Age, Gender Prediction and Emotion recognition using Convolutional Neural Network

Arjun Singh[1], Nishant Rai[2], Prateek Sharma[3,*], Preeti Nagrath[4] and Rachna Jain[5]

[1, 2, 3, 4, 5] *Computer Science and Engineering Department, Bharati Vidyapeeth College of Engineering, New Delhi*

[1] arjunsingh.cse1@bvp.edu.in

[2] nishantrai.cse1@bvp.edu.in

[3] prateeksharma.cse1@bvp.edu.in

[4] preeti.nagrath@bharatividyapeeth.edu

[5] rachna.jain@bharatividyapeeth.edu

Abstract: *Automated age and gender detection has been generally used in our daily lives that we come across, majorly in a person to computer interaction, visual surveillance, biometric analysis, electronics and other applications of commercial use. By recognizing the emotions of a person, we can improve the recommendation system. The existing methods have quite satisfying performance on real-world images if facial expressions of input image is neutral or calm, it lacks significantly in age prediction when facial expressions are altered. The dataset was obtained from IMDb-WIKI for age-gender classification and Fer2013 dataset for emotion recognition from kaggle. This project has two models, one for age-gender prediction using wide resnet architecture and the other model is trained for emotion recognition using conventional CNN architecture. An improvement in the performance of these tasks was observed by using the convolutional neural-nets (CNN).The accuracy of the wide-resnet model is 96.26% and for the emotion recognition model, accuracy achieved is 69.2%.*

Keywords: Adam · Convolutional Neural Network · Haar Cascades classifier · IMDB-WIKI · Fer2013 · Tensorflow · Keras · OpenCV · ReLU · Softmax · Wide ResNet

1 Introduction

Owing to the surge in multimedia resources these days, intelligent methods are required to process them. What is touted as one of the most demanding works in Computer Vision is analysing the face image to predict the age of the person. A typical person has a variety of features on his face like age, expression, shape and gender. As humans are the most evolved and the most advanced of species, they can identify those features with considerable ease. For instance, the majority of people can identify fundamental features such as age, gender, race by telling if the person is female or male; they can also tell if the other person is young or old by

just looking at him/her through their naked eyes. In order to detect a face through a single image, a plethora of methods have been proposed, which include knowledge-based approaches, invariant methods like face texture, complexion, template, feature matching methods, namely predefined and changeable, appearance-based approach by way of neural networks.

It is known to us that emotions play a very important and useful role in our life as humans. Our face reflects how we feel or what mood we are in at different moments and situations. During communication, we have the capability to produce thousands of facial reactions and actions which differ in meaning and degree of complexity and depth. There are many ways to bring improvements in the system of recommendation. One of them is identifying the emotions of people. Even then, if the facial expressions are calm or neutral in the input image, the effectiveness of presently used methods on real world pictures is outstanding and very satisfying. However, when the expressions are changed, there are errors in predicting the age of the person correctly.

In this research paper, we endeavour to study about age prediction, emotion and gender by using face images and suggest an effective method and significantly optimized neural architecture for the cause. This project uses CNN to propose a method capable of gender and age prediction besides emotion recognition. We try to explore the best possible architecture for estimating gender, age and emotion recognition through facial images.

The IMDB-WIKI[1] stock of data comprises 523,051 images in Wikipedia3 and IMDB2. The above said stock consists of date of birth, data acquisition, face score and the gender of face detector. The dataset fer2013[2] is used for the purposes of data recognition and includes face images having different emotions.

The rest of the paper is organized as follows. Section 2 discusses the ‘Related work and data,’’ and Section 3 describes the data collection and cleaning techniques. This section also discusses the methodology and describes the architecture of Wide Resnet. In Sect. 4, the results are discussed and compared with other related works on usability feature selection. Section 5 concludes the paper and suggests further work.

2 Related Work and Data

CNN architectures are really popular because of their ability to automatically detect the important features and their computational efficiency. Deep CNNs have been used successfully in applications that include human pose measurement [3], facial parsing [4], facial keypad detection [5], speech recognition [6] and action separation [7]. We will now use in-depth neural networks of age, gender and emotional recognition.

