Transfer Learning based Gender Identification using Arbitrary Celebrity Image Sets

Erick Alejandro Borda Mercado
Northumbria University London Campus UK
Email: erick.borda@northumbria.ac.uk

Sonjoy Ranjon Das
Dep. of Comp. Science & Eng., *Global Banking School UK*
Email: sdas@globalbanking.ac.uk

Bilal Hassan
School of Computing & Digital,
London Metropolitan University, UK
Email: b.hassan@londonmet.ac.uk

Preeti Patel
School of Computing & Digital,
London Metropolitan University, UK
Email: p.patel@londonmet.ac.uk

***Abstract*— Gender Identification is important for security, personalization, and social media analysis, where accurate gender identification enhances the system performance. In this study, we investigated the use of transfer learning for Gender identification from the perspective of minimizing accuracy and efficiency loss. The models, pre-trained on ImageNet weights, were fine-tuned on 300 celebrity images to analyze their performance with varied data constraints. The models were evaluated based on the model architecture and hyperparameters, such as batch size and data split. VGG16 and VGG19 worked impressively with a combined performance level of 98% (97% for female samples and 99% for male samples). However, ResNet50 and ResNet101 showed fluctuating levels of performance, attaining the best accuracy levels of 77.5% and 79.5%, respectively. The results indicate that less complex models, such as VGG16 and VGG19, outperform their more complex versions, such as ResNet50 and ResNet101, on smaller datasets because of their higher efficiency and suitability. More complex models require more sophisticated fine-tuning procedures but tend to have lower performance levels than less complex models. The research identifies that transfer learning significantly eliminates the necessity for longer retraining and customization of models, especially when adapting them to similar tasks. Additionally, considerable performance differences between the male and female categories were identified, highlighting the necessity of balanced datasets and model training to accurately reflect varied gender expressions.**

Keywords*— *Transfer learning, gender identification, CNNs, arbitrary images, celebrity.

I. INTRODUCTION

Soft biometrics, such as Gender identification, are concerned with the analysis of probabilistic and nonintrusive attributes. The application of pre-trained models to large and diverse data has been found to provide considerable benefits, particularly when data availability is poor or training resources are scarce [1]. This has enabled the development of highly accurate gender-classification systems that can be applied to a wide range of images. However, the deployment of these sophisticated techniques requires a thorough evaluation of their scientific and social implications. Ethical use demands attention to the accuracy and reduction of biases, which can reinforce stereotypes. Advances in such technologies should be secured with technical soundness and ethical fairness to foster equitable applications. Recent advances in Computer Vision (CV) and Machine Learning (ML) have significantly enhanced gender recognition using digital images [2]. However, the correct classification of global attributes, such as gender, age, and ethnicity, is not easy due to bias in the training data. This study attempts to solve such issues by developing a dataset of 300 celebrities representing a wide spectrum of genders, ethnicities, and ages. The construction of such a dataset was imperative for ascertaining the strength and variety of transfer learning models in the classification of gender across demographic categories. Preprocessing included resizing the images to uniform sizes and normalizing the pixel values to perform uniform training and testing. The data aimed at capturing human diversity in its totality, maximizing the aim of the study, and providing insight into anticipated expansions. Females and males were the indicated sex categories, while Asian, black, white, mixed, and others represented the ethnicity categories. The age groups were represented by five categories, which–5-15, 16-30, 31-45, 46-60, and 61+ years.

The motivating factor for this research was to make Gender identification technologies more inclusive, accurate, and ethically responsible. Acknowledging the limitations of previous methods, this study utilized state-of-the-art model-adaptation techniques to optimize both fairness and accuracy. The overall goal is to create artificial intelligence models that are diversity-aware, bias reducing, and equity promoting in digital gender recognition. For this purpose, this study examined the applicability of transfer learning in CV and soft biometrics based on a celebrity dataset to analyze the performance of various pre-trained deep learning models in Gender identification across various ethnic and age groups. This study specifically aimed to apply transfer learning to transfer pre-trained deep learning models to Gender identification based on the Celebrity Image Dataset. This study adds to the literature on gender recognition using deep learning in four primary aspects, one of which is the introduction of a varied dataset comprising celebrity images to augment the diversity in AI-powered Gender identification.

The subsequent sections present the methodology, findings, and implications of Gender identification based on transfer learning. This study begins with a survey of the existing literature on computer vision, soft biometrics, deep learning, and transfer learning in gender identification, highlighting the key gaps that this study seeks to fill. The methodology outlines the formal procedure adopted to obtain the objective of the study, followed by an explanation of data preprocessing, model establishment, and optimization through various parameter values. The results and evaluation outline the results, model performance analysis, and findings that formed the final conclusions. This paper concludes with an overview of the recommendations, challenges encountered, and guidelines for follow-up research studies.

Convolutional Neural Networks (CNNs) have become a key technology in supervised deep learning, resulting in phenomenal progress in activities such as computer vision, speech recognition, and image classification [3]. Architectures such as ResNet50 have been very successful in the field of soft biometrics, as demonstrated by Md. Islam et al.[4] . The initial boost in Gender identification was provided by binary classifiers, which reduced the gender identification task to two distinct classes, Male and Female. The approach described was constructed by training image classifiers to identify sex-specific features without manual annotation [5]. Soft biometrics targets perceptually visible attributes, enabling the semantic description of an individual. Gender and overall demographic features, such as age and ethnicity, were analyzed in this study using a descriptive approach to improve recognition accuracy [6]. Categorical, comparative, and hybrid methods have been employed to identify soft biometrics. Unlike the comparative method, which is based on a comparison with a reference database, the categorical method categorizes individuals based on apparent features. Quantitative methodologies overcome the constraints associated with conventional recognition technologies and enhance their accuracy and reliability [7]. Comparative studies on the performance of humans and algorithms confirm that, although humans are better at gender discrimination tasks, algorithms yield better outcomes in age estimation. The fusion of soft biometrics and traditional recognition approaches marginally boosts the performance, reflecting the potential to reduce false matches and becoming increasingly comparable to the performance of humans [8]. In gender recognition, Transfer Learning (TL) plays a vital role in alleviating issues in training complex deep-learning models. TL enables the models to benefit from pre-trained weights and filters, thereby eliminating the need for large amounts of labeled data and computational resources [9]. Pinto et al. [10] formally defined TL in terms of three primary elements: (1) a domain D with a feature space χ and a marginal probability distribution P(X); (2) a task T with a label space Y and a predictive function f, where f learns to estimate probability P(X) in order to forecast new instances, and (3) given a source domain D_S with task T_S and a target domain D_T with task T_T, TL enhances learning in D_T by transferring knowledge from D_S, where D_T ≠ D_S or T_T ≠ T_S. In TL, pre-trained models are tuned to new tasks by modifying some network layers, which enables faster training and improved accuracy [11].

