JOURNAL OF BASIC SCIENCE AND ENGINEERING

Secure and Scalable Home Monitoring A Federated Learning Architecture

Avijit Bose[1], Dipannita Ghosh Sneha[2], Ahana Mukherjee[3], Mrinal Mondal[4], Ipsita Dutta[5], Satyajit Chakrabarti[6]

[1] Institute of Engineering and Management, University of Engineering and Management Kolkata, India,

[2] Institute of Engineering as Student, Kolkata, Salt Lake, India,

[3] Institute of Engineering as Student, Kolkata, Salt Lake, India,

[4] Institute of Engineering as Student, Kolkata, Salt Lake, India,

[5] Institute of Engineering as Student, Kolkata, Salt Lake, India,

[6] Institute of Engineering and Management, University of Engineering and Management Kolkata, India

DOI: https://doie.org/10.0307/Jbse.2025203228

ABSTRACT

Video Surveillance is a well-established monitoring technology which has applications in a myriad of sensitive and complex domains. In this following context, we are presenting a novel home surveillance system on pose estimation, fog computing, federated learning and image processing for real time surveillance of small children and old people keeping the sensitivity of personal information in mind. The said system would help in preventing child abuse and monitoring of old age people when no-one is present in the house. It is implemented using areal time analysis algorithm which uses a unique kind of data-preprocessing. We also propose a unique two-level encryption method which would prevent personal data from being divulged and helps in maintaining the privacy of the concerned individual, and as such our solution has an impact on the present society which demands both real-time surveillance and privacy-security. We also demonstrate how our solution can be adapted to the next generation wireless architecture i.e. 5G and 6G when fully deployed.

Keywords: Federated Learning, Home Surveillance, Deep Learning, Pose Estimation.

1. INTRODUCTION

Over the past few years, the number of working people in a family has been increasing [1], because of which children and senior people have to stay at home alone without some- one to look after them. As a result, at present many work- ing couples are looking for alternatives to look after their children and also their aged parents which includes appoint- ing a caretaker or some sort of video surveillance. Typical Video surveillance systems are difficult, time-consuming, and costly to install and operate. The general wireless based solutions allow remote access for users for viewing, review- ing stored information, and controlling the system's compo- nents, done primarily via wireless connection to a remote controller computer (Internet) or via cellular path. How- ever, they do very rarely incorporate measures to secure the streamed data from wireless interception or Internet enabled interception. Apart from video surveillance, the option of manual observance i.e. appointing a caretaker may also lead to many problems. In recent times, the number of child abuse cases and the bad take-care of old aged people have increased considerably. For all such reasons, most people nowadays are trying to move towards modern solutions, which include intelligent surveillance systems to meet their needs.

The increase in the computational power and the advance-
ment of computer vision systems, the research and the ap- plications of surveillance systems has developed a lot. In [2],
[3] and [4], the reader can find the evolution of video surveillance in recent decades. In [5], the author has de- signed a robust video surveillance technique for fall detec-tion of elder people by analyzing the human shape defor- mation. The innovation in computer vision for human behavior understanding is considered as an advancement in the video analysis (including video surveillance) field. Along with video analysis, computer vision techniques havealso evolved in other tasks which include object tracking [6],hand gesture [7], autonomous vehicles [8] among many others. It will be unfair if the mention of deep learning is not mentioned in this context. Deep learning techniques have given the research in computer vision [9] [10] [11] a new perspective. The deep neural network architectures devel- oped in recent years helped in making huge progress in the above mentioned fields. Some of the notable and benchmark architectures are VGG16, VGG19 [12], Inception [13] [14] [15], Res-Net [16] and Spine-Net [17]. Graph is said to be a great tool to build the combinatorial relationship between entities. Relationship between features in an input dataset is also combinatorial in nature - hence an ideal candidate for being modelledas a graph [**??**]. This not only helps in the visualization of feature-to-feature association, but also trigger the important thoughts regarding how to select a subset which is nearly optimal.

The remaining part of this paper are organized as follows:

- Section 2 lays down all the related works.
- Section 3 set the context for the proposed system and details on datasets used, the benchmarking of our dataset with different algorithms.
- Section 4 describes our on premise and federated de- ployment of our model.
- Security and Encryption of data is very important in any surveillance system and Section 5 is dedicated to it.
- Section 6 is all about conclusion.

2. SECTION TWO

2.1 Related Work

With the progress of Deep learning and increase in the com- pute power of edge devices, there has been a lot of devel- opment in the genres of Video surveillance and monitoring in latest years. Karen Simonyan and Andrew Zisserman, in
[18] released a Two Stream Convolutional network that can be used to perform Action recognition in video sequences. One CNN is used to extract spatial information such as the information about the scenes and the objects present in the video sequence, while another CNN, that has been trained on multi-frame dense optical flow vectors performs the tem- poral feature extraction. The SoftMax scores of each CNN stream is then combined by late fusion. Jeff Donahue et al. in
[19] developed an end-to-end trainable recurrent convolu- tional architecture that allows the model to learn long-term dependencies and model complex temporal dynamics. Jun Liu et al. in [20] introduced a new type of LSTM [21] net- work called Global Context-Aware Attention LSTM (GCA- LSTM), which performed 3D action recognition on skeleton data by selectively focusing on the most informative joints in the action sequence. Unlike the regular LSTM, it evalu- ates the global context and iteratively improves the perfor- mance by using a recurrent-attention mechanism. S¸ eymanur Aktı et al. in [22] used Xception for the feature extraction and trained a slightly modified Xception based CNN model for Fight detection. The classification part consists of a Bi- LSTM model with attention modules[[23].

Shu Zhang et al in [24] developed a method to summa- rize entire video sequences by finding the most information segments using a novel Context-aware video summarization (CAVS) framework. A dictionary of video features and spa-

tiotemporal feature correlation graphs are learned which in- dicates how the motions correlate with each other in a global context.

Changchun Long et al in [25] noted the bandwidth star- vation problem that arises with sending entire videos from edge devices to remote IoT servers, and proposed a novel edge computing framework in which the edge devices with their abundant computation capabilities will transmit only the video features to the servers. The proposed framework has three main architectural components. First, The Camera node which can be any static device fixed atop street lamps etc will record a video, break them into chunks and transmit them to edge devices via D2D communication. The Edge nodes, which consist of mobile devices with abundant com- putation capabilities, form a co-operative group, processes the video chunks to perform detection and feature extraction. It then sends the extracted data to the server thus minimizing the bandwith requirement. The last component is the Server. It has powerful computational capabilities and receives data from different Edge groups and performs further video pro- cessing tasks.

R. Cucchiara, et al in [26] developed a ingenious system that used a distributed camera system to track people's move- ment in a house. Their tracking module incorporated a pos- ture detector and could tackle occlusions and shadow-related problems.

Our system is developed around a similar ideology, dif- fering in our method using a more recent state-of the-art ap- proach for Pose Estimation, along with using a new novel en- cryption method and the concept of Federated learning [27].

2.2 Proposed System

We have discussed in the previous section the works related to home surveillance using computer vision systems. In the present society, as per the need of the time, an ideal surveil- lance system should contain the following features:

2.2.1 It should have extremely low-latency for an efficient realtime feedback.

2.2.2 Data should be anonymized and privacy should be maintained.

2.2.3 Bandwidth Optimization.

2.2.4 No effect of lost connectivity.

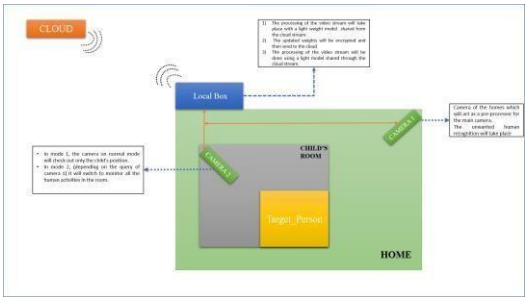
We have designed our system keeping in mind the above four points. Our work has been done using multi-CCTV sys- tems. We have designed a Smart Box which is installed in the home. The cameras are connected to the Smart Box us- ing a wired connection for maximum speed but can be con- figured wireless with a trade-off of latency. The smart box is connected to the cloud through internet connectivity. The typical prototype of the smart box is shown in the figure.

The local connection between the CCTV Cameras and the Smart Box is responsible for low latency. The privacy of personal data is maintained by maintaining a Federated Learning architecture. The efficient bandwidth optimization is achieved by the intelligence of the local hub. Our system will even work in case of lost internet connection because the prediction and the warning system is done completely offline. Due to the federated architecture the local hub does not re- quire any heavy computational hardware which makes our system even cost-effective. The visualization of the whole system can be illustrated from figure:

3. PRIMARY MODEL DEPLOYMENT

In this section the generalized training of the model in the cloud is discussed which will help for identification of abuse and unwanted scenes in the concerned room. There can be various ways to classify the frames but we chose to use a combination of Image Feature Extraction Model and Pose Estimation model for best results. The basic block diagrams

demonstrating the working is shown in Fig. 1, Fig. 2 and Fig.
3. Fig 1. represents how the setup will be made in the house, Fig 2. corresponds to multiple such devices installed in var- ious homes and Fig 3. corresponds to the internal schematic of the Smart-Box.



Vol. 22, No. 1, (2025)
ISSN: 1005-0930

Figure 1. Block Diagram of a home

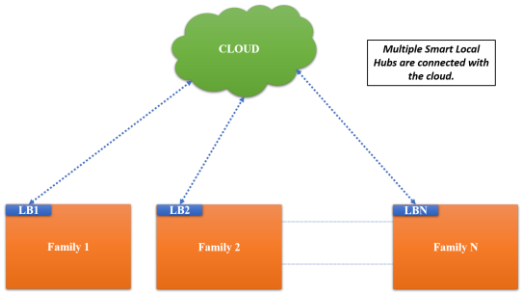


Figure 2. Representation of Multiple Smart-Box

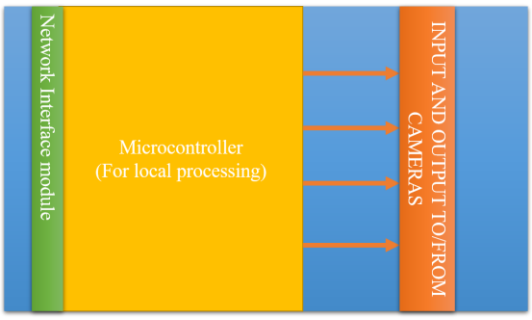


Figure 3. Schematic of the Hardware in Smart Box

3.1 Dataset

We collected about 200 images each belonging to two cat- egories – Fighting and Non-Fighting. Each image contains two humans. The images are of medium to high resolution, since OpenPose doesn't run well if the images are too noisy or of very low resolution. The images have been collected using Scraping, and from various Stock Image websites, and then manually cropped and cleaned so that no image contains no more or less than two humans.

Data Augmentation

Our dataset contains 200 images each of Fighting and Non Fighting classes. In order to get more variety, we per- formed Image Augmentation by performing various opera- tions suck as skewing, flipping, rotating etc. We also tweaked the brightness, contrast and color saturation at random to make the data as varied as possible. We generated a total of 3000 images maintaining an approximate 50-50 distribution among the two categories.

3.1.1 Estimating Pose

After augmenting the data, we ran OpenPose on each of the images. We stored the generated pose points i.e. their x and y co-ordinates along with their confidence scores in a data frame.

For the task of estimating the Pose of various humans present in the images, we used OpenPose[28], a position of the art Pose Estimation Model. They used a novel bottom-up approach in which Part Affinity Fields (PAFs) were used to match identified body parts with corresponding individuals in the image. They note that PAF refinement is critical for maximizing accuracy, while body part prediction refinement is not that important, while increasing the network depth.

