
    gray = cv2.cvtColor(img,cv2.COLOR_BGR2GRAY)
    ret, thresh = cv2.threshold(gray,0,255,cv2.THRESH_BINARY_INV+cv2.THRESH_OTSU)
[] # noise removal
    kernel = np.ones((3,3),np.uint8)
    opening = cv2.morphologyEx(thresh,cv2.MORPH_OPEN,kernel, iterations = 2)
    # sure background area
    sure_bg = cv2.dilate(opening,kernel,iterations=3)
    # Finding sure foreground area
    dist transform = cv2.distanceTransform(opening,cv2.DIST L2,5)
    ret, sure_fg = cv2.threshold(dist_transform,0.7*dist_transform.max(),255,0)
    # Finding unknown region
    sure_fg = np.uint8(sure_fg)
    unknown = cv2.subtract(sure_bg,sure_fg)
[] # Marker labelling
    ret, markers = cv2.connectedComponents(sure_fg)
    # Add one to all labels so that sure background is not 0, but 1
    markers = markers+1
    # Now, mark the region of unknown with zero
    markers[unknown==255] = 0
[ ] markers = cv2.watershed(img,markers) #Segmentation
    img[markers == -1] = [255,0,0]
```

Figure 11 Watershed Transform for Image Segmentation

5 Modelling

[] base_dir = '/content/drive/My Drive/Data

```
#Train, Validation & Testing Directory
    train_dir = os.path.join(base_dir, 'Training')
    validation_dir = os.path.join(base_dir, 'Validation')
    test dir = os.path.join(base dir, 'Testing')
    #Train Data
    train_abnormal_dir = os.path.join(train_dir, 'Abnormal')
    train_normal_dir = os.path.join(train_dir, 'Normal')
    #Validation Data
    validation_abnormal_dir = os.path.join(validation_dir, 'Abnormal')
    validation normal dir = os.path.join(validation dir, 'Normal')
    #Test Data
    test_abnormal_dir = os.path.join(test_dir, 'Abnormal')
    test_normal_dir = os.path.join(test_dir, 'Normal')
[ ] print("Total training abnormal images:", len(os.listdir(train_abnormal_dir)))
    print("Total training normal images:", len(os.listdir(train_normal_dir)))
    print("Total validation abnormal images:", len(os.listdir(validation abnormal dir)))
    print("Total validation normal images:", len(os.listdir(validation normal dir)))
    print("Total testing abnormal images:", len(os.listdir(test_abnormal_dir)))
    print("Total testing normal images:", len(os.listdir(test_normal_dir)))
➡ Total training abnormal images: 970
    Total training normal images: 230
    Total validation abnormal images: 246
    Total validation normal images: 54
    Total testing abnormal images: 304
    Total testing normal images: 72
```

Figure 12 Data loaded into Directories



Figure 13 Image Data Generators

5.1 Support Vector Machine (SVM)

The implementation of SVM with Grid Search Cross Validation is illustrated.



Figure 14 Implementation of SVM

5.2 Convolution Neural Network (CNN)

The implementation of CNN built from scratch is illustrated. Early Stopping and Reduce Learning Rate on Plateau features are used as well.