Many methods for extracting facial features to predict age have been used and are described in [8] and [9]. In the same manner, a survey describing various gender classification approaches can be found in [10, 11]. In the paper [12], they used deep CNN based WideResNet architecture

for age and gender prediction. The results of WideResNet are superior to other conventional Convolutional Neural Networks methods and acts as the best method for age and gender prediction. In this architecture classification is used to estimate gender, and age is treated as a regression problem. Also multi task learning reduces infrastructure and computational complexity. In the paper [13], they utilised WideResNet for age estimation with optimizations to make the architecture more efficient. Test accuracy was improved by efficient gradient optimization on loss function. Their model performed well in assessment contrasted with conventional profound Convolutional Neural Networks, for example VGG-16, when assessed on IMDB-WIKI and APPA-REAL dataset. In contrast to the VGG-16 model, Mean Apparent Error on clear, visible images and real pictures were significantly decreased.

There are a variety of feature extraction techniques like geometric based, appearance based, template based and other techniques based on color. A method called segmented detection was used in [14] and SVM classifiers were used for classification processes. [15] presents a very dependable method of facial identification and extraction utilizing genetic algorithms technology and the Eigenface process. [16] brings out the conclusion that feature detection approaches can be categorized as: A feature based approach and a picture based process. Features of a human face are utilised for detection processes in feature based techniques.

3 Methodology

The Convolutional layer becomes an integral and basic part of every convolution neural network (CNN) [20-23]. Large datasets can be processed by a convolution network which has the capability to read an impressive number of equations. The core function of the Convolutional Neural Network (CNN) is its capability to read independently with a huge number of filters; Hence, video and image processing often comprises a reliable neural network that works on a picture composed of a large quantum of data points. This section describes the proposed approach using CNN's age and gender prediction as well as emotional recognition. We describe the details of the network structure and how the Wide ResNet used in the proposed way in the following.

3.1 Data Collection

The IMDB-WIKI dataset consists of 523,051 images in IMDb2 and Wikipedia3. This dataset includes: 1) acquisition data 2) gender 3) birth date(DOB) 4) face score(FS).

For emotion recognition, fer2013 dataset from kaggle[2] is used which consists of 35,890 face images with 7 types of emotions : 1)Angry 2)Disgust 3)Fear 4)Happy 5)Sad 6)Surprise 7)Neutral.

3.2 Dataset Formation

The rotated and cropped images in the dataset were left out before the process of augmentation. Non-facial images, improperly generated images and images having multiple faces with no primacy were deleted from the dataset which resulted in face detection failure. Examples of these images from the dataset can be seen in Fig 2. Some images with their respective face score are shown in Fig 1. Augmentation is used to reduce the error along with random erasing and mix-up generator[17] before feeding it in WideResNet, that is computationally efficient too owing to the property of res-net. Augmentation parameters are shown in Table 1. Unwanted behaviors were reduced namely sensitivity and memorisation. Mix-up generators were used to train the neural networks on convex groupings of example pairs and the image labels. Due to this, the mixup regularly organized the neural-net to maintain linear operation among training data.. The Fer2013 dataset was split in the ratio of 80:10:10 for training, validation and testing respectively. Image Augmentation was performed only on the training dataset to increase the size of the dataset as well as to improve the performance of the model. The augmentation parameters are the same as in Table 1.

Table 1: Augmentation parameters

Width shift(range)	0.1
Height shift(range)	0.1
Horizontal flip	Yes

FS<1.0, images: 1319 FS<0.0, images: 1801 FS>100, images:493 **Fig 1.** Face score (FS) of images in the dataset with number of image

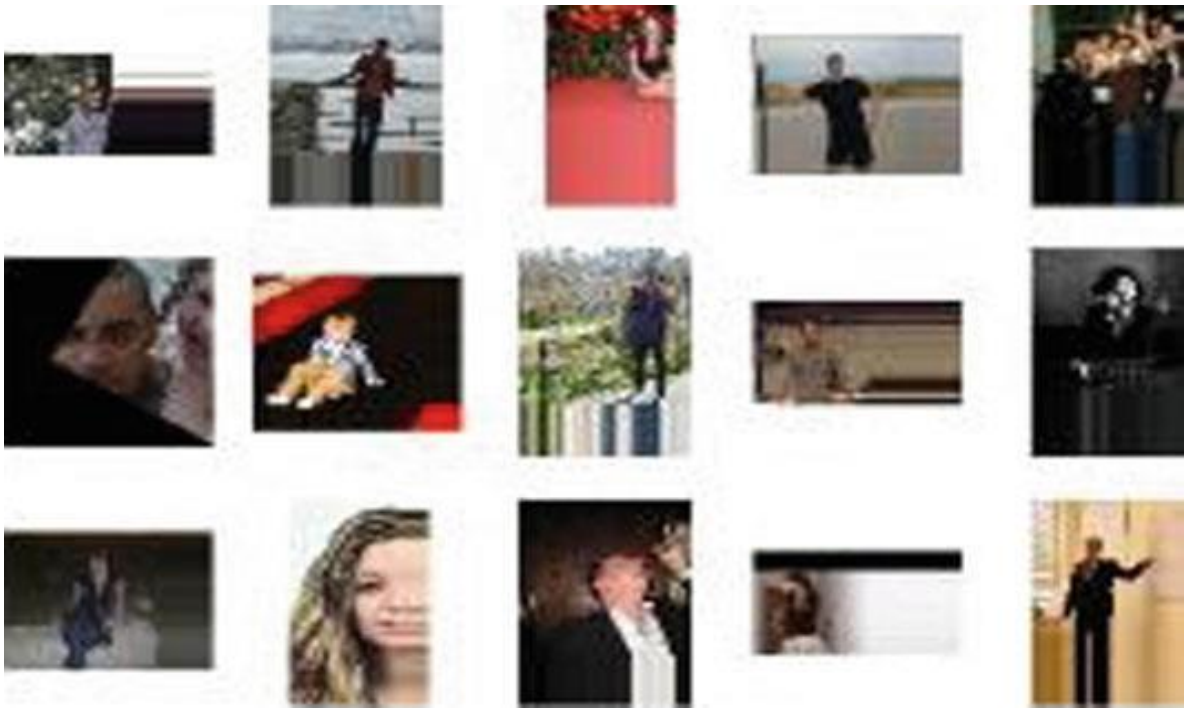

Fig. 2 Ignored faces

3.3 Model Architecture

3.3.1 Wide Residual Network

The Wide Residual Network is a later enhancement on the original Deep Residual Networks[18]. It was concluded that the network does not rely on increasing the network depth to improve its accuracy, but can be flat and wide without affecting performance. The idea was put forward in the paper Wide Residual Networks[19], published in 2016 (and was later updated in 2017 by Sergey Zagoruyko and Nikos Komodakis).

Network degradation and exploding or vanishing gradients are the major challenges faced by a deep network. Even a deep residual network does not guarantee that the contribution to extract useful information is made by all the residual blocks. As all residual blocks are not used , thus a small portion of them can be skipped. It was proven that Wide residual network can perform better than a deep one, by the authors of Wide ResNet.

Wide Resnet Architecture

A Wide ResNet has a group of ResNet-blocks which are stacked together, where every ResNet block sticks to the BatchNormalization-ReLU-Conv structure. This structure is represented as follows:

group name	output size	block type = $B(3,3)$
conv1	32×32	[3×3, 16]
conv2	32×32	$\begin{bmatrix} 3\times3, 16\times k \\ 3\times3, 16\times k \end{bmatrix} \times N$
conv3	16×16	$\begin{bmatrix} 3\times3, 32\times k \\ 3\times3, 32\times k \end{bmatrix} \times N$
conv4	8×8	$\begin{bmatrix} 3\times3, 64\times k \\ 3\times3, 64\times k \end{bmatrix} \times N$
avg-pool	1×1	[8 × 8]

Fig 3. Wide Resnet Structure

A Wide ResNet consists of five groups. The residual block is referred as B(3,3). The first convolution group, conv1, is the same for all types of networks, and the groups conv2, conv3, conv4 differ according to the value of k, the width of the network. An average pool layer and a classification layer are included at the end of convolution groups.