A major contribution of OpenPose is their first combined body and foot keypoint detector, which greatly reduced the inference time compared to running both these detectors se- quentially, without any decrease in accuracy. Their project gave rise to the first multi-person real-time 2D pose detection system, including body, foot, hand and facial keypoints. A few examples of the OpenPose results on our dataset have been shown in Fig. 4 and Fig. 5.

3.1.2 Pre-processing

The images in our dataset did not always contain the en- tire body of person or even if it did, some body parts were occluded. As a result, the pose estimation model rarely re- turned all 18 points of the human body. Our data frame thus contained empty cells denoting "no point" of a particular body-part/id being generated by OpenPose for that particu- lar image. To deal with these missing data, we filled them with zeroes, since using any other method such as mean, me- dian etc

weren't logical, and would instead introduce irrel- evant points that weren't actually there. Since we were to pass these points through Support Vector Machines, Logis- tic Classifiers and Neural Networks, we standardized them to have a mean of zero and standard deviation of one.

3.1.3 Training

Our motive was to test and review how Pose Estimation can be used to classify actions. To that end, we decided to train four models - a Support Vector Machine, a Logistic Regres- sor, a basic Dense Neural Network and a Convolutional Neu- ral Network. The first three used only the Pose points as in- put, and was trained solely on it. The last model included the CNN layers of VGG16, which were attached to a Average Pooling Layer and then flattened before being concatenated

Figure 4. OpenPose Results on Fighting Images from our dataset

with output of a second Input layer through which the Pose Point values were passed. The Concatenate layer was then connected to a line of Fully Connected Dense layers, with the Output layer containing a single perceptron with *S igmoid* as the activation function. The structure of our CNN-based model is shown in Fig. 6. Our idea was to see how extracting the image features would help classifying the image along- side the Pose points.

We performed a 65%-35% training and validation split onour augmented set of 3000 images. After training all the four models, we see that the CNN model has the highest accu- racy, most probably because it takes into account both the Pose Points and the Feature vectors extracted from the im- ages. The Accuracy and Loss plots of the Logistic Regressor,Neural Network and CNN-based model have been shown in

Fig. 10.

Figure 5. OpenPose Results on Not-Fighting Images from our dataset

A comparison of all the four models have been shown in Table 1.

4. ON PREMISE MODEL

The on-premise model focuses on using the light-weight ver- sion of the model trained on the cloud which can be used for predicting the live-video streams. The two stage learning in

Model	Input	Training Accuracy	Validation Accuracy
Support Vector Machine	Pose Points	95.21±0.34	88.96±0.27
Logistic Classifier	Pose Points	85.70±0.5	83.12±0.74
Neural Network	Pose Points	95.95±0.62	89.07±0.44
Conv. Neural Network	Pose Points + Image	98.34±0.67	98.99±0.42

Table 1. Comparsion of the SVM, Logistic Regressor, Neural Network and CNN models

this problem was advantageous in many cases it helped in prevention of high computational resources and the concept of centralized training was changed.It is here the frames of the video streams are processed and any anomalous activity is detected.

5. FEDERATED ARCHITECTURE AND IT'S NECESSITY

Federated learning (FL) is a machine learning technique that upskills an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchang- ing them. The approach is different to other traditional intelligent dataset based algorithms where all the data are up- loaded to one common server as well as to some decen- tralized methods where the data are identically distributed among the servers.

FL allows multiple users to design and implement a com- mon machine learning model without compromising the privacy of the data and also an added advantage of heteroge- neous data (data from different organizations will ensure robustness), thus also ensuring the preservation of data access rights.

In any home surveillance technique, it is always necessary to maintain the privacy of the data. We, hence found the necessity of federated learning in this context because federated learning means learning at the edge which ensures data is not sent anywhere. We ensure that using this the personal infor- mation is not shared to any third-party person. FL has a two way communication architecture where it will receive reg- ular updates from the on-cloud model and a feed-back will also be sent to the cloud in the form of encrypted weights. Hence one can say the model is regularly trained without any need of exhaustive dataset. Hence, this architecture solves the problem of dataset collection because datasets for sensi- tive issues has always been a problem. The smart boxes are behaving as the client and the cloud is acting at the server.

