COVID-19: Improving the accuracy using data augmentation and pre-trained DCNN Models

Saif Hassan[1], Abdul Ghafoor[2], Zahid Hussain Khand[3], Zafar Ali[4], Ghulam Mujtaba[5], Sajid Khan[6]

Center of Excellence for Robotics, Artificial Intelligence, and Blockchain,
Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

Summary
Since the World Health Organization (WHO) has declared COVID-19 as pandemic, many researchers have started working on developing vaccine and developing AI systems to detect COVID-19 patient using Chest X-ray images. The purpose of this work is to improve the performance of pre-trained Deep convolution neural nets (DCNNs) on Chest X-ray images dataset specially COVID-19 which is developed by collecting from different sources such as GitHub, Kaggle. To improve the performance of Deep CNNs, data augmentation is used in this study. The COVID-19 dataset collected from GitHub was containing 257 images while the other two classes normal and pneumonia were having more than 500 images each class. There were two issues whike training DCNN model on this dataset, one is unbalanced and second is the data is very less. In order to handle these both issues, we performed data augmentation such as rotation, flipping to increase and balance the dataset. After data augmentation each class contains 510 images. Results show that augmentation on Chest X-ray images helps in improving accuracy. The accuracy before and after augmentation produced by our proposed architecture is 96.8% and 98.4% respectively.

Keywords:
COVID-19, data augmentation, Deep CNNs, Chest X-ray dataset

1. Introduction

Coronavirus is a recently discovered an infectious disease [1]. First case of COVID-19 reported in December 2019 in Wuhan, China and rapidly spread all over the globe [7]. World health Organization (WHO) announced COVID-19 as a pandemic in March of 2020. The virus responsible for COVID-19 disease given official names: Coronavirus disease (COVID-19) for Disease by WHO and International Classification of Diseases (ICD) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) for virus by International Committee on Taxonomy of Viruses (ICTV) [3]. As of July 19, 2020 13876441 confirmed case of covid19 and 593087 death reported around world with approximately 4.27 death rate [2]. To slowdown the impact of disease, countries around globe use different strategies such as: lockdown cities, travel ban and peoples also avoided to go outside. These practice somehow helped to decrease the impact but create many other economics disasters. COVID – 19 devastated health as well as many other economic sectors globally and leave millions of people jobless. According to [5], top affected sectors are Manufacturing, Travel and transportation, Retail, Energy & resources and High tech & telecommunication. To fill these losses, countries around globe announce the economics packages for industries as well as individuals [4].

COVID-19 patients have been reported with many symptoms from mild to severe illness. These symptoms can appear between 2 to 14 days after contacting to COVID-19 [6]. These symptoms incudes: Illness or chills, cough, breathing problems, tiredness, pain in body, less smell feelings, Aching, rhinitis, vomiting and Diarrhea [6]. To date no medicine is available for COVID-19 but many vaccines are being tested in medical trials [6]. In current situation experts suggest the more and more testing and proper screening of infected people [8]. Identify the stage of COVID 19 patients they are in mild or severe stage. The basic testing tool used to detect the COVID-19 patient is reverse transcriptase–polymerase chain reaction (RT-PCR) [9]. This method uses the nasopharyngeal swab for testing.

Another method has been applied to diagnose COVID-19 is radiography images (CXR or chest X-ray and CT computerized tomography) [8]. In this method Radiologist look for the SARS-CoV-2 visuals in CT OR CXR report. The study [10] has reported that 41 laboratory confirmed patient were admitted in hospital and abnormalities diagnosed in chest CT images of all patients. Another study has analyzed the chest CT of CONVID-19 patients and study found that Pleural effusions and lymphadenopathy were absent in all patients [11]. Many other studies [12, 13, 14, 15, 16] have concluded that

Manuscript received July 5, 2024
Manuscript revised July 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.7.20

abnormalities have detected in radiography images of most of the laboratory confirmed coronavirus patient. Some studies have also suggested that radiography images method should be used for COVID-19 Screening [13].

CXR or CT has many advantages over RT-PCR method such as: heavily affected place or countries face lack of testing kits and testing laborites but these countries has radiography laborites. [8]. RT-PCR sample have been the taken from the person so there are chances of virus transmission from person to person but we can set the Isolation room fix CT scanner on right location and can take images without contacting with a person [14]. Main problem with these CXR or CT method is need of radiologist's experts to understand the images. This problem can be solved by developing the deep learning based systems to interpret CT or CXR report and classify them that person has COVID-19 or not [8]. This model can be used as screening tool for COVID-19.

