Advanced Convolutional Neural Network Paradigms-Comparison of VGG16 with Resnet50 in Crime Detection

Taiwo. M. Akinmuyisitan, John Cosmas

Abstract**—**This paper practically demonstrates the theories and concepts of an Advanced Convolutional Neural Network in the design and development of a scalable artificial intelligence model for the detection of criminal masterminds. The technique uses machine vision algorithms to compute the facial characteristics of suspects and classify actors as criminal or non-criminal faces. The paper proceeds further to compare the results of the error accuracy of two popular custom convolutional pre-trained networks, VGG16 and Resnet50. The result shows that VGG16 is probably more efficient than ResNet50 for the dataset we used.

Keywords**—**Artificial intelligence, convolutional neural networks, Resnet50, VGG16.

I. Introduction

ACROSS the world, there is an increasingly growing insecurity concern. Traditional policing has been a subject of public discourse due to the ever-growing insecurity concern. The evaluation of current global security has called for an urgent increase in security personnel to protect community dwellers and valuable properties in diverse places. While some [7] opined that an increase in the community police force is not guaranteed to improve the security situation, researchers are considering the need for an innovative solution. An attempt at using machine vision for crime prediction was first carried out in 1998 [1]. However, nothing much was achieved due to the computational power such an intelligence system required.

Training an artificial intelligence (AI) model that could deliver the tasks of who, where, and when crimes would occur is inherently very difficult because it is sensitive to the highly complex distributions of crimes in space and time [2]. Literature work suggested that crime incidents exhibit spatial and temporal dependencies with dynamic social environments [3]. In [4], the authors posited that through facial expressions, a lot of things could be determined including an inherent possibility of a person committing a crime.

The remainder of this paper is organized as follows. Section II presents a review of the related theories and methods that have been used to predict crime. Section III discusses the materials and methods while the presentation of our findings, results, and the discussion is contained in Section IV. The study's conclusion and recommendations are in Section V.

II. Related Work

Before the advancement in deep neural networks, methods such as naïve Bayes and kernel density estimations among others had been used [5]. The author's methods appear to work over certain extended periods but recently it has been adjudged not to be a good predictor of future crime and criminals. In another research conducted by [6], the authors used multivariate regression. At the end of their experiments, [6] discovered that the success of their multivariate method relied on the use of the correct features.

Reference [7] worked on a crime prediction experiment using a shallow feed-forward network, with the aim to predict burglary-type crime hot spots every month. The study used datasets that comprised records of crimes with locations as well as the events that preceded them. The study of [8] used two approaches to extract facial expressions. The authors believed that through facial expressions, it was possible to determine criminal masterminds. Thus, a pre-trained VGG-16, VGG-19, inceptionV3, and the standard Convolutional Neural Network (CNN) approach was used. A CNN in another research was however used to predict dangerous weapons in a crime scene [9]. However, it failed to detect when a person holds a knife in their hand.

Although the research conducted by [8] claims an accuracy of 99.5%, there is scepticism about this result in reality. Such result in AI could be realized only where there are data leakages. Optimizing models is ideally done on the validation set. So, if an AI model was evaluated on test data, after a series of iterations and hyperparameter tuning, the model would learn all the features of the training data. Consequently, it will generalize well only on the test set but when introduced to a new dataset, may perform poorly.

This paper aims at practical developments of advanced CNN and facial recognition algorithms that compare the accuracies and losses of VGG16 with Resnet50 for the classification of actors into criminals and non-criminals. The study is unique because of the absolute training and retraining of the classifier output that eventually achieved validation accuracy of 96.3% and training accuracy of 98.1% for VGG16 at the 30th epoch, with no overfitting problem, as shown in Fig. 4.

T. M. Akinmuyisitan is with the Department of Electronic and Computer Engineering, Brunel University London. Kingston Lane. Uxbridge, Middlesex. UB83PH London, UK (phone: +447424570091; e-mail: 2037437@brunel.ac.uk).