6. TWO-STEP ENCRYPTION

Encryption is one of the most essential components in any cloud based working model. In this project we have em-ployed a two-step encryption model to encrypt the weights before they are sent to the cloud for processing. In the first step/layer of encryption, we first extract the fractional part of the weight after the decimal point and then code it by *bitwise left shift(<< operator)* which multiplies the number with any power of 2 as one may desire; preferably not more than 2 to avoid making it large. This is repeated for the integral part of the weight (without the sign) but with a different powerof 2 preferably not more than 4. In the second step/layerof encryption, we convert the coded integral and fractional part obtained in the first step to a single string literal along with the information of the sign it carries. We first appendedthe coded fractional part to *D* indicating the decimal part and then depending on the sign we appended *S1N* or *S0P* for neg- ative and positive values respectively followed by the coded integral part. The *S1N* indicates the sign bit being 1 for nega- tive and *S0P* indicates the sign bit being 0 for positive. After the two steps of encryption, the coded form of a single weight stands as:-

D(coded fractional part)(S1N/S0P)(coded integral part)

In the decryption portion, we start unwinding the code in the reverse order. We first extract the fractional and integral parts separately and then we use the *bitwise right shift(>> operator)* which divides the number by the same power of 2 we used in the encryption. Then depending on the sign we create the original number in the string form first and then convert it to a numerical datatype.

7. DISCUSSION

In this paper we have discuss and analyzed how federated learning can be used for surveillance systems. One of the disadvantage of Federated learning is that it requires many users so that the training and update is more robust. Our sys- tem also requires when there is multiple no. of users. High internet connectivity is also required for better results but in- stead of these constraints our system holds an important rev- olution in home surveillance.

8. CONCLUSION

We have demonstrated how pose points of humans in images can be used to an extent to classify the actions in said im- ages and how federated learning can be used to establish a secure surveillance system in the home. Although, it alone does not produce any state of the art result, it opens up new possibilities and ways to improve existing models. We plan to extend our implementation to work on videos by using a combination of Convolutional and Sequence Models.

9. ACKNOWLWDGEMENT

I want to appreciate my supervisor Prof. Dr. Satyajit Chakraborti for his continuous guidance and support. Without his tremendous encouragement it would be impossible for me tocomplete my research on time.

REFERENCES

9.1 Catalyst: Workplaces that work for Women. Working parents: Quick take, 2017.

9.2 Niels Haering, Peter Venetianer, and Alan Lipton. The evolution of video surveillance: An overview. *Mach. Vis. Appl.*, 19:279–290, 10 2008.

9.3 Roberto Vezzani and Rita Cucchiara. Video surveil- lance online repository (visor): an integrated frame- work. *Multimedia Tools Appl.*, 50:359–380, 11 2010.

9.4 Ying Li, Lisa Brown, Arun Hampapur, Max Lu, An- drew Senior, and Chiao-Fe Shu. Ibm smart surveil-lance system (s3): Event based video surveillance sys- tem with an open and extensible framework. *Mach. Vis. Appl.*, 19:315–327, 10 2008.

9.5 C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Robust video surveillance for fall detection based on human shape deformation. *IEEE Transactions on Cir- cuits and Systems for Video Technology*, 21(5):611– 622, 2011.

9.6 Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *CoRR*, abs/1905.05055, 2019.

9.7 H. Cheng, L. Yang, and Z. Liu. Survey on 3d hand gesture recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(9):1659–1673, 2016.

9.8 Joel Janai, Fatma Gu¨ney, Aseem Behl, and An- dreas Geiger. Computer vision for autonomous vehi- cles: Problems, datasets and state-of-the-art. *CoRR*, abs/1704.05519, 2017.