The work done by Prabira et al [20] concluded that the CNN model based on ResNet50 among other pre-trained CNN architectures outperformed on Chest X-ray images collected from GitHub and Kaggle. A network designed by Maghdid et al. [17] produced good accuracy but the dataset contains less amount of data, this network consists of only 16 layers to predict COVID19 using Chest X-ray images as well as CT Scans. Many researchers have worked on either of the datasets (Chest X-ray or CT Scans) such as work done by Shuai et al. [18] designed transfer learning model based on Inception and predicted with 89.5% accuracy on CT Scans dataset. Apostolopoulos et al. [19] trained pre-trained architectures on Chest X-ray images to predict normal, covid19 and pneumonia cases and produced accuracy of 96.78% but the data was imbalanced like COVID19, normal and pneumonia images were 224, 504 and 700 respectively.

The work in this study outperforms the literature by using data augmentation to increase and balance the dataset for all classes.

2. Methodology

Deep Learning approaches works better on large datasets rather than small datasets. So for applying deep learning approaches such as convolutional neural networks, dataset with large instances is required. We know that there are large number of COVID19 patients globally but the data for those patients is not available publically. We hardly found COVID19 dataset with less instances, so we performed data augmentation techniques including rotation, flipping. Testing

2.1 Dataset

In this paper, Chest X ray images dataset is used. The dataset is developed by collecting from different sources such as github and kaggle. The dataset contains three classes, COVID19, pneumonia and normal. We found the dataset publically available for COVID19 with 257 images on GitHub [21] till April, 2020. Kaggle Dataset for normal and pneumonia patients contains 5k images for each but for balancing the dataset, we sampled 257 images for each. Figure 1 represents sample image from each class.

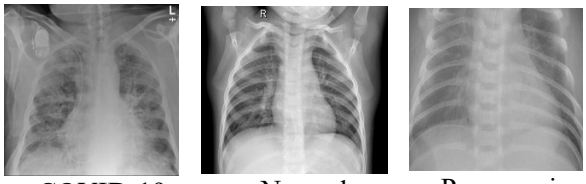


Fig. 1. Sample image from each category

Original COVID19 dataset contains 257 images only. As deep convolutional neural nets requires lot of data for achieving good results, we performed data augmentation as the work [22] suggests that data augmentation in x-ray images approach outperforms.

2.2 Data Augmentation

Data Augmentation is a technique to artificially produce more instances of data from existing instances in the dataset. As discussed above, COVID19 Chest X-rays [21] is dataset having less images. So data augmentation is used for expanding the size of data. Data Augmentation is used primarily before training the model. We have expanded the size of COVID19 Chest X-ray dataset by augmenting existing images. After augmentation, COVID19 class contains 510 images. Augmentation is performed for COVID19 class because other two classes, pneumonia and normal already contain more about 5k images, so we picked 510 images from both classes.

Data augmentation is performed in a way that, we iterate over COVID19 class images and randomly choose any existing image. Then more images are generated by applying rotation and flipping on chosen image.

Image Augmentation through Rotation

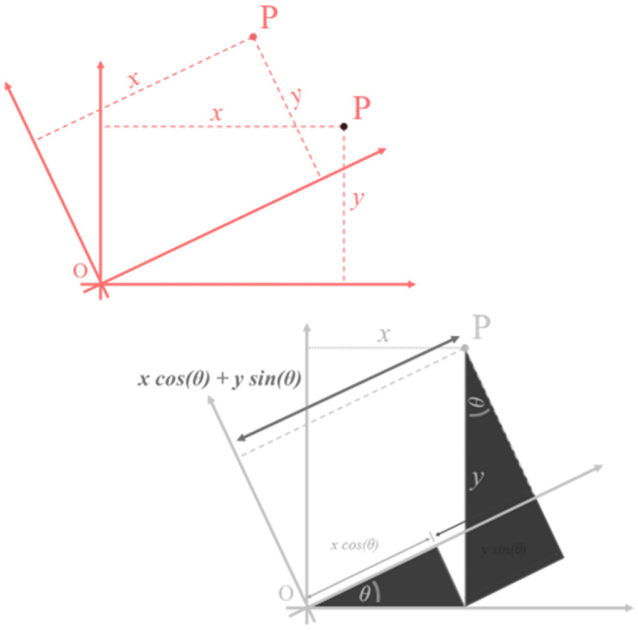
Rotation [28] is one of the Data augmentation technique which is done by rotating the image between 1 to 360 degrees. In this work, image rotation is applied only

by rotating between -30 to 30 degree, as we observed after many experiments, this could be useful for Chest .X-ray Images.