Nguyen et al. [3] point out that TL facilitates the rapid adaptation of already acquired knowledge to new tasks at lower computational expense. Its success is based on the degree of alignment between the source model training and target tasks. Pre-trained models are treated as feature extractors; therefore, no manual feature engineering is required, and the performance of new data is improved [12]. The adaptation of pre-trained models involves either initializing the model with some parameters when there is some sample data present or selectively fine-tuned layers based on the amount of data and model complexity. This improves the adaptability and efficiency of TL for deep learning applications [13].

The gender recognition pipeline comprises data collection, preprocessing, and adaptation of the CNN model. The dataset was labeled carefully for correct classification. Preprocessing involves normalizing the image sizes, orientations, and color profiles to ensure consistency with the CNN architecture [14]. Pre-trained models such as DenseNet and ResNet, which are trained on large datasets such as ImageNet, have the advantage of utilizing the learned features for Gender identification [15]. CNNs leverage convolutional layers for visual feature extraction and fully connected layers for final classification. Transfer learning optimizes such models by maintaining convolutional layers and fine-tuning fully connected layers to recognize sex-specific features. Thus, the model can maintain generic visual representations while specializing in gender-classification tasks. Fine-tuning also optimizes the performance with guaranteed feature extraction and classification efficacy [16].

Gender identification systems are integral to security, marketing, healthcare, and social-science applications. Such applications have enabled video surveillance, targeted marketing, medical diagnoses, and policymaking. Some issues related to biased training datasets, variations between cultures, lack of model interpretability, and ethical concerns related to privacy and data abuse must be resolved [17]. This study examines how TL enhances Gender identification by leveraging pre-trained models to enhance classification performance while transcending dataset limitations. Soft biometrics plays a critical role in the development of Gender identification by incorporating contextual features. By leveraging TL, Gender recognition models become more flexible, efficient, and ethically responsible, thereby ensuring fairness across different applications [18]. This study underscores the effectiveness of TL for sex classification and its broader ramifications in AI-powered biometric technology.

II. DATASET

The dataset utilized for this research consisted of 300 celebrity images gathered from publicly accessible media outlets and celebrity photograph archives, thereby ensuring that there was total adherence to legal use requirements. To provide an equal distribution of global characteristics, a very important factor in training models with good performance across various demographic groups, images were carefully selected [6][19]. Such a balance not only reduces bias, but also makes the model fairer. Moreover, utilizing celebrity faces provided a dataset with well-known people featuring high variability in their presentation, context, and style, thereby enhancing the generalizability of the model.

Nevertheless, for consistency and reliability in further analyses, preprocessing was essential for tidying up the dataset[20]. Because the cleanliness of the dataset has direct implications for the efficiency and accuracy of the model, consistency in the formats of image files must be observed. Different formats require different processing techniques, which can complicate the data management during training[21]. For simplicity, the project employed only a single JPEG format, as shown in Fig. 1. After standardization, JPEG alone was used in file format, as shown in Fig. 2a.

In addition to standardizing file formats, image size analysis and standardization were also required to bring about consistency in dimensions for model input. This uniformity improves the learning efficiency, resulting in improved model performance. The inconsistency in image sizes before standardization is shown in Fig. 2b, while Fig. 3a shows the resizing process, where all images were altered to a size of 224×224 pixels to meet the input requirements of the model. Fig. 3b provides a visual comparison of the images before and after resizing, highlighting the importance of this process. Finally, dataset splitting is a crucial step in the development of robust models that can learn effectively and generalize new data. To achieve the best performance and testing, the dataset was split strategically based on the defined split ratios, as presented in Table 1.

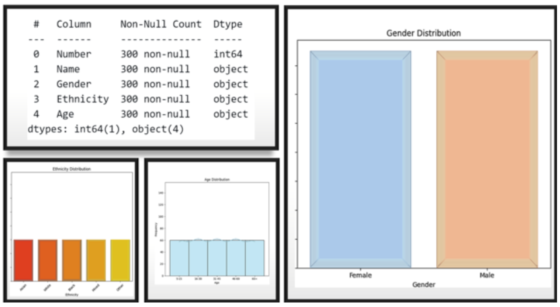

Fig. 1. Data distribution for the primary dataset.

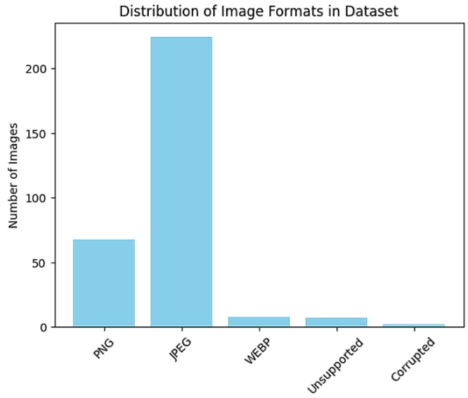

Fig. 2a. Image format distribution.

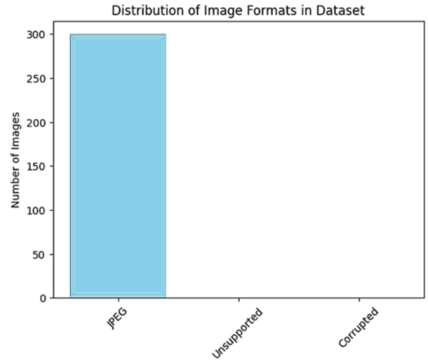

Fig. 2b. Standardized image format distribution.

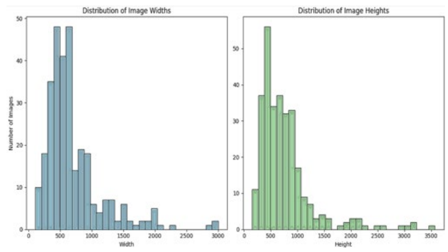

Fig. 3a. Image size distribution.

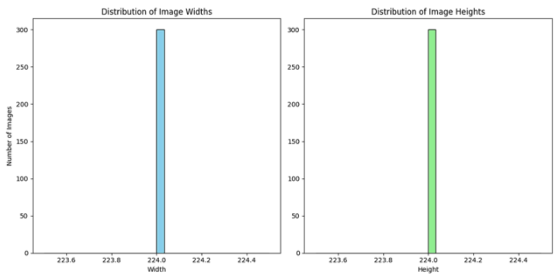

Fig. 3b. Standardized image size distribution.

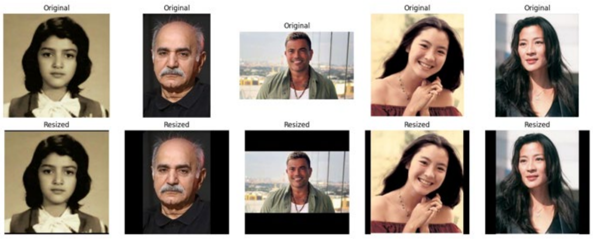

Fig. 4. Original vs. resized images.

TABLE 1 DATASET SPLIT STRATEGY

Split Ratio	Training Set (%)	Validation Set (%)	Test Set (%)	Rationale
60/20/20	60	20	20	Equal emphasis on validation and testing to ensure robustness in model performance and parameter tuning. This split provides a balanced approach to model development and evaluation, especially useful for comprehensive testing and validation.
70/15/15	70	15	15	More data allocated to training to leverage learning from a larger set, reducing the proportion for validation and testing. Suitable for scenarios where increased training data significantly aids model performance.
80/10/10	80	10	10	Maximizes the training dataset to improve model learning capabilities, with minimal data for validation and testing. Ideal for complex models requiring extensive training data or when underfitting is a concern.

III. MTHODOLOGY

In the present study, we employed four transfer learning methods VGG16, VGG19, ResNet50, and ResNet101 chosen because of their performance, architectural variety, and applicability to our research interest [23]. These models are examples of varying depths and design philosophies within convolutional neural networks, enabling us to make an educated comparison regarding their relative performance. The motivation for choosing these methods is the need to fill identified gaps in the existing literature, specifically the absence of direct comparisons between these architectures under comparable environments. Through an in-depth examination of their weaknesses and strengths, we seek to establish the best approach, understand the conditions under which each model performs best, and offer insights that can inform model choice and optimization decisions. This comparative study strengthens both the theoretical understanding and practical advancement in the field of transfer learning applications.

A. VGG16 with ImageNet

This section explores the application of the VGG16 model pre-trained on ImageNet weights for gender identification tasks. The process began with the configuration of computational resources, ensuring that GPU allocation was optimized for training. Following the configuration, we loaded the pre-trained weights from the ImageNet dataset for the VGG16 model. These pre-trained weights enabled us to leverage the network's learned features for robust gender classification. To prevent the model from updating the pre-trained layers during the initial training phase, we applied the transfer learning technique by freezing these layers. Additionally, we added two dense layers for gender classification (male and female). The training process for VGG16 involved setting an 80/10/10 data split, adjusting batch sizes (16 and 32), and experimenting with epochs (20 and 30). The results were evaluated using metrics such as training accuracy, validation accuracy, and training time. The number of images in each training and validation split with 80% for training and 20% for testing. After training the model, Table 2 was created, which represents the performance of the training and validation of the model.

TABLE 2 TRAINING AND VALIDATION PERFORMANCES FOR VGG16

Model	Split ratio	batch	epoch	Training accuracy (%)	Validation accuracy (%)	Training time (seg)
VGG16	60/20/20	16	10	90.59	85	102.75
		16	20	99.76	81	208.16
		16	30	98.59	83	348.92
		32	10	85.66	81.67	118.32
		32	20	99.16	81.67	220.52
		32	30	100	83.33	426.08
		64	10	90.34	75	104.19
		64	20	94.96	81.67	199.37
		64	30	97.46	80	325.77
VGG16	70/15/15	16	10	92.58	84.44	102.44
		16	20	98.86	86.67	239.85
		16	30	95.12	82.22	328.2
		32	10	86.57	80	107.46
		32	20	98.24	84.44	207.25
		32	30	99.13	86.67	319.03
		64	10	86.99	80	102.82
		64	20	95.78	84.44	216.03
		64	30	98.47	86.67	307.2
VGG16	80/10/10	16	10	96.03	90.32	132.53
		16	20	99.63	83.87	283.87
		16	30	100	93.55	454.87
		32	10	92.23	90.32	140.56
		32	20	99.37	93.55	311.32
		32	30	99.91	93.55	479.85
		64	10	83.84	74.19	156.51
		64	20	95.33	83.37	321.06
		64	30	97.46	93.55	453.91

Fine-tuning was conducted based on the results obtained from different model configurations, as presented in Table 2, with optimal parameters including an 80/10/10 data split, batch sizes of 16 and 32, and epochs set to 20 and 30. A reduced learning rate was implemented to enhance the performance and mitigate overfitting. Additionally, data augmentation techniques using "ImageDataGenerator" were applied to improve generalization by incorporating rescaling, geometric transformations, and brightness adjustments. Early termination was introduced to prevent unnecessary training cycles. Despite these optimizations, no significant improvements were observed in model performance, as indicated by the results in Table 3.

TABLE 3 TRAINING AND VALIDATION PERFORMANCE AFTER FINE TUNING FOR VGG16

Model	Split ratio	batch	epoch	Training accuracy (%)	Validation accuracy (%)	Training time (seg)	Configuration
VGG16	80/10/10	16	20	99.19	90.32	230.13	Basic
		16	20	97.76	93.55	259.59	Finetune
		16	30	100	87.1	336.57	Basic
		16	30	96.17	90.32	408.2	Finetune
		32	20	99.37	93.55	311.32	Basic
		32	20	96.29	87.1	246.72	Finetune
		32	30	99.91	93.55	479.85	Basic
		32	30	98.66	93.55	377.65	Finetune
		64	20	95.33	83.37	321.06	Basic
		64	20	90.93	93.55	263.17	Finetune
		64	30	97.46	93.55	453.91	Basic
		64	30	96.79	93.55	434.43	Finetune

B. VGG19 with ImageNet

In this section, we describe the application of the VGG19 architecture pre-trained on ImageNet weights. The objective is to fit this model via transfer learning mechanisms to classify genders. The minimal setup of the VGG19 model with pre-trained ImageNet weights enables the use of a network that has already learned from a rich variety of images. We froze the pre-trained layers to preserve the learned features and prevent them from being updated during the initial phase of training. The addition of two dense layers for Gender identification in this case, male and female. Training was conducted by experimenting with different data-splitting ratios, batch sizes, and epochs, and results were evaluated based on training accuracy, validation accuracy, and training time. Table 4 presents the training and validation performance for VGG19.

TABLE 4 TRAINING AND VALIDATION PERFORMANCE FOR VGG19

Model	Split ratio	batch	epoch	Training accuracy (%)	Validation accuracy (%)	Training time (seg)
VGG19	60/20/20	16	10	93.26	83.33	170.08
		16	20	97.86	83.33	330.02
		16	30	100	81.67	466.45
		32	10	92.22	80	149.38
		32	20	97.3	80	362.11
		32	30	99.37	85	494.13
		64	10	62.48	75	149.29
		64	20	96.2	76.67	229.97
		64	30	96.66	81.67	442.63
VGG19	70/15/15	16	10	93.56	77.78	121.72
		16	20	97.66	77.78	291.95
		16	30	99.84	77.78	449.64
		32	10	88.91	82.22	136.4
		32	20	98.17	77.78	262.36
		32	30	99.24	84.44	388.7
		64	10	78.04	68.89	140.96
		64	20	93.67	80	293.96
		64	30	96.33	88.89	464.73
VGG19	80/10/10	16	10	96.73	87.1	169.49
		16	20	97.9	90.32	309.1
		16	30	99.21	83.87	558.26
		32	10	84	90.32	182.57
		32	20	97.47	83.87	352.02
		32	30	98.84	87.1	522.85
		64	10	82.66	83.87	168.52
		64	20	93.68	90.32	323.67
		64	30	96.95	87.1	503.32

Fine-tuning was selectively applied to the models that demonstrated the best performance, ensuring optimization only when significant improvements were achieved. For VGG19, the optimal parameters identified included a splitting ratio of 80/10/10, batch sizes of 32 and 64, and 20 and 30 epochs. A similar configuration was implemented to fine-tune and refine the model and enhance its learning capabilities. The impact of this fine-tuning on VGG19's performance is detailed in Table 5, which presents the post-optimization training and validation results.

TABLE 5 TRAINING AND VALIDATION PERFORMANCE AFTER FINE-TUNING VGG19

Model	Split ratio	batch	epoch	Training accuracy (%)	Validation accuracy (%)	Training time (seg)	Configuration
VGG19	80/10/10	32	20	97.47	83.87	352.02	Basic
		32	20	86.15	80.65	302.03	Finetune
		32	30	98.84	87.1	522.85	Basic
		32	30	89.48	87.1	363.72	Finetune
		64	20	93.68	90.32	323.67	Basic
		64	20	83.93	90.32	308.07	Finetune
		64	30	96.95	87.1	503.32	Basic
		64	22/30	86.91	90.32	313.85	Finetune

C. ResNet50 with ImageNet

We also explored the ResNet50 model, which has a more complex architecture compared to the VGG models. The general configuration for deploying ResNet50 with ImageNet weight. The training strategy for ResNet50 replicates the proven methodologies used in previous models with adjustments for its distinct architecture. The strategy then uses different configuration parameters such as the splitting ratio, batch size, and epochs, as shown in the Table 6.

TABLE 6 RESNET50 PERFORMANCE FOR TRAINING AND VALIDATION

Model	Split ratio	batch	epoch	Training accuracy (%)	Validation accuracy (%)	Training time (seg)
ResNet50	60/20/20	16	10	59.92	55	96.7
		16	20	63.72	56.67	205.11
		16	30	59.96	65	223.7
		32	10	51.63	55	103.44
		32	20	67.65	56.67	196.53
		32	30	79.43	60	218.68
		64	10	52.63	58.33	98.95
		64	20	55.06	56.67	153.26
		64	30	77.02	61.67	206.88
ResNet50	70/15/15	16	10	63.39	44.44	87.9
		16	20	59.23	51.11	161.29
		16	30	58.29	55.56	259.57
		32	10	55.71	62.22	100.9
		32	20	65.28	62.22	170.68
		32	30	69.43	60	228.43
		64	10	50.24	55.56	103.48
		64	20	61.18	57.78	209.43
		64	30	59.09	55.56	348.89
ResNet50	80/10/10	16	10	61.1	67.74	101.95
		16	20	60.66	45.16	220.9
		16	30	71.89	64.52	260.04
		32	10	60.16	77.42	101.89
		32	20	53.78	41.94	178.04
		32	30	64.54	61.29	299.34
		32	50	76	64.52	456.19
		64	10	53.12	64.52	93.43
		64	20	59.74	54.84	167.65
		64	30	59.12	70.97	230.35
		64	50	66.02	61.29	442.18

By evaluating ResNet50, a model recognized for its depth and complexity, its performance was analyzed across three dataset split ratios: 60/20/20, 70/15/15, and 80/10/10. This analysis aimed to identify the model's strengths and limitations in Gender identification tasks under varying training parameters while utilizing its basic configuration. The insights derived from this evaluation are summarized in Table 7.

D. ResNet101 with ImageNet

Building on our analysis of the ResNet50 model, we turned to ResNet101, a model known for its even more complex and deeper architectural framework. This section explores how ResNet101 is configured and evaluated using pre-trained ImageNet weights, specifically targeting the nuances of Gender identification tasks. Based on the training methodologies refined through earlier models, this section introduces a similar strategy tailored for ResNet101 to maximize its performance in Gender identification. Table 7 synthesizes our tailored training parameters, demonstrating how we adapted and scaled our strategies to harness this advanced model with its basic configuration.

TABLE 7 RESNET101 PERFORMANCE FOR TRAINING AND VALIDATION

Model	Split ratio	batch	epoch	Training accuracy (%)	Validation accuracy (%)	Training time (seg)
ResNet101	60/20/20	16	10	62.43	67.74	184.71
		16	20	71.99	54.84	270.07
		16	30	80.17	45.16	402.27
		32	10	60.4	74.19	79.87
		32	20	68.99	71.12	161.41
		32	30	74.12	70.97	228.64
		64	10	55.62	41.94	78.66
		64	20	58.99	61.29	156.83
		64	30	65.4	74.19	241.6
ResNet101	70/15/15	16	10	56.4	48.89	171.84
		16	20	70.69	62.12	320.51
		16	30	66.41	62.22	495.98
		32	10	63.66	62.22	190.88
		32	20	80.72	64.44	301.72
		32	30	65.81	62.22	471.43
		64	10	56.09	46.67	153.97
		64	20	56.67	68.89	275.8
		64	30	67.13	62.22	397.12
ResNet101	80/10/10	16	10	52.15	70.97	183.34
		16	20	72.34	67.74	329.55
		16	30	82.78	74.19	476.53
		32	10	57.82	70.97	157.79
		32	20	69.16	74.19	332.36
		32	30	73.28	67.74	456.42
		64	10	54.32	41.94	156.78
		64	20	70.05	60.33	283.24
		64	30	74.43	58.06	445.22

IV. RESULT EVALUATION

The evaluation of the VGG16 model showed that the optimal configuration was achieved with an 80/10/10 data split, batch size of 16, and 20 epochs, resulting in 98% accuracy. Increasing the batch size to 32 achieved 97.5% accuracy and reducing the training data led to poorer performance (88.5% for a 60/20/20 split). These findings indicate the importance of data distribution, with insufficient training samples deterring the learning. Additionally, longer training times improved accuracy but led to increased computational resource demands, underscoring the need to balance training duration and resource efficiency. Further, male/female identification potential disparities must be analyzed to determine bias and ensure equivalent, uniform performance in all groups; the optimal VGG16 model performed best with 80/10/10 split, as presented in Table 8.

TABLE 8 VGG16 PERFORMANCE FOR TRAINING AND VALIDATION

Model	Split ratio	batch	epoch	Recall for Female	Precision for Female	F1-Score for Female	Recall for Male	Precision for Male	F1-Score for Male	Female Prediction accuracy (%)	Male Prediction accuracy (%)	General Prediction accuracy (%)
VGG16	60/20/20	16	10	0.63	1	0.78	1	0.73	0.85	63	100	81.5
		16	20	0.89	0.98	0.93	0.98	0.9	0.94	89	98	93.5
		16	30	0.95	0.93	0.94	0.93	0.95	0.94	95	93	94
		32	10	0.89	0.97	0.93	0.97	0.9	0.93	89	97	93
		32	20	0.95	0.92	0.94	0.92	0.95	0.94	95	92	93.5
		32	30	0.93	0.95	0.94	0.95	0.93	0.94	93	95	94
		64	10	0.92	0.86	0.89	0.85	0.91	0.88	92	85	88.5
		64	20	0.85	0.98	0.91	0.99	0.87	0.93	85	99	92
		64	30	0.89	0.97	0.93	0.97	0.9	0.93	89	97	93
VGG16	70/15/15	16	10	0.95	0.92	0.94	0.92	0.95	0.94	95	92	93.5
		16	20	0.93	0.99	0.96	0.99	0.93	0.96	93	99	96
		16	30	0.98	0.9	0.94	0.89	0.98	0.93	98	89	93.5
		32	10	0.86	0.95	0.9	0.95	0.87	0.91	86	95	90.5
		32	20	0.94	0.97	0.95	0.97	0.94	0.95	94	97	95.5
		32	30	0.97	0.95	0.96	0.95	0.97	0.96	97	95	96
		64	10	0.87	0.96	0.91	0.97	0.88	0.92	87	97	92
		64	20	0.85	0.97	0.91	0.97	0.87	0.92	85	97	91
		64	30	0.97	0.94	0.95	0.93	0.97	0.95	97	93	95
VGG16	80/10/10	16	10	0.93	0.98	0.96	0.98	0.94	0.96	93	98	95.5
		16	20	0.97	0.99	0.98	0.99	0.97	0.98	97	99	98
		16	30	0.98	0.99	0.98	0.99	0.98	0.98	98	96	97
		32	10	0.94	0.95	0.94	0.95	0.94	0.94	94	95	94.5
		32	20	0.96	0.99	0.98	0.99	0.96	0.98	96	99	97.5
		32	30	0.96	0.99	0.98	0.99	0.96	0.98	96	99	97.5
		64	10	0.68	1	0.81	1	0.76	0.86	68	100	84
		64	20	0.95	0.9	0.93	0.9	0.94	0.92	95	90	92.5
		64	30	0.96	0.99	0.98	0.99	0.96	0.98	96	99	97.5

As previously discussed, the model with the optimal configuration utilized a splitting ratio of 80/10/10. *Table 9* lists the data obtained using the proposed model.

TABLE 9 METRICS AFTER FINE-TUNING OF VGG16.

Model	Split ratio	batch	epoch	Recall for Female	Precision for Female	F1-Score for Female	Recall for Male	Precision for Male	F1-Score for Male	Female Prediction accuracy (%)	Male Prediction accuracy (%)	General Prediction accuracy (%)	Configuration
VGG16	80/10/10	16	20	0.97	0.99	0.98	0.99	0.97	0.98	97	99	98	Basic
		16	20	0.97	0.99	0.98	0.99	0.97	0.98	97	99	98	Finetune
		16	30	0.98	0.99	0.98	0.99	0.98	0.98	98	96	97	Basic
		16	30	0.95	0.99	0.97	0.99	0.96	0.97	95	99	97	Finetune
		32	20	0.96	0.99	0.98	0.99	0.96	0.98	96	99	97.5	Basic
		32	20	0.97	0.92	0.94	0.92	0.97	0.94	97	92	94.5	Finetune
		32	30	0.96	0.99	0.98	0.99	0.96	0.98	96	99	97.5	Basic
		32	30	0.95	0.99	0.97	0.99	0.96	0.97	95	99	97	Finetune
		64	20	0.95	0.9	0.93	0.9	0.94	0.92	95	90	92.5	Basic
		64	20	0.91	0.98	0.94	0.98	0.92	0.95	91	98	94.5	Finetune
		64	30	0.96	0.99	0.98	0.99	0.96	0.98	96	99	97.5	Basic
		64	30	0.94	0.99	0.97	0.99	0.94	0.97	94	99	96.5	Finetune

The VGG19 model demonstrated similar performance trends as VGG16, with an optimal configuration of an 80/10/10 split, batch size of 32 and 64, and 20 and 30 epochs. Table 10 presents the performance metrics for the basic configuration of VGG19.

TABLE 10 PERFORMANCE METRICS FOR VGG19

Model	Split ratio	batch	epoch	Recall for Female	Precision for Female	F1-Score for Female	Recall for Male	Precision for Male	F1-Score for Male	Female Prediction accuracy (%)	Male Prediction accuracy (%)	General Prediction accuracy (%)
VGG19	60/20/20	16	10	0.88	0.96	0.92	0.96	0.89	0.92	88	96	92
		16	20	0.91	0.95	0.93	0.95	0.92	0.93	91	95	93
		16	30	0.92	0.95	0.93	0.95	0.92	0.93	92	95	93.5
		32	10	0.84	0.93	0.88	0.94	0.85	0.9	84	94	89
		32	20	0.92	0.92	0.92	0.92	0.92	0.92	92	92	92
		32	30	0.91	0.96	0.94	0.97	0.92	0.94	91	97	94
		64	10	0.96	0.75	0.84	0.68	0.94	0.79	96	68	82
		64	20	0.89	0.91	0.9	0.91	0.89	0.9	89	91	90
		64	30	0.9	0.94	0.92	0.95	0.9	0.93	90	95	92.5
VGG19	70/15/15	16	10	0.98	0.81	0.89	0.77	0.97	0.86	98	77	87.5
		16	20	0.97	0.9	0.93	0.89	0.96	0.93	97	89	93
		16	30	0.97	0.94	0.95	0.93	0.97	0.95	97	93	95
		32	10	0.93	0.92	0.92	0.91	0.93	0.92	93	91	92
		32	20	0.98	0.86	0.92	0.84	0.98	0.9	98	84	91
		32	30	0.97	0.94	0.95	0.94	0.97	0.95	97	94	95.5
		64	10	0.63	0.97	0.76	0.98	0.72	0.83	63	98	80.5
		64	20	0.95	0.87	0.91	0.86	0.95	0.9	95	86	90.5
		64	30	0.95	0.97	0.96	0.97	0.95	0.96	95	97	96
VGG19	80/10/10	16	10	0.97	0.93	0.95	0.93	0.97	0.95	97	93	95
		16	20	0.99	0.95	0.97	0.95	0.99	0.97	99	95	97
		16	30	0.98	0.91	0.95	0.91	0.98	0.94	98	91	94.5
		32	10	0.9	0.95	0.92	0.95	0.91	0.93	90	95	92.5
		32	20	0.96	0.94	0.95	0.94	0.96	0.95	96	94	95
		32	30	0.97	0.97	0.97	0.97	0.97	0.97	97	97	97
		64	10	0.95	0.84	0.89	0.81	0.95	0.87	95	81	88
		64	20	0.97	0.93	0.95	0.93	0.97	0.95	97	93	95
		64	30	0.95	0.98	0.97	0.98	0.95	0.97	95	98	96.5

The fine-tuned section is incorporated based on the performance of the optimal model, as indicated in Table 11.

TABLE 11 PERFORMANCE METRICS AFTER FINE-TUNING VGG19

Model	Split ratio	batch	epoch	Recall for Female	Precisio n for Female	F1-Score for Female	Recall for Male	Precisio n for Male	F1-Score for Male	Female Predictio n accuracy (%)	Male Predictio n accuracy (%)	General Predictio n accuracy (%)	Configuratio n
VGG1 9	80/10/1 0	32	20	0.96	0.94	0.95	0.94	0.96	0.95	96	94	95	Basic
		32	20	0.86	0.97	0.91	0.97	0.87	0.92	86	97	91.5	Finetune
		32	30	0.97	0.97	0.97	0.97	0.97	0.97	90	97	93.5	Basic
		32	30	0.9	0.97	0.93	0.97	0.91	0.94	90	97	93.5	Finetune
		64	20	0.97	0.93	0.95	0.93	0.97	0.95	97	93	95	Basic
		64	20	0.91	0.94	0.93	0.94	0.92	0.93	91	94	92.5	Finetune
		64	30	0.95	0.98	0.97	0.98	0.95	0.97	95	98	96.5	Basic
		64	22/30	0.95	0.88	0.91	0.87	0.95	0.91	95	87	91	Finetune

As in the preceding models, a table was constructed based on the model output, as presented in *Table 12*.

