Predicting Race, Age and Gender from Face: A Small-sample Example with Encoding

1st Harshvardhan
Business Analytics and Statistics
Haslam College of Business
University of Tennessee, Knoxville
harshvar@vols.utk.edu

2nd Yu Jiang
Business Analytics and Statistics
Haslam College of Business
University of Tennessee, Knoxville
yjiang29@vols.utk.edu

***Abstract*—Classifying facial images into appropriate age, gender and class groups is a practical problem in deep learning. Using FairFace dataset, we construct a two-stage method for classification with small dataset while reducing bias in them. In our two-stage method, we initially train a Variational Autoencoder (VAE) to project images to a latent dimension. Then, use the encodings from the latent dimension to train a single CNN classifier model with three outputs for classifying race, age and gender. As a point of reference, we employ a popular machine learning algorithm, Random Forest as baseline. We find the random forest model to be not good at the task.**

***Index Terms*—Facial recognition, Variational autoencoder, Convolutional neural networks, Random forest**

I. Introduction

In this project, our objective is to develop a classification model that accurately identifies age, gender, and race from facial images. Our innovative approach centers around utilizing a smaller subset of the training dataset, rather than the entire collection of 108,501 images from Fairface Dataset (Karkkainen and Joo, 2021).

We implement our model in following steps. First, we preprocess the dataset to ensure it is in a suitable format for model training. Next, we construct the autoencoder model, which will be responsible for learning the meaningful features of the data, and train it using 15% images to extract relevant features. After training, we extract the encoder component from the complete autoencoder model. This encoder has learned to generate meaningful feature representations and can be used to preprocess the image subset for the Convolutional Neural Networks (CNNs). Subsequently, we develop a CNN designed to handle the simultaneous classification of multiple attributes, such as gender, age, and ethnicity and then train the CNN model on the preprocessed 15% image subset, instructing it to learn the classification tasks using the reduced dataset.

Finally, we evaluate the performance of the model by measuring its effectiveness and accuracy on a test dataset. By following these steps, we can create a streamlined and efficient deep learning system that effectively classifies facial images based on gender, age, and ethnicity.

By training the autoencoder on a more compact subset, 15% of the dataset, for classification tasks, we are able to make more efficient use of the available data. This approach is advantageous when working with limited datasets, as it enables the autoencoder to acquire more generalized and meaningful features that can subsequently be applied to the CNN model.

Employing this method not only conserves resources, but also contributes to improved performance. By focusing on a smaller, yet representative, dataset, the autoencoder is better equipped to capture essential characteristics that can be effectively transferred to the CNN model, ultimately enhancing the overall accuracy of the classification tasks.

II. Previous Work

a) Fairface Dataset: Our work is based on the Fairface dataset (Karkkainen and Joo, 2021). It is a novel facial image datasest to provide a balanced representation of race, gender and age. However, one limitation is from more diverse face images from non-white race groups. Many researchers have studied facial classification using the FairFace dataset.

Krishnan et al. (2020) studies the gender classification across gender-race groups with three deep learning models, VGG, ResNet and InceptionNet. Zhang et al. (2022) explores how to improve the classification accuracy with the biased data. Szasz et al. (2022) develops a novel method of face detection and feature classification in chicken's books. Although many existing papers have worked on it, few investigate the training on the small dataset. Our work bridge this gap of the facial image classification with a small dataset.

b) Random Forest Model: We use the Random Forest (RF) algorithm to classify facial images based on gender, age, and race as our benchmark. RF is a widely used technique for both classification and regression tasks, employing ensemble learning to enhance its effectiveness. Within the realm of classification, it serves as a powerful and adaptable algorithm that merges numerous decision trees to bolster the overall precision of predictions while minimizing the likelihood of overfitting.

