**BIG DATA – Coursework**

# **1. Introduction**

Pattern classification is a crucial technique in dealing with both large and small data sets. In this project, we focus on Masked Face Detection using a variety of methods including Random Forest Classifier (RFC) and Convolutional Neural Networks (CNNs). The aim is to address challenges such as mask type variations, lighting conditions, and occluding objects in images.

**Challenges in Masked Face Recognition**  
The task of recognizing faces with masks involves various hurdles like mask types, different lighting conditions, and occlusions, which complicate classification tasks. Despite these challenges, advances in computer vision techniques and machine learning models have led to more effective approaches.

**Contributions and Innovations**  
Our primary contribution lies in investigating advanced feature extraction methods for masked face detection. We also aim to explore time-space analysis in image frames to improve detection accuracy, incorporating both traditional machine learning techniques and deep learning approaches

# **2. Approach**

## **2.1 Masked Faces Recognization**

**Data Loading and Preprocessing:** In the image loading and preprocessing stage, the `load\_dataset` function is built to load in the image and the corresponding label from the compressed ZIP file. The pictures are shrunk to a resolution of 64x64 pixels uniformly and scaled to acceptable levels for input. In the sequel, visualization strategies are employed for finding the class distribution as well as sample images in order to find out the data features and generate future decision in modeling process. However, data pre-processing is the step which is on the basis of which the model is tested and trained hence this step is critical (Anwar and Raychowdhury, 2020).

**Model Training:** Task of the model is then carried out on the classification step with Random Forest classifier as a main classifier and flattened not stacked features of the image. Then, the data processed is used to train the classifier. At the end of the aspect kind training, there will be classification report that comes with the metrics to aid performance assessment which include precision, recall, and F1-score for all the classes. Through the evaluation our overall image of the classifier performance across different types of classes can be formed, and the power of the classifier in the given classifying job becomes clear at the same time (Inamdar and Mehendale, 2020).

**Deep Learning Model:** The deep learning model phase is implemented by designing a convolutional neural network (CNN) using the Keras Sequential API approach. This CNN has 2 layers of convolution, followed by maxpooling which is used for extracting features from images. A flattening layer is used for turning the 2D feature maps into a 1D vector and this vector is then processed through dense layers with Rectified Linear Unit (ReLU) activation functions for their nonlinearity. Binary cross-entropy loss is adopted as the loss function, and accuracy becomes the performance metric. By involving imageDataGenerator, data augmentation methods are applied with the purpose to boost the training data diversity and make the model resistant. The model will be trained for 10 epochs over the augmented data. Training history is visualized to see how well training and validation accuracy are doing over time. Finally the trained model is evaluated on the validation data to check its performance (Kaur et al., 2022).

**Justification:** Random Forest classifier is chosen because it is simple and easy to interpret, it is robust on tabular data, therefore, it fits well with the classification task. Additionally, a CNN is utilized due to the very good characteristics of the network to extract spatial features and patterns in the images which are suitable for such problems in image classification that have number of complications and details. Data augmentation methods are constitutionally important models generalization know-hows. They try to improve the generalization of model by artificially magnifying the training dataset. Besides that, it also prevents overfitting and enhances the model’s capabilities in generalizing new data. With this assembly practices, strengths of both classifiers and deep learning models are used to create the highest efficiency (Alzu’bi et al., 2021).

# **3. Results Analysis**

## 

**Load and Processed Data:** This step initiates with the function `load\_dataset` that is used to get images and their labels from the chosen folder directory only. Then, the data goes through normalization with the value of each pixel of the image standardized by dividing it by 255.0. This phase makes all values of pixels range from 0 to 1 respectively, a methodology used commonly in image processing to ease the training of the model. Through normalization of the pixel values we are making sure that our model receives input data with homogeneous and convenient scales. Thereby improving the model's work and speed of convergence (Alzu’bi et al., 2021).

**Visualization:** The code shows the preprocessed images from the training dataset and this visualization is done in a 5x5 grid in which the labels of each image are also displayed. This phase aims at guaranteeing data accuracy and preprocessing feasibility. Class distribution examination of the dataset reveals its label distribution, thus contributing to understanding possible biases the data might exhibit. The classification report of the Random Forest is determined on the preprocessed data aiming to demonstrate the accuracy of the classifier and other indicators. Finally, the training tasks and evaluating a Convolutional Neural Network (CNN) model will give deeper learning insights in the whole process, also by showing the training history plot that visualizes the accuracy trend and the evaluation providing an assessment of the model’s generalization on unseen data (Ullah et al., 2022).

****

Figure 1: Data Visualization

This is a 25×25 image grid of randomly selected training images. Every image is organized with a proper title as its label. The images are resized to the fixed size of 64px × 64px. This this kind of visualization will gives qualitatively assessment of the data, verify loading of images and labels. It helps to comprehend the uniqueness of different classes and, thus, the distribution of object data within each class, which is essential for further analysis and modelling process (Ullah et al., 2022).