In our implementation, we have used Multi Task Learning i.e. the age and gender are simultaneously estimated using a single CNN instead of independent CNN for each one of them.

Multi Task Learning

Multi-task learning is a method of training in more than one task using a shared structure. Layers at the start of the network will learn a combined generalized representation, avoiding overfitting to a particular task that might contain noise.

Owing to training with a multi-task network, the network can be trained in parallel on both the tasks. As a result, the infrastructure complexity is reduced to only one training pipeline. Moreover, the computation needed for training is reduced as both tasks are trained simultaneously.

In single task learning, an independent feature extraction is carried out for every task. However, in Multi task learning both age and gender share a set of layers, which simultaneously extract features for both of them and then the features output is input to the independent fully connected layers for age and gender.

A general Multi task learning architecture can be represented as follows :

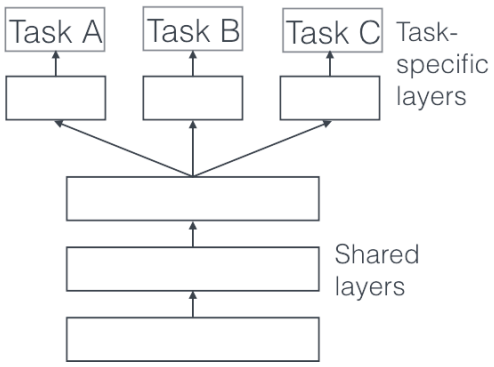

Fig 4. Representation of Multi Task learning architecture

3.3.2 Emotion Model Architecture

The CNN network of models is divided into 5 blocks, each with some number of convolution layers. All the convolution layers have activation function Relu. Batch Normalization is applied to every convolution layer to normalize the outputs by adjusting and scaling the activations. This increases the stability of the neural network and also reduces overfitting. At the end of each block, Max Pooling is done and a Dropout layer is added. Max pooling is performed to reduce overfitting as well as the computational cost, and it helps in extracting the sharpest features of an image. Dropout is another regularization technique to prevent neural networks from overfitting. The CNN takes grayscale images as input of size (48, 48).

The first block consists of two convolution layers , each with kernel size 3 and 64 filters. There is a Batch normalization layer after each convolution layer. And at the end of the block there is a max pooling layer and a dropout layer. Similarly, in the second , third, fourth and fifth block there are three(128 filters each) , four(256 filters each), four(256 filters each), four(512 filters each) convolution layers respectively. The output of the fifth block is flattened using a flatten layer and is fed to the final output layer. The final output layer uses a softmax function. Softmax function gives the probabilities for each class, and the sum of probabilities of all classes equals one. From the final output we can select the class with maximum probability as our predicted output.

4 EXPERIMENTAL RESULTS

4.1 Age and Gender Model

The accuracy achieved for age and gender classification model is 96.26% with the wide ResidualNet design with the implementation of erasing techniques on images and augmentation.

The model worked well as test images with a slight decrease in loss and the number of epochs to specify age and gender in this dataset.