J. Cosmas is with the Department of Electronic and Computer Engineering, Brunel University London. Kingston Lane. Uxbridge, Middlesex.UB83PH London (e-mail: john.cosmas@brunel.ac.uk)

III. Material and Method

In this section, the evaluation of the research methods used focused solely on CNN, Computer Vision, and Deep Machine Learning ethodologies. Specifically, the method used in this research is an advanced pre-trained VGG-16 and Restnet-50 CNN for feature extraction on the popular National Institute of Standard and Technology (NIST) datasets. The dataset images are male and the model extracts criminal traits and features. The results were analysed, and the best-performing model was determined between ResNet50 and VGG16 as shown in Section IV.

A. Convolutional Neural Network

Convnets is one of the most popular neural networks commonly used for image classifications in deep learning. It involves the task of performing the function:

$$R^{H*W*K} \rightarrow \{0,1\}^{C}$$

where K = the number of input channels; C = Number of class labels.

ResNet was the winner of the 2015 Image Classification Challenge. It was proposed by a Microsoft team [22].

The crucial feature of ResNet is the use of residual block, modeled as:

$$X_{l+1} = \varphi(x_l + f_{l(}x_1)$$

It is called residual block because f_1 needs to learn the easily comprehended residual or the difference between the input and output of the layer. CNN is a special multilayer neural network that entirely comprises conv2d and max-pooling layers used for spatial data. The architecture of CNN is inspired by the visual perception of living beings [10]. CNNs became popular after they achieved nearly 98% accuracy on AlexNet [11] in 2012. The foundation of the CNN started with the discovery of Hubel and Wiesel in 1959 [12].

Neocognitron proposed by Kunihiko Fukushima is considered the first theoretical model for CNN [13]. In 1990, LeCun et al. developed the modern framework of CNN called LeNet-5 [14] to recognize handwritten digits. Training by backpropagation [15] algorithm helped LeNet-5 in recognizing visual patterns from raw images directly without using any separate feature engineering.

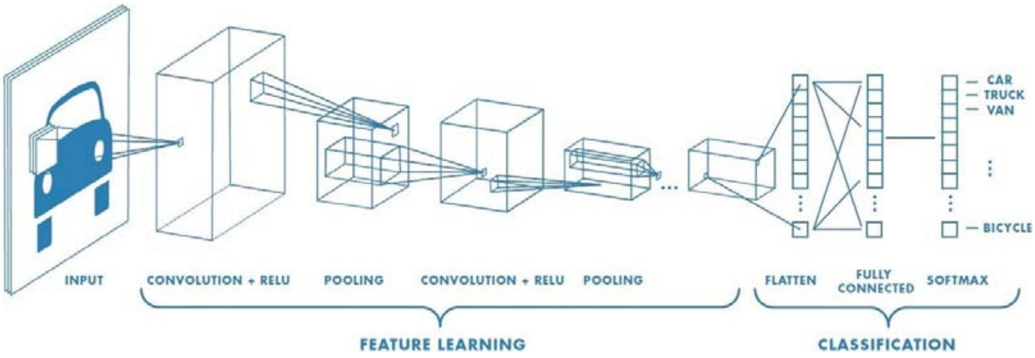


Fig. 1 CNN Architecture [16]

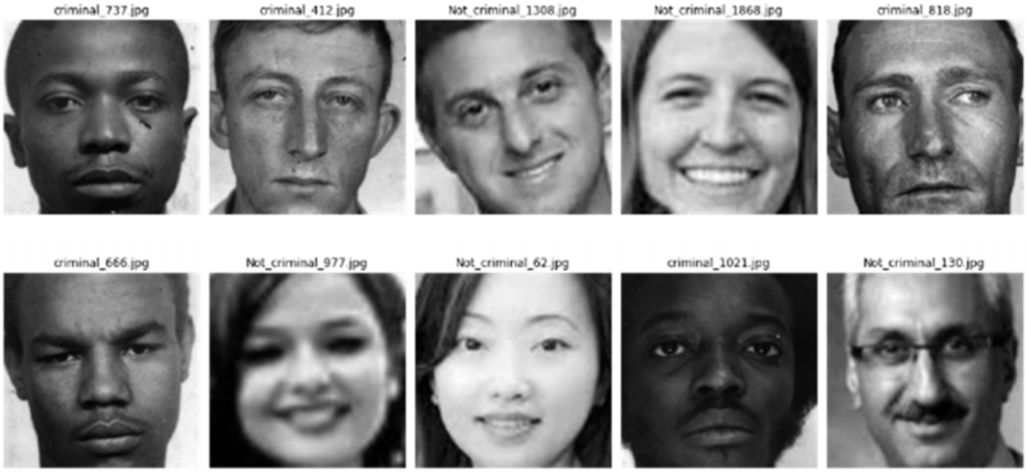


Fig. 2 Cross section of the dataset images

B. Datasets

The model was trained using VGG-16 and Restnet-50 on the NIST datasets for image recognition. These datasets contained 3248 mugshot data of criminals in PNG format [17]. A total of 3249 front and side views were present. However, it was the 1349 images containing only the front views that were used for the study. This was because the research aims to establish the possibility of detecting criminals through facial expressions. So, the Haar cascade algorithm was used to extract the frontal view of all the images. The dataset for this research also includes 2,013 images of non-criminals that were sourced on the internet [19]. The mugshot special NIST dataset and the non-criminal openly sourced together put the total datasets for this research to 3,362. Fig. 2 shows random images in our dataset.

C. Data Pre-processing

The paper allows the use of two separately stored datasets. The NIST special dataset is stored in 8-bit grayscale [17]. The study requires a frontal face, and to obtain them, Haar cascade [18] was used for data extraction. Images that were incorrectly extracted were manually cleaned in the process. Since the non-criminal dataset was larger than the criminal dataset and stored in RGB, it was converted to grayscale, and the image dimensions were reduced to 128*128 pixels using OpenCV. As such, all the duplicates were removed. Additional cleaning was performed on the non-criminal datasets as they contained a mixture of classes other than human faces, which were needed for the experiments. Similarly, Haar cascade was used to extract the frontal face as was applied on the criminal dataset. However, since both the criminal and non-criminal datasets were stored in PNG and JPEG formats, they were converted to JPEG for unit uniformity. These series of activities shall be referred to in this paper as preliminary data pre-processing as further data analysis saw that the data:

(a) Reads from our local storage,
(b) Decodes the JPEG file to grayscale grids of pixels that range only between 0 and 1,
(c) Is converted to floating point tensors,
(d) Is normalized by rescaling to intervals 0 and 1.

All these steps were handled by Keras modules named “ImageDataGenerator”.

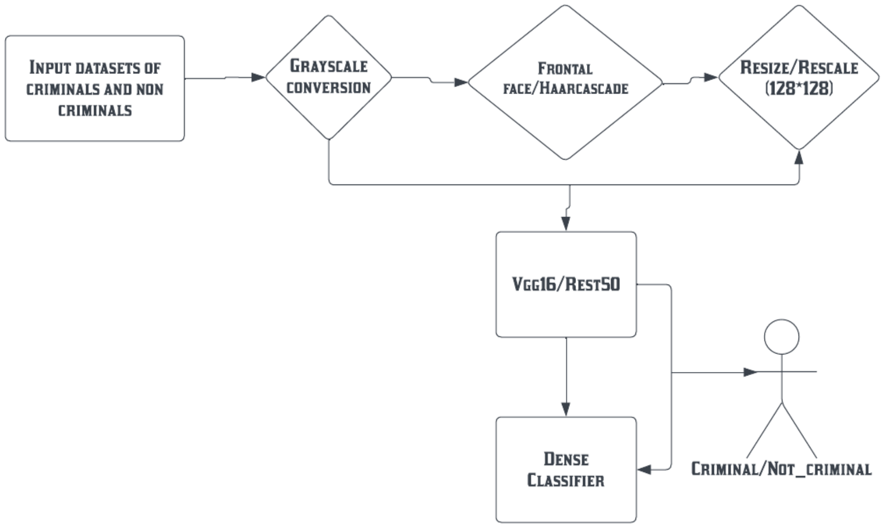


Fig. 3 System diagram

D. Haar Cascade

Haar cascade is a computer vision algorithm used in the detection of objects in images and video [18]. The algorithm was developed by Viola-Jones and can be used in plate number recognition, facial and eye recognition, emotion detection, etc.