TABLE 12 PERFORMANCE METRICS FOR RESNET50

Model	Split ratio	batch	epoch	Recall for Female	Precision for Female	F1-Score for Female	Recall for Male	Precision for Male	F1-Score for Male	Female Prediction accuracy (%)	Male Prediction accuracy (%)	General Prediction accuracy (%)
ResNet50	60/20/20	16	10	0.99	0.54	0.7	0.15	0.96	0.25	99	15	57
		16	20	0.96	0.58	0.73	0.31	0.89	0.46	96	31	63.5
		16	30	0.9	0.68	0.77	0.57	0.85	0.68	90	57	73.5
		32	10	0.97	0.55	0.7	0.2	0.86	0.32	97	20	58.5
		32	20	0.98	0.57	0.72	0.25	0.93	0.4	98	25	61.5
		32	30	0.95	0.59	0.73	0.35	0.88	0.5	95	35	65
		64	10	0.97	0.57	0.72	0.27	0.89	0.41	97	27	62
		64	20	0.99	0.54	0.7	0.15	0.96	0.26	99	15	57
		64	30	0.41	0.87	0.55	0.94	0.61	0.74	41	94	67.5
ResNet50	70/15/15	16	10	0.01	1	0.01	1	0.5	0.67	1	100	50.5
		16	20	0.31	0.87	0.46	0.95	0.58	0.72	31	95	63
		16	30	0.92	0.65	0.76	0.5	0.86	0.63	92	50	71
		32	10	0.99	0.53	0.69	0.13	0.95	0.23	99	13	56
		32	20	0.99	0.53	0.69	0.13	0.95	0.22	99	13	56
		32	30	0.99	0.56	0.71	0.21	0.97	0.35	99	21	60
		64	10	0.29	0.77	0.42	0.91	0.56	0.7	29	91	60
		64	20	0.63	0.69	0.66	0.72	0.66	0.69	63	90	76.5
		64	30	0.41	0.87	0.56	0.94	0.62	0.74	41	91	66
ResNet50	80/10/10	16	10	0.95	0.58	0.72	0.3	0.87	0.45	95	30	62.5
		16	20	0.06	0.9	0.11	0.99	0.51	0.68	6	99	52.5
		16	30	0.95	0.67	0.79	0.53	0.92	0.67	95	53	74
		32	10	0.85	0.64	0.73	0.51	0.78	0.62	85	51	68
		32	20	0.03	1	0.05	1	0.51	0.67	3	100	51.5
		32	30	0.95	0.61	0.75	0.4	0.9	0.55	95	40	67.5
		32	50	0.62	0.89	0.73	0.93	0.71	0.8	62	93	77.5
		64	10	0.71	0.59	0.65	0.51	0.64	0.57	71	51	61
		64	20	0.23	0.83	0.36	0.95	0.55	0.7	23	95	59
		64	30	0.76	0.72	0.74	0.71	0.75	0.73	76	71	73.5
		64	50	0.96	0.64	0.77	0.46	0.92	0.61	96	46	71

This model represents an expanded version of ResNet101, demonstrating that larger models exhibit suboptimal performance on small datasets even when employing transfer learning techniques. This observation was substantiated by the data presented in Table 13.

TABLE 13 PERFORMANCE METRICS FOR RESNET101

Model	Split ratio	batch	epoch	Recall for Female	Precision for Female	F1-Score for Female	Recall for Male	Precision for Male	F1-Score for Male	Female Prediction accuracy (%)	Male Prediction accuracy (%)	General Prediction accuracy (%)
ResNet101	60/20/20	16	10	1	0.52	0.68	0.07	1	0.14	100	7	53.5
		16	20	0.45	0.94	0.61	0.97	0.64	0.77	45	97	71
		16	30	0.21	0.97	0.35	0.99	0.56	0.71	21	99	60
		32	10	0.86	0.62	0.72	0.47	0.77	0.58	86	47	66.5
		32	20	0.57	0.88	0.69	0.92	0.68	0.78	57	92	74.5
		32	30	0.87	0.74	0.8	0.69	0.84	0.76	87	69	78
		64	10	0.12	0.72	0.21	0.95	0.52	0.67	12	95	53.5
		64	20	1	0.52	0.68	0.07	1	0.14	100	7	53.5
		64	30	0.71	0.84	0.77	0.87	0.75	0.8	71	87	79
ResNet101	70/15/15	16	10	0.13	0.95	0.22	0.99	0.53	0.53	13	99	56
		16	20	0.99	0.6	0.75	0.34	0.96	0.5	99	34	66.5
		16	30	0.88	0.78	0.82	0.75	0.86	0.8	88	75	81.5
		32	10	0.97	0.55	0.7	0.21	0.86	0.34	97	21	59
		32	20	0.57	0.8	0.67	0.86	0.67	0.75	57	86	71.5
		32	30	0.71	0.82	0.76	0.84	0.75	0.79	71	84	77.5
		64	10	0.03	1	0.05	1	0.51	0.67	3	100	51.5
		64	20	0.91	0.63	0.74	0.45	0.84	0.59	91	45	68
		64	30	0.63	0.79	0.7	0.83	0.69	0.75	63	83	73
ResNet101	80/10/10	16	10	0.97	0.58	0.73	0.31	0.9	0.47	97	31	64
		16	20	0.69	0.87	0.77	0.9	0.74	0.81	69	90	79.5
		16	30	1	0.56	0.72	0.23	1	0.37	100	23	61.5
		32	10	0.99	0.54	0.7	0.14	0.95	0.24	99	14	56.5
		32	20	0.97	0.61	0.75	0.37	0.93	0.53	97	37	67
		32	30	0.83	0.77	0.8	0.75	0.81	0.78	83	75	79
		64	10	0.16	0.8	0.27	0.96	0.53	0.69	16	96	56
		64	20	0.97	0.55	0.7	0.2	0.88	0.33	97	20	58.5
		64	30	0.55	0.88	0.67	0.93	0.67	0.78	55	93	74

This study assessed the effectiveness of multiple models, including VGG and ResNet, with varying structures and capacities for classifying gender, using a dataset of 300 celebrity images. The examination of these models showed their adaptability and efficiency under limited data conditions with different configurations, as well as the correlation between model complexity and dataset size. The results revealed that more compact models, such as VGG16 and VGG19, generally achieved higher accuracy and consistency in the test scenarios than their larger counterparts. This less complex architecture proved to be effective, given the constrained data, showing minimal overfitting while still extracting valuable information. In contrast, the more extensive models, ResNet50 and ResNet101, tended to overfit, despite thorough parameter adjustments and increased computational resources, leading to suboptimal results.

