For University Teams' Perusal:

Team Name	**Data Kaarigars**
Academic Institution	**Name: Habib University**
	City: Karachi
	Country: Pakistan

Problem Statement:

Cheque fraud has been one of the largest challenges faced by banks and financial institutions. Its numbers have been increasing at an alarming rate. More and more criminals defraud their victims, including the banks, financial institutions or businesses that issue and accept cheques, as well as customers, either the account holder or the recipient. In most cases, cheque fraud begins with the fraudsters stealing a genuine cheque before they alter some or even all the information on the cheque to their own benefits.

The problem that we will be working on is the detection of forged signatures on the cheque and along with that, their classification in comparison to the genuine signatures.

Category:

Image and video analytics.

Brief Model Description:

We have used two approaches to solve the above mentioned problem. Firstly, we have used CNNs, for which our model accuracy was around 90%. To further optimize the accuracy, we've applied the technique of transfer learning using the VGG16 model, through which the accuracy achieved is 96%.

List down the dataset(s) used along with their sources/ links:

Dataset	Source
Signature_Verification_Dataset	https://www.kaggle.com/datasets/robinreni/signature-verification-dataset

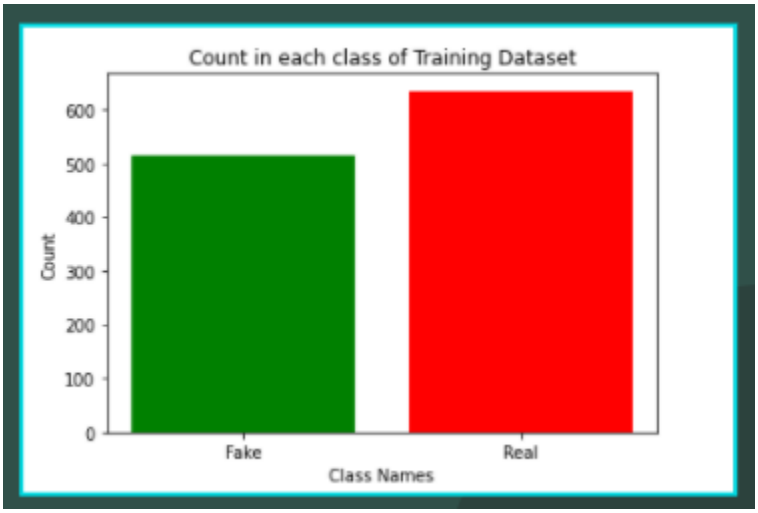
How was the data pre-processed? (steps/ software/ techniques/ etc.)

1) **Data Acquisition - The data was acquired from the above mentioned source**
2) **Importing Libraries to be Used:**
    a) **Keras**
    b) **NumPy**
    c) **Tensorflow**
    d) **Matplotlib**
    e) **CV2**
    f) **OS**
    g) **Sklearn**
3) **Importing and loading dataset using OS library.**
4) **The data was then rearranged into Train and Test data to make it ready to be used in Keras and read the class names with ease. 1149 images belonged to the train data set while 500 belonged to the test data set. 256 by 256 pixel images were preprocessed using Keras built-in library for CNN model and 224 by 224 pixel images were preprocessed using Keras built-in library for Transfer learning (this is because of the base model we have used for transfer learning, detailed discussion is made in later sections).**

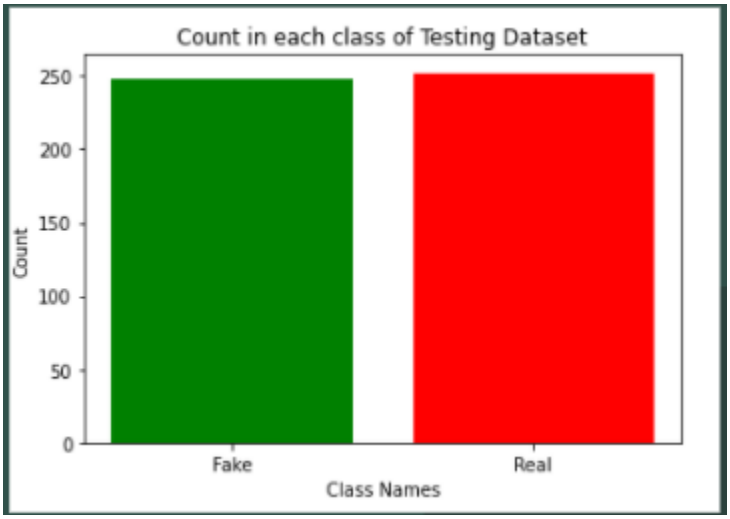
What patterns/ anomalies/ outliers were identified during Exploratory Data Analysis (EDA), if any?