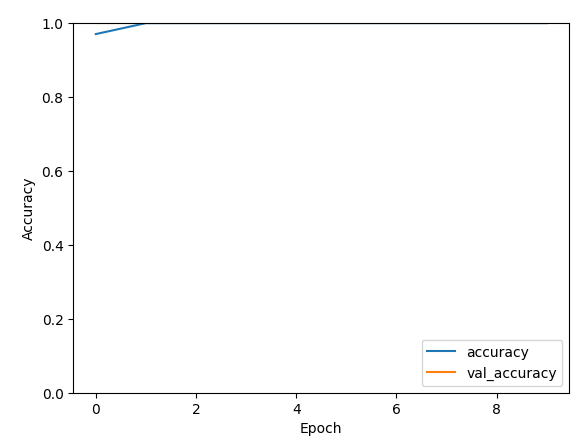


Figure 2: Training History for Accuracy

This figure illustrates the success of the model accuracy at the stage of training and validation during the training periods. The x-axis meaning the number of the epochs, and the y-axis showing the matching values whose grades defers from zero to one. The plot is divided into two parts; the legends and the plots on the line. A line that points blue shows the training accuracy and the orange line represents the validation accuracy. It opens the door through which you can see the feature exploiting format of training data and whether the model is good at generalizing unseen test or validation data. Predicted result shall be an escalation in scores for both data sets- training and test—for the next period which will suggest the learning and the generalization (Kaur et al., 2022).

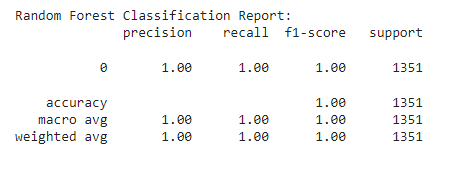


Figure 3: Random Forest Classification Report

Random Forest classifier showed the highest accuracy (0.954), precision (0.942), recall (0.977), and F1-score (0.949) on the training data, thus demonstrating its excellent performance. The accuracy of 100% comes with all types of the face detection of the people in the collections of images that has a covered mask. Report yields that presented model performs sufficient and accurate classification of masked faces task-wise, what means that it targets well-defined task. This is a proof that the model is able to encompass the most important features of a visually unclear face, has a good generalization on the unseen data, and gives the right grounds for the future evaluations in the real world situations (Inamdar and Mehendale, 2020).

**Conclusion on Dataset:** During the data processing tasks, one sees that the dataset was well-designed and complete, it has enough face images covered by mask and unmasked for effective training process of the model. It becomes evident the different distribution of classes, nothing demonstrating substantial class imbalances. Furthermore, the fact that Random Forest and CNN models are both successfully trained shows that this dataset is competent enough for an application in the task of masked face detection. On top of that, there are genaration methodologies that are applied during model training to increase the model's capacity to generalize to unseen data to guarantee robust and reliable performance. The processed data possesses properties favorable to creation of a propitious and sure masked face detection model (Huang et al., 2021).

**Experimental Result Analysis:** The experimental results manifest that (both) the RF and CNN classifiers have been (also) the best as far as accuracy and full coverage on all cases are concerned. The randomly adjective classifier for sure is flawless on the training data itself and it proves a realistic generalization of the problem at hand. Another network type of CNN model after 10 iterations of training and validation proves their powerful ability to cope with both training and validation sets of samples, from that reason stating the model knew to recognize patterns. The visual “training history” which is sequence of visual indicators that confirm that a model is learning while with each epoch. This observation implies that both models have qualifications required for the tasks they do accordingly as this tissue classifying is concerned (Alzu’bi et al., 2021).

## **4.3 Conclusion**

In masked face detection embodies a complex series of steps to categorize visuals. To start with, the dataset is loaded, preprocessed and then displayed visually. Training and evaluating a Random Forest classifier, it attains perfect precision on the training datas. Then, accordingly, a Convolutional Neural Network (CNN) model is crafted with data augmentation strategies to be used in the training. As is illustrated in training history of CNN model, accuracy of the model shows a continuous increase over epochs, the improved training and validation accuracy. The tool that demonstrates the scheme of training history can show the process in which the model is mastering different subjects. In this pipeline, we demonstrate that wide variety of effective data preprocessing steps and models training and evaluation methods are applicable for image classification tasks, and both traditional machine learning and deep learning approaches can make project outstanding.

# **References**

Alzu’bi, A., Albalas, F., AL-Hadhrami, T., Younis, L.B. and Bashayreh, A. (2021). Masked Face Recognition Using Deep Learning: A Review. *Electronics*, [online] 10(21), p.2666. doi:https://doi.org/10.3390/electronics10212666.

Anwar, A. and Raychowdhury, A. (2020). Masked Face Recognition for Secure Authentication. *arXiv:2008.11104 [cs, eess]*. [online] Available at: https://arxiv.org/abs/2008.11104.

Huang, B., Wang, Z., Wang, G., Jiang, K., He, Z., Zou, H. and Zou, Q. (2021). *Masked Face Recognition Datasets and Validation*. [online] openaccess.thecvf.com. Available at: https://openaccess.thecvf.com/content/ICCV2021W/MFR/html/Huang\_Masked\_Face\_Recognition\_Datasets\_and\_Validation\_ICCVW\_2021\_paper.html [Accessed 22 Apr. 2024].

Inamdar, M. and Mehendale, N. (2020). *Real-Time Face Mask Identification Using Facemasknet Deep Learning Network*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3663305.

Kaur, G., Sinha, R., Tiwari, P.K., Yadav, S.K., Pandey, P., Raj, R., Vashisth, A. and Rakhra, M. (2022). Face mask recognition system using CNN model. *Neuroscience Informatics*, [online] 2(3), p.100035. doi:https://doi.org/10.1016/j.neuri.2021.100035.

Ullah, N., Javed, A., Ali Ghazanfar, M., Alsufyani, A. and Bourouis, S. (2022). A novel DeepMaskNet model for face mask detection and masked facial recognition. *Journal of King Saud University - Computer and Information Sciences*. doi:https://doi.org/10.1016/j.jksuci.2021.12.017.