IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 11, 2022, accepted February 25, 2022, date of publication March 8, 2022, date of current version March 16, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3157316

A Hybrid VDV Model for Automatic Diagnosis of Pneumothorax Using Class-Imbalanced Chest X-Rays Dataset

TAHIRA IQBAL[1], ARSLAN SHAUKAT[1], MUHAMMAD USMAN AKRAM[1], ABDUL WAHAB MUZAFFAR[2], ZARTASHA MUSTANSAR[3], AND YUNG-CHEOL BYUN[4]

[1]Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
[3]Research Centre for Modelling and Simulation (RCMS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[4]Department of Computer Engineering, Jeju National University, Jeju 690756, South Korea

Corresponding authors: Arslan Shaukat (arslanshaukat@ceme.nust.edu.pk) and Abdul Wahab Muzaffar (a.muzaffar@seu.edu.sa)

ABSTRACT Pneumothorax, a life-threatening disease, needs to be diagnosed immediately and efficiently. The prognosis, in this case, is not only time-consuming but also prone to human errors. So, an automatic way of accurate diagnosis using chest X-rays is the utmost requirement. To date, most of the available medical image datasets have a class-imbalance (CI) issue. The main theme of this study is to solve this problem along with proposing an automated way of detecting pneumothorax. To find the optimal approach for CI problem, we first compare the existing approaches and find that under-bagging method (referred as data-level-ensemble formed by creating subsets of majority class and then combining each subset with all samples of minority class) outperforms other existing approaches. After selection of best approach for CI problem, we propose a novel framework, named as VDV model, for pneumothorax detection from highly imbalance dataset. The proposed VDV model is a complex model-level ensemble of data-level-ensembles and uses three convolutional neural networks (CNN) including VGG16, VGG-19, and DenseNet-121 as fixed feature extractors. In each data-level-ensemble, features extracted from one of the pre-defined CNN architectures are fed to support vector machine (SVM) classifier, and output is calculated using the voting method. Once outputs from the three data-level-ensembles (corresponding to three different CNN architectures as feature extractor) are obtained, then, again, the voting method is used to calculate the final prediction. Our proposed framework is tested on the SIIM ACR Pneumothorax dataset and Random Sample of NIH Chest X-ray dataset (RS-NIH). For the first dataset, 85.17% Recall with 86.0% Area under the Receiver Operating Characteristic curve (AUC) is attained. For the second dataset, 90.9% Recall with 95.0% AUC is achieved with a random split of data while 85.45% recall with 77.06% AUC is obtained with a patient-wise split of data. The comparison of our results for both the datasets with related work proves the effectiveness of proposed VDV model for pneumothorax detection.

INDEX TERMS Class-imbalance, chest X-rays, classification, deep learning, ensemble, machine learning, pneumothorax, under-bagging.

I. INTRODUCTION

Pneumothorax can be interpreted as a life-threatening condition which occurs due to the collapse of the respiratory system. The disease occurs as the air present inside the lungs is leaked to the space between chest walls and lungs. Due to this air, pressure is exerted on the lungs, and it becomes difficult for the person to breathe as the lungs cannot expand properly, thus the respiratory system collapses. Symptoms include shortness of breath and sudden pain in the chest. In some cases, these symptoms can be deadly, so it is very important to get them diagnosed in time [1]. The most common way of diagnosis is Radiographs while other diagnosis techniques include Chest Tomography (CT) scans, ultrasound, and Magnetic Resonance Imaging (MRI). Because of the cheap cost and availability of Chest X-ray

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

(CXR) machines almost everywhere, doctors prefer to recommend X-rays instead of CT scans [2]. However, identifying chest diseases from radiographs can be a challenging task for the radiologist because of the overlapping structure of the thoracic region. Hence, a computer-aided diagnostic system is needed for the automatic detection of pneumothorax from CXR images which can assist the radiologists.