```
[ ] model = Sequential()
model.add(Conv2D(16, (3,3), activation='relu', input_shape=(299, 299, 3)))
model.add(MaxPooling2D(2,2))
model.add(Conv2D(32, (3,3), activation='relu'))
model.add(MaxPooling2D(2,2))
model.add(Conv2D(64, (3,3), activation='relu'))
model.add(MaxPooling2D(2,2))
model.add(Flatten())
model.add(Dense(512, activation='relu',kernel_regularizer=regularizers.l2(0.01)))
model.add(Dense(2, activation='sigmoid'))
```

Figure 15 Building CNN Model



Figure 16 Compiling CNN Model

[]	STEP_SIZE_TRAIN = train_generator.n//train_generator.batch_size STEP_SIZE_VALID = valid_generator.n//valid_generator.batch_size print(STEP_SIZE_TRAIN,STEP_SIZE_VALID)
C→	37 9
[]	<pre>history_finetunning = model.fit(train_generator,</pre>
₽	Epoch 1/10 37/37 [
	Epocn 10/10 37/37 [====================================

Figure 17 CNN Model Fitting

5.3 Inception-v4

The implementation of Inception-v4 based on transfer learning is illustrated. The pre-trained weights used are "inception-v4_weights_tf_dim_ordering_tf_kernels_notop.h5" ⁶.

⁶ https://github.com/kentsommer/keras-inceptionV4/releases

Figure 18 Building Inception-v4 model

Figure 19 Compiling Inception-v4 model

history_finetunning = m	<pre>del.fit(train_generator,</pre>
C+ Epoch 1/10 37/37 [= Epoch 2/10 37/37 [= Epoch 3/10 37/37 [= 37/37 [= Epoch 00003: early stop	<pre></pre>

Figure 20 Inception-v4 Model Fitting

5.4 DenseNet-169

The implementation of DenseNet-169 based on transfer learning is illustrated. The pretrained weights used are "DenseNet-BC-169-32-no-top.h5"⁷.

Figure 21 Building DenseNet-169 Model

<pre>[] history_finetunning = model.fit(train_ge</pre>	nerator,
	<pre>steps_per_epoch=STEP_SIZE_TRAIN, validation_data=valid_generator, validation_steps=STEP_SIZE_VALID, epochs=10, callbacks=callback_list, verbose=1)</pre>
C→ Epoch 1/10 37/37 [======] -	1762s 48s/step - loss: 0.5539 - accuracy: 0.8031 - val_loss: 0.4837 - val_accuracy: 0.9306
Epoch 2/10 37/37 [=====] - Epoch 3/10	1682s 45s/step - loss: 0.3627 - accuracy: 0.9426 - val_loss: 0.3376 - val_accuracy: 0.9757
	1682s 45s/step - loss: 0.2750 - accuracy: 0.9666 - val_loss: 0.2415 - val_accuracy: 0.9792
37/37 [======] -	1680s 45s/step - loss: 0.1919 - accuracy: 0.9846 - val_loss: 0.1880 - val_accuracy: 0.9826
37/37 [] -	1686s 46s/step - loss: 0.1421 - accuracy: 0.9863 - val_loss: 0.1444 - val_accuracy: 0.9826
37/37 [=====] -	1670s 45s/step - loss: 0.1185 - accuracy: 0.9897 - val_loss: 0.1131 - val_accuracy: 0.9896
Bpoch 7/10 37/37 [] -	1672s 45s/step - loss: 0.0952 - accuracy: 0.9940 - val_loss: 0.0961 - val_accuracy: 0.9896
Epoch 8/10 37/37 [] - 37/37 [] - Epoch 00008: early stopping	ETA: 0s - loss: 0.0821 - accuracy: 0.9966 Restoring model weights from the end of the best epoch. 1674s 45s/step - loss: 0.0821 - accuracy: 0.9966 - val_loss: 0.0751 - val_accuracy: 0.9896

Figure 22 DenseNet-169 Model Fitting

5.5 Cross-Validation

⁷ https://www.kaggle.com/xhlulu/densenet-keras?select=DenseNet-BC-169-32-no-top.h5

The implementation of K-Fold Cross-Validation is illustrated. It is used to check the variance of the models and avoid the overfitting.

```
#K-fold Cross Validation
VALIDATION_ACCURACY = []
VALIDAITON_LOSS = []
save_dir = '/saved_models/'
fold_var = 1
for train_index, val_index in kf.split(np.zeros(n),Y):
 training_data = train_data.iloc[train_index]
 validation_data = train_data.iloc[val_index]
 # CREATE NEW MODEL
 model = create_new_model()
 # COMPILE NEW MODEL
 model.compile(loss='categorical crossentropy',
          optimizer=opt,
          metrics=['accuracy'])
 # CREATE CALLBACKS
 checkpoint = tf.keras.callbacks.ModelCheckpoint(save_dir+get_model_name(fold_var),
              monitor='val_accuracy', verbose=1,
              save_best_only=True, mode='max')
 callbacks_list = [checkpoint]
 # There can be other callbacks, but just showing one because it involves the model name
 # This saves the best model
 # FIT THE MODEL
 history = model.fit(train data generator,
          epochs=num_epochs,
          callbacks=callbacks_list,
          validation data=valid data generator)
 # LOAD BEST MODEL to evaluate the performance of the model
 model.load_weights("/saved_models/model_"+str(fold_var)+".h5")
 results = model.evaluate(valid_data_generator)
 results = dict(zip(model.metrics_names,results))
 VALIDATION ACCURACY.append(results['accuracy'])
 VALIDATION_LOSS.append(results['loss'])
 tf.keras.backend.clear_session()
 fold_var += 1
```

Figure 23 K-Fold Cross Validation