# Age estimation with mask

李宣緯*
陳敬元
王駿豪

## Abstract

We discuss a research study on age estimation with masks. With the COVID-19 pandemic, traditional age prediction methods have faced challenges due to people wearing masks. To address this issue, the researchers propose a novel approach of cropping images to retain only the portion above the nose, allowing the model to focus on facial features above the mask and eliminate interference from the mask itself. The study utilizes existing age prediction methods to evaluate the accuracy of age prediction for masked faces. The results of this research contribute to understanding the feasibility and accuracy of age prediction for faces wearing masks.

## I. INTRODUCTION

In today's fast-paced world, where security threats are becoming increasingly sophisticated, it's imperative for companies to stay one step ahead. Traditional methods of access control, such as key cards and PIN codes, are no longer sufficient to ensure the safety of employees, visitors, and sensitive data(Lee, 2020; Rameswari, 2021). This is where facial detection systems step in, offering a cutting-edge solution that combines advanced technology with enhanced security protocols.

Facial detection systems provide a higher level of security compared to traditional access control methods. Unlike key cards or passwords, facial features are unique to each individual, making it virtually impossible to impersonate someone else (Hadid, 2004). This inherent security feature significantly reduces the risk of unauthorized access, ensuring that only authorized personnel can enter designated areas.

Because of Covid-19, people started to be in the wake of the ongoing global health crisis, the demand for contactless solutions has surged. Facial detection systems offer a hands-free and hygienic alternative to traditional access control methods. Employees and visitors can gain access without physically touching any surfaces, minimizing the risk of germ transmission. Additionally, this technology eliminates the inconvenience of forgotten key cards or passwords, streamlining the access process for everyone involved.

Modern facial detection systems are highly scalable and can be seamlessly integrated with existing security infrastructure. Whether you're a small startup or a large corporation, these systems can be tailored to meet your specific needs. They can easily integrate with other security measures, such as surveillance cameras and alarms, creating a comprehensive security ecosystem that enhances overall protection.

In an era where security threats are constantly evolving, companies cannot afford to rely on outdated access control methods. Facial detection systems offer a state-of-the-art solution that combines enhanced security, convenience, and efficiency. By making the switch to facial recognition technology, companies can bolster their security measures, protect sensitive information, and provide a safer environment for employees and visitors alike. Embracing this cutting-edge technology is not just a step towards a more secure future; it's a leap into a new era of access control.

With the outbreak of the COVID-19 pandemic, people have started wearing masks, this will increase the difficulty of confirming people’s identities (Ekundayo,2019). And we found out that in the facial detection research field, age prediction for faces has achieved significant accuracy[14] . However, if faces were covered with masks, the accuracy would drop, and we want to find out the result and do some experiments to reduce the decreasing accuracy due to the masks. Therefore, we aim to investigate whether accurate age prediction can still be achieved for masked faces.

To address this issue, we propose a novel approach of cropping the images to retain only the portion above the nose. This allows the model to focus on the facial features above the mask and eliminate interference from the mask itself.

We utilize existing age prediction methods to predict the ages based on these cropped images. By comparing the predicted ages with the actual ages, we can evaluate the accuracy of age prediction under the condition of wearing masks. The results of this research will contribute to understanding the feasibility and accuracy of age prediction for faces wearing masks.

We use the IMDb dataset and delete the upper half of the face in the picture to simulate the situation of wearing a mask, then use CGAN to generate the situation of wearing a mask, then use CGAN to generate to corresponding lower half of the face based on the age label, and use pre-trained ShuffleNet and MobileNet to predict the real age of the person in the picture.

# II. LITERATURE REVIEW

## 2.1 Facial Recognition

Early facial recognition systems were primarily based on traditional image processing techniques and methods. These systems usually relied on feature extraction and pattern recognition technologies, such as Haar features, Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) (Bruce & Young, 1986; Barnouti, Al-Fabbagh, & Matti, 2016; Bhele & Mankar, 2012). These techniques aimed to capture the basic geometric structures and texture information of the face, thereby identifying and classifying faces in datasets. However, these methods often performed poorly in complex real-world applications, such as under varying lighting conditions, pose changes, and facial expressions. Additionally, these technologies had significant limitations in handling facial obstructions like glasses, hats, or masks.