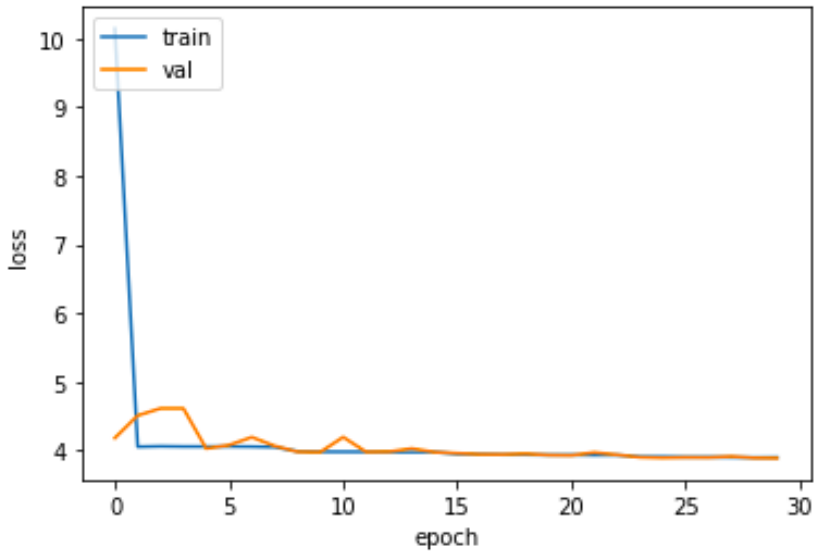

Fig 5. loss vs epochs for Age Prediction

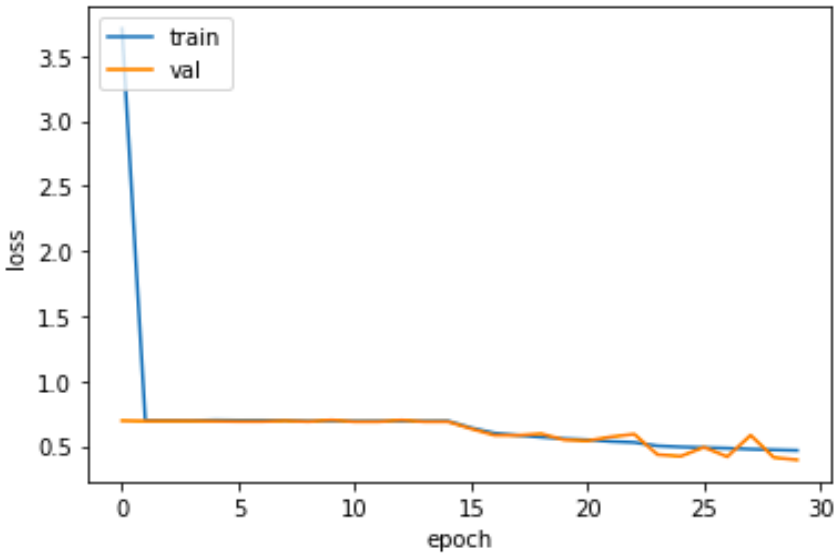

Fig 6. loss vs epochs for Gender prediction

4.2 Emotion Recognition Model

The model for emotion recognition has achieved the accuracy of 69.5%. The confusion matrix of the model with different emotions and the accuracy of the model with increasing epochs is shown in fig 6 and 7 resp.

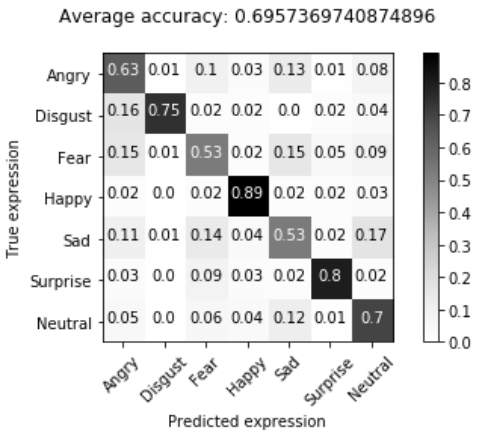

Fig 7. Confusion matrix

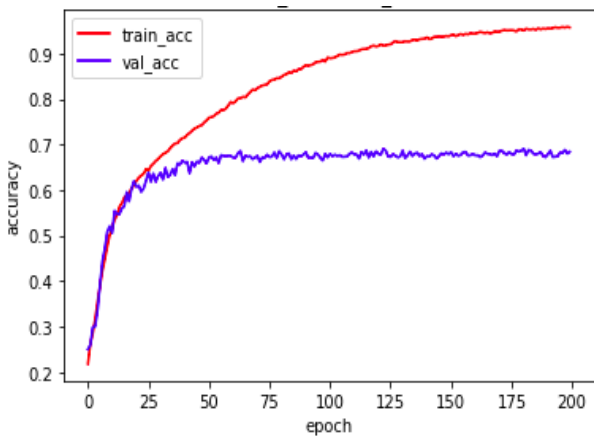

Fig 8. Accuracy plot

4.3 Live Predictions

Fig 9. Output of the models with age, gender and emotion mentioned on the top of the box.

