

**COLLEGE OF COMPUTING AND INFORMATION SCIENCE**

**BACHELOR OF SCIENCE IN SOFTWARE ENGINEERING YR 2**

**(BSSE II)**

**RECESS PROJECT MODEL 1: GENDER CLASSIFICATION MACHINE LEARNING BASED ON**

**FINGERPRINT DATASET**

**06/08/2023**

**GROUP O**

|  |  |  |
| --- | --- | --- |
| **NAME** | **STUDENT No.** | **REGISTRATION No.** |
| SSEKITOOLEKO PETER CLAVER | 2200706919 | 22/U/6919 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

TABLE OF CONTENTS

[Abstract: 2](#_Toc142236037)

[Terms and definitions: 2](#_Toc142236038)

[1. Introduction: 3](#_Toc142236039)

[2. Background and Related Work: 4](#_Toc142236040)

[3. Methodology: 4](#_Toc142236041)

[a) Model structure 4](#_Toc142236042)

[b) Model Compilation and Training Setup 5](#_Toc142236043)

[c) Summary 5](#_Toc142236044)

[d) Data augmentation 6](#_Toc142236045)

[e) Oversampling 6](#_Toc142236046)

[4. Data Overview: 8](#_Toc142236047)

[5. Experiments: 8](#_Toc142236048)

[6. Model Evaluation: 9](#_Toc142236049)

[a) Easy Difficulty Level: 9](#_Toc142236050)

[b) Medium Difficulty Level: 10](#_Toc142236051)

[c) Hard Difficulty Level: 10](#_Toc142236052)

[7. Analysis and Discussion: 10](#_Toc142236053)

[a) Initial Model vs. Modified Model: 10](#_Toc142236054)

[b) Easy, Medium, and Hard Difficulty Levels: 10](#_Toc142236055)

[c) Mitigating Imbalance and Overfitting: 11](#_Toc142236056)

[8. Recommendations: 11](#_Toc142236057)

[9. Conclusion: 11](#_Toc142236058)

[10.Appendix 11](#_Toc142236059)

LIST OF FIGURES

[Figure 1 CNN model with regularization 5](#_Toc142236088)

[Figure 2 Optimized model with L2regularization and Dropout regularization 7](#_Toc142236089)

[Figure 3 Shows data augmentation 7](#_Toc142236090)

[Figure 4 Data representation before oversampling 8](#_Toc142236091)

[Figure 5 Data representation after oversampling 8](#_Toc142236092)

[Figure 6 Training log showing the last training epoch 9](#_Toc142236093)

[Figure 7 Showing Classification report after regularization 9](#_Toc142236094)

[Figure 8 Showing Classification report after augmentation 10](#_Toc142236095)

[Figure 9 Showing Classification report after oversampling 10](#_Toc142236096)

[Figure 10 Training logs for easy altered dataset 13](#_Toc142236097)

[Figure 11 Training logs for altered medium 14](#_Toc142236098)

[Figure 12 Training logs for hard altered 15](#_Toc142236099)

# Abstract:

This report presents a comprehensive analysis of a machine-learning project aimed at gender classification using altered fingerprint images. The project involves data preprocessing, model training, and evaluation. Various strategies, including regularization, data augmentation, and random oversampling, were employed to address challenges posed by class imbalance and overfitting.

# Terms and definitions:

* **Validation Accuracy:**

The accuracy of the model's predictions on a separate validation dataset. It indicates the model's performance on unseen data and its ability to generalize.

* **Validation Loss:**

A measure of the difference between the predicted values and the actual values in the validation dataset. It helps in assessing the model's generalization capability.

* **Training Accuracy:**

The accuracy of the model's predictions on the training dataset. It shows how well the model fits the training data and learns from it.

* **Training Loss:**

A measure of the difference between the predicted values and the actual values in the training dataset. It helps in understanding how well the model is learning the training data.

* **F1 Score:**

A metric that combines both precision and recall. It provides a balanced measure of a model's accuracy by considering both false positives and false negatives.

* **Precision:**

The ratio of correctly predicted positive observations to the total predicted positives. It measures how accurate the positive predictions are.

* **Recall:**

The ratio of correctly predicted positive observations to all observations in actual class. It measures the model's ability to identify all relevant instances.

* **Dataset:**

A collection of data used to train, validate, and test machine learning models. It forms the foundation for training and evaluating the model's performance.