The risk error here is put as:

$$E_{Risk}\,(w) = \frac{1}{N}\sum L\big(y_i, f(x_i, w)\big) \qquad (1)$$

The loss function L measures the cost of predicting an output $f(x_i, w)$ for input x_i and the model parameter w, when the corresponding target is $F_i{}^2$. E = Expected risk in the classification.

The cost (penalty) could be a simple quadratic or robust function difference between the desired output (y_i) and the output predicted by the AI model given as: $f(x_i;\ w)$.

Probability Generative Classification

In the experimental test for the model, the above theory was one of the techniques that could help to estimate the probability distribution of the feature vectors for the criminal class:

$$Pk = P(c_H|x) = \frac{P(x \mid C_H)\,p(C_H)}{\sum j\,P(x \mid C_H)\,p(C_H)} = \frac{expLk}{\sum j\,expLj} \qquad (2)$$

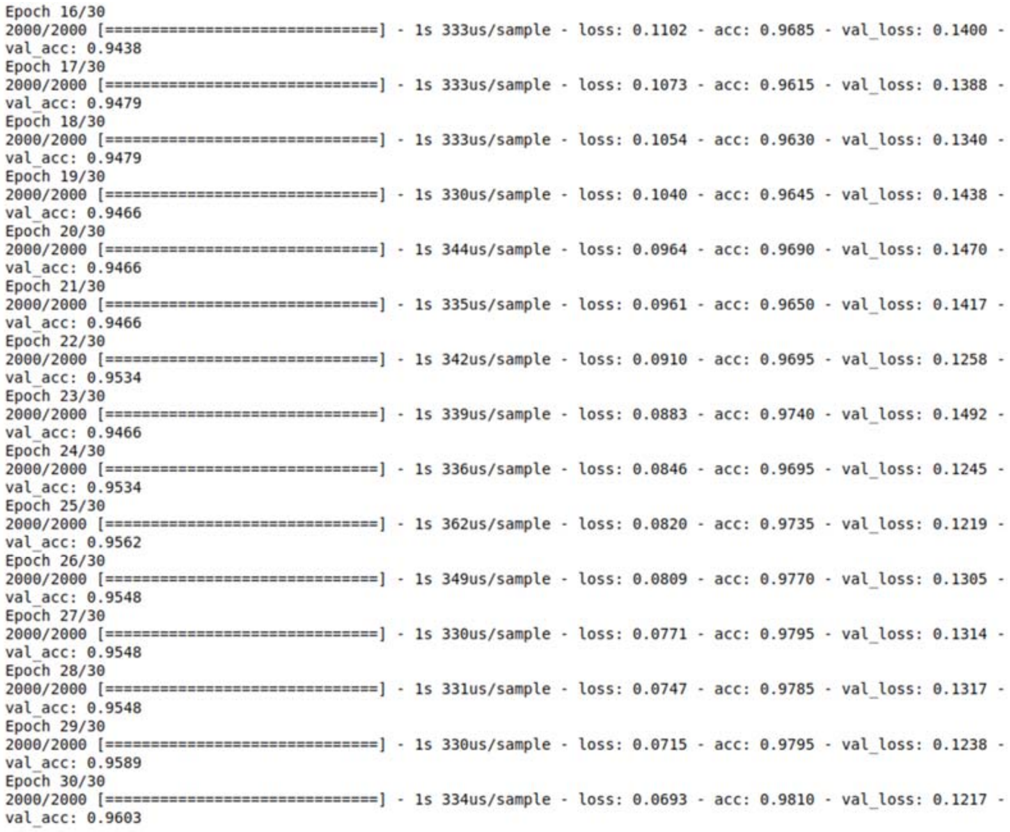
Here, the expected function is the normalized exponential.

$$lk = \log P(x \mid C_H)\ +\ logP(C_H) \qquad (3)$$

Lk = likelihood of sample X being from class C_H. C_H = Criminal class in a dataset.