9.9 Theodoros Georgiou, Youfang Liu, Wei Chen, and Michael S. Lew. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Mul- timedia Information Retrieval*, pages 1 – 36, 2019.

9.10 Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, and Matti Pietik¨ainen. Deep learning for generic object detection: A survey. *CoRR*, abs/1809.02165, 2018.

9.11 Shengyong Chen, Yingkun Xu, Xiaolong Zhou, and Fenfen Li. Deep learning for multiple object tracking: A survey. *IET Computer Vision*, 13, 01 2019.

9.12 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recogni- tion. *arXiv 1409.1556*, 09 2014.

9.13 Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Ser- manet, Scott E. Reed, Dragomir Anguelov, Dumitru Er- han, Vincent Vanhoucke, and Andrew Rabinovich. Go- ing deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

9.14 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

9.15 Christian Szegedy, Sergey Ioffe, and Vincent Van- houcke. Inception-v4, inception-resnet and the im- pact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

9.16 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

9.17 Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghi- asi, Mingxing Tan, Yin Cui, Quoc V. Le, and Xiaodan Song. Spinenet: Learning scale-permuted backbone for recognition and localization. *ArXiv*, abs/1912.05027, 2019.

9.18 Karen Simonyan and Andrew Zisserman. Two- stream convolutional networks for action recognition invideos. *CoRR*, abs/1406.2199, 2014.

9.19 Jeff Donahue, Lisa Anne Hendricks, Sergio Guadar- rama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and de- scription. *CoRR*, abs/1411.4389, 2014.

9.20 J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3671– 3680, 2017.

9.21 Sepp Hochreiter and Ju¨ rgen Schmidhuber. Long short- term memory. *Neural computation*, 9:1735–80, 12 1997.

9.22 S¸ eymanur Aktı, Go¨ zde Ays¸e Tatarog˘ lu, and Hazım Ke- mal Ekenel. Vision-based fight detection from surveil- lance cameras. *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2019.

9.23 Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Ben- gio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

9.24 S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury. Context- aware surveillance video summarization. *IEEE Trans- actions on Image Processing*, 25(11):5469–5478, 2016.

9.25 Changchun Long, Yang Cao, Tao Jiang, and Qian Zhang. Edge computing framework for cooperative video processing in multimedia iot systems. *IEEE Transactions on Multimedia*, 20:1126–1139, 05 2018.

9.26 Maria Valera and Sergio Velastin. Intelligent distributed surveillance systems: A review. *Vision, Image and Sig- nal Processing, IEE Proceedings* -, 152:192 – 204, 05 2005.

9.27 H. Brendan McMahan, Eider Moore, Daniel Ram- age and Blaise Agu¨ era y Arcas. Federated learn- ing of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

9.28 Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018.