An Image has pixels and every pixel has x and y values showing its position on two orthogonal axis from origin O. We rotate an image over this origin.

What we have to do first is to pick x and y values for each pixel and rotate on the degree between -30 and 30. The new values for x and y will be placed on new location.

So we rotate an image by angle θ. Values for θ are used between -30 and 30.



The matrix for rotation of an image is shown below.

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Image Augmentation through Flipping

Flipping augmentation is of two types vertical and horizontal — in this work only horizontal flip is used as researchers [23] has proved this useful on many of the datasets such as ImageNet comparing to vertical.

Horizontal flip is applied on the original covid19 images; some examples are shown in figure 2. The matrix for horizontal flip, which is applied on dataset is as follows:

$$Horizontal\ Flip\ Matrix: \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

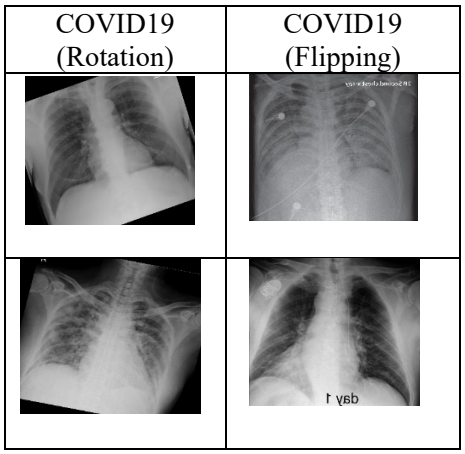


Fig. 2. Sample COVID19 generated images after augmentation

After Data Augmentation process, COVID19 class in dataset increased from 257 images to 510 images total. The number of total instances in dataset are 1530.

The following graph in figure shows the number before and after augmentation.

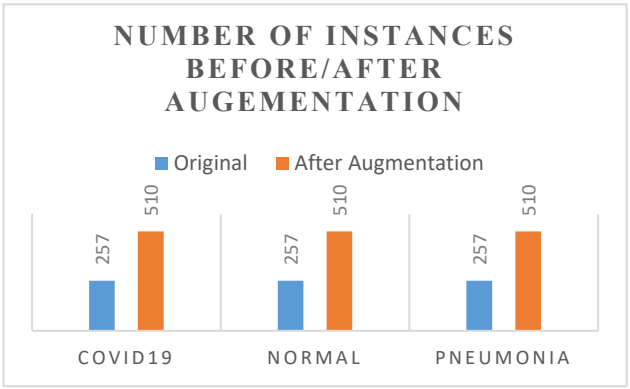


Fig. 3. Graph for number of before and after augmentation

2.3 Network Configuration

In this study, two datasets are developed, one is original data collected from different sources mentioned above and other is with augmentation. Figure 4 shows architecture of our model in which we used simple CNN and pre-trained deep CNN architectures on both the datasets. We first performed augmentation only on COVID19 Chest X-ray (CXR) images to increase the

dataset then trained on pre-trained CNN models such as VGG16, DenseNet121, MobileNet and NASNetMobile.

Overall the model works like an image is fed to network which will classify as either of three classes (COVID19, Pneumonia or Normal).

