**Aims and objectives**

This report summarizes the performance and effectiveness of two machine learning methods for image classification of different species of animals and then further considers the reliability of the applicability of the machine learning methods on the dataset.

**Background**

The data this report uses to explore the machine learning performance contains pictures of animals such as birds, incest or mammals grouped into different datasets with different data distributions which means that some datasets are more balanced, and some are less. The dataset used is the iNat2017 dataset. (iNaturalist, 2021)

One of the main projects iNaturalist is known for is the social network app for naturalists, citizens scientists and biologists which allows them to identify the plants and animals while having access to a large community of 400000 scientists. (About, iNaturalist, 2021)

**Methods**

The methods selected to evaluate the machine learning performance on image classification for the inat2017 dataset are the supervised **CNN** (convolution neural network) and the unsupervised **Autoencoders Method**.

To be able to better evaluate the effectiveness of ML image classification I have chosen one dataset which is balanced (**Most common ten species(excluding birds)**) and one dataset which is unbalanced **Distinct bird species ( filtered species).**

Each dataset was tested with both the CNN and Autoencoder methods.

**THE CNN METHODS MODELS:**

For the CNN method, I started with the base model and then applied different techniques to improve the performance and learning rate of the models.

The CNN models explored:

1. The base model

2. Extra layers and residual links

3. Data augmentation and normalisation layer

Jupyter notebook: 2021J\_EMA.ipynb

|  |  |  |  |
| --- | --- | --- | --- |
| METHOD USED: **CNN** | | **THE BASE MODEL** | |
| **Text  Description automatically generated** | | | |
| **MODEL NAME:**  **CNN\_base\_most\_common\_not\_aves.h5** | | **PERFORMANCE MOST COMMON NOT AVES:**  **CNN\_base\_most\_common\_not\_aves.h5**  **accuracy: 0.7867 -precision: 0.8324 - recall: 0.7578** | |
| **CNN\_unbalanced\_filtered\_confusing\_species.h5** | | **PERFORMANCE FILTERED CONFUSING SPECIES:**  **CNN\_unbalanced\_filtered\_confusing\_species.h5**  **loss: 0.0000e+00 - accuracy: 0.6770 - precision: 0.7466 - recall: 0.6237** | |
| **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: FILTERED CONFUSING SPECIES** | | | **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: PERFORMANCE MOST COMMON NOT AVES:** |
| **Chart, line chart  Description automatically generated****Qr code  Description automatically generated** | | | Chart, line chart  Description automatically generated |
| **Evaluation comment:** | Based on the performance evaluations, the model performs better on the Most Common Not Aves dataset with 78% accuracy, and 83% precision which means that there are 83% true positives/ 17% false positives and recall at 75% meaning there is 75% true positives/25% false negatives.  Based on the scattered points and the training history, both models start to overfit after 5 epochs. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| METHOD USED: **CNN** | | **THE RESIDUAL LINKS MODEL** | |
| **A picture containing text  Description automatically generated** | | | |
| **CNN\_reslinks\_most\_common\_not\_aves.h5** | | **PERFORMANCE MOST COMMON NOT AVES:**  **CNN\_reslinks\_most\_common\_not\_aves.h5**  **loss: 0.0000e+00 - accuracy: 0.7535 - precision: 0.7756 - recall: 0.7391** | |
| **CNN\_weighted\_reslinks\_filtered\_confusing\_species.h5**  Due to the dataset being  unbalanced, this model was trained on weighted train\_dataset\_w . | | **PERFORMANCE FILTERED CONFUSING SPECIES:**  **CNN\_weighted\_reslinks\_filtered\_confusing\_species.h5**  **loss: 0.0000e+00 - accuracy: 0.5796 - precision: 0.6632 - recall: 0.5118** | |
| **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: FILTERED CONFUSING SPECIES** | | | **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: PERFORMANCE MOST COMMON NOT AVES:** |
| **Qr code  Description automatically generated** | | |  |
| **Evaluation comment:** | To improve the base model's performance, I have added residual links end two extra 258 filters convolution layers.  Based on the performance results for this model, the performance dropped, by 3 % overall for the Most Common Not Aves dataset and by an overall 10% for the Filtered Confusing dataset.  Training history shows that the overfitting slightly improved for the Filtered Confusing Species dataset but decreased for the Most Common Not Aves dataset which is visible in the scattered plot which now has more false-negative predictions. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| METHOD USED: **CNN** | | **THE AUGMENTED MODEL with NORMALISATION LAYER** | |
| Text  Description automatically generated | | | |
| **CNN\_norm\_agn\_most\_common\_not\_aves.h5** | | **PERFORMANCE MOST COMMON NOT AVES:**  **CNN\_norm\_agn\_most\_common\_not\_aves.h5**  **loss: 0.0000e+00 - accuracy: 0.8283 - precision: 0.8717 - recall: 0.7969** | |
| **CNN\_w\_agn\_norm\_filtered\_confusing\_species.h5**  Due to the dataset being  unbalanced, this model was trained on weighted train\_dataset\_w . | | **PERFORMANCE FILTERED CONFUSING SPECIES:**  **CNN\_w\_agn\_norm\_filtered\_confusing\_species.h5**  **loss: 0.0000e+00 - accuracy: 0.7019 - precision: 0.7815 - recall: 0.6332** | |
| **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: FILTERED CONFUSING SPECIES** | | | **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY:PERFORMANCE MOST COMMON NOT AVE** |
| **Chart, qr code  Description automatically generated** | | | Chart, scatter chart  Description automatically generated |
| **Evaluation comment:** | To reduce overfitting even more I have now decided to do data augmentation and include a normalization layer.  This seems to have significantly boosted the performance and also got rid of the overfitting.  The Accuracy for the Most Common, not Aves has risen to a high 82 % with even higher precision at 87% which means that there were 87% True positives / 18% False positives which is very good. And the recall also rose to 79% which means there were 79% true positives /21% only False negatives.  The accuracy for the Filtered Confusing Dataset also rose to a high 70% with 78% precision and 63% recall, which is not as good as the Most Common Not Aves dataset suggesting that it is a more challenging dataset. | | |