* **Augmentation:**

A technique that involves artificially increasing the diversity of the dataset by applying transformations to the original data, such as rotations, flips, and noise addition.

* **Oversampling:**

A method to address the class imbalance by increasing the number of instances in the minority class. It involves creating synthetic instances to balance class distribution.

* **Preprocessing:**

The steps are taken to prepare and clean the raw data before feeding it into a machine-learning model. This includes tasks like data normalization, resizing, and cleaning.

* **Regularization:**

Techniques are employed to prevent overfitting by adding penalties or constraints to the model's parameters. It helps the model generalize better to unseen data.

* **Overfitting:**

A scenario where a model learns the training data too well, capturing noise and irrelevant patterns. This can lead to poor performance on new data.

* **Deep Learning:**

A subfield of machine learning that involves the use of neural networks with multiple layers. It's particularly effective for complex tasks like image and speech recognition.

* **Neural Networks:**

Computational models inspired by the human brain's structure and function. They consist of interconnected nodes or "neurons" that process and transmit information.

* **Layering:**

Organizing neural networks into sequential layers, each performing specific operations. These layers include input, hidden, and output layers.

# 1. Introduction:

Gender classification using fingerprint images is a challenging problem with applications in security and identification systems. This report outlines our approach to tackling this problem using machine learning techniques. We employed some techniques from various fields of machine learning like deep learning, neural networks, and layering. This required usage of several libraries to assist in the study and analysis.

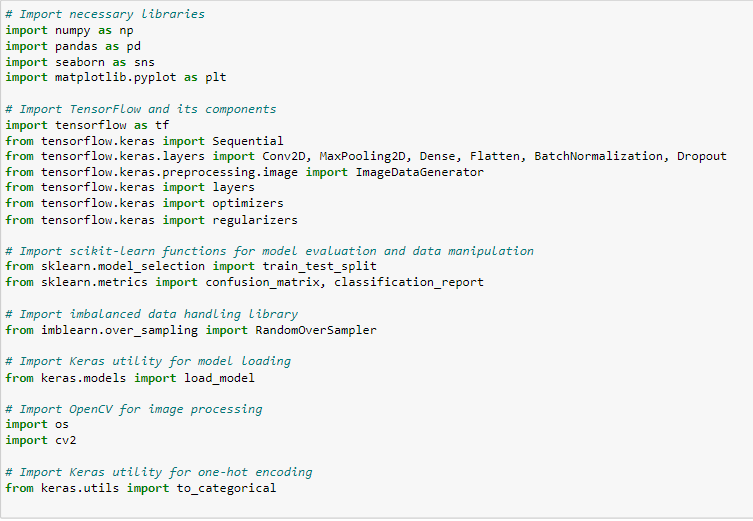


Figure 1 CNN model with regularization

# 2. Background and Related Work:

The use of fingerprint recognition for gender classification has gained attention due to its potential applications. These applications include biometric security systems and biometric user authentication systems. With the addition of the gender classification models to the fingerprint recognition models results show a better performance of the overall model in terms of F1 score, recall, and accuracy. Previous studies have focused on addressing class imbalance and overfitting, which are common challenges in such tasks.

# 3. Methodology:

## Model structure

Our approach involved the implementation of a Convolutional Neural Network (CNN) architecture for feature extraction from fingerprint images. To combat overfitting, we incorporated L2 regularization for Dense layers and Dropout regularization for Convolutional layers.

1. **Input Layer:**
   1. Conv2D Layer: 32 filters, kernel size 3x3, 'same' padding, ReLU activation, and 'he\_uniform' kernel initializer.
   2. Input Shape: (96, 96, 1) - representing a grayscale image with a resolution of 96x96 pixels.
2. **MaxPooling2D Layer:**
   1. Pooling size: 2x2.
   2. Dropout Layer: Dropout rate of 0.25.
   3. Purpose: Introduces pooling to downsample the spatial dimensions and reduce computational complexity. Dropout helps mitigate overfitting.
3. **Convolutional Layer:**
   1. 32 filters, kernel size 3x3, 'same' padding, ReLU activation, and 'he\_uniform' kernel initializer.
   2. Purpose: Extract complex features from the downsampled representation.
4. **MaxPooling2D Layer:**
   1. Pooling size: 2x2.
   2. Dropout Layer: Dropout rate of 0.25.
   3. Purpose: Further downsample the feature maps while introducing regularization through dropout.
5. **Flatten Layer:**
   1. Converts the 2D feature maps into a 1D feature vector, preparing for fully connected layers.
6. **Dense Layer:**
   1. 128 units, ReLU activation, 'he\_uniform' kernel initializer.
   2. L2 Regularization: Regularizes the layer with L2 regularization strength of 0.01.
   3. Purpose: Adds a dense, fully connected layer with L2 regularization to prevent overfitting and improve generalization.
7. **Dense Layer:**
   1. 1 unit with sigmoid activation.
   2. Purpose: The final layer produces a binary classification output, indicating whether the fingerprint image is male or female.