Note: Please attach code/ visuals to EDA.

Class imbalance Detection was performed to check whether there is a need of performing undersampling or oversampling in both test and train dataset.



The above plot shows the distribution ratio of real:fake = 55:45 in the train dataset. This shows slight imbalance hence we can safely proceed with training our model.

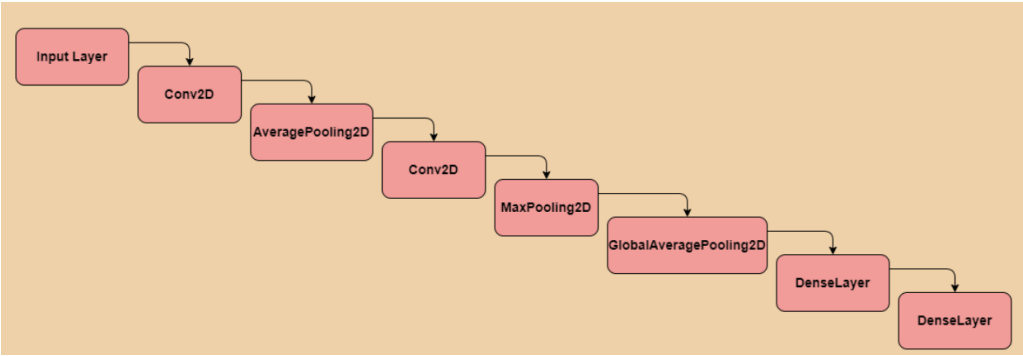


The above plot shows the distribution ratio of real:fake = 51:49 in our test data set.

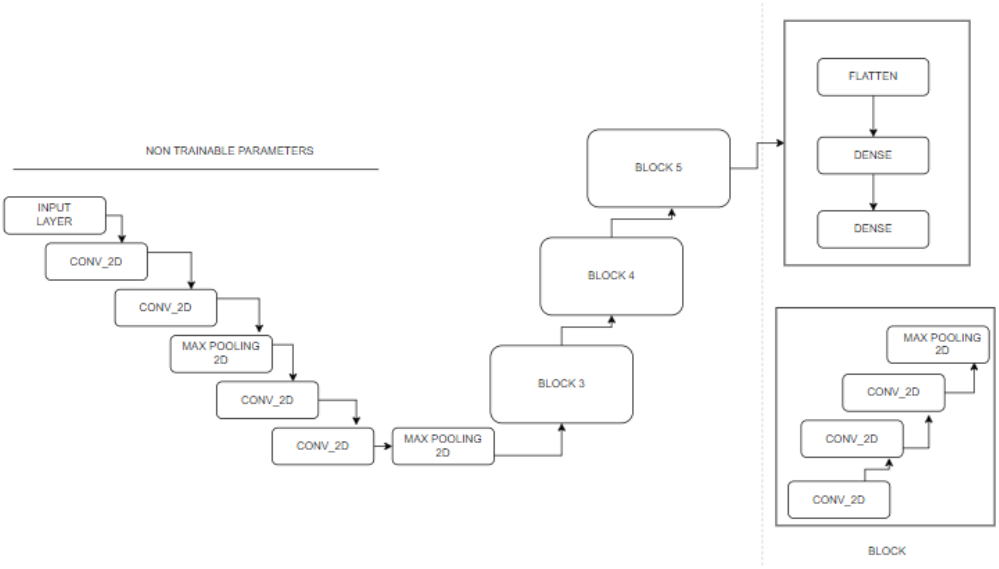
How was the dataset modeled (machine learning)?

The dataset was modeled using two different models:

1) **CNN**



2) **Transfer Learning - VGG16**



The above mentioned images are of models we have designed for the solution to the problem statement mentioned above. The results (accuracy scores) obtained from these models are discussed in the next section.