Extensive research has been done in the medical field after the emergence of machine learning techniques including skin cancer detection [3], detection of arrhythmia [4], and detection of diabetic retinopathy [5]. Recently, the detection of diseases from chest radiographs has become a hot topic. In [6], deep learning-based frameworks have been proposed to detect lung nodules. Outstanding work has been done in [7], by proposing a 121-layered dense network to detect pneumonia with radiologist-level performance along with predicting multiple thoracic diseases.

However, class imbalance (CI) is a massive problem in most of the medical image datasets [8]. In literature, mostly resampling techniques have been used, which could be disadvantageous as it can lead to the removal of samples which might be important for training the model or it may lead to overfitting. So, there is a need to propose a model which tackles the CI problem efficiently along with predicting the presence of disease. In this research, we aim to find an automated way to detect pneumothorax from images of chest x-rays by utilizing deep learning techniques along with comparing different existing approaches to tackle the CI issue. Our proposed approach has been tested on two publicly available chest X-ray datasets.

The remaining paper is arranged as follows. Existing work for the detection of pneumothorax from chest x-rays and the different approaches in the literature for the CI issue (experimented in our research) are discussed in Section II. Section III describes the proposed VDV ensemble model for pneumothorax detection. Section IV describes the datasets used for our research purpose. Results and discussion are provided in Section V. Section VI describes the limitations of this study. In Section VII, the conclusion and future work are mentioned.

II. LITERATURE REVIEW

As stated earlier, several Artificial Intelligence (AI) based techniques have been implemented for the segmentation and automatic diagnosis of lung diseases. In [9], the researchers used a local dataset containing 32 pneumothorax x-ray images and 52 normal cases and identified the presence of pneumothorax by extracting features with Local binary patterns and using SVM as a classifier. The mean accuracy achieved was 82%. In [10], two different machine learning approaches including the bag of features (BoF) and CNN were experimented with the intention of differentiating between normal and abnormal CXR. The research covered 5 different thoracic pathologies including pneumothorax. The experiments were performed on animal X-ray images. A total of 78 images were used for the pneumothorax detection and it was found that features extracted from CNN architecture outperformed BoF in all 5 different cases. Park [11] trained YOLO Darknet-19 pre-trained model for automatically diagnosing pneumothorax, using a dataset containing 1596 pneumothorax and 11137 normal X-ray images, which were acquired from tertiary hospitals. In [12], classification was performed using a dataset of 13,292 DICOM images, and training was done using several CNN architectures including VGG-16, VGG-19, Inception, and Xception. In [13], chest CT scans were used for automatic detection of pneumothorax, where a total of 280 CT scans were used to extract features from trained CNN architecture, and then SVM was used as a classifier. Thoracic Ultrasound images were used in [14] to train a model for distinguishing between normal and pneumothorax cases. Image preprocessing techniques were implemented for the removal of textural information from the Ultrasound images and image enhancement. Model accuracy was increased by the application of transfer learning and fine-tuning techniques on pre-trained CNN architectures. In [15], pixel-classification based CNN approach was used for pneumothorax detection using a training set of 117 CXRs. 95% Area under the Receiver Operating Characteristic curve (AUC) was achieved when evaluated on a test set of 86 CXRs. Texture analysis-based technique was combined with supervised learning technique (KNN) to detect pneumothorax and this proposed framework was tested on a dataset of 108 CXRs, giving the performance of 81% and 87% in terms of sensitivity and specificity respectively [16]. Jakhar *et al.* [17] used the SIIM ACR Pneumothorax dataset, for segmentation of the region of pathology from the chest x-ray images, while making use of U-Net architecture with ResNet encoder.

From the literature, it has been observed that in most cases, the dataset is either small or there exists a problem of class imbalance (CI) i.e., an unequal number of samples in different classes of the dataset. As per our findings, mostly researches carried out utilizing imbalanced dataset have used a single approach to solve the problem of CI [18]. In [19], the researchers proposed an *Under-Bagging* based ensemble model for the said problem, where several subsets of majority class were created which were combined with minority class samples. Salehinejad *et al.* [20] used GANs intending to increase the minority class samples.