V. Conclusion & Recommendation

Conclusion

In conclusion, the model performance depends heavily on data quality and preprocessing. Effective data handling and preparation guarantee consistency and reliability, which significantly *affects* the performance of models in practical applications. Model selection should be informed by the size of the dataset; small models are typically best for small datasets, whereas large models require extensive fine-tuning to achieve the best performance. This research accentuates the advantage of transfer learning, demonstrating that pre-trained models can be fine-tuned to gender-classification tasks to enhance the overall performance. Simpler models, that is, VGG16 and VGG19, have been effective with smaller datasets, recording up to 98% accuracy rates, whereas deeper models, that is, ResNet50 and ResNet101, require additional fine-tuning. Learning rate adjustment is important when fine-tuning small datasets, and rates that are too low can result in under-fitting. Ultimately, although male predictions performed better, the challenges experienced in female predictions highlight the necessity of balanced data and model tuning for fairness and enhanced accuracy for both genders.

Recommendations

To enhance the performance of the gender recognition model, targeted improvements should focus on addressing dataset limitations and model optimization specific to Gender identification, while additional advanced data augmentation techniques, including facial occlusion, lighting variations, and aging transformations, will enhance the robustness of the model, especially in low-represented gender classes. Data diversity enrichment with well-balanced male, female, and non-binary samples will assist in reducing bias and generalization. A dynamic dataset-complexity-based architecture selection approach can offer an optimum architecture with a growing data size. Layer freezing and selective fine-tuning of transfer learning must be attempted to maintain salient gender-distinguishing features, while reducing overfitting. Exploring domain-specific pretraining, that is, models pre-trained on facial recognition or human attribute datasets, can yield better Gender identification performance. Periodic bias tests such as gender misclassification for various age ranges and ethnicities can promote fairness. Supplementing with global features, such as facial structure and expression variations, will enable the systems to be robust. Supplementing datasets with balanced demographic representations from datasets, such as UTKFace or FairFace, will make the models fairer. Finally, utilizing high-performance computing facilities, such as cloud-based GPUs or TPUs, will enable training on large Gender identification datasets effectively.

VI. REFERENCES

[1] V. Mishra and L. Kumar, “A survey of designing convolutional neural networks using evolutionary algorithms,” *Artif Intell Rev*, vol. 56, no. 6, pp. 5095–5132, 2023.

[2] P. Smith and C. Chen, “Transfer learning with deep CNNs for gender,” in *Proceedings of the IEEE International Conference on Big Data (Big Data)*, 2018.

[3] C. T. Nguyen and others, “Transfer learning for wireless networks: A comprehensive survey,” in *Proceedings of the IEEE*, 2022, pp. 1073–1115.

[4] M. M. Islam, N. T., and J.-H. Bae, “Human Gender Identification using transfer learning via Pareto frontier CNN networks,” *Inventions*, vol. 5, no. 2411–5134, pp. 2–4, 2023.

[5] N. Le, V. S. R. Kanjarla, and Y. K. Liu, “Deep reinforcement learning in computer vision: A comprehensive survey,” *Artif Intell Rev*, vol. 55, pp. 2733–2819, 2021.

[6] M. A. Morid, A. Ben Abdessalem, and G. D. Foran, “A scoping review of transfer learning research on medical image analysis using ImageNet,” *Computational Biology and Medicine*, vol. 128, 2021.

[7] M.-H. Guo and others, “Attention mechanisms in computer vision: A survey,” *Computational Vision and Media*, vol. 8, pp. 331–368, 2022.

[8] B. Hassan, M. Fiaz, H. H. R. Sherazi, and U. J. Butt, “Annotated Pedestrians: A Dataset for Soft Biometrics Estimation for Varying Distances,” *IEEE J Sel Top Signal Process*, vol. 17, no. 3, pp. 699–707, 2023, doi: 10.1109/JSTSP.2023.3234494.

[9] R. F. Ali, A. Shehzadi, H. Jahankhani, and B. Hassan, *Emerging trends in cloud computing paradigm: An extensive literature review on cloud security, service models, and practical suggestions*. 2024.

[10] A. Pinto and others, “Transfer learning for gender recognition,” *Journal of Artificial Intelligence Research*, vol. 58, pp. 1–20, 2024.

[11] A. K. Jain and U. Park, “Facial marks: Soft biometric for face recognition,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, 2009, pp. 37–40.

[12] A. Othman and A. Ross, “Privacy of facial soft biometrics: Suppressing gender but retaining identity,” in *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, Zurich, Switzerland, 2014, pp. 682–696.

[13] A. K. Jain, S. C. Dass, and K. Nandakumar, “Soft biometric traits for personal recognition systems,” in *Proceedings of the International Conference on Biometric Authentication*, Hong Kong, China, 2004, pp. 731–738.

[14] A. Dantcheva, J.-L. Dugelay, and P. Elia, “Soft biometrics: A survey,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 1, pp. 49–64, 2016.

[15] H. Ajmal and others, “Convolutional neural network-based image segmentation: A review,” in *Proceedings of Pattern Recognition Track XXIX*, 2018, pp. 191–203.

[16] Q. Liu and others, “A review of image recognition with deep convolutional neural networks,” in *Proceedings of the International Conference on Intelligent Computing (ICIC)*, Liverpool, UK, 2017, pp. 69–80.

[17] S. Das, A. Kruti, R. Devkota, and R. Bin Sulaiman, “Evaluation of Machine Learning Models for Credit Card Fraud Detection: A Comparative Analysis of Algorithmic Performance and their efficacy.,” *FMDB Transactions on Sustainable Technoprise Letters*, vol. 1, no. 2, pp. 70–81, 2023.

[18] S. R. Das, A. Salih, R. Bin Sulaiman, and M. Farhan, “Enhancing Lung Cancer Classification with MobileNetV3 and EfficientNetB7: A Transfer Learning Approach,” in *2024 International Conference on Computer and Applications (ICCA)*, 2024, pp. 1–8. doi: 10.1109/ICCA62237.2024.10927970.

[19] A. Dantcheva, C. Velardo, A. D’angelo, and J.-L. Dugelay, “Bag of soft biometrics for person identification,” *Multimed Tools Appl*, vol. 51, no. 2, pp. 739–777, 2011.

[20] M. Awais and others, “Foundational models defining a new era in vision: A survey and outlook,” *IEEE Trans Pattern Anal Mach Intell*, vol. PP, no. 99, pp. 1–20, Jan. 2025.

[21] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, “Facial soft biometrics for recognition in the wild: recent works, annotation, and COTS evaluation,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, 2018.