With the rise of artificial intelligence technology and the successful application of neural networks in data processing, facial recognition technology has undergone revolutionary changes. AI-based on machine learning can extract and learn complex and abstract features from a large volume of data (Nawaz, 2020), greatly enhancing the accuracy and reliability of recognition. Apart from these emerging technologies, this improvement is also largely due to the extensive databases of facial images (Sharma, Bhatt, & 2020). Combining various machine learning methods, specialized facial recognition models can be trained. These models can not only capture specific or diverse subtle facial features according to requirements (Vaishnavi & Achwani, 2018) but can also effectively handle various feature variations.

## 2.2 Artificial Intelligence

The exact definition of Artificial Intelligence (AI) tends to vary depending on the field of application. One of the earliest interpretations can be traced back to 1942, when Isaac Asimov, a Jewish science fiction author, proposed a framework for the thinking patterns of machines in his work, giving a rational explanation for the differences and similarities between machines and human beings. The term "Artificial Intelligence" was first officially introduced by McCarthy and others (1955) at the Dartmouth Conference. The concept included: 1. The ability to simulate the human brain's functions; 2. The capability to use language and text; 3. The idea of forming neural networks by arranging neurons; 4. Establishing standards and theories to measure computational efficiency and assess complexity; 5. The ability for self-improvement.

Today, the most widely known definition comes from Russel & Norvig (2010), describing AI as an agent that can perceive its environment and take actions based on this information to achieve the best expected outcome in uncertain situations.

In the field of facial recognition, the application of AI has attracted extensive attention. Thanks to various technologies serving as tools and aids, significant progress has been made in feature extraction. Traditional facial recognition used Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), or Local Binary Patterns (LBP) for feature extraction. However, these methods were often limited by lighting, angles, age, and other minute, uncontrolled changes (Wang & Deng, 2021). With the development and deployment of neural network-based deep learning, networks with multi-layered structures can extract features in a more detailed and hierarchical manner (Coşkun et al., 2017). This enables systems to be more tolerant of minor variations in input data (Li et al., 2020), and they can still recognize or predict faces even if parts of the face are obscured or deformed, using large datasets for learning as a foundation. Furthermore, with the expansion of system architecture, some AI systems based on deep learning models are capable of performing multiple tasks related to facial recognition simultaneously, such as age estimation, emotion recognition, and extracting various other features. This has also expanded the systems' multifunctionality and application scope (Parkhi, Vedaldi, & Zisserman, 2015).

### 2.3 Machine Learning

Machine learning can be considered an extension of the field of artificial intelligence, mainly involving "teaching" machines to learn, rather than instructing them to execute tasks based on input commands (Prüfer, 2020). Instead, it involves improving performance in future tasks through training experiences acquired during task execution (Jordan & Mitchell, 2015). Machine learning is generally classified into supervised learning, unsupervised learning, and reinforcement learning.

Under supervised learning, machines collect labeled data and are able to recognize these labels (Metcalf, Askay, & Rosenberg, 2019). After training on the data, machines' predictions of labels are compared with the actual labels to assess the training outcomes, continuing until the training results sufficiently match real-world results. In facial and image recognition, supervised learning models use sample learning data like SuffleNet and MobileNet to track facial contours and continuously update themselves (De-la-Torre, 2015), achieving strong predictive and recognition capabilities for facial models (Lin & Wu, 2021).

Unsupervised learning implies that there is no "supervision" from humans; no additional labels or instructions are added. Machines learn to identify patterns and similarities in the data and categorize and recognize observations based on these (Overgoor, 2019). This training mode can flexibly handle large and diverse data sets, avoids human errors that might occur in data labeling, and has a high tolerance for anomalies (Datta, Sharma, & Jawahar, 2018). It can recognize more rare or special facial features, such as scars, artistic tattoos, or decorative items.

Reinforcement learning emerges as another branch of machine learning with a unique method of learning. It differs from the previous two methods in that it focuses on interacting with the environment to learn, identifying the best strategies through repeated trial and error (Arulkumaran et al., 2017). In the reinforcement learning framework, neural network models interact as much as possible with the environment, obtaining positive and negative feedback from these interactions and continuously training to maximize rewards obtained from the environment (Sewak, 2019). In facial recognition, reinforcement learning can continuously adjust recognition strategies, adapting to environmental changes more effectively than the other two learning models. It can also adjust the accuracy of recognition for specific individuals or groups based on strategy directives, creating a personalized recognition system (Liu et al., 2019).