## Model Compilation and Training Setup

* After defining the architecture, the model is compiled using the Adam optimizer with a learning rate of 1e-3 and binary cross-entropy loss function. The chosen optimizer is known for its efficiency in training deep neural networks.
* To prevent overfitting and ensure the model generalizes well, an Early Stopping callback is employed. This callback monitors the validation loss and halts training if the loss does not improve over a certain number of epochs (patience set to 10).

## Summary

* The model aims to classify fingerprint images into male or female categories.
* The model architecture incorporates convolutional layers to learn hierarchical features from fingerprint images. MaxPooling layers downsample the feature maps, and dropout layers introduce regularization to mitigate overfitting. A fully connected layer with L2 regularization ensures better generalization.
* This architecture's combination of convolutional, pooling, and regularization layers creates a balance between capturing intricate features and preventing overfitting. The subsequent compilation and training setup enhance the model's convergence and generalization abilities.



Figure Optimized model with L2regularization and Dropout regularization

## Data augmentation

We also employed data augmentation to solve the issue of class imbalance as the female fingerprints were significantly lower than the male fingerprints.

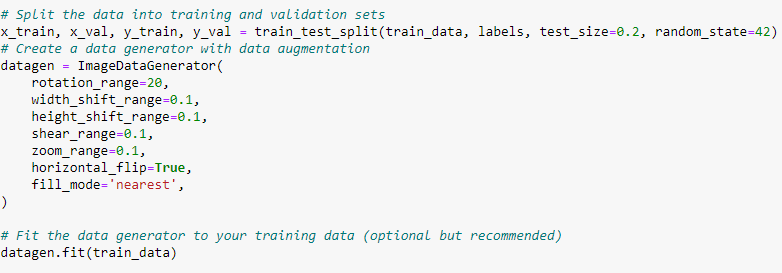


Figure Shows data augmentation

## Oversampling

We implemented the Random Oversampling technique during training. This approach rebalanced the gender distribution within each difficulty level, enhancing the model's ability to learn from both genders.

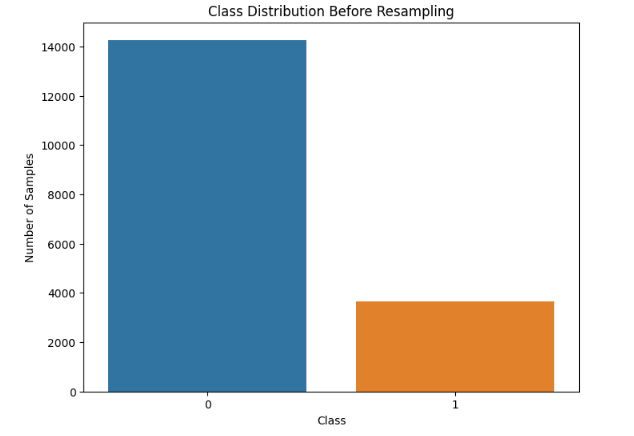


Figure 4 Data representation before oversampling

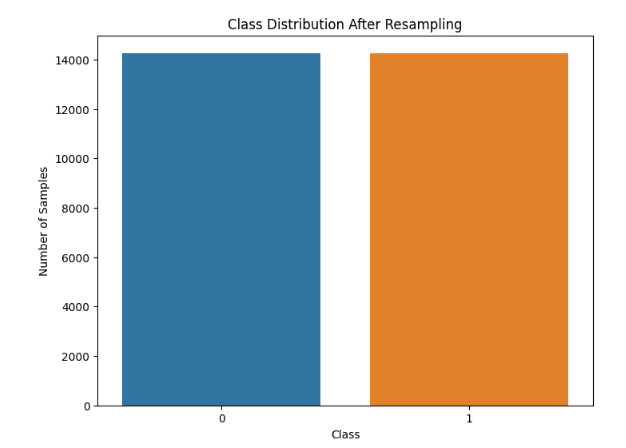


Figure 5 Data representation after oversampling

# 4. Data Overview:

The training dataset consisted of altered fingerprint images categorized into three difficulty levels: easy, medium, and hard.