**THE AUTOENCODER METHOD MODELS**

The autoencoder unsupervised method uses an autoencoder and decoder, where image inputs are fed to the encoder which then converts the data to features and then the decoder will take the features from the encoder and reconstructs them to the intended output. (USING AUTOENCODERS FOR IMAGE CLASSIFICATION, 2022)

The learned weights in the autoencoders encoder to decoder process are saved and then used in the CNN method for the image classification part.

In my investigation I have created 3 models:Model has 14 classes

1. Base model

2. Model with dropout and normalisation layers

3. Model with dropout and normalisation layer but slower learning rate.

Jupyter notebook: 2021J\_EMA.ipynb

|  |  |  |  |
| --- | --- | --- | --- |
| METHOD USED: **AUTOENCODERS** | | **THE BASE MODEL** | |
| The autoencoder model (encoder and decoder to learn weights) | | | |
| A picture containing text  Description automatically generated | | | |
| The encoder for image classification with the fully connected layers in the connecting method. | | | |
|  | | | |
| **AUTOEN\_base\_most\_common\_not\_aves.h5** | | **PERFORMANCE MOST COMMON NOT AVES:**  **AUTOEN\_base\_most\_common\_not\_aves.h5**  **loss: 0.0000e+00 - accuracy: 0.7841 - precision: 0.7990 - recall: 0.7739** | |
| **AUTOEN\_weighted\_base\_filtered\_confusing\_species.h5** | | **PERFORMANCE FILTERED CONFUSING SPECIES:**  **AUTOEN\_weighted\_base\_filtered\_confusing\_species.h5**  **loss: 0.0000e+00 - accuracy: 0.6507 - precision: 0.7008 - recall: 0.6116** | |
| **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: FILTERED CONFUSING SPECIES** | | | **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY:PERFORMANCE MOST COMMON NOT AVE** |
| Chart  Description automatically generated  **Qr code  Description automatically generated** | | | Chart, scatter chart  Description automatically generated |
| **Evaluation comment:** | The accuracy of 78% for most common, not Aves dataset is already quite high and precision as well at 78% which means that 78% of prediction was true positives 22% was False positives.  The accuracy is worse for the Filtered Confusing Species at 65% and precision at 70 %, which is like the performance of the CNN base model  The scatter points and training history suggests overfitting therefore the hyperparameters need to be tweaked again. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| METHOD USED: **AUTOENCODERS** | | **MODEL WITH EXTRA DROPOUT AND NORMALISATION LAYERS** | |
| The autoencoder model(encoder and decoder to learn weights) | | | |
| A picture containing table  Description automatically generated | | | |
| The encoder for image classification with the fully connected layers in the connecting method. | | | |
| Text  Description automatically generated with low confidence | | | |
| **AUTOEN\_dropout\_most\_common\_not\_aves.h5** | | **PERFORMANCE MOST COMMON NOT AVES:**  **AUTOEN\_dropout\_most\_common\_not\_aves.h5**  **loss: 0.0000e+00 - accuracy: 0.8173 - precision: 0.8413 - recall: 0.8020** | |
| **AUTOEN\_w\_drop\_filtered\_confusing\_species.h5** | | **PERFORMANCE FILTERED CONFUSING SPECIES:**  **AUTOEN\_w\_drop\_filtered\_confusing\_species.h5**  **loss: 0.0000e+00 - accuracy: 0.7558 - precision: 0.7858 - recall: 0.7285** | |
| **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: FILTERED CONFUSING SPECIES** | | | **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY:PERFORMANCE MOST COMMON NOT AVE** |
|  | | |  |
| **Evaluation comment:** | After applying extra dropout and normalisation layers, both models' performance increased by around 10% overall and the overfitting significantly reduced too.  The Most Common Not Aves dataset accuracy increased to 81 % with 84 % precision meaning 84 % of predictions was true positives and 16 % are false positive, 80% recall meaning that 80 % are true positives and 20 % are false negatives which are quite good even though the training significantly slowed down after 70%.  The Filtered Confusing Species also scored higher at 75% accuracy, 78% precision and 72% recall.  The results of scattered points and training history both match the metrics evaluations on both datasets show a reduction in overfitting and an increase in the learning rate. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| METHOD USED: **AUTOENCODERS** | | **MODEL WITH SLOWER LEARNING RATE** | |