### 2.4 Summary

In this chapter, we reviewed a range of traditional and emerging technologies. The cultural practice of wearing masks, emerging in recent years due to environmental demands, has presented significant challenges to traditional facial recognition technologies. The extensive coverage and changes in facial features caused by masks make feature extraction much more difficult. However, by integrating new technologies endowed with artificial intelligence and machine learning, known for their strong adaptability, novel facial recognition systems are gaining enhanced analytical, personalized, and adaptive capabilities, rapidly finding practical applications across various fields.

## III. EXPERIMENT DESIGN

All figures and tables should be numbered consecutively and captioned, and be placed in text as close to the reference as possible.

### 3.1 Dataset

We use the IMDb-WIKI dataset, IMDb dataset is a large collection of data related to movies and television shows on IMDb, commonly used in various research and machine learning tasks. We use the IMDb-crop face-only dataset and wiki-crop face-only dataset for our age prediction task, IMDb Face Crop is a dataset specifically focused on facial analysis and recognition. It consists of images extracted from the movie posters and associated metadata available on IMDb. The dataset contains about 400,000 facial images and divides them into 100 categories, representing face pictures from 0 to 100 years old.

And we used FGNET to evaluate the model's performance, the FGNET is a dataset for age estimation and face recognition across ages. It is composed of a total of 1,002 images with 82 people aged 0 to 69. It is often used for face verification across large age gaps. The dataset contains images ranging from child/young to adult/old.

### 3.1.1 Data cleaning

We use OpenCV (Open-Source Computer Vision Library) and Dlib library to do data cleaning, OpenCV is an open-source computer vision and machine learning library. It provides various functions and algorithms for image and video processing, including face detection. It offers image preprocessing, object detection, and tracking. And Dlib is a powerful C++ library that is widely used for machine learning and computer vision tasks, It includes a variety of algorithms and tools for image processing, numerical computations, and machine learning applications. One of the key features of Dlib is its face-detection capabilities. We use the face detection function to filter our data set, the function we use will detect the image and check if there is only one person's face in the picture, this will filter out unclear, too far, or blurry to detect a human face of the picture. After filtering, there are 280,000 images left in this dataset for our age prediction task.

### 3.1.2 Data augmentation

Data augmentation is a technique used in machine learning and deep learning, specifically in the field of computer vision. Its primary objective is to increase the size and diversity of the dataset to improve the performance of predictive models, reduce overfitting, and enhance the model's ability. Data augmentation creates new training samples by applying various transformations to the original data. For example, rotation, scaling, flipping, translation, brightness changes, cropping, and so on, also improving its robustness and generalization capabilities.

## 3.2 Experiment Flow

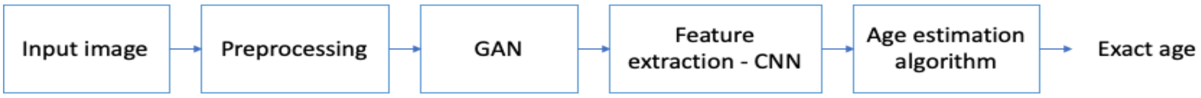

Figure 1. Experiment flow.

### 3.2.1 Preprocessing

(a) Filtering one-face-detected images

First, we only use those photos with only one face detected, since the model won't be confused about which faces were the age labels for. After the preprocessing work, we got 287,479 images, and the age distribution is like Figure 2.