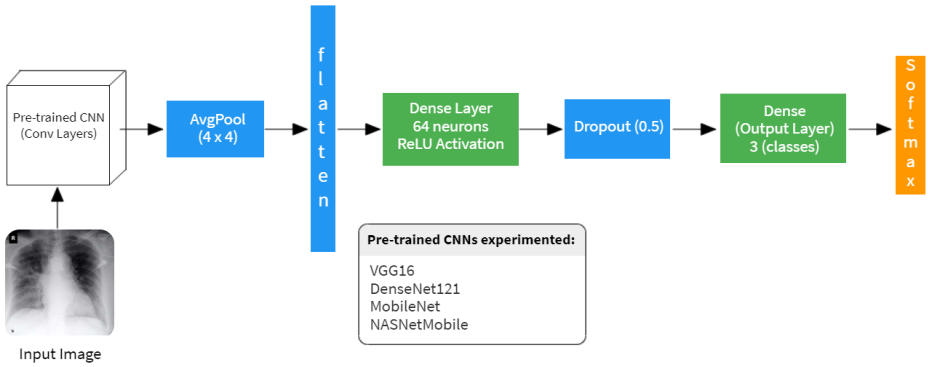


Fig. 4. Architecture of our model

Using keras library, we trained VGG16, DenseNet121, MobileNet and NASNetMobile, ensuring the head FC layers' sets are left off by setting value false to param include_top. Images are resized to 224 x 224 with three channels. Next we constructed customized model that will be placed on the base model (VGG16, DenseNet121 and etc.). Next we constructed Fully-connected layer head that consists of average pooling of 4x4, Flatten, Dense with 64 neurons followed by relu activation function, dropout of 0.5 and finally dense output layer with softmax activation function, hyper-params details are given in Table I. Then we iterated loop over all layers in VGG16, DenseNet121, MobileNet121 and NASNetMobile so that only fully-connected layer head will be trained. After designing the model, model is compiled with optimizer as **Adam** and learning rate decay of **1e-3** and used loss as "**categorial cross_entropy**" for more than two classes.

Finally, we call keras fit generator method to train our model by setting epochs to 25, BatchSize to 8. We performed experiments shown in experiments section.

Table. I. Details of neural net model hyper-parameters

Type	Hyper-parameters	Value selected
Architec-ture	Hidden Layers	Pre-trained + avg_pool + FC
	Dropout	0.5
	Filter Size	3x3
	Pooling Layer	Max-Pooling (4x4)
Training	Activation Functions	Softmax (output layer)
	Batch Size	8
	Epochs	25
	Steps per epoch	train_size / batch_size
Learning	Optimizer	Adam
	Learning rate	1e-3

2.4 Experiments and Results

Dataset Split

Dataset is divided as traditional train, val and test with 80, 10 and 10 ratios respectively. Table II shows the number of instances in each class train, val and test for original dataset and after augmentation. Dataset is increased almost double of original after augmentation.

Table. II. Number of train, val and test instances

Dataset	Train	Val	Test	Total
Before Augmentation	205	26	26	257
After Augmentation	408	51	51	510

Accuracy/Loss for Training and Validation

After many experiments, it is noticed that MobileNet along with fully-connected layer outperforms on original dataset (before augmentation) with 96.8% accuracy as training and 97.2% as validation as shown in Table III. On the other end, the training and validation accuracy on dataset after augmentation is this and that respectively, achieved by VGG16.

Table. III. Accuracy of models before and after augmentation

Model	Before Augmentation		After Augmentation	
	Training Accuracy	Val Accuracy	Training Accuracy	Val Accuracy
Simple CNN Model	92.6%	94.8%	90.4%	89.2%
VGG16	93.3%	89.7%	93.1%	92.4%
DenseNet121	93.8%	91.0%	95.6%	92.2%

MobileNet	**96.8%**	**97.2%**	**98.4%**	**98.1%**
NASNetMobile	96.3%	93.5%	95.7%	94.0%

Looking at Figure 5, graphs of loss/accuracy for comparing training and validation, we can easily observe almost all the CNN models performed better but the MobileNet outperforms all others on both datasets (before and after augmentation).