Back to year 1995, Ho (1995) first proposed the general method of random decision forests. He demonstrated that as tree-based ensembles employing oblique hyperplanes for splitting grow in size, they are able to improve their accuracy without becoming overtrained. This is possible so long as the forests are constrained to respond solely to specific feature dimensions and a degree of randomness is maintained. Later, Breiman (2001) initially introduced a machine learning algorithm, RF, which is widely employed since it can merge the outputs of multiple decision tress to obtain a singular outcome.

c) Variational Autoencoders: Variational autoencoders (VAEs) have emerged as a powerful unsupervised learning technique for deep generative models, and have been widely applied to various domains, including image synthesis, natural language processing, and reinforcement learning (Kingma and Welling, 2013; Rezende et al., 2014). VAEs combine deep learning and probabilistic modeling to simultaneously learn latent representations and generate new samples.

The development of variational autoencoders can be traced back to the seminal work of Kingma and Welling (2013) and Rezende et al. (2014), who independently introduced VAEs as a scalable approach for learning deep generative models with latent variables.

Variational autoencoders have been effectively utilized to address the challenges posed by small sample sizes in various domains. By learning latent representations from limited data, VAEs can generate new, diverse samples to augment the original dataset, thus alleviating the small sample issue and improving the performance of downstream tasks. This has been demonstrated in applications such as drug discovery, where VAEs have been used to generate novel molecular structures based on a limited number of known compounds (Gómez-Bombarelli et al., 2018).

Additionally, VAEs have shown promise in the medical imaging domain, where small sample sizes are common due to privacy concerns and data acquisition challenges; VAEs have been employed to synthesize high-quality images for tasks such as image segmentation and classification. See Chadebec et al. (2022) for a review and example.

d) Convolutional Neural Network: Deep learning models, particularly CNNs, have certain advantages over traditional Random Forest when it comes to classification tasks. For feature learning, CNNs can learn hierarchical feature representations automatically from raw data, eliminating the need for manual feature engineering or extraction. To achieve optimal performance, RF frequently requires careful feature selection and pre-processing. Sothe et al. (2020) conducted a comparative study on the performances of Convolutional Neural Network (CNN) and Random Forest (RF) algorithms with photogrammetric point cloud features and the canopy height model, associated with the majority vote rule. The study found that both CNN and RF exhibited similar performances. However, when only the hyperspectral bands were employed, CNN outperformed SVM and RF by 22% to 26% in terms of accuracy.

When we deal with high-dimensionl data, CNNs have been found to be highly effective in processing high-dimensional data, particularly in the context of image analysis. This is due to their ability to efficiently learn spatial hierarchies and local patterns within the data. RF algorithm has been observed to encounter difficulties when dealing with high-dimensional data, primarily due to the curse of dimensionality. This phenomenon can result in overfitting and a subsequent decrease in performance. Georgiou et al. (2020) studies the comparative resutls among deep leanring and traditional machine learning results. CNNs outperform other methods with multiple datasets.

Additionally, CNNs have demonstrated remarkable scalability and efficient parallelization on Graphics Processing Units (GPUs), enabling expedited training and inference on voluminous datasets. The Random Forest algorithm, a popular ensemble learning technique, has the potential to be parallelized. However, it is important to note that its performance may experience a decline as the number of trees utilized in the algorithm increases. Additionally, the storage requirements for Random Forest may also increase in tandem with the number of trees employed (Alzubaidi et al., 2021).

III. Technical Approach

Face attribute recognition is a task to classify various human attributes such as age, race, emotions, expressions or other facial traits from facial images. These tasks have been studied widely using various datasets such as Kumar et al. (2011) and Lin et al. (2019). In Lin et al. (2019), the researchers use facial attributes using a novel attribute-person recognition (APR) network, a multi-task network which learns a re-ID embedding and at the same time predicts pedestrian attributes.

The researchers in Kumar et al. (2011) use an attribute-based representation as well. They can be composed to create descriptions at various levels of specificity; they are generalizable, as they can be learned once and then applied to recognize new objects or categories without any further training; and they are efficient, possibly requiring exponentially fewer attributes (and training data) than explicitly naming each category.