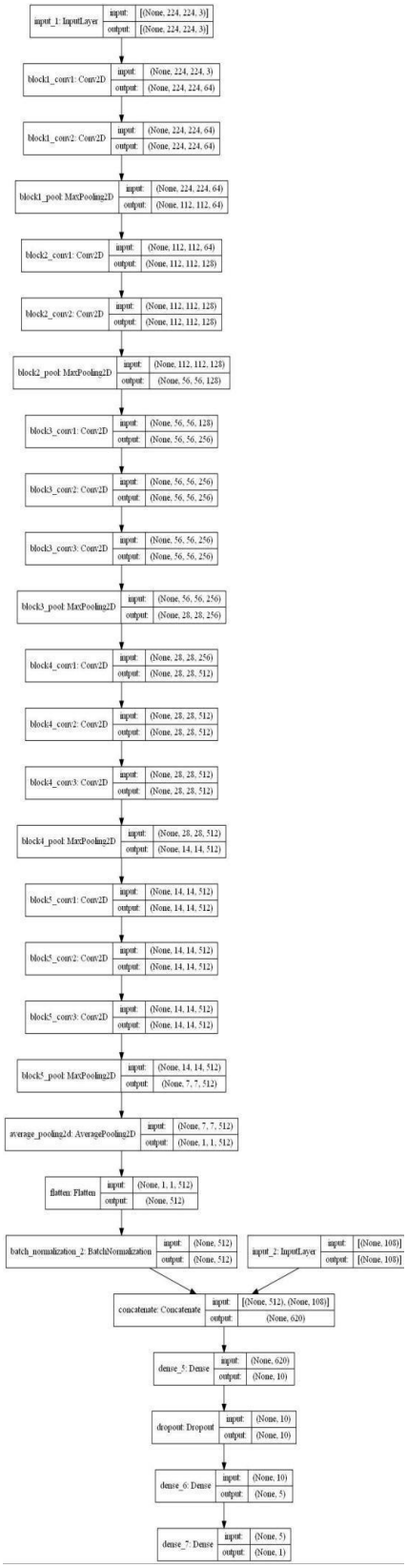
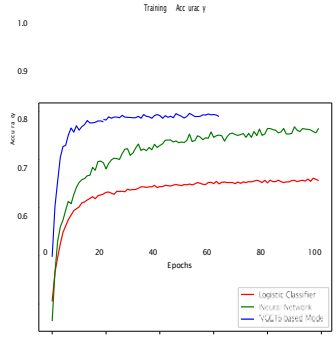
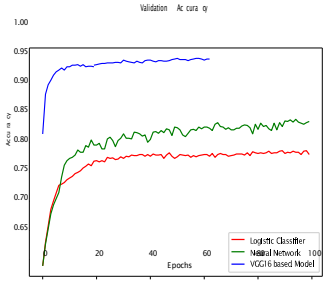
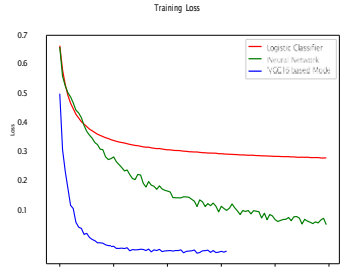


Figure 6. Our VGG16 based CNN model

Vol. 22, No. 1, (2025)
ISSN: 1005-0930







Vol. 22, No. 1, (2025)
ISSN: 1005-0930

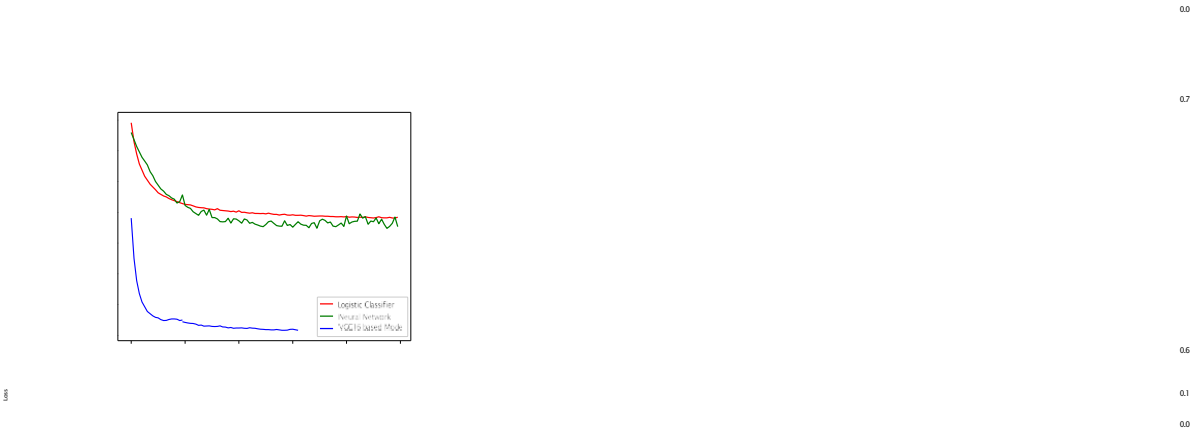


Figure 7. Accuracy and Loss plots of the Logistic Regressor, Neural Network and CNN model

Vol. 22, No. 1, (2025)
ISSN: 1005-0930