Each level exhibited a significant gender imbalance, with more male samples compared to females.

* Easy Dataset: 3665 F, 14266 M
* Medium Dataset: 3426 F, 13641 M
* Hard Dataset: 2746 F, 11526 M

The testing dataset consisted of real fingerprint images which also exhibited a significant gender imbalance

* Test Dataset (Real Fingerprints): 1230 F, 4770 M

***Note:*** *unlike the training dataset, the testing dataset wasn’t exposed to random oversampling or data augmentation.*

# 5. Experiments:

Our model underwent training on all difficulty levels. The model first underwent training without the regularization techniques with the easily altered dataset where upon analysis of the training logs we noticed a very high training accuracy and fluctuating validation and training loss which we concluded that there must be overfitting.



Figure Training log showing the last training epoch

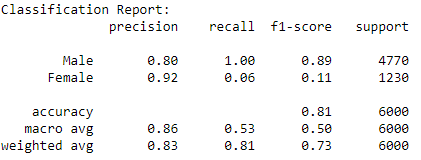
We employed L2 regularization for Dense layers and Dropout regularization for Convolutional layers to mitigate overfitting to make the model more robust as visualized in Figure 1. We then generated a report to show the model’s performance.  


Figure Showing Classification report after regularization

The report showed a struggle by our model to correctly identify female fingerprints so we first employed data augmentation as a means to solve the issue at hand and generated the report below.

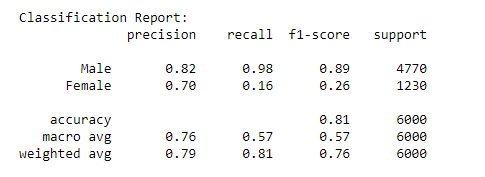


Figure Showing Classification report after augmentation

The report showed that the model hadn’t improved in the direction of detecting female fingerprints correctly. Upon that conclusion, we went ahead to try random oversampling and we generated this report.

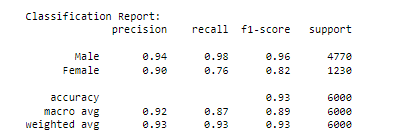
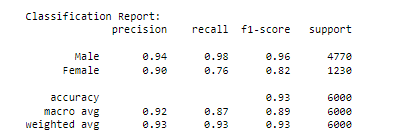


Figure Showing Classification report after oversampling

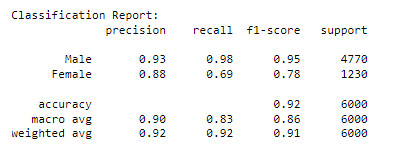
The results showed a great improvement in our model’s overall performance and more so being able to correctly distinguish and identify the female fingerprints.

# 6. Model Evaluation:

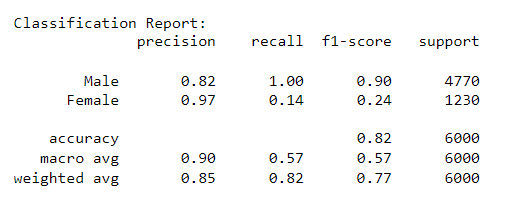
## Easy Difficulty Level:



## Medium Difficulty Level:



## Hard Difficulty Level:



# 7. Analysis and Discussion:

## Initial Model vs. Modified Model:

* The initial model suffered from overfitting, as evidenced by a widening gap between training and validation accuracy.
* The modified model, with regularization techniques, demonstrated improved generalization, as indicated by narrowed accuracy discrepancies.

***Note: The training logs can be viewed from the appendix***

## Easy, Medium, and Hard Difficulty Levels:

* The model excelled in distinguishing genders in easy and medium difficulty levels, showcasing balanced precision and recall.
* We still notice a small fall in the precision recall and F1 score which is attributed to the greater distortion in the medium-level fingerprints compared to the altered easy fingerprints.
* Challenges emerged in hard cases, particularly in accurately classifying female fingerprints as we see a significant fall in the F1 score and recall.
* This was partially attributed to the reduction in the training data at the altered hard level (from 17,000-14,000).  
  It is also attributed to the greater distortion of the fingerprints at the hard level compared to the medium level.