| Encoder and decoder to learn weights | | | |
| A picture containing table  Description automatically generated | | | |
| The encoder for image classification with the fully connected layers. | | | |
| Text  Description automatically generated with low confidence | | | |
| **AUTOEN\_slowlr\_most\_common\_not\_aves.h5** | | **PERFORMANCE MOST COMMON NOT AVES:**  **AUTOEN\_slowlr\_most\_common\_not\_aves.h5**  **loss: 0.0000e+00 - accuracy: 0.7943 - precision: 0.8402 - recall: 0.7641** | |
| **AUTOEN\_w\_slow\_rate\_filtered\_confusing\_species.h5** | | **PERFORMANCE FILTERED CONFUSING SPECIES:**  **AUTOEN\_w\_slow\_rate\_filtered\_confusing\_species.h5**  **loss: 0.0000e+00 - accuracy: 0.6436 - precision: 0.7383 - recall: 0.5554** | |
| **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY: FILTERED CONFUSING SPECIES** | | | **SCATTERED PLOT POINTS AND TRAINING**  **HISTORY:MOST COMMON NOT AVE** |
|  | | |  |
| **Evaluation comment:** | Applying a slower learning rate on the models reduced overfitting even more on both datasets but for the cost of reduced learning performance.  The accuracy of recalls and precision dropped by 10 % for the Filtered Confusing Species dataset but the learning stayed consistent and slowed down after 10 epochs with very small signs of overfitting.  The accuracy of the model for the Most common not Aves dataset dropped only by 2 % which is great with the same precision as the previous model had.  The scatter points suggest that the model is very confident in predicting the images considering most points in true positive and true negative areas of the graph and a few of the points in false positive and false negative parts matching the recall score of 76%.  This cannot be said for the Filtered Confusing Species dataset which is not very confided and there are quite a lot of points scattered in the false-negative part which is matching the low recall score of only 55% | | |

**Findings**

**The Most Common Not Aves dataset**

Based on the models, the CNN method is slightly more efficient in image classification compared to the semi-supervised method which uses Autoencoders.

If I was to explore other methods to evaluate the datasets I would choose the k-nearest neighbours' algorithm which can be used for image classification by using the proximity to make classifications and predictions about the data points.

The Most Common Not Aves dataset all models performance and scattered points:

|  |
| --- |
| Table  Description automatically generated |
| Qr code  Description automatically generated |

Based on this training history and performance evaluation data, we can say that the best performing model for the Most Common Not Aves dataset is the CNN model with augmented data and an extra normalisation layer with 82% test accuracy, and the best performing model which uses Autoencoders is the model which extra normalisation layers and drop out layers at 81% test accuracy.

The table comparing the performance between the best CNN method and the best Autoencoders method for the Most Common Not Aves dataset:

|  |  |
| --- | --- |
| Classification report | |
| The normalisation/augnumented CNN model  'CNN\_norm\_agn\_most\_common\_not\_aves.h5' | The extra normalisation/dropout Autoencoders  AUTOEN\_dropout\_most\_common\_not\_aves.h5 |
| Table  Description automatically generated | Table  Description automatically generated |
| Confusion metrixes | |
| Calendar  Description automatically generated | Calendar  Description automatically generated with medium confidence |

The confusion matrix shows that the models were mostly consistent at correctly predicting all the classes, except the class 3 the Sceloporus occidentalis was more challenging for the models to predict correctly.

Both models have eliminated the overfitting and reached an accuracy score above 80%.

**The filtered confused species dataset:**

Table

Description automatically generatedI have chosen this dataset because it should be more challenging for the machine learning methods to learn as it has confusingly similar images and it is an imbalanced dataset as shown in the screenshot below:

Due to the dataset being unbalanced, I have trained some of the models on the weighted dataset to increase the weights of the classes with a smaller number of samples and I have under-sampled class 6 which has a significantly larger number of samples compared to the rest of the classes.