IV. Results and Discussion

The research was conducted on core i7 HP Pavilion installed with 8 GB RAM, 2 GB Nvidia dedicated graphics card, CUDA 10.1, and Cudnn7.6 on Ubuntu system distribution. The system was implemented using a modern CNN for the prediction of criminals, and was trained and evaluated on the NIST mugshot database. As discussed in the previous sections, it comprises entirely of images of real-time criminals. Also, the effective training of the model required augmenting with non-criminal images which were sourced from the internet. The non-criminal images were first converted to grayscale images after being kept in the same directory as that of the NIST mugshot data that were already in grayscale. The datasets were accurately labelled and resized to 128*128 pixels. The performance summary for VGG16 between the 16th and 30th epochs is shown in Fig. 4.



```
Epoch 16/30
2000/2000 [==============================] - 1s 333us/sample - loss: 0.1102 - acc: 0.9685 - val_loss: 0.1400 -
val_acc: 0.9438
Epoch 17/30
2000/2000 [==============================] - 1s 333us/sample - loss: 0.1073 - acc: 0.9615 - val_loss: 0.1388 -
val_acc: 0.9479
Epoch 18/30
2000/2000 [==============================] - 1s 333us/sample - loss: 0.1054 - acc: 0.9630 - val_loss: 0.1340 -
val_acc: 0.9479
Epoch 19/30
2000/2000 [==============================] - 1s 330us/sample - loss: 0.1040 - acc: 0.9645 - val_loss: 0.1438 -
val_acc: 0.9466
Epoch 20/30
2000/2000 [==============================] - 1s 344us/sample - loss: 0.0964 - acc: 0.9690 - val_loss: 0.1470 -
val_acc: 0.9466
Epoch 21/30
2000/2000 [==============================] - 1s 335us/sample - loss: 0.0961 - acc: 0.9650 - val_loss: 0.1417 -
val_acc: 0.9466
Epoch 22/30
2000/2000 [==============================] - 1s 342us/sample - loss: 0.0910 - acc: 0.9695 - val_loss: 0.1258 -
val_acc: 0.9534
Epoch 23/30
2000/2000 [==============================] - 1s 339us/sample - loss: 0.0883 - acc: 0.9740 - val_loss: 0.1492 -
val_acc: 0.9466
Epoch 24/30
2000/2000 [==============================] - 1s 336us/sample - loss: 0.0846 - acc: 0.9695 - val_loss: 0.1245 -
val_acc: 0.9534
Epoch 25/30
2000/2000 [==============================] - 1s 362us/sample - loss: 0.0820 - acc: 0.9735 - val_loss: 0.1219 -
val_acc: 0.9562
Epoch 26/30
2000/2000 [==============================] - 1s 349us/sample - loss: 0.0809 - acc: 0.9770 - val_loss: 0.1305 -
val_acc: 0.9548
Epoch 27/30
2000/2000 [==============================] - 1s 330us/sample - loss: 0.0771 - acc: 0.9795 - val_loss: 0.1314 -
val_acc: 0.9548
Epoch 28/30
2000/2000 [==============================] - 1s 331us/sample - loss: 0.0747 - acc: 0.9785 - val_loss: 0.1317 -
val_acc: 0.9548
Epoch 29/30
2000/2000 [==============================] - 1s 330us/sample - loss: 0.0715 - acc: 0.9795 - val_loss: 0.1238 -
val_acc: 0.9589
Epoch 30/30
2000/2000 [==============================] - 1s 334us/sample - loss: 0.0693 - acc: 0.9810 - val_loss: 0.1217 -
val_acc: 0.9603
```

Fig. 4 VGG16 performance between the 16th and 30th epochs

As shown in Fig. 4, the training and validation accuracy of the model increases as the log loss decreases. The system was recalled at the weight where the maximum error accuracy was recorded.

Table I shows the data representation for the proposed approach.

TABLE I
DATASET REPRESENTATION

	Criminal	Not_Criminal
Training	1000	1000
Validation	234	500
Test	262	513

After processing the frontal faces containing both criminal and non-criminal images, the datasets were trained with two of the most modern advanced CNN: VGG16 and Resnet50 pre-trained CNN. The feature extraction in both cases used the representation learned by the VGG16 and Resnet 50 paradigms to extract features of criminals and non-criminals from the training dataset. The new features extracted were trained on a densely connected classifier. It may be of interest to note that although a custom CNN was used, the classifier was trained from the beginning. For VGG16 and Resnet50, only a subset of its convolutional layers was used. These layers were the trained with a new densely connected classifier.