The resampling methods (i.e., under-sampling and oversampling) can become a reason for the loss of important data or overfitting. Additionally, none of the existing techniques can be declared as the best one to solve the CI problem as most researchers had adopted a single approach for the research purpose without comparing multiple techniques. Comparison of multiple approaches to solve class imbalance (CI) had been made by some of the researchers using commonly available datasets like MNIST, CIFAR, etc. Buda *et al.* [21] compared the performance of CNNs using multiple approaches including oversampling, under-sampling, and thresholding, and evaluated their results on MNIST and CIFAR-10 datasets. However, to the best of our knowledge, such comparison has not yet been done using a medical image dataset. Such a comparison is important as the CI approaches are domain-dependent [22]. In our research, we have made a comparison of different existing approaches to tackle this problem using publicly available pneumothorax dataset, along with proposing an ensemble-based framework for the automatic diagnosis of pneumothorax from the CXRs. Our proposed model has been tested on two openly available datasets.

A. COMPARISON OF DIFFERENT METHODS FOR IMBALANCED DATASET

Mainly two different approaches are available for an imbalanced dataset [21]. The first one, known as data-level methods, deals with altering the original dataset such that each class contains the same number of samples. The second one is a classifier-level method in which the algorithms are adjusted to solve the said issue. The different approaches experimented in our research are explained below.

1) WEIGHT BALANCING (CLASSIFIER-LEVEL METHOD)

It is one of the classifier-level techniques to solve the class-imbalance (CI) problem [23]. In this technique, the whole training set (as provided in the original dataset) is used, however different weights are assigned to the majority and minority class in the training set. The class weights are assigned according to the formula given below:

$$class\ weight = \frac{n_{samples}}{n_{classes} * np.bincount\ (y)} \tag{1}$$

In (1), $n_{samples}$ refer to the total size of the training set, the total number of classes is represented by $n_{classes}$, and $np.bincount\ (y)$ is a function which counts the frequency of each element in y array (i.e. count the frequency of 0 and 1 class' elements separately).

2) UNDER-SAMPLING (DATA-LEVEL METHOD)

As the name suggests, a subset of samples is randomly selected from the majority class so that we have an equal number of observations in both classes [24]. Although in this approach an enormous number of samples are discarded, still it has been found that in some situations, under-sampling works better than the other approaches [25].

3) OVER-SAMPLING (DATA-LEVEL METHOD)

In this approach, the sample size of the minority class is increased so that it becomes equal to the sample size of majority class [26]. Some of the oversampling techniques include SMOTE [27], Cluster-based oversampling [28], and DataBoost-IM [29]. In case of an image dataset, another way to generate more samples is Data Augmentation [30]. For experimentation purpose, we generated synthetic sample of minority class using data augmentation technique.

4) ENSEMBLE (HYBRID APPROACH)

This approach combines multiple techniques from both or one of the above-mentioned approaches. In case of using the under-sampling method, EasyEnsemble and BalanceCascade are used to train multiple classifiers [31]. The data-level-ensemble approach in [32] describes the way to create subsets of whole data, assuming an identical distribution of observations, thus creating subsets of data that contain the same ratio of samples in each class as present in the original dataset. The data-level-ensemble model experimented in our research finds its root from [33] which utilizes the idea of the *under-bagging method*. According to this, subsets of the majority class are created in such a manner that the sample size of each subset of the majority class is the same as the total sample size of the minority class. These class-balanced (i.e., equal sample size in every class) subsets of training data (containing a subset of majority class combined with all samples of minority class) are then utilized for training a classifier. In this research, the data-level-ensemble (under-bagging method) experimented for the sake of comparison of existing approaches utilizes VGG-16 as fixed-feature extractor from each subset and Linear SVM as a classifier, while final output is based on voting method, (i.e., maximum occurring class is selected as final output) [34].

III. MATERIAL AND METHODOLOGY

This section explains the proposed framework along with describing the feature extractors and the classifier used in this research. Section 3-A explains the different CNN architectures used in our experiments. Section 3-B explains the SVM classifier. Section 3-C explains the proposed VDV model for the classification of pneumothorax from CXR images.