## Mitigating Imbalance and Overfitting:

* Random oversampling successfully balanced the gender distribution during training, mitigating the effects of class imbalance.
* L2 regularization for Dense layers and Dropout regularization for Convolutional layers curbed overfitting, yielding better generalization.

# 8. Recommendations:

* Further fine-tuning and feature engineering could be explored to improve performance in hard difficulty levels.
* Analyzing misclassified examples from challenging cases can provide insights into specific patterns causing difficulties.
* Considering advanced techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to tackle class imbalance more effectively.

# 9. Conclusion:

Our model's performance in easy and medium difficulty levels is promising, demonstrating its ability to learn features effectively. However, challenges persist in correctly classifying female fingerprints in complex cases. Iterative refinement and exploration of advanced techniques are recommended to bridge this gap.

# 10.Appendix

* **NumPy** (**import numpy as np**): NumPy is a popular library for numerical computing in Python. It provides support for multi-dimensional arrays and mathematical functions to operate on these arrays efficiently.
* **matplotlib.pyplot** (**import matplotlib.pyplot as plt**): Matplotlib is a widely used library for creating visualizations and plots in Python. The **pyplot** module within Matplotlib provides a simple interface to create various types of plots.
* **keras** and **tensorflow** (**import keras** and **import tensorflow as tf**): Keras is a high-level neural networks API written in Python. It allows easy and fast prototyping for building and training deep learning models. TensorFlow, on the other hand, is a popular deep-learning framework developed by Google. Keras is often integrated with TensorFlow to leverage its computational capabilities.
* **seaborn** (**import seaborn as sns**): Seaborn is a Python data visualization library based on Matplotlib. It provides an interface for creating more attractive and informative statistical graphics.
* **layers** and **Model** from **keras**: These are specific components from Keras used to define and build neural network architectures.
* **shuffle** from **sklearn.utils** (**from sklearn.utils import shuffle**): **shuffle** is a utility function from scikit-learn (sklearn) used for randomly shuffling data.
* **train\_test\_split** from **sklearn.model\_selection** (**from sklearn.model\_selection import train\_test\_split**): This function from scikit-learn is used to split the data into training and testing sets for model evaluation.
* **iaa** from **imgaug** (**from imgaug import augmenters as iaa**): ImgAug is a library for image augmentation, which is a technique used to increase the diversity of the training data by applying random transformations to the images.
* **os** (**import os**): The **os** module provides a way to interact with the operating system. It is commonly used to handle file and directory operations.
* **cv2** (**import cv2**): OpenCV (Open Source Computer Vision) is a popular computer vision library that provides tools for image and video processing.
* **random**: The **random** module is used for generating random numbers or making random selections.
* **from tensorflow.keras import Sequential**: This imports the **Sequential** class from TensorFlow's Keras API. **Sequential** is a linear stack of layers that can be used to build neural network models.
* **from tensorflow.keras.layers import Conv2D, MaxPooling2D, Dense, Flatten, BatchNormalization, Dropout**: These are various types of layers that will be used to build the CNN model.
* **from tensorflow.keras import layers**: This imports additional layers from TensorFlow's Keras API.
* **from tensorflow.keras import optimizers**: This imports the optimizers module from TensorFlow's Keras API, which contains different optimization algorithms like Adam, SGD, etc.
* **from sklearn.metrics import confusion\_matrix**: This imports the **confusion\_matrix** function from scikit-learn. The confusion matrix is used to evaluate the model's performance in a classification problem.