The Confusing Filtered species dataset models performance and scattered points:

|  |
| --- |
| Table  Description automatically generated |
| Qr code  Description automatically generated |

The training history and performance show that the best CNN model for the Filtered Confusing Species dataset is the weighted augmented method (CNN\_weighted\_agn\_filtered\_confusing\_species.h5) and the best Autoencoder model is the weighted model with drop/normalisation layers (AUTOEN\_w\_drop\_filtered\_confusing\_species.h5)

|  |  |
| --- | --- |
| Classification report | |
| The weighted augnumented CNN model  CNN\_weighted\_agn\_filtered\_confusing\_species.h5 | The weighted training dataset, dropout, and normalisation  AUTOEN\_w\_drop\_filtered\_confusing\_species.h5 |
| Table  Description automatically generated | Table  Description automatically generated |
| Confusion matrixes | |
| Calendar  Description automatically generated with medium confidence | Table  Description automatically generated with medium confidence |

The confusion matrix shows that the Autoencoder method was more successful at predicting the classes correctly with more correct predictions per class.

Based on the evaluations of both datasets, I can say that both methods used, were quite successful at learning but the main difference was shown when it was applied to a different dataset.

When the methods were applied to the dataset with the unbalanced data with very similar features like the Filtered Confusing Species dataset, the performance dropped.

The models trained on augmented data improved their training by around 10 % and the training stayed consistent without any overfitting suggesting the datasets are unbalanced and need more data.

Below are two images comparing the performance of image classification of the two best performing models on test data.

|  |
| --- |
| The Most Common Not Aves dataset with augmented data and extra normalization with 24 correct predictions out of 25. **CNN\_norm\_agn\_most\_common\_not\_aves.h5** |
| A picture containing different  Description automatically generated |

|  |
| --- |
| The Filtered Confusing species dataset with extra dropout and normalisation  layer with 2 wrong predictions out of 25:  **AUTOEN\_w\_drop\_filtered\_confusing\_species.h5** |
|  |

Based on the pictures above we can say that the classification is reliable already at 75/85% accuracy.

**Discussion of wider implications**

The possible issue with the iNaturalist dataset would be the geographical bias.

Specifically, the problem is that the datasets are updated with pictures taken by the public which means that only the animals in places which are easily accessible to the public can be properly documented therefore there will be a lot of other animals which will not be included in the data. (Hu, 2018)

People usually choose to document images of animals they find most interesting, and a lot of time ignore the other, which could lead to some of the species overlooked.

This presents possible ethical implications, such as the fact that people who choose to document the wildlife do not necessarily know how to do it without disturbing the natural environment of the animals, and in some situations they could cause more damage than good. For example by causing the animals stress, making them migrate or even damaging/disturbing their natural habitats.

Due to very large amount of animal species and the fact that some of them look very similar, there is also a risk of bias where the models could get confused due to similarities between some species.

If the machine learning models were used on data which they were not trained on, for example, either a picture of an animal which looks similar but belongs to a different part of the earth, it would be mistakenly classified as the wrong species which would sabotage the reliability of machine learning models to classify the images correctly.

If the model was trained very accurately, and presented with some never seen data, this could be potentially used as a technique for discovery of new never seen animals’ species.

The possible approaches to minimise the risk of misclassification would be to train the models only on very balanced data, which contains all the possible classes needed for that specific classification.

**Conclusion**

Based on the results of the machine learning methods and their models used on the iNaturalist dataset in this report, which resulted in 82 % accuracy for the predictions, we can say that the Machine learning methods could be reliably used for image classification.

The investigation also proved that to be able to achieve high performance it is very important to use correct hyperparameters and techniques for the training and even more importantly it is important to have a balanced dataset with adequate quality of the pictures.

If even higher accuracy performance was reached, the ML models could be a great technique for analysing a large amount of data which could help to learn new information, find new species, potentially rediscover species already classed as extinct and enable people to learn in a new way.

**(2496 WORDS)**

**References:**

Charterglobal.com. 2022. *USING AUTOENCODERS FOR IMAGE CLASSIFICATION*. [online] Available at: <https://www.charterglobal.com/using-autoencoders-in-ai-for-image-classification/#:~:text=Autoencoders%20are%20neural%20networks%20that,image%20from%20the%20encoded%20image.> [Accessed 15 May 2022].

Hu, C., 2018. *Open data is a blessing for science—but it comes with its own curses*. [online] Popular Science. Available at: <https://www.popsci.com/technology/open-data-scientific-research> [Accessed 23 May 2022].

iNaturalist. 2021. *iNaturalist*. [online] Available at: <https://www.inaturalist.org> [Accessed 23 May 2022].

iNaturalist. 2022. *About · iNaturalist*. [online] Available at: <https://www.inaturalist.org/pages/about> [Accessed 23 May 2022].