The images for models attached above are self explanatory as it shows the number of total convolutional layers, pooling and dense layers used in total for each model.

For our initial model that is basically a model based on CNN, we have used ReLU and Sigmoid activation function with batch size of 30. The sparse categorical cross entropy loss has been used as loss function and Adams optimizer is used for optimization of parameters and features extraction in each epoch. Coming to layers, you may refer to the image attached above. The accuracy score obtained from this model is discussed in the next section. Note that these ML models are supposed to be optimized to every possible extent via optimization techniques. Considering this let us discuss our next optimized approach of transfer learning, following this CNN approach.

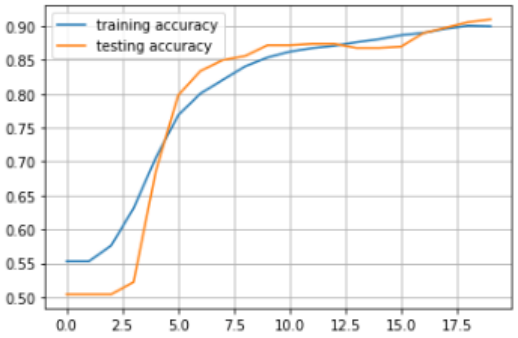
The image for the model shows that we have used the VGG16 model as our base model for implementation of transfer learning. Note that VGG 16 model is a CNN model for image classification that turns rgb images into features for pattern recognition. Particularly for the transfer learning-VGG16 model, note that non-trainable parameters indicate that we have freezed the trainable parameters in the first five layers making them non trainable. The rest of the layers remain untouched. After this step we have topped the base model (VGG16) with a Sequential model that consists of one flatten layer and two dense layers. This is for model designing, then the model was compiled and trained on the dataset and then tested on the testing data set.

What were the evaluated results to reject or fail to reject the proposed hypothesis?

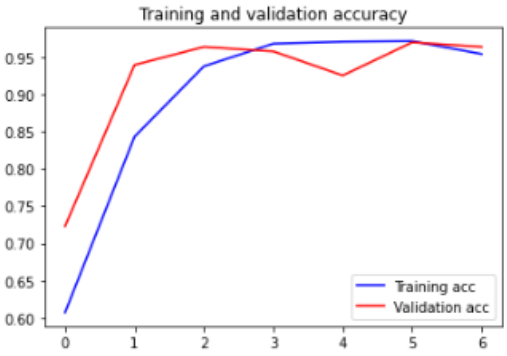
The above-mentioned model description suggests that initially our hypothesis was based on analysis of the two approaches we have opted to optimize the accuracy score of prediction our model is making.

After achieving an accuracy score of 90% using CNN, we hypothesized that the technique of transfer learning can increase and optimize the accuracy score. In order to test for the mentioned hypothesis we used a pre-trained model of VGG16.

Following are the obtained plots for accuracy score.



CNN



Transfer Learning - VGG16

These plots satisfy and prove our hypothesis as we see that after implementation of transfer learning, the curve for validation accuracy reached 0.96 which is 96%. Note that before implementation of transfer learning we had accuracy of only 90% using CNN only.

Now that we have improved our prediction score, we can claim that our model is able to recognize and learn patterns of genuine and fake signatures more accurately and can be integrated with bank systems practically and optimally.

***Note: The working behind obtaining the mentioned plots and scores has been submitted in the form of .ipynb (notebook) files. It also contains the plots for loss (validation and training) which is very less and is tolerable.**

How is your model a significant value-addition towards UBL or banking industry in general?

- **Complete automated forged signature detection system with automated scanning and verification process.**
- **Detection of fraudulent cheques at point of presentation with high accuracy.**
- **Avoiding potential collusion between tellers and culprits.**
- **Improve customer experience.**
- **Save banks from loss of millions that happens due to cheque forgery.**
- **Minimize account misuse.**

Links for Notebooks (prototype):

For CNN model:

https://www.kaggle.com/sanafatima123/cnn-for-signature-verificationn

For Transfer Learning VGG16 model:

https://www.kaggle.com/sanafatima123/transfer-learning-for-signature-verification for transfer learning

For EDA: