ISSN No:-2456-2165

Development of a Convolutional Neural Network-based Ethnicity Classification Model from Facial Images

Segun Aina, Mosunmola Oluwabusola Adeniji, Aderonke Rasheedat Lawal, Adeniran Isola Oluwaranti

Abstract:- The human face is considered to be the seat of man's identity and information such as age and ethnicity are often automatically deduced from the face by people. However, deducing the same information by a computing system is not a straight forward process and have in recent years be powered by Convolutional Neural Networks (CNN). CNN can automatically extract hidden patterns in data. These hidden patterns are often complex to represent using hand-crafted representation methods. Although automated classification of demographic traits such as age, gender and ethnicity is a well-studied research problem, it is still far from being considered a solved problem for Nigerian ethnic groups. In this paper, a CNN model for ethnicity classification of Nigerians from facial images is proposed based on transfer learning techniques conducted on VGG-16 architecture. The model is evaluated on a dataset consisting of facial images of Yoruba, Hausa and Igbo ethnic groups of Nigeria. The developed VGG-16 based ethnicity classification model had an overall accuracy of 92.86%, with the precision, sensitivity and specificity shedding more light on the model's performance.

I. INTRODUCTION

The face is a prominent part of the human body that is often considered to be the seat of a person's identity (Gudi, 2014). It is home to some demographic information. In its active and passive states, it constantly conveys certain information about an individual(Over and Cook, 2018). Also, judgments about people such as age and ethnicity are often made based on their faces.

In recent years, computer vision researches have given tremendous attention to identifying human demographic traits like age, gender and ethnicity. These traits play significant roles in different applications such as surveillance, biometrics, video conferencing systems, facial reconstruction, image processing, security and telecommunication (Ng et al, 2015).

Ethnicity classification from facial images involves the extraction and processing of relevant data from images of the human face to get information about the person's ethnicity as shown in Figure 1. In recent years, there has been a global interest in automatic ethnicity classification from facial images.

Fig. 1: Examples of facial demographic estimation

Source: Han *et al*. (2013)

Conventionally, a two-stage method was used to solve classification problems. The first stage involved the use of feature extractors such as local binary pattern, Haar-wavelets and histogram of oriented gradients among others to extract handcraft-ed features from images.The extracted features served as input to a classifier at the second stage. The major drawback of this method was that the accuracy of the classification task greatly depended on the design of the feature extraction stage. In addition, the system was being told what features to look for.For majority of the extractors, performance reduces with degradation in image quality and with substantial increase in the number of classes to be classified. Deep learning algorithms have since addressed this limitation. Deep learning is a subset of machine learning which has shown momentous performance in computer vision significantly outperforming previous traditional classification techniques. Deep learning eliminates the need for feature definition and extraction by performing end-to-end learning. A type of deep learning algorithm that has achieved cutting-edge results in computer vision and particularly in image classification tasks are Convolutional Neural Networks (CNN) also known as ConvNets.

ISSN No:-2456-2165

Nigeria is a multiethnic society, like a number of other African countries, it has over three hundred different ethnic groups. The three largest ethnic groups according to geopolitical locations are Hausa/Fulani located in the Northern part of the country making up 29% of the population, followed by Yorubas located in the South West with about 21% of the population while Igbos located in the South East makes up 18% of the population (Kaplan, 2012). According to Boski (1983) the three largest major ethnic groups of Nigeria, Igbo, Hausa and Yoruba as illustrated in Figure 2. can be regarded as potential nations in terms of population, cultural heritage and social structure with considerable differences existing among them. These ethnic groups are also reported to possess defined distinctions which are possibly more sharply defined than in any other sister society (Rao *et al*., 2011).

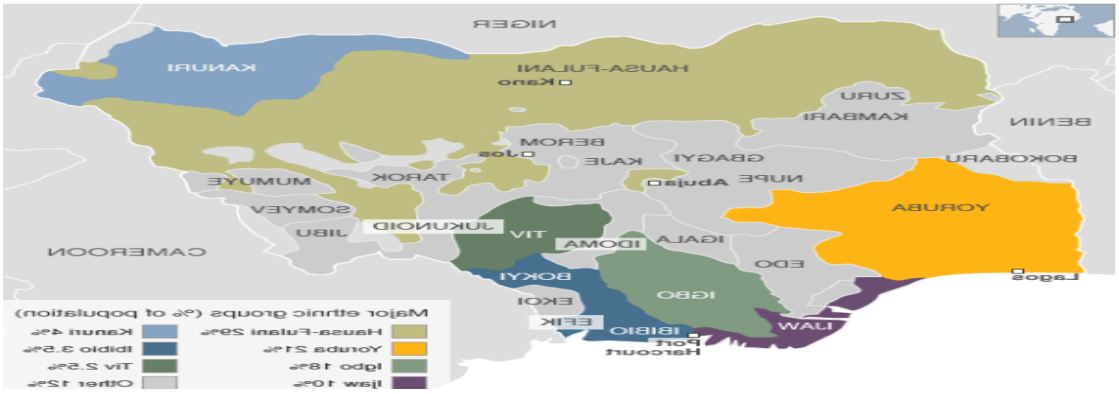

Fig. 2: Map of Nigeria showing Ethnic distributions

Source: Kaplan, (2012)

II. RELATED WORKS

Deep learning is a subset of supervised machine learning which uses several combinations of multiple layers to automatically learn efficient representation from data as opposed to being hand crafted by humans. It evolved out of the need to address challenges of traditional approaches to artificial neural networks (Goodfellow*et al*.,(2016).The use of hand-engineered features used to be the standard practice of finding patterns in computer vision tasks. It was however plagued with the limitation of needing expert knowledge in designing efficient feature descriptors. The designed feature descriptors often turned out to be task and domain specific and as such were difficult to generalize to other domains (Long *et al.*, 2012). Deep learning has attracted a lot of attention in recent years due to its remarkable performance on many applications including language translation, making medical diagnosis from x-ray, recognizing objects to help with self-driving cars and winning the best players on some games to mention a few.Other areas of applications such as in medical diagnosis is as presented by (Qing *et al.*, 2019). Early works on computer vision tasks using deep learning were typically trained on private datasets. Facebook's Deepface (Taigman *et al.*, 2014) model was trained on four million images of four thousand people; Google's FaceNet (Schroff *et al*., 2015) was trained on two hundred million images of three million people; DeepID serial models (Sun *et al*., 2014); (Sun *et al*., 2014); (Sun *et al*., 2015) were trained on two hundred thousand images of ten thousand people. Although they reported groundbreaking performance, researchers are not able to accurately reproduce or compare their models without public training datasets of equivalent magnitudes. As such, it was almost impossible for small datasets in the order of hundreds to benefit from deep learning. A technique was developed to cater to this challenge called transfer learning. Transfer learning is a machine learning technique that focuses on utilizing knowledge gained in solving one or more problems to improve learning in a different but related problem. A new model is pre-trained with a large dataset to learn features that are generalizable to new task and subsequently trained with application specific data, thereby benefitting from features similar to the pre-trained ones. In transfer learning tasks, publicly released models pre-trained on ImageNet dataset (Deng *et al*., 2009) (in most cases) are adapted to new tasks. The aim of transfer learning is to leverage knowledge gained from a source task to improve learning in a target task as a step towards making machine learning as efficient as human learning (Torrey and Shavlik, 2010).

Among earlier works on transfer learning is Ahmed *et al*. (2008), in which resultant prior knowledge from some pseudo-tasks were moved to Deep Convolutional Neural Networks (DCNN) through transfer learning resulting in improved performance of the DCNN. It was thus concluded that transfer learning could be applied to diverse computer vision tasks including object, gender and ethnicity identification.

Transfer learning using supervised features has been successful in several computer vision applications (Sharif-Razavian *et al*., 2014) including face recognition (Taigman *et al*., 2014) and visual question answering (Antol *et al*., 2015), where image features trained on ImageNet and word embeddings trained on large unsupervised corpora were combined. Figure 3 illustrates the process of transfer learning where a pre-trained network is re-trained for a similar task.

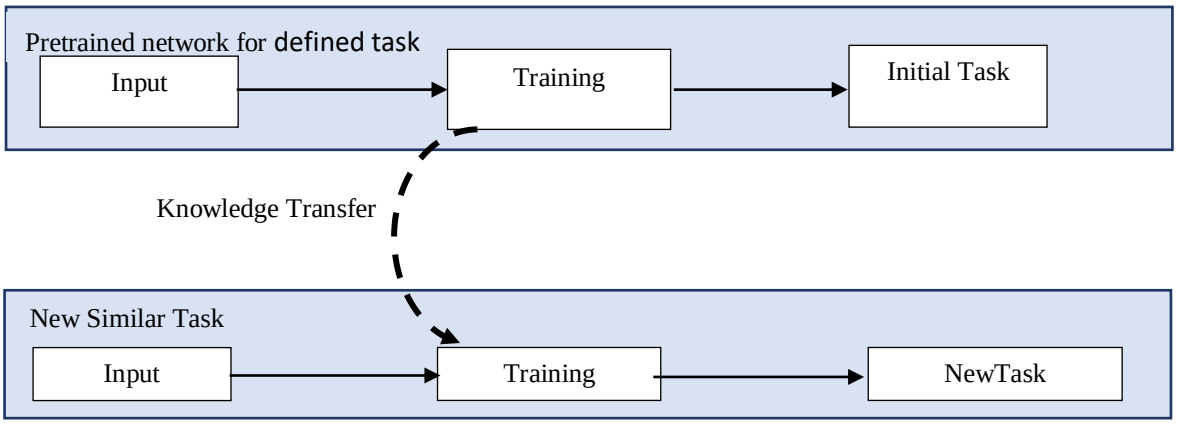

Fig. 3: Illustration of Transfer learning concept

Transfer learning in computer vision is made possible by the fact that low-level features such as edges, corners, shapes, intensity etc. is common to most tasks and can therefore be shared. For problems where there is paucity of data, transfer learning can enable the development of skillful models that would have been otherwise impossible in the absence of transfer learning. In transfer learning, the final layers of models that had been pre-trained on millions of images are retrained with new usually much smaller dataset to identify features specific to that dataset while taking advantage of the generic features already identified by the lower layers of the pre-trained model.

III. METHODOLOGY

Following the identified need for an automated ethnicity classification model in Nigeria, raw data intended for use in the experiment were preprocessed. During preprocessing, the images were scaled and normalized these processes are depicted in Figure 4. Subsequently, the VBEC model was developed by conducting transfer learning on the VGG-16 architecture. The developed model is then trained by part of the image dataset and evaluated. The result of the model evaluation is used to retrain the model in a loop. When the model has been well trained, a separate image dataset is used to evaluate the model.

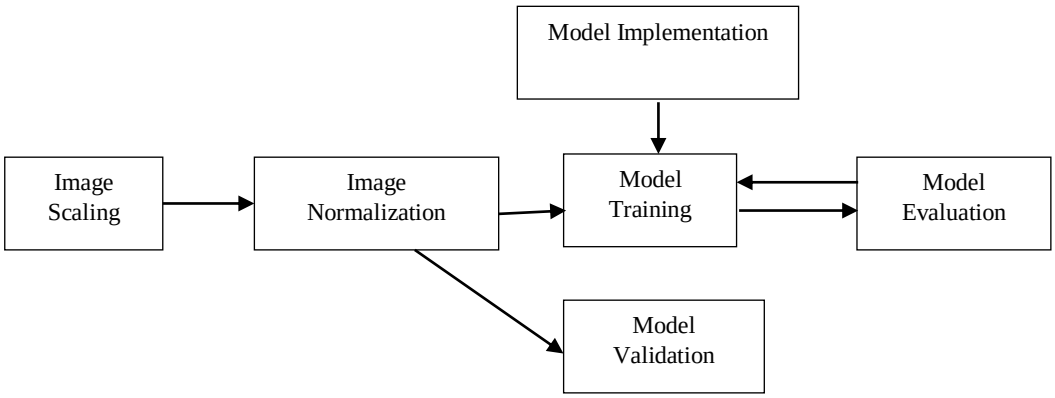

Fig. 4: Illustration of experimental process for ethnicity classification

A. *Data Preprocessing*

Collected data was preprocessed because the raw images contained a lot of redundancies as is typical of raw image data. Direct usage would have been computationally expensive as well as yielding undesirable results. The following pre-processing steps were applied to the images.

B. *Image scaling*

There is a need to scale the pixel values in an image before supplying it as an input to a deep learning neural network. In this study, image scaling was done to reduce the physical size of the image by changing the number of pixels it contains. The images were scaled to ensure uniform pixel size along the width and height. The input images for this model were rescaled to 150 × 150 pixels to reduce the complexity of training and testing. Scaling was done manually using "Paint" image processing application on Windows operating system.

ISSN No:-2456-2165

C. Image normalization

Normalization refers to regulating the data dimensions of images so that they end up within a certain range. It involves changing the range of pixel intensity values are. This important step ensures that each input parameter (pixel) has similar distribution which ensures faster convergence during network training. Normalizing image input is performed to take pixel values from the range of 0 – 255 to the preferred range of 0 – 1 for neural network models. Normalization ensures that gradients do not go out of control during network optimization. Neural networks process inputs using small weight values, and inputs with large integer values can disrupt or slow down the learning process. As such, it is good practice to normalize the pixel values so that each pixel has a value between 0 and 1. Equation 3.1was used to normalize the images in the YIH dataset.

$$x_{new} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x_{new} is the new pixel intensity value, x_i is the original pixel intensity value, x_{min} is 0 and x_{max} is 255.

D. Model Development

To develop the ethnicity classification model named VBEC Model, (acronym for VGG-16 Based Ethnicity Classification Model), VGG-16 (a CNN based architecture) was employed. To compensate for the relatively low number of images collected, application of transfer learning on the data collected after it has been preprocessed was carried out.

E. Developed VBEC model

The VBEC model was developed by conducting transfer learning on the VGG-16 architecture. In using transfer learning, the convolutional blocks of VGG-16 architecture and weights were used as the base convolutional layers of the developed VBEC model. Retaining only the convolutional blocks of VGG-16 offers the flexibility of using a different input size as opposed to the 224 x 224 of the original model. The convolution blocks were kept frozen meaning the weights at those layers were prevented from being updated during model development. The model architecture was configured as shown in Figure 5.

New layers were added onto the base layers, initialized with the VGG-16 weights and trained with the new YIH dataset. A global average pooling layer was stacked immediately after the base convolutional layers followed by three fully connected layers. Dropout was introduced in between the fully connected layers and the 1000 class softmax layer wasalso replaced with a new softmax layer with 3 neurons to suit the 3 ethnic groups to be classified. The output of the model is given by the softmax layer as a probability distribution summing up to 1.

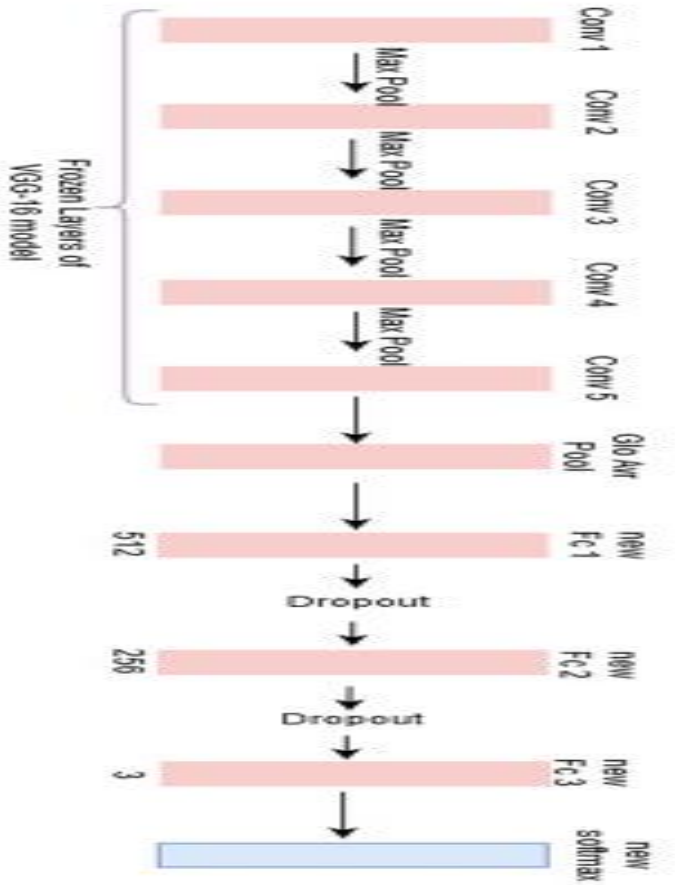

Fig. 5: The Developed VBEC Model's Architecture

Source: Field work, (2019)

ISSN No:-2456-2165

Global average pooling is an operation that calculated the average output of each feature map in the previous layer significantly reducing data in preparation for classification. It served to down sample the output of the base convolutional layers. It was introduced to reduce the total number of parameters of the model. It was introduced to reduce the total number of parameters of the model. It reduces a convolutional layer output of $h \times w \times d$ dimension to $1 \times 1 \times d$ dimension by reducing each feature map of $h \times w$ dimension to a single number by taking the average of all h and w values.

Dropout probability of 0.2 was used (as that was found out to be the optimal value) to mitigate overfitting to the training dataset. After the design of the model's architecture, the model was compiled with both the optimization algorithm and loss function specified as cross entropy loss function and Root Mean Square prop (RMSprop) optimization algorithm.

Cross entropy loss is calculated using Equation 2. Categorical cross entropy loss

$$D\ (S, L) = -\sum_i L_i \log(S_i) \qquad (2)$$

where S is the softmax output which is the model prediction and L is image label which is the correct prediction.

RMSprop operates by calculating gradient descent with an extra factor known as momentum. Momentum restricts the rate of increase of gradient descent in the vertical direction while it progresses in the horizontal direction. To optimize using RMSprop:

On each iteration t, compute the partial derivatives of weights and bias ∂w and ∂b for current batch and calculate $V_{\partial w}$ and $V_{\partial b}$ using Equation 3 and Equation 4:

$$V_{\partial w} = \beta V_{\partial w} + (1 - \beta)\partial w^2 \qquad (3)$$

$$V_{\partial b} = \beta V_{\partial b} + (1 - \beta)\partial b^2 \qquad (4)$$

Update weight and bias using Equation 5 and Equation 6:

$$W = W - \alpha \frac{\partial w}{\sqrt{V_{\partial w}}} \qquad (5)$$

$$b = b - \alpha \frac{\partial b}{\sqrt{V_{\partial b}}} \qquad (6)$$

where α is the learning rate, β is a parameter that controls the exponentially weighted average. It is a scalar value usually less than 1. It was set to 0.9, serving to restrict rate of acceleration of descent.W is weight, b is bias and$V_{\partial w} V_{\partial b}$ are momentum terms with respect to weight and bias.

During network training, the objective is to find an optimal set of weights W that produce class scores most consistent with the ground truth labels of the input training set. The loss function, also known as objective function is the function used to measure the consistency of the class scores to the ground truth. The goal during training is therefore to find a set of weights W that causes the loss to be minimized. A high loss means the computed class scores are inconsistent. A low loss value means the computed class scores are in harmony with the ground truth labels of the input training data. The resulting error is used to modify the weights of each neuron by multiplying it with the learning rate and subtracting the result from the current value. The learning rate is then adapted during training to become smaller leading to less modification of the parameters thereby achieving convergence.

F. Model validation for performance evaluation

The entire dataset is divided into training and validation or test set. Due to its size, 90% and 10% were used for training and testing respectively.

IV. RESULTS AND DISCUSSION

	Accuracy	Precision	Sensitivity	Specificity
VBEC	92.86%	94.87%	91.67%	95%
Baseline Model	89.29%	91.67%	89.68%	94.67%
ResNet-50	32.14%	10.71%	33.33%	66.67%
Inception-V3	64.29%	62.5%	59.60%	81.33%

Table 1: Summary of Performance Evaluation of All Models

The developed VBEC model was evaluated based on accuracy, precision, sensitivity and specificity and the results were reported in Section 4.4. The result of three other architectures (Baseline, Resnet-50 and Inception-V3) on the YIH dataset were also obtained as reported in Section 4.5. While the developed VBEC model had an overall accuracy of 92.86%, the precision and sensitivity shed more light on the model's performance.

From the results obtained, Table 1 shows that the ResNet-50 architecture performed worst of all achieving an accuracy of 32.14%, precision of 10.71% and sensitivity of 33.33% respectively. This is implicative of the fact that this model while being much deeper with 50 layers, is not best suited for transfer learning on image classification tasks of very small dataset as negative transfer was obtained.

This study shows a contrast to the work of Wang *et al.* (2016) which achieved a higher result of 75% accuracy with Resnet-50 than with shallower models in classifying Chinese, Koreans and Japanese. This study however corroborates the work of Masood *et al.* (2018) in that a VGG-16 based model performs better in ethnicity classification involving very small dataset. Masood *et al.* (2018) classified three very distinct ethnicities (Mongolian, Negroid and Caucasian) obtaining an accuracy of 98.6%

using 447 images evenly distributed among the three ethnic groups.

V. CONCLUSION

In conclusion, the study showed that VBEC model is best suited for ethnicity classification tasks in the presence of an extremely small and imbalanced (unevenly distributed) dataset, as was the case in this study. Furthermore, while it is generally implied that accuracy increases with depth, from this study, it can be concluded that there is a limit to the model depth that will perform well on very small image datasets as negative transfer sets in at some point.

REFERENCES

[1.] Gudi, A. (2014). Recognizing Semantic Features in Faces using Deep Learning. Retrieved August 21, 2018 from https://arxiv.org/abs/1512.00743

[2.] Over, H., and Cook, R. (2018). Where do spontaneous first impressions of faces come from?*Cognition*, 170, 190–200.

[3.] Ng, C., Tay, Y., and Goi, B. (2015). A review of facial gender recognition. *Pattern Analysis and Applications*. https://doi.org/10.1007/s10044-015-0499-6

[4.] Han, H., Otto, C., and Jain, A. K. (2013). Age estimation from face images: Human vs. machine performance. *2013 International Conference on Biometrics (ICB)*, 1–8. https://doi.org/10.1109/ICB.2013.6613022

[5.] Kaplan, S. (2012). *Nigeria's Potential for Sectarian Conflict*. [Online] Fragile States Forum. Available at: https://www.fragilestates.org/2012/01/29/nigerias-potential-for-sectarian-conflict/ [Accessed 9 Sep. 2019].

[6.] Boski, P. (1983). Egotism and evaluation in self and other attributions for achievement related outcomes. *European Journal of Social Psychology*, *13*(3), 287–304. https://doi.org/10.1002/ejsp.2420130307

[7.] Rao, D., Paul, M., Fink, C., Yarowsky, D., Oates, T., and Coppersmith, G. (2011). Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 598–601.

[8.] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[9.] Long, F., Wu, T., Movellan, J. R., Bartlett, M. S., and Littlewort, G. (2012). Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. Neurocomputing, 93, 126-132.

[10.] Taigman, Y., Yang, M., Ranzato, M. A., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).

[11.] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 815-823.

[12.] Sun, Y., Chen, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems* (pp. 1988-1996).

[13.] Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1891-1898).

[14.] Sun, Y., Liang, D., Wang, X., & Tang, X. (2015). Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*.

[15.] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

[16.] Torrey, L., and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242-264). IGI Global.

[17.] Sharif-Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 806-813).

[18.] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* 2425-2433.

[19.] Wang, Y., Liao, H., Feng, Y., Xu, X., and Luo, J. (2016). *Do They All Look the Same? Deciphering Chinese, Japanese and Koreans by Fine-Grained Deep Learning.*

[20.] Masood, S., Gupta, S., Wajid, A., Gupta, S., and Ahmed, M. (2018). Prediction of Human Ethnicity from Facial Images Using Neural Networks.In *Data Engineering and Intelligent Computing*, 217-226, Springer, Singapore.

[21.] Qing G., Yunjun W., Bo P., Duanshu L., Jiajun D., Yu Q., Hongtao L., Xiaochun W., and Jun X. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. *Journal of Cancer* vol 10(20): (pp.4876-4882)