A. CONVOLUTIONAL NEURAL NETWORKS

Single or multiple convolutional layers are arranged in a particular manner to create a neural network named Convolutional Neural Network (CNN). CNN requires a huge amount of data to train itself which can then be used for supervised or unsupervised decision making. It does so by extracting features from the input data and adjusting weights of the neurons by forward and back-propagation [35]. There are many different CNN architectures available, trained on the ImageNet dataset and their weights can be used as initial weights for any classification problem.

In any CNN architecture, the convolutional layers are used for feature extraction from the input while the last fully connected (FC) dense layers act as a classifier. One of the ways to utilize the pre-trained CNNs is known as *"Fixed Feature extractor"*. In this method, the CNN architectures trained on large datasets, like the ImageNet dataset, are used as feature extractors by removing the fully connected dense layers and features are extracted from the remaining CNN architecture. The extracted features can be fed to any classifier like SVM or softmax classifier [36], [37].

As we have used CNN architectures as fixed feature extractors, so we did not train the CNN models on our dataset, hence the training options like learning rate, optimizer, and number of epochs are not specified. However, Section 4-B describes the number of extracted features from each CNN architecture and the number of layers from which features are extracted. We selected VGG and DenseNet as fixed feature extractor and these networks were selected based on the fact that they are most commonly used CNN architectures in CAD systems and have given promising results [57].

1) VGG-16

One of the CNN architectures proposed by Simonyan *et al* is known as the VGG network. The architecture is composed of sixteen layers which include twelve convolutional layers. These layers are the predecessor of three fully connected dense layers. The convolutional layers use 3 × 3 filters, stride and padding of 1. Followed by some of the convolutional layers is the 2 × 2 maximum pooling layer (stride of 2). There are 4096 neurons each in the first two dense layers. The third layer is meant for classification thus it contains 1000 channels. After the fully connected dense layers, there is a soft-max activation layer. This CNN architecture takes an RGB image as input with the default size of 224 × 224. In VGG-16, the total number of parameters is 14,714,688 [38]. The break-down structure of VGG-16 architecture is shown in Table 1.

TABLE 1. VGG-16 architecture.

Layer	Operation
Input Layer	
Convolution	$[3x3\ conv]\ x\ 2$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 2$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 3$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 3$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 3$
Pooling	$2x2\ max\ pool, stride\ 2$
Classification	$4096D\ fully\ connected$ $4096D\ fully\ connected$ $1000D\ fully\ connected$ $softmax$

2) VGG-19

This model is the extension of VGG-16, except that it comprises 19 layers, out of which there are 16 convolutional layers and three FC layers. The architecture arrangement is the same as VGG-16. There are 20,024,384 parameters in VGG-19. Like VGG-16, it takes 224 × 224 RGB images as its input. To use this architecture for the classification problems, the last fully connected dense layer with 1000 neurons/channels is replaced by a dense layer containing neurons equal to the number of classes in the classification problem [38], [39]. The architecture of VGG-19 is shown in Table 2.

TABLE 2. VGG-19 architecture.

Layer	Operation
Input Layer	
Convolution	$[3x3\ conv]\ x\ 2$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 2$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 4$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 4$
Pooling	$2x2\ max\ pool, stride\ 2$
Convolution	$[3x3\ conv]\ x\ 4$
Pooling	$2x2\ max\ pool, stride\ 2$
Classification	$4096D\ fully\ connected$ $4096D\ fully\ connected$ $1000D\ fully\ connected$ $softmax$

3) DENSENET-121

DenseNet-121 is a CNN architecture with 121 layers. It has a total of four dense blocks and a transition layer is present between every consecutive dense block. Every dense block consists of many convolutional layers and the transition layers consist of batch normalization, a convolution layer with 1 × 1 kernel, and an average pooling layer of size 2 × 2. At the end of the architecture, there is a fully connected layer with a softmax activation function. It has 1000 neurons referring to the total number of classes in the ImageNet dataset on which it is trained. It takes an RGB image with a default input size of 224 × 224. There are 7,037,504 parameters in DenseNet121. As opposed to traditional CNNs, here every layer has a connection with all the other layers and direct access to loss functions and original input is given to every layer. The feature maps extracted from all the previous layers are concatenated and fed as input to the next layer. This special design enhances the flow of information throughout the network and also minimizes the vanishing gradient problem [40]. The detailed structure of DenseNet-121 is shown in Table 3.

