**Additional Data**

**Data Motivation**

The original dataset did not provide the accuracy that was strived for in the model, with a 66.22% mean validation accuracy with a 72.40% mean validation loss. The original dataset was images derived from image scraping mushroomworld.com for edible and poisonous mushrooms. With deeper insight and exploration into the original dataset, it was derived that the initial dataset was small, consisting of 415 mushroom-images: 291 images of edible mushrooms & 124 images of poisonous mushrooms.

The size of the dataset affected the accuracy as it had to be fractionized further into their respective training and testing sets using the 80:20 split ratio. This ratio was chosen based of a study conducted by Rácz et al. concluded: as the dataset size increased so did the classification accuracy. The study also concluded an increase in the classification accuracy increased as the train/test ratio increased from 50:50 to 80:20.

The ratio of model was selected to ensure the model does not overfit the training data. Overfitting is when the neural network model learns the specific detail and noise in the input data, due to having limited data or having too much data in the training set, relative to the test set. Overfitting is a problem, as the model has learnt and adjusted its weights & biases to the ‘noisy’ training set; this as a result, negatively impacts the model’s performance in its predictions of new data (Rácz et al. 2021).

The incorporation of new data will increase the model’s accuracy within its predictions. Additional data is a predominant method to increase the accuracy. The increase comes from the capability of increasing the size of the training and testing sets. Theoretically, the increase of size of the dataset should have a positive effect on the model’s accuracy. The positive effect can be derived to the training of the model, the increase of the training set allows the model to increase its capabilities via deriving new features (Luo et al., 2018)

The increase of the size of the dataset ensures the model does not rely on assumptions and weak correlations, thus theoretically allowing the models mean validation accuracy to increase.

**Data preparation**

The mushroom classification is a novel problem and required a custom solution to be implemented to achieve the necessary size of images for the mushroom dataset. The research was conducted into the classification of a mushroom that was either edible or poisonous, and the usability of the source websites.

Python scripts was implemented for the collection of images which are a mixture of edible or poisonous mushrooms. The python scripts use web scraping to collect images for the data set then writes the images into a folder created by the script in the operating system. The original scraped images are in different sizes as the source websites use community photographs for classification. The image scrapers derived the additional images from Wikipedia and foragingguide.com. With the use of the additional scrapers, the dataset size increased to 1184 mushroom images; 871 edible & 315 poisonous, still maintaining the 80:20 train, test split.

The model uses TensorFlow’s Keras API (TensorFlow, no date). The API allows input images to be augmented and preprocessed before being used as an input in the neural network model. The images used after the preprocessing stage were of a constant size of 180 x 180 pixels. The Keras API also performs scaling & translation on the image as a part of the preprocessing. Due to the nature of the original images, the original data needed to be centered, resized and as they were a range of different photos from different sources. Data preparation is important as it is good practice for a model to have a fixed input size, as the size of the kernel in the model is fixed on one size (Zheng, L. et al., 2016).

**Impact report**

The project strived towards an accurate model which could safely predict the safety of consumption due to the increase of outdoor activity due to covid-19. The initial dataset provided a classification accuracy which was unsatisfactory in respect to the scope of the project. Upon deriving the models lack of performance to the dataset size; Additional web scrapers were implemented to increase the dataset size.

Additional data has proved beneficial in numerous studies in the machine learning domain. Luo et al., studied the effect of dataset size on image-based classification using a Convolutional Neural Network. They derived that there was a positive correlation between dataset size and classification accuracy. This paper supports our findings. The additional data increases the mean classification accuracy from 66.22% to 77.20% and a decrease on mean validation loss from 72.40% to 58.88% as the data set increased by 769 mushroom images, using the implemented web-scrapers to derive images from Wikipedia and foragingguide.com.

The original dataset lacking in mean validation accuracy as only 332 mushroom images were being used to train the data compared to the 947 mushroom images used to train the model. The 615 mushroom image increase yielded a 10.98% in the model’s mean validation accuracy. The increase warrants further research of increasing the training data size in the Iterative development with experimentation of data augmentation.

**References**

TensorFlow (no date) *tf.keras.preprocessing.image.ImageDataGenerator.* Available at: <https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator> (Accessed: 15/04/2021)

Rácz, A., Bajusz, D. and Héberger, K., 2021. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules*, 26(4), p.1111.

Luo, C., Li, X., Yin, J., He, J., gao, D. and Zhou, J., 2019. How Does the Data Set and the Number of Categories Affect CNN-based Image Classification Performance?. *Journal of Software*, pp.168-181.

Wang, J. and Perez, L., n.d. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. Available at: http://cs231n.stanford.edu/reports/2017/pdfs/300.pdf (Accessed: 15/04/2021).