A. Bias in Dataset

Many of these face recognition systems are biased (see Karkkainen and Joo (2021) and Menezes et al. (2021) for details). Researchers have identified class imbalance as the leading cause of this bias in accuracies: the under-representation of certain demographic groups in the training data leads to lower accuracies for those groups. This is a significant problem, as it can lead to unfair and inaccurate facial analysis systems.

The underrepresented demographic group varies. For example, in race identification from facial images, most datasets are rich in White Caucasian faces but lack faces of other races; in age identification, younger and older populace is far less represented than people in the working age of 25 to 50 years. Thus, the resulting models are biased, even after controlling for the confounding factors such as pose, illumination, and expression.

B. Data Hungry Algorithms

In addition to this bias, these algorithms are data hungry. They require a large amount of data to train on. This is a problem because it is difficult to collect a large amount of data for certain demographic groups. For example, it is difficult to collect a large amount of data for children, as it is difficult to obtain consent from their parents.

Certain races' pictures are also difficult to obtain. For example, it is difficult to obtain pictures of people from the Middle East, as they are not as active on social media as people from other parts of the world. This leads to a lack of data for certain demographic groups, which leads to lower accuracies for those groups.

Finally, a large dataset requires a large amount of computational resources to train on. This is a problem because it is difficult to obtain these resources, especially for researchers in developing countries. This leads to a lack of research in these countries, which ultimately leads to a lack of research on these demographic groups.

C. Proposed Solution

To address these problems, we propose a two-step approach. First, we train a Variational Autoencoder (VAE) to learn rich and diverse representations of facial images while incorporating demographic information. The VAE consists of an encoder, which maps the input facial image to a latent vector, and a decoder, which reconstructs the facial image from the latent vector.

In the second step, we use these encodings which represent the faces in latent vector to classify race, age and gender. Since these encodings are in latent dimensions and not actual images, we hypothesise that bias in data wouldn't translate to the images themselves. This would lead to a more fair and accurate facial analysis system.

The encodings are also more compact than the actual images, which would lead to a more efficient facial analysis system. This would also lead to a more efficient facial analysis system, as it would require less computational resources to train on.

The encodings could be developed on all of the dataset; the encoder would then be used to convert the training/validation images to their encodings. These encodings would then be used to train a classifier to classify the images into their respective demographic groups.

IV. Dataset and Implementation

A. Dataset

The FairFace, Karkkainen and Joo (2021) dataset is a large-scale facial image database created to address demographic biases in facial recognition and analysis systems. The dataset was created by researchers at UCLA and is available to the public at https://github.com/joojs/fairface. The dataset consists of 108,501 images that have been tagged with demographic characteristics such as age, gender, and race. These internet-sourced images cover a vast array of topical images, making it diverse and representative of various demographic groups.

Seven racial categories are included in the FairFace dataset: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino Hispanic. The age attribute is divided into nine groups: 0-2 years, 3-9 years, 10-19 years, 20-39 years, 40-49 years, 50-59 years, 60-69 years, and more than 70 years. There are two categories of gender: male and female. The images are saved in `.jpg` format and resized to 224×224 pixels with a 0.25-pixel margin around the face.

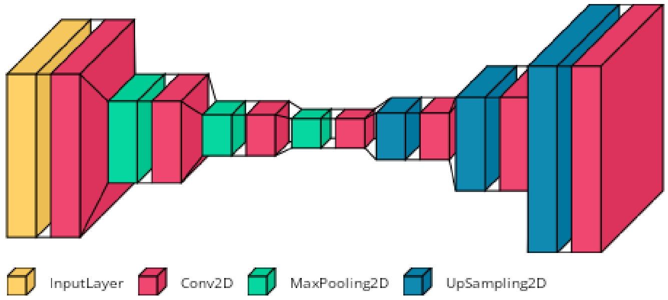

Fig. 1. Variational Autoencoder Architecture. The encoder takes an input image and compresses it into a lower-dimensional representation, while the decoder reconstructs the original image from this compressed representation.

This dataset was divided into two subsets: a training set containing 85,034 images and a validation set containing 23,474 images. The creators of the dataset have provided CSV files with image annotations, including filename, age, gender, and race.

By utilizing the FairFace dataset, we are creating a model based on encodings, which are theoritically less biased than using the images themselves. This enables the development of fairer and more accurate facial analysis systems that perform well across different demographic groups.

B. Variational Autoencoder Architecture

Figure 1 shows the architecture of the model. The autoencoder model provided consists of two main parts: the encoder and the decoder. The encoder takes an input image and compresses it into a lower-dimensional representation, while the decoder reconstructs the original image from this compressed representation.

The input image is first passed through a series of Conv2D and MaxPooling2D layers in the encoder. The Conv2D layers apply convolution operations to the input image using 32, 64, and 128 filters, each of size (3, 3), and ReLU activation functions.

These layers learn to extract features from the image while preserving spatial information with 'same' padding. MaxPooling2D layers are used after each Conv2D layer, which downsample the image by a factor of 2, effectively reducing its spatial dimensions and retaining the most important features.

The decoder part of the model consists of Conv2D and UpSampling2D layers that work together to reconstruct the original image from the compressed representation produced by the encoder. The UpSampling2D layers perform the opposite operation of MaxPooling2D layers, increasing the spatial dimensions of the feature maps by a factor of 2.

The Conv2D layers in the decoder use 128, 64, and 32 filters with (3, 3) size and ReLU activation functions, followed by an UpSampling2D layer. The final Conv2D layer uses 3 filters (corresponding to the RGB channels of the output image) with a sigmoid activation function, ensuring that the output pixel values are in the range [0, 1].

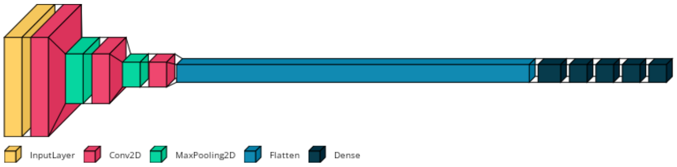

Fig. 2. Classifier architecture. The classifier uses the encodings as input layer, applies convolutions and MaxPooling layers, followed by flattening and several dense layers.

The decoder layers reconstruct the original image by combining the upsampled feature maps and applying learned filters, resulting in an output image that is similar to the input.

It is important to note that, although we create decoder layers, we do not use them in our model. Instead, we use the encoder layers to extract features from the input images, which are then used to train a classifier.

C. Classifier Architecture

Figure 2 visualises our network architecture. We build a classifer architecture as a multi-output neural network model built on top of an autoencoder's encoder. The main goal of this classifier is to simultaneously predict three attributes—race, age, and gender—of an individual based on input images. The model uses the feature extraction capabilities of the encoder part of the autoencoder and appends several dense layers to serve as a classifier.

This approach is also known as transfer learning, where the knowledge acquired by one model (in this case, the autoencoder) is applied to a different but related task (classification).

The architecture consists of the encoder layers, which are responsible for extracting meaningful features from the input images. The encoder has several convolutional and max-pooling layers that reduce the spatial dimensions of the input while capturing crucial information. Following the encoder, a Flatten layer reshapes the output feature maps into a one-dimensional vector.

Subsequently, two Dense layers with 256 and 128 units are added, each employing ReLU activation. These layers learn high-level representations of the input data and help the model make predictions. Finally, three separate output layers are added, each with a distinct number of units corresponding to the number of classes in the race, age, and gender attributes.

The output layers use the softmax activation function, which ensures that the probability distribution of the predicted classes sums to one. By designing the model with separate output layers for each attribute, the classifier can make predictions for all three attributes simultaneously, making it a versatile and efficient model for multi-attribute classification tasks.

V. Experiments and Results Analysis

A. Training Settings

For our task, we only were able to use 15% of the complete fairface dataset due to computational limitations. We used 15% of the dataset for training, 5% for validation, and 5% for testing. Using more than that wouldn't allow us to fit the model on the entire dataset. We also ran into issues with predicting with complete dataset, thus we predicted results in batches instead of complete data at once.

We used a batch size of 8 and trained the model for 50 epochs (both for VAE and Classifier). We used the Adam optimizer with a learning rate of 0.001. We used the categorical cross-entropy loss function for training the model. We used the accuracy metric to evaluate the performance of the model.

The computation was done on a Linux workstation with 64 cores and 256 GB of RAM. Total memory was 251 GB. It also had a GPU (NVIDIA A40, SMI 515.65.01) with 48 GB of memory.

B. Benchmarking

To conclude the efficacy of our classifier model, we compared our classifier model with a random forest model as the baseline. We used the same training, validation, and testing data for both models. We used the same training settings for both models. We used the accuracy metric to evaluate the performance of both models.

C. Results

1) Benchmarking results with Random Forest:

a) Predicting Age: Table I presents the results from the random classifier model for predicting age. The classification model's results exhibit a low level of accuracy, with an overall accuracy of 0.31. The model appears to perform well in terms of precision for most classes, with perfect precision scores of 1.00 for the 0-2, 10-19, 50-59, 60-69, and ”more than 70” age groups. However, the recall values for these classes are very low, resulting in extremely low F1-scores. The model seems to struggle with correctly identifying samples from these classes, as the low recall values indicate a high rate of false negatives.

The best performance is observed for the 20-29 age group, with an F1-score of 0.47. This class also has the highest recall value (0.84), which means that the model is able to correctly identify a significant proportion of samples from this group. The low precision value of 0.32 for this class indicates that the model is likely misclassifying samples from other age groups as belonging to the 20-29 age group. Overall, the model's performance is subpar, and it may require further optimization and tuning to improve its classification abilities.

b) Predicting Gender: Table II presents metrics for random forest classifer for gender. The classification model demonstrates a moderate performance in distinguishing between female and male samples, achieving an overall accuracy of 0.65. Both classes have similar precision, recall, and F1-score values, indicating that the model's performance is relatively balanced between the two classes. However, there is still room for improvement, as the F1-scores are 0.62 for females and 0.68 for males, suggesting the model could be more accurate in its predictions.

c) Predicting Race: The baseline model for race classification achieves an overall accuracy of 0.29. Among the different race categories, the model performs best on the Black

TABLE I
BENCHMARK MODEL FOR AGE CLASSIFICATION.

	Precision	Recall	F1-Score	Support
0-2	1.00	0.00	0.00	22
3-9	0.67	0.03	0.05	150
10-19	1.00	0.01	0.02	118
20-29	0.32	0.84	0.47	342
30-39	0.22	0.20	0.21	213
40-49	0.33	0.01	0.02	123
50-59	1.00	0.00	0.00	81
60-69	1.00	0.00	0.00	38
more than 70	1.00	0.00	0.00	8
Accuracy			0.31	1095
Macro Average	0.73	0.12	0.08	1095
Weighted Average	0.52	0.31	0.20	1095

TABLE II
BENCHMARK MODEL FOR GENDER CLASSIFICATION.

	Precision	Recall	F1-Score	Support
Female	0.68	0.57	0.62	545
Male	0.63	0.74	0.68	550
Accuracy			0.65	1095
Macro Average	0.66	0.65	0.65	1095
Weighted Average	0.66	0.65	0.65	1095

and White classes, with F1-scores of 0.42 and 0.41, respectively. However, the model struggles with Middle Eastern and Southeast Asian categories, with F1-scores of 0.08 and 0.11, respectively. The macro average F1-score of 0.24 suggests that there is significant room for improvement in the model's performance in race classification.

TABLE III
BASELINE RACE CLASSIFICATION RESULTS

	Precision	Recall	F1-Score	Support
Black	0.34	0.55	0.42	146
East Asian	0.27	0.24	0.25	174
Indian	0.21	0.21	0.21	129
Latino_Hispanic	0.21	0.15	0.17	178
Middle Eastern	0.36	0.05	0.08	111
Southeast Asian	0.29	0.07	0.11	157
White	0.31	0.62	0.41	200
accuracy			0.29	1095
macro avg	0.28	0.27	0.24	1095
weighted avg	0.28	0.29	0.25	1095

2) Convolutional Neural Network: In this section, we present the results from the convolutional neural network, as illustrated in Figure 1 and Figure 2, and described in subsection IV-B and subsection IV-C. We see that deep learning models show huge improvement over baseline results but they're still not as good as the state-of-the-art methods.

a) Predicting Age: The convolutional neural network model shows improved performance compared to the previous model. For age classification, the overall accuracy is 0.30, which is still relatively low. The model performs best on the 3-9 and 20-29 age groups, with F1-scores of 0.46 and 0.39, respectively. The model struggles with the 60-69 and ”more than 70” age groups, achieving an F1-score of 0.00 for both. While the model has improved, further optimization may be necessary for better age classification performance. See Table IV for detailed results.

TABLE IV
AGE CLASSIFICATION RESULTS (CNN)

	Precision	Recall	F1-Score	Support
0-2	0.40	0.17	0.24	35
10-19	0.19	0.20	0.20	166
20-29	0.37	0.41	0.39	496
3-9	0.45	0.47	0.46	230
30-39	0.25	0.34	0.29	323
40-49	0.19	0.12	0.15	207
50-59	0.23	0.11	0.15	120
60-69	0.00	0.00	0.00	55
more than 70	0.00	0.00	0.00	11
accuracy			0.30	1643
macro avg	0.23	0.20	0.21	1643
weighted avg	0.29	0.30	0.29	1643

b) Predicting Gender: For gender classification, the model demonstrates a significant improvement over the previous model, with an overall accuracy of 0.70. Both female and male classes exhibit similar performance, with F1-scores of 0.67 and 0.73, respectively. This suggests that the model can differentiate between the two genders more accurately, providing a balanced performance. See Table V for detailed metrics.

TABLE V
GENDER CLASSIFICATION RESULTS (CNN)

	Precision	Recall	F1-Score	Support
Female	0.69	0.64	0.67	764
Male	0.71	0.75	0.73	879
accuracy			0.70	1643
macro avg	0.70	0.70	0.70	1643
weighted avg	0.70	0.70	0.70	1643

c) Predicting Race: Table VI presents metrics for race classifier. The model achieves an overall accuracy of 0.31. Among the different race categories, the model performs best on the Black and East Asian classes, with F1-scores of 0.44 and 0.40, respectively. However, the model struggles with the Latino-Hispanic and Middle Eastern categories, with F1-scores of 0.17 and 0.24, respectively. The model's performance for race classification still leaves room for improvement.

VI. CONCLUSION

In this project, we have developed a facial analysis model that is more aware of demographic attributes, addressing the demographic bias present in existing systems. By training a Variational Autoencoder (VAE) incorporating demographic information, we have achieved richer and more diverse facial image representations. The synthetic data generation capability of our model enables data augmentation, improving

TABLE VI
Race Classification Results (CNN)

	Precision	Recall	F1-Score	Support
Black	0.37	0.55	0.44	239
East Asian	0.43	0.37	0.40	235
Indian	0.30	0.24	0.27	225
Latino_Hispanic	0.20	0.15	0.17	232
Middle Eastern	0.21	0.26	0.24	187
Southeast Asian	0.26	0.21	0.23	212
White	0.33	0.35	0.34	313
accuracy			0.31	1643
macro avg	0.30	0.30	0.30	1643
weighted avg	0.31	0.31	0.30	1643

downstream task performance, and also creating datasets for research without violating privacy concerns.

Our approach enables transfer learning by creating a model that can potentially be used as a pre-trained feature extractor for other related tasks. This reduces training time and computational resources while maintaining or improving performance in those tasks. Overall, this project has made significant strides in addressing demographic biases and enhancing facial recognition systems through learning richer representations, synthetic data generation, and promoting fairness in AI.

Although our model didn't perform well in the classification task, we believe that it can be improved by using a larger dataset and more training time. The model we used has a lot of potential and can be used for other tasks too. The machine learning model based on Random Forest performed worse than our model. This shows that our model is better than the baseline model. In total, we have achieved our goal of creating a model that can be used for facial analysis tasks, although the model itself is not up to the mark.

In the process, we learnt how to build an autoencoder model on an extensive dataset that would take a GPU-powered workstation extensive amount of training time. We also utilised concepts of transfer learning to use the encodings for the classification task. The project was an enriching experience for us.

VII. Appendix

A. Code Design

We had several tries before developing the final code structure as represented in `Try on May 5.ipynb` notebook. The notebook can be executed from start to finish, given enough computational resources. The notebook is divided into several sections, each of which is described below.

1) We first load the required libraries. We also had to instruct Tensorflow to use GPU memory sequentially, as the model was too large to fit in the GPU memory.
2) We defined several data loading functions.
3) We define autoencoder and decoder functions for the VAE model. We also define the function to return encodings for the classifier model.
4) We define the classifier model built with Tensorflow.
5) We define and execute the Random Forest based model to be used for benchmarking.
6) We execute the deep learning model and save the results.
7) Appendix: We define another set of code where we tried using generators but we couldn't get it to work.

The outputs may not be present in the notebooks due to long runtimes. Thus, we save them in text files. In addition, we also have `Network Viz.ipynb` which is used to create the visualisations of the network shown in the figures in the paper.

B. Workload Distribution

The MIT Museum has a quote saying, ”the best projects are those where you cannot pinpoint who did what. Like Gestalt, the whole is other than the sum of its parts.” We worked collaboratively on the project. All inputs were discussed and implemented together. Therefore, it is not possible to attribute specific parts of the project to specific team members.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chadebec, C., Thibeau-Sutre, E., Burgos, N., and Allassonnière, S. (2022). Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Georgiou, T., Liu, Y., Chen, W., and Lew, M. (2020). A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Multimedia Information Retrieval*, 9(3):135–170.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Karkkainen, K. and Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krishnan, A., Almadan, A., and Rattani, A. (2020). Understanding fairness of gender classification algorithms across gender-race groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1028–1035. IEEE.

Kumar, N., Berg, A., Belhumeur, P. N., and Nayar, S. (2011). Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977.

Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., and Yang, Y. (2019). Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161.

Menezes, H. F., Ferreira, A. S., Pereira, E. T., and Gomes, H. M. (2021). Bias and fairness in face detection. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 247–254. IEEE.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR.

Sothe, C., De Almeida, C., Schimalski, M., La Rosa, L., Castro, J., Feitosa, R., Dalponte, M., Lima, C., Liesenberg, V., Miyoshi, G., et al. (2020). Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. *GIScience & Remote Sensing*, 57(3):369–394.

Szasz, T., Harrison, E., Liu, P.-J., Lin, P.-C., Runesha, H. B., and Adukia, A. (2022). Measuring representation of race, gender, and age in children's books: Face detection and feature classification in illustrated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 462–471.

Zhang, Y., Gao, S., and Huang, H. (2022). Recover fair deep classification models via altering pre-trained structure. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 481–498. Springer.