B. SUPPORT VECTOR MACHINE

A machine learning algorithm that can be utilized for classification as well as regression. Here, mapping of the data points takes place in n-dimensional feature space, where n refers to the total number of features. For classification, the hyperplane is created in such a way that it best separates the two classes and maximizes the margin. In our work, we have used both the linear SVM and polynomial kernel SVM. In practice, the SVM algorithm is implemented using a kernel. Data space containing input data points is transformed into higher dimensional space using kernel tricks. This is done to convert the non-separable classification problem into a separable problem. The linear kernel can be implemented as the normal dot product between x and x_i, where x is the input vector and x_i refers to each support-vector. It is implemented using the following equation:

$$f\ (x,x_i) = B\ (0) + sum\ (a_i * (x,x_i)) \qquad (2)$$

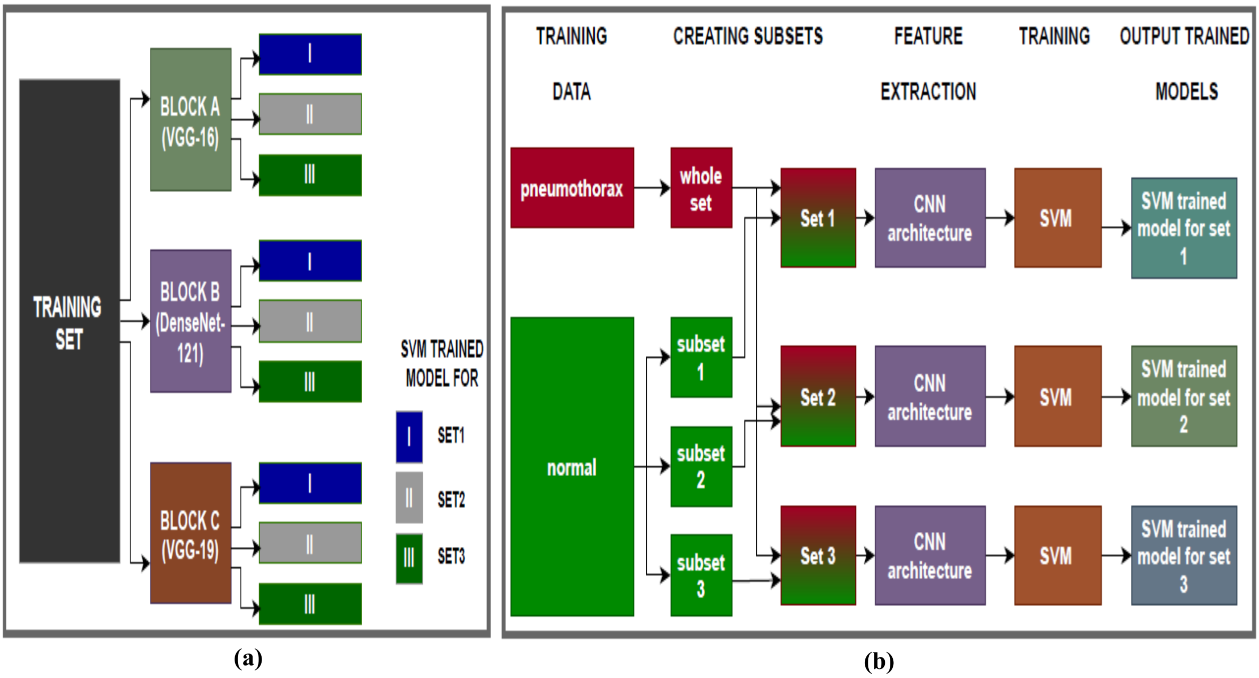

FIGURE 1. Training module of our proposed VDV model in shown above. a) Shows the Block diagram for the Training Module. Here each block uses different CNN architecture for feature extraction and outputs three trained SVM models with respect to each mini-training-set. b) Shows the Internal Working of each Block in Training Module. After creating subsets of training data, features are extracted from each mini-training set using one of the three CNN architectures, then SVM model is trained on these extracted features separately, thus generating SVM trained model with respect to each mini-set of training data.