5 Conclusions

This paper proposed an unique compilation of both age and gender prediction & Emotion Recognition. We have implemented two models, one for Age and Gender prediction, and the other one for Emotion recognition. For Age and Gender prediction we have used a wide-resnet based CNN approach, and for Emotion Recognition we have used conventional CNN.

In the past many projects, research papers have been put forward in this field, using various datasets from various sources, none of the research papers that i came through had complied both age, gender & emotions together.

Future work: This model might be additionally improved and the extent of future work incorporates including entity recognition, the residual architecture can be more efficient by hyperparameter tuning, addition and widening of more convolutional layers per block.

References

[1] https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

[2] https://www.kaggle.com/deadskull7/fer2013

[3] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In Proc. Conf. Comput. Vision Pattern Recognition, pages 1653–1660. IEEE, 2014.

[4] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In Proc. Conf. Comput. Vision Pattern Recognition, pages 2480–2487. IEEE, 2012.

[5] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In Proc. Conf. Comput. Vision Pattern Recognition, pages 3476–3483. IEEE, 2013.

[6] A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 6645–6649. IEEE, 2013

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proc. Conf. Comput. Vision Pattern Recognition, pages 1725–1732. IEEE, 2014

[8] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. Trans. Pattern Anal. Mach. Intell., 32(11):1955–1976, 2010.

[9] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In Biometrics (ICB), 2013 International Conference on. IEEE, 2013.

[10] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. Trans. Pattern Anal. Mach. Intell., 30(3):541–547, 2008.

[11] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. Machine learning: theory and applications. Elsevier, pages 327–352, 2013.

[12] K. Ito, H. Kawai, T. Okano, and T. Aoki, “Age and Gender Prediction from Face Images Using Convolutional Neural Network,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2018, pp. 7–11.

[13] R. Debgupta, B. B. Chaudhuri, and B. K. Tripathy, “A Wide ResNet-Based Approach for Age and Gender Estimation in Face Images,” in *International Conference on Innovative Computing and Communications*, 2020, pp. 517–530.

[14] Serban, Ovidiu, et al. “Fusion of Smile, Valence and NGram features for automatic affect detection.“ Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013.

[15] Wong, Kwok-Wai, Kin-Man Lam, and Wan-Chi Siu. “An efficient algorithm for human face detection and facial feature extraction under different

[16] Bakshi, Urvashi, and Rohit Singhal. “A survey on face detection methods and feature extraction techniques of face recognition.“ International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 3.3 (2014).

[17] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random Erasing Data Augmentation,” *arXiv [cs.CV]*, Aug. 16, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016, [Online]. Available: http://arxiv.org/abs/1605.07146.

[20] I. V. Pustokhina et al., "Automatic Vehicle License Plate Recognition Using Optimal K-Means With Convolutional Neural Network for Intelligent Transportation Systems," in IEEE Access, vol. 8, pp. 92907-92917, 2020, doi: 10.1109/ACCESS.2020.2993008.

[21] Shankar, K., Lakshmanaprabu, S. K., Khanna, A., Tanwar, S., Rodrigues, J. J., & Roy, N. R. (2019). Alzheimer detection using Group Grey Wolf Optimization based features with convolutional classifier. Computers & Electrical Engineering, 77, 230-243.

[22] Anupama, C.S.S., Sivaram, M., Lydia, E.L. et al. Synergic deep learning model–based automated detection and classification of brain intracranial hemorrhage images in wearable networks. Pers Ubiquit Comput (2020). https://doi.org/10.1007/s00779-020-01492-2

[23] R. J. S. Raj, S. J. Shobana, I. V. Pustokhina, D. A. Pustokhin, D. Gupta and K. Shankar, "Optimal Feature Selection-Based Medical Image Classification Using Deep Learning Model in Internet of Medical Things," in IEEE Access, vol. 8, pp. 58006-58017, 2020, doi: 10.1109/ACCESS.2020.2981337.