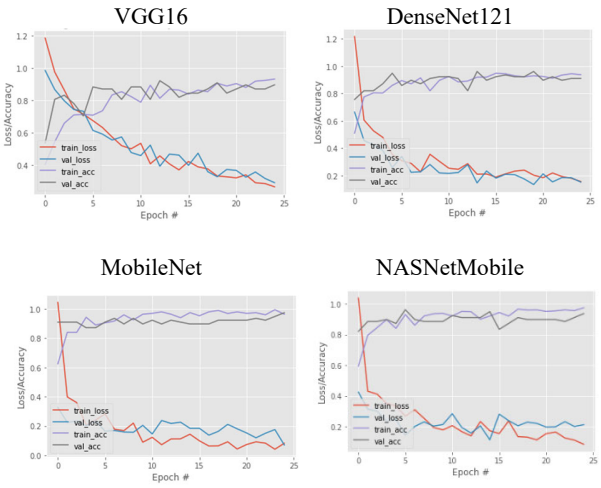
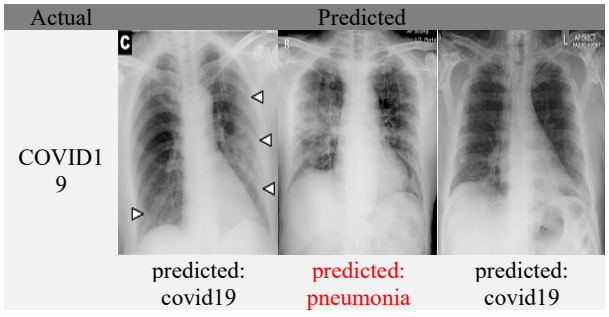


Fig. 5. Training/Validation Accuracy and Loss Graphs on Original dataset

Prediction on unseen data

After training the model, it has been tested on unseen data. Predictions for the test data on Chest X-ray images are shown in Figure 6 that shows some sample results predicted by proposed model. Image is fed to trained model, model returns probability mapped with each class, because we use *softmax* as activation function on output layer. Finally, class with highest probability is retrieved using *argmax*.



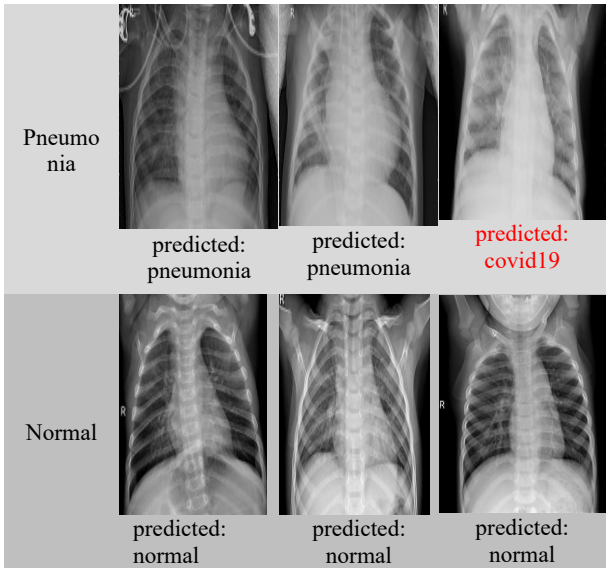


Fig. 6. Results on test data

3. Conclusion

This study focuses on improving the accuracy on Chest X-ray images dataset using data augmentation and Deep CNNs. Results clearly show that the work in this paper has achieved the promising results on classifying the COVID-19, Pneumonia and Normal patients based on CXR images.

The augmentation included rotation on angle θ where θ is (-30, 30) and horizontal flip. Proposed method achieves the accuracy of 98.1% on augmented dataset. Future work includes increasing the dataset and deploying this system with the concern of Doctors.

References

[1]. https://www.who.int/health-topics/coronavirus
[2]. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200718-covid-19-sitrep-180.pdf?sfvrsn=39b31718_2
[3]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it
[4]. Fernandes, Nuno. "Economic effects of coronavirus outbreak (COVID-19) on the world economy." *Available at SSRN 3557504* (2020).
[5]. https://www.statista.com/statistics/1106302/coronavirus-impact-index-by-industry-2020/
[6]. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[7]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses#:~:text=protect
[8]. Wang, Linda, and Alexander Wong. "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images." *arXiv preprint arXiv:2003.09871* (2020).
[9]. Wang, Wenling, et al. "Detection of SARS-CoV-2 in different types of clinical specimens." *Jama* 323.18 (2020): 1843-1844.
[10]. Huang, Chaolin, et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." *The lancet* 395.10223 (2020): 497-506.
[11]. Ng, Ming-Yen, et al. "Imaging profile of the COVID-19 infection: radiologic findings and literature review." *Radiology: Cardiothoracic Imaging* 2.1 (2020): e200034.
[12]. Guan, Wei-jie, et al. "Clinical characteristics of coronavirus disease 2019 in China." *New England journal of medicine* 382.18 (2020): 1708-1720.
[13]. Ai, Tao, et al. "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases." *Radiology* (2020): 200642.
[14]. Rubin, Geoffrey D., et al. "The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society." *Chest* (2020).
[15]. Nair, A., et al. "A British Society of Thoracic Imaging statement: considerations in designing local imaging diagnostic algorithms for the COVID-19 pandemic." *Clinical Radiology* 75.5 (2020): 329-334.
[16]. Adam Jacobi, A. B., Michael Chung & Eber, C. Portable chest x-ray in coronavirus disease-19 (covid 19): A pictorial review. Clin. Imaging (2020).
[17]. Maghdid, Halgurd S., et al. "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms." *arXiv preprint arXiv:2004.00038* (2020).
[18]. Wang, Shuai, et al. "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)." *MedRxiv* (2020).
[19]. Apostolopoulos, Ioannis D., and Tzani A. Mpesiana. "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks." Physical and Engineering Sciences in Medicine (2020): 1.
[20]. Sethy, Prabira Kumar, and Santi Kumari Behera. "Detection of coronavirus disease (covid-19) based on deep features." Preprints 2020030300 (2020): 2020.
[21]. https://github.com/ieee8023/covid-chestxray-dataset
[22]. Madani, Ali, et al. "Chest x-ray generation and data augmentation for cardiovascular abnormality classification." Medical Imaging 2018: Image Processing. Vol. 10574. International Society for Optics and Photonics, 2018.
[23]. Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of Big Data 6.1 (2019): 60

Saif Hassan received the B.S degree in Computer Science from Sukkur IBA University, Pakistan in 2016 and M.S degree in Computer Science from Mohammad Ali Jinnah University, Karachi, Pakistan (Gold Medalist in MS). He is currently teaching at Sukkur IBA University. He has teaching experience of more than 4 years. His specialization areas include Deep Learning and Computer Vision. Mr. Saif's awards include PM ICT (R&D) Fund Scholarship for Four Years B.S Program, Gold Medalist in Masters, online certifications and many appreciation shields and certificates for conducting workshops/seminars on different technologies including python, machine learning, game development and deep learning.

Abdul Ghafoor is MS Research Scholar at the Department of Computer Science, Sukkur IBA University, Pakistan. He is also working as a Subject Specialist, Computer Science at IBA-IET Khairpur, Pakistan. He received her Bachelor's degree in Computer Science from Sukkur IBA University, Pakistan in 2016. His research interests include Deep Learning, NLP and Computer Vision.

Zahid Hussain Khand is currently working as Registrar, Sukkur IBA University. He has been associated with Sukkur IBA University since 2003. His field of research includes Information and Communication Technology, Agri-tech, and Smart-tech. Mr. Khand teaches various courses such as Network Security, Computer Networks, Data Communication, Internet of Things, and Research Methods. He has authored or co-authored several articles in academic journals indexed in well-reputed databases.

Zafar Ali is currently working as an ERP Manager at Sukkur IBA University. He has been associated with Sukkur IBA University since 2008. His field of research includes Information and Communication Technology, ERP, Visual Programming, Databases, Information Retrieval, and Distributed Databases.

Ghulam Mujtaba is currently working as an Associate Professor at the Department of Computer Science, Sukkur IBA University. He is also working as the Director of the Center of Excellence for Robotics, Artificial Intelligence, and Blockchain (CRAIB). He has been associated with Sukkur IBA University since 2006. He received his Doctorate in the field of

Computer Science from the University of Malaya, Kuala Lumpur, Malaysia in 2018. His field of research includes artificial intelligence, machine learning, online social networking, text mining, text classification, image classification, and deep learning. Dr. Mujtaba teaches various courses such as Computer Programming, Object-Oriented Programming, Data Science, Machine Learning, Natural Language Processing, Deep Learning, and Advanced Research Methods. He has authored or co-authored several articles in academic journals indexed in well-reputed databases.

Sajid Khan is currently working as an Assistant Professor at the Department of Computer Science, Sukkur IBA University He received the B.S. degree in telecom engineering from FAST-NUCES University, Pakistan, in 2011, and the M.S. leading to Ph.D. degrees in electronics and communication engineering from Hanyang University, Ansan, South Korea.