TABLE 3. Densenet-121 architecture.

Layer	Operation
Input Layer	
Convolution	$[7x7\ conv]$
Pooling	$3x3\ max\ pool, stride\ 2$
Dense Block	$\begin{bmatrix}1x1\ conv\\3x3\ conv\end{bmatrix} x\ 6$
Transition Layer	$[1x1\ conv]$ $2x2\ average\ pool, stride\ 2$
Dense Block	$\begin{bmatrix}1x1\ conv\\3x3\ conv\end{bmatrix} x\ 12$
Transition Layer	$[1x1\ conv]$ $2x2\ average\ pool, stride\ 2$
Dense Block	$\begin{bmatrix}1x1\ conv\\3x3\ conv\end{bmatrix} x\ 24$
Transition Layer	$[1x1\ conv]$ $2x2\ average\ pool, stride\ 2$
Dense Block	$\begin{bmatrix}1x1\ conv\\3x3\ conv\end{bmatrix} x\ 24$
Classification	$7x7\ \ average\ pool$ $1000D\ fully\ connected,$ $softmax$

For every input, training data is used for calculating $B(0)$ and a_i using the learning algorithm.

The SVM with a polynomial kernel can distinguish the non-linear input space. It is expressed as follows:

$$f(x, x_i) = 1 + sum(x * x_i)^d \tag{3}$$

where d represents the degree of the polynomial [41].

C. PROPOSED MODEL

Among the different approaches that we have experimented to solve the class imbalance (CI) problem, the data-level-ensemble (DLE) (i.e., *under-bagging method*) performs better than other approaches. The superiority of an *under-bagging-based data-level-ensemble*, which is created by making several class-balanced subsets of data is proven in [53], [54]. The results in [42] also show that the *model-level-ensemble (MLE)* created by training different classifiers on the data separately and later combining the results of individual classifiers (either by averaging or voting method) gives better performance. Thus, uniting these two ideas (i.e., MLE and under-bagging-based DLE), we present a novel framework VDV which is MLE of multiple DLEs.

The data-level-ensembles are designed as explained in Section II-A-4. It utilizes three different CNN architectures as fixed feature extractor and polynomial kernel SVM as classifier [43]. The selected CNN architectures for feature extraction purposes are VGG16, VGG19, and DenseNet121, thus the proposed framework is named as VDV model. In all these CNN architectures, the last fully connected layers are removed, and the features extracted from the architecture are sent to the classifier. Like any other machine learning based CAD system, our proposed framework comprises two parts, i.e., training and testing.

1) TRAINING MODULE

The block diagram for the training module is shown in Fig. 1. Basically, the training process makes use of three different