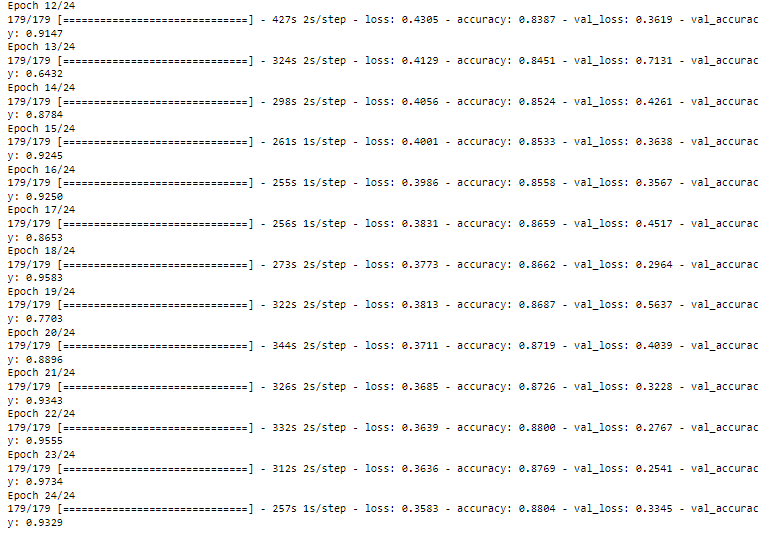


Figure Training logs for easy altered dataset

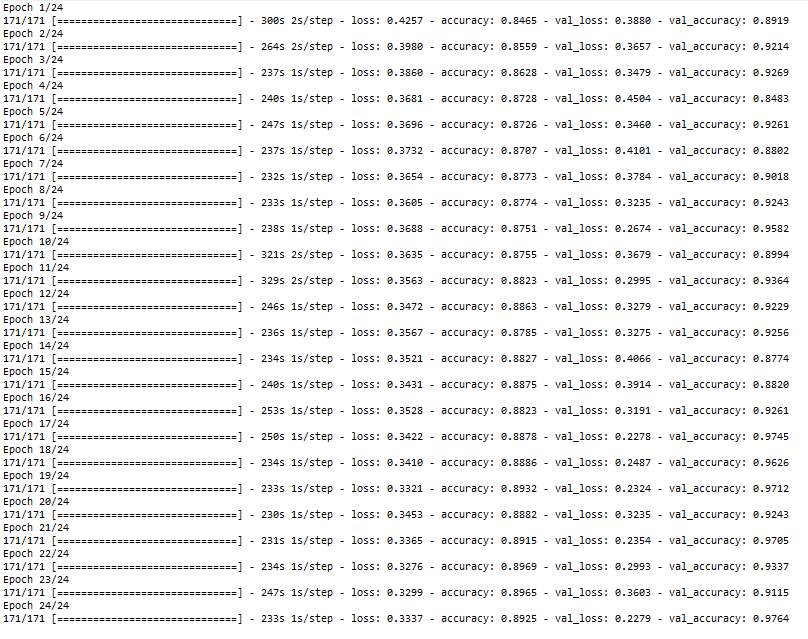


Figure Training logs for altered medium

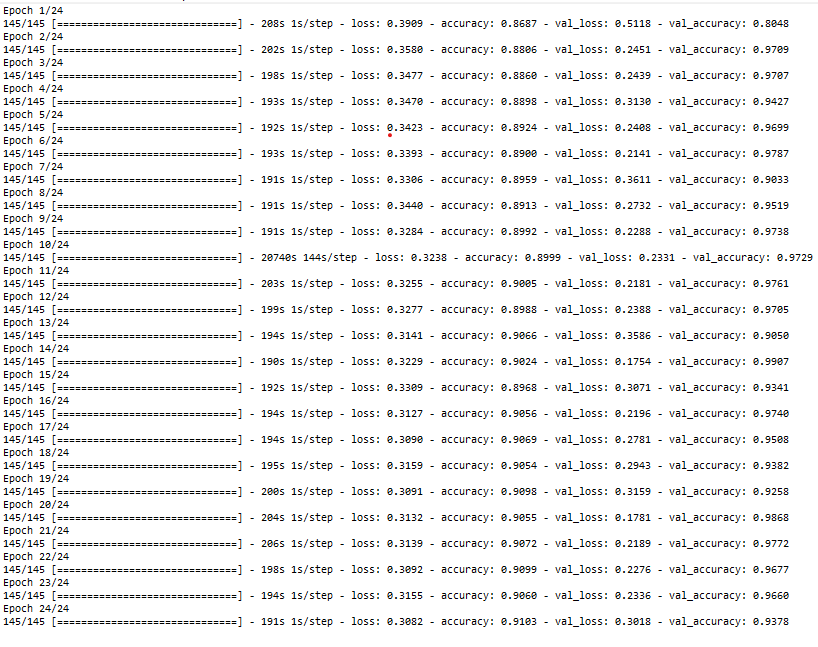


Figure Training logs for hard altered