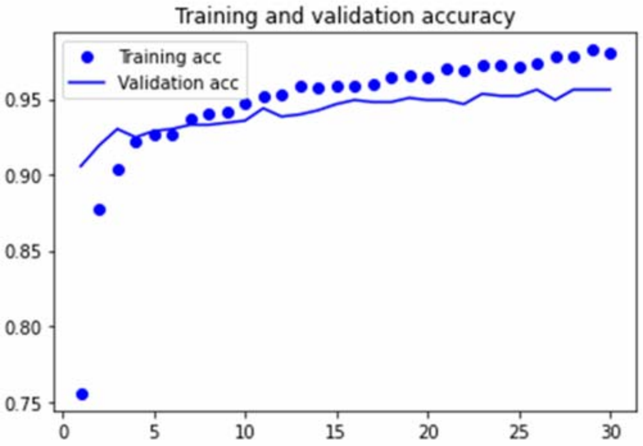


Fig. 5 VGG16 Training and validation accuracy

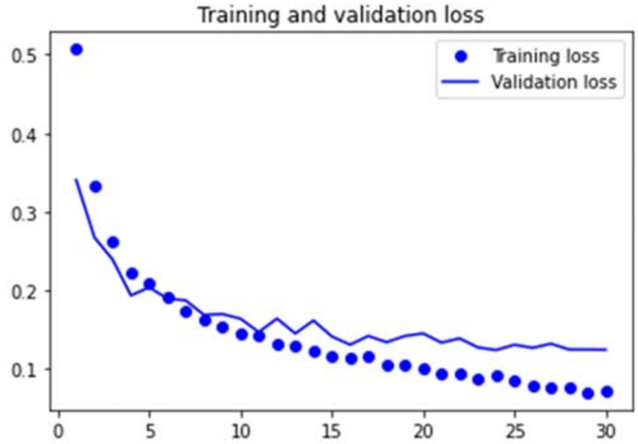


Fig. 6 VGG16 Model training and validation loss

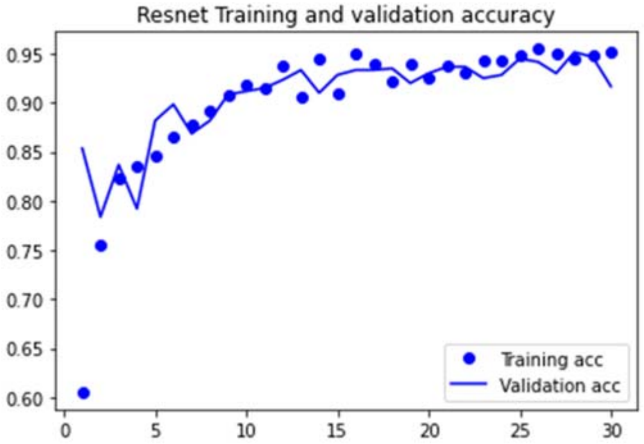


Fig. 7 Resnet50 model training and validation accuracy

The densely connected classifier was trained on top of the VGG16 pretrained model using an RMSprop optimizer with a learning rate of 0.0001. Given that this research is a classification problem, the sigmoid function was used as an output activation function.

The model training and validation accuracy and losses for both models are shown in Figs. 5 and 6 respectively. From the Fig. 5, it could be seen that the model trained well and did not overfit the training data (the model generalized very well). The validation accuracy as shown means that the trained model generalizes very well on an entirely new dataset. At 30 epochs, the training accuracy reaches its maximum of 98.10% and the validation reaches a maximum of 96.03%. Meanwhile, the maximum validation suggests that the model learns all the necessary features of a criminal and a non-criminal without necessarily having to overfit. From the findings obtained in the paper, the validation losses reach their minimum at almost 30 epochs as shown in Fig. 6.