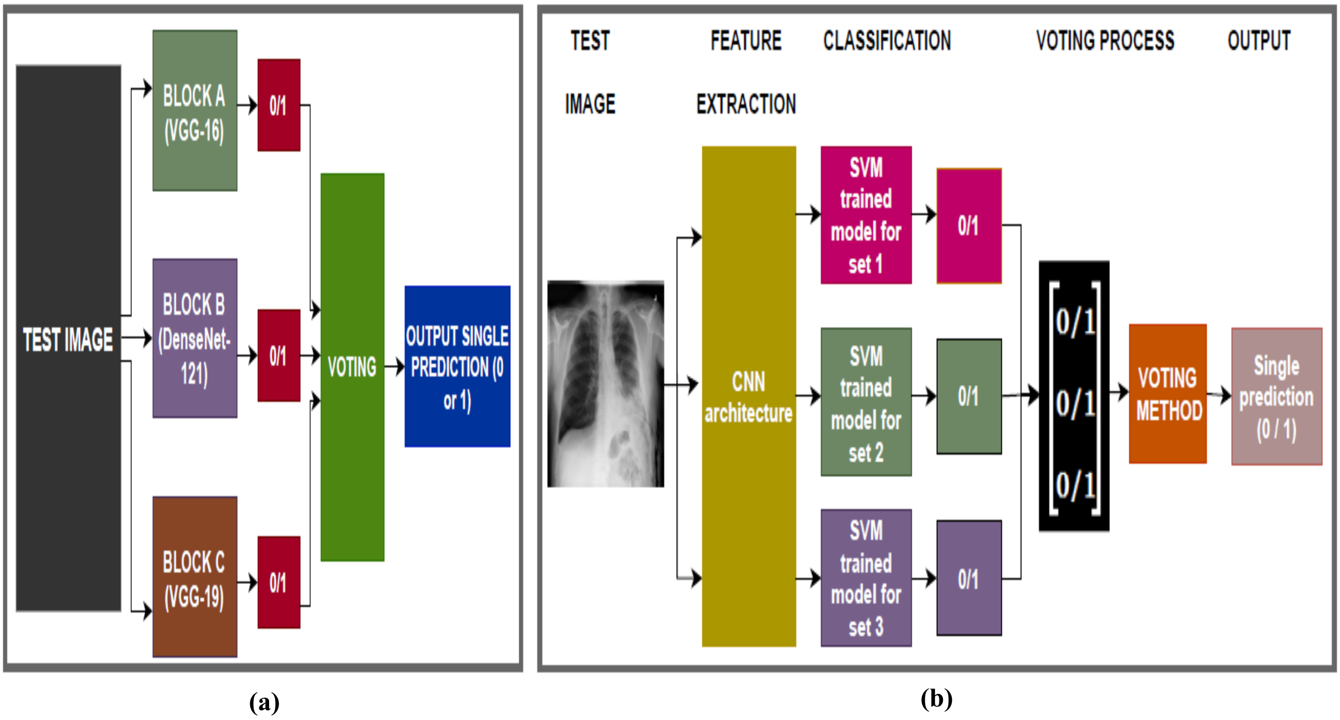

FIGURE 2. Test module of our proposed VDV model is shown above. a) Shows the Block diagram for Test module. Here each block uses different CNN architecture for feature extraction from test image and outputs predicted class. These predictions are sent to voting unit which outputs the maximum occurring class. b) Shows the generic Internal working of each Block in Test Module, in which trained SVM model with respect to each mini-training-set predicts the class of the test sample based on features extracted by respective CNN architecture. These predictions are combined using the Voting method to obtain a single prediction.

CNN architectures as shown in Fig. 1a. Block A refers to VGG-16, block B refers to DenseNet121, and block C refers to VGG-19, used as feature extractors. The training set is sent to each block separately and each block generates three SVM trained models. These trained models are later used for predicting the class of test samples. The internal working of each block is the *same* except for using different CNN architectures (as fixed feature extractor).

Fig. 1b explains the generic working of each block, in which under-bagging-based ensembles are created. Thus, for creating a data-level-ensemble, first we create the mini-training set, by dividing the majority class samples into multiple subsets. Here we have two classes, Normal (negative class or class 0), and Pneumothorax (positive class or class 1). In our dataset, class 0 has the majority number of samples, so it is divided in such a way that each subset has n number of samples of class 0, where n is same as the sample size of minority class (i.e., class 1). This way, we have N subsets of class 0, where N is equal to the imbalance ratio between class 0 and class 1. As in the SIIM dataset, the imbalance ratio is around 3.49:1, thus we have 3 subsets of class 0 which are referred to as subset1, subset2, and subset3 in Fig. 1b. Then each subset of class 0 is combined with the whole minority class (i.e., class 1) thus creating mini-training sets (referred as set1, set2, and set3). Features are extracted from each mini-training set and sent to the SVM classifier for training purposes, which generates SVM trained model for every mini-training set. Note that this process is repeated for all three blocks which are shown in Fig. 1a. Hence, we have three SVM trained models for each block. As we have three blocks, so each block has three SVM trained models each, which will be used for testing purposes.

2