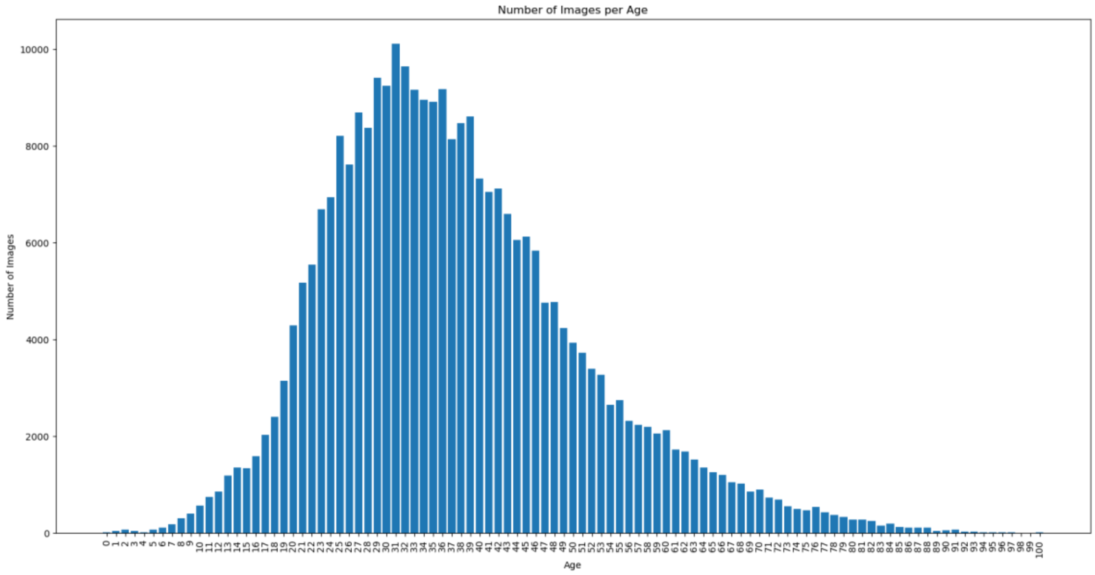

Figure 2.

(b) Simulating the face with a mask

In order to simulate wearing a mask for prediction, we use face detection methods to crop the face of each photo, leaving only the upper part of the face above the nose. We recognized the face in the image then kept the face in a frame, and only kept the upper face of the frame. The cropped image is like Figure. 2.

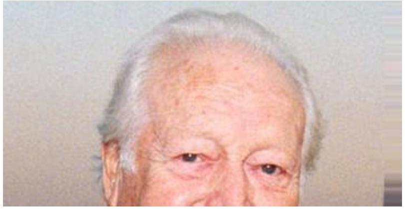

Figure 3. Cropped image.

(c) Image transforming

The face image is standardized to have zero mean and variance one and scaled to 256 × 128 pixels.(original picture was 256x256)

(d) Data augmentation

We use the "imgaug" library for image augmentation. Each augmentation is applied with a specific probability specified by the "sometimes" function. The augmentation methods we use include: rotation, left-right flip, Gaussian blur, multiplication, contrast normalization, cropping and padding. We didn't do much data augmentation because we cropped the image, leaving only the top half of the face for our task, which could greatly reduce our accuracy in age prediction. The application probability of each data enhancement method is 10%, which means that each picture has a 10% chance of being rotated, a 10% chance of flipping left and right, a 10% chance of Gaussian blur, and so on.

### 3.2.2 Conditional Generative adversarial network

Conditional Generative Adversarial Networks (CGANs) are an extension of the traditional Generative Adversarial Networks (GANs)(不詳; Wang 等, 2020), it allows for the conditional generation of data. In a CGAN, both the generator and discriminator networks are conditioned on additional input information. In the process of training CGAN, we used the age labels and face pictures in the IMDb-crop data set as training data. A total of 280,000 face pictures were used to train CGAN for a total of 20 epochs. In our task, we use CGANs to generate the lower part of the face according to the age label and combine it with the upper part of the face, and then let the pre-trained model make the age predictions. Each picture output is 128*128 pixels. The purpose is to compare whether the prediction performance can be improved if we use CGAN to generate the lower part of the image and then predict it when the output is the upper part of the face image (wearing a mask). Based on this idea, we trained a CGAN to generate pictures. The generated image shown in Figure 4.

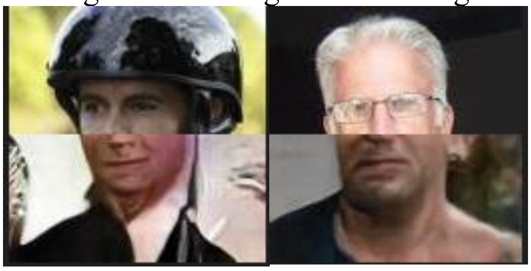

Figure 4. Generative picture.

## 3.3 Predicted Model

### 3.3.1 ShuffleNet with standard cross entropy

We used the 'shufflenet_v2_x2_0' version of ShuffleNetV2 in PyTorch and made modifications based on the referenced paper. ShuffleNetV2(Ma 等, 2018), have significantly accelerated the deployment of neural network models on mobile terminals. Although these lightweight networks have far fewer parameters compared to heavyweight networks, they possess strong feature extraction capabilities, and the training time for these networks is also relatively short.

We modified the model by replacing the final output layer with a classification layer consisting of 101 categories corresponding to ages ranging from 0 to 100 years. We then selected the class with the highest probability from the classification layer as the final output. During training, we employed the cross-entropy loss function and utilized Stochastic Gradient Descent (SGD) as the optimizer. The training was conducted for a total of 10 epochs.

3.3.2 MobileNetv2

We test the MobileNetv2 with the custom loss to find out the performance of the model on our dataset. MobileNet is a family of architectures proposed specifically with the efficiency of execution on mobile and embedded hardware in mind. The basic idea, introduced with the first generation of the model is the use of depthwise separable convolutions – the regular convolutions are replaced by a series of two computation-ally lighter convolutions: the spatial 3 × 3 convolution applied to each input channel separately, followed by a 1×1 cross-channel convolution. The MobileNetV2 (Howard ,et. al, 2017), used in our experiments, improves the basic building block module by using linear bottlenecks and inverted residuals. This combination is proposed by the paper; however, our performance is much lower even in the situation where we used the cropped images for training.

3.3.3 MobileNetv3

After tested MobileNetv2, we tested the latest version of MobileNet and compared the performance. In most cases, MobileNetv3 outperforms MobileNetv2 in terms of both efficiency and performance. MobileNetv3 introduce the lightweight attention called squeeze-and-excite blocks and a new nonlinear function called hard-swish that makes MobileNetv3 outperformance than MobileNetv2 and also reduces the computational complexity of the model. Besides, the internal structure of MobileNetv3, the 1x1 expansion layer from the last stage has been moved beyond the pooling layer, and this allows the 1x1 layer to operate on a 1x1 feature map instead of 7x7, to reduce the latency and computation time.(Kavyashree & El-Sharkawy, 2021)

## IV. EXPERIMENT RESULT

Table1. is our experiment result, and we find out that ShuffleNet performs better than MobileNetV2 and MobileNetV3, and with our custom loss we got significant improvements, and our best model is ShuffleNet with custom loss, which got MAE=22.01 on the FGNET dataset. The Mean Absolute Error (MAE) gap between facial images with and without masks is significant, indicating the challenges people face when trying to estimate age. This phenomenon is also observed in models, highlighting the increased difficulty in age estimation tasks caused by the presence of masks. It shows that the research on improving facial feature extraction on faces with masks is meaningful, causing the situation to become more common in the future. We know that this MAE is barely satisfactory, however, the recent results show that we have already made progress, and we believe that the more high-quality training dataset and some modifications on models, like increasing the epochs, might get a better result. There are still a lot of improvements and other models that might outperform in future work.

Table 1. MAE Result.

| Model | MAE |
|---|---|
| ShuffleNet with standard cross entropy | 66.53 |
| ShuffleNet with custom loss | 22.01 |
| MobileNetv2 with standard cross entropy | 78.65 |

| | |
|---|---|
| MobileNetv2 with custom loss | 65.47 |
| MobileNetv3 with standard cross entropy | 73.53 |
| MobileNetv3 with custom loss | 61.12 |

## V. CONCLUSION

In our analysis of the IMDb-wiki dataset, we noticed that using ShuffleNet and MobileNet for face recognition with masks significantly improved accuracy after training, especially in age differentiation. This advancement fosters an optimistic outlook for future applications of machine learning in corporate facial recognition access control systems, as the variety and volume of facial samples with masks are likely to increase. This will undoubtedly aid in further research extensions.

However, at present, we still face some limitations and propose several suggestions for future research. Firstly, we observed many images of the elderly in the dataset, which typically don't contribute significantly to access control systems. This indicates a need for cleaning the dataset to ensure the accuracy of prediction models. Secondly, regarding the combination of the upper half of the face image with those generated by our CGANs, we found that although this method enhanced model performance, the results were still not satisfactory. This might be a fitment issue between CGANs and the IMDb-crop dataset. While CGANs generate new images based on age labels, IMDb-crop encompasses diverse features like gender, race, and hair color. This leads to a focus on the entire face rather than specific regions (such as areas not covered by masks). We suggest using a dataset emphasizing facial regions (as shown in Figure.5) or considering more potent generative models in the future, which should help accelerate training outcomes. Thirdly, the improvement in model performance was not as significant as expected, possibly due to constraints in training time and cycles, limiting opportunities for fine-tuning experiments to adjust the parameters of the loss function, or using more powerful and resource-intensive CNN architectures. Nevertheless, as mentioned earlier, with the increasing variety and volume of masked facial samples and the continuous evolution of machine learning models and artificial intelligence technologies, we remain hopeful and encouraged about the future research and development in this field. We also urge more researchers to join and expand the knowledge frontier in this domain.

## REFERENCES

1. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 34(6), 26-38.

2. Barnouti, N. H., Al-Dabbagh, S. S. M., & Matti, W. E. (2016). Face recognition: A literature review. International Journal of Applied Information Systems, 11(4), 21-31.

3. Bhele, S. G., & Mankar, V. H. (2012). A review paper on face recognition techniques. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(8), 339-346.

4. Bruce, V., & Young, A. (1986). Understanding face recognition. British journal of psychology, 77(3), 305-327.

5. Coşkun, M., Uçar, A., Yildirim, Ö., & Demir, Y. (2017, November). Face recognition based on convolutional neural network. In 2017 International Conference on Modern Electrical and Energy Systems (MEES) (pp. 376-379). IEEE.

6. Datta, S., Sharma, G., & Jawahar, C. V. (2018, May). Unsupervised learning of face representations. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 135-142). IEEE.

7. De-la-Torre, M., Granger, E., Radtke, P. V., Sabourin, R., & Gorodnichy, D. O. (2015). Partially-supervised learning from facial trajectories for face recognition in video surveillance. Information fusion, 24, 31-53.

8. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

9. Li, L., Mu, X., Li, S., & Peng, H. (2020). A review of face recognition technology. IEEE access, 8, 139110-139120.

10. Lin, C. H., & Wu, B. F. (2021, September). Domain adapting ability of self-supervised learning for face recognition. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 479-483). IEEE.

11. Liu, B., Deng, W., Zhong, Y., Wang, M., Hu, J., Tao, X., & Huang, Y. (2019). Fair loss: Margin-aware reinforcement learning for deep face recognition. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10052-10061).

12. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. 1955. A proposal for the Dartmouth summer research project on artificial intelligence. Available at: https://doi.org/10.1609/ aimag.v27i4.1904

13. Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping humans in the loop: pooling knowledge through artificial swarm intelligence to improve business decision making. California management review, 61(4), 84-109.

14. Nawaz, N. (2020). Artificial intelligence applications for face recognition in recruitment process. Journal of Management Information and Decision Sciences, 23, 499-509.

15. Overgoor, G., Chica, M., Rand, W., & Weishampel, A. (2019). Letting the computers take over: Using AI to solve marketing problems. California Management Review, 61(4), 156-185.

16. Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association.

17. Prüfer, J., & Prüfer, P. (2020). Data science for entrepreneurship research: studying demand dynamics for entrepreneurial skills in the Netherlands. Small Business Economics, 55(3), 651-672.

18. Russell, S. J., & Norvig, P. (2010). Artificial intelligence a modern approach. London.Sewak, M. (2019). Deep reinforcement learning. Singapore: Springer Singapore.

19. Sharma, S., Bhatt, M., & Sharma, P. (2020, June). Face recognition system using machine learning algorithm. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1162-1168). IEEE.

20. Vaishnavi, K. A., & Achwani, S. (2018). A study on use of artificial intelligence in human resource management. Advance and Innovative Research, 199.

Wang, M., & Deng, W. (2021). Deep face recognition: A survey. Neurocomputing, 429, 215-244.

21. Lee, H., Park, S. H., Yoo, J. H., Jung, S. H., & Huh, J. H. (2020). Face recognition at a distance for a stand-alone access control system. Sensors, 20(3), 785.

22. Rameswari, R., Kumar, S. N., Aananth, M. A., & Deepak, C. (2021). Automated access control system using face recognition. Materials Today: Proceedings, 45, 1251-1256.

23. Ekundayo, O., & Viriri, S. (2019, March). Facial expression recognition: a review of methods, performances and limitations. In 2019 Conference on Information Communications Technology and Society (ICTAS) (pp. 1-6). IEEE.

24. Hadid, A., Heikkilä, M., Ahonen, T., & Pietikäinen, M. (2004, June). A novel approach to access control based on face recognition. In Proc. Workshop on Processing Sensory Information for Proactive Systems (PSIPS) (pp. 68-74).

25. P. S. P. Kavyashree and M. El-Sharkawy, "Compressed MobileNet V3:A Light Weight Variant for Resource-Constrained Platforms," *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, NV, USA, 2021, pp. 0104-0107, doi: 10.1109/CCWC51732.2021.9376113.