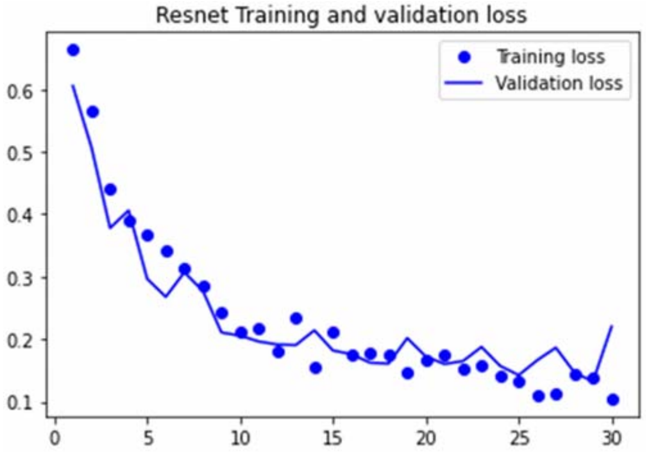


Fig. 8 Resnet50 model loss

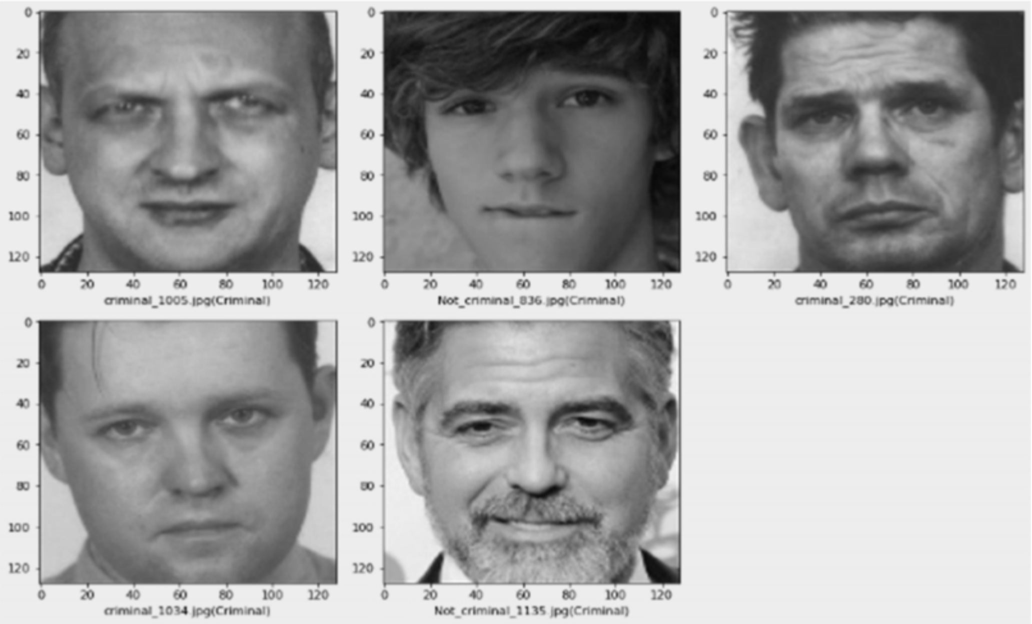

criminal_1005.jpg(Criminal)
Not_criminal_836.jpg(Criminal)
criminal_280.jpg(Criminal)
criminal_1034.jpg(Criminal)
Not_criminal_1135.jpg(Criminal)
0
20
40
60
80
100
120

Fig. 9 Model prediction result - criminal/not-criminal

On the other hand, upon training using Resnet50, a feature extraction method was used. But unlike VGG16 which achieves 98.10% accuracy with a 0.0001 learning rate, Resnet50 on training accuracy reaches its maximum 95.05% between 22-25 epochs and reaches its minimum losses also in 25 epochs. Therefore, from the findings, it could be deduced that VGG16 could be a better model compared to Resnet50 for this research and similar image classification problems.

TABLE II
COMPARISON OF TRAINING AND VALIDATION ACCURACIES FOR VGG16 AND RESNET50 MODELS

Model	Max Training accuracy	Max Validation Accuracy	Minimum Loss Functions
VGG16	98.10	96.03	12.17
Resnet50	95.05	89.5	10.09

As shown in Table III, the results of the study have been compared with some results from related works by other researchers. So, with the 98.10% training accuracy recorded by VGG16, using the feature extraction approach, and with the limited dataset used for the study, we can conclude that the paper is an improvement of some previous works on the use of CNN for criminal face detection.

TABLE III
ERROR ACCURACY COMPARED WITH RELATED WORKS

Author	Max Training Accuracy	Weights	Convnet. Layers
Our work	98.10%	ImageNet	13
Mikolov et al. [20]	97%	Random	2
Navalgund et al. [21]	92%	ImageNet	16
Harsh et al. [8]	99.5%	ImageNet	13

V. CONSEQUENCE

The results above may be contentious or lead to injustice if the system is overestimated. System overestimation is a significant ethical concern and is sensitive in AI research of this type. Conversely, underestimation could result in innocent individuals with facial expressions mistakenly considered as having criminal attributes, making the system less effective.

VI. CONCLUSION

Image classification requires a substantial amount of computational power and energy. With the proposed model, we have been able to successfully demonstrate that it is possible to detect criminals and non-criminals through facial expression or facial recognition. Additionally, the paper has practically demonstrated how CNN, using VGG16, could be effective in criminal detection based on facial expression.

REFERENCES

[1] Alexendar Stec, Diego Klabjan. Forecasting Crime with Deep Learning. https://arxiv.org/abs/1806.01486, 2018.

[2] Lian Duan, Tao Hu, et. al. Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction. Proceed of International Proceedings of the International Conference on Information and Knowledge Engineering (IKE); Athens, 2017.

[3] Jerry Ratcliffe. Crime Mapping: Spatial and Temporal Challenges. 2010.

[4] LA Zebrowitz and JM Montepare, "Social psychological face perception: why appearance matters", Soc Personal Psychol Compass., vol. 2, no. 3, pp. 1497-517, 2008.

[5] Eck, John, Chainey, Spencer, Cameron, James, and Wilson, Ronald. Mapping crime: understanding hotspots. National Institute of Justice special report, 2005.

[6] Gorr Wilpen and Olligschlaeger Andreas. Crime hot spot forecasting: Modeling and comparative evaluation, summary. 2002. URL https://www.ncjrs.gov/pdffiles1/nij/grants/195168

[7] Akinyemi, Olabambo Evelyn. "Community policing in Nigeria: implications for national peace and security." *International Journal of Management, Social Sciences, Peace and Conflict Studies* 4.1 (2021): 469-488.

[8] Harsh Vermal, Siddharth Lotia, Anurag Singh. "Convolutional Neural Network Based Criminal Detection." 2020 IEEE Region 10 Conference, 2020

[9] Mohammed Nakib, et. al. Crime Scene Prediction by Detecting Threatening Objects Using Convolutional Neural Network. https://ieeexplore.ieee.org/document/8465583/, 2018.

[10] Andreas Olligschlaeger. Artificial neural networks and crime mapping. Crime mapping and crime prevention, pages 313–348, 1997.

[11] Saad Albawi and Tareq Abed. Understanding Convolutional Neural Network. ICET, 2017.

[12] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology (London), 195:215–243, 1968.

[13] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193–202, Apr 1980.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, Nov 1998.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. Mcclelland, editors, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations, pages 318–362. MIT Press, Cambridge, MA, 1986.

[16] Sumit Saha. A Comprehensive Guide to Convolutional Neural Networks-The ELI5 Ways, url: https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/, 2018.

[17] NIST Special Database 18. 2010. https://www.nist.gov/srd/nist-special-database-18. Accessed 16 July 2023.

[18] Abhishek Jaiswal. Guide to HaarCascade Algorithm with Object Detection Example, 2022.

[19] Ziwei Liu, Ping Luo, et. al. Large-scale CelebFaces Attributes (CelebA) Dataset. url: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, 2021.

[20] Mikolov, Tomas, et al. "Extensions of Recurrent Neural Network Language model." 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2011.

[21] Navalgund, Umadevi V., and K. Priyadharshini. "Crime Intention Detection System Using Deep Learning." 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET). IEEE, 2018

[22] Murphy Kelvin P. Probabilistic Machine Learning: An Introduction. USA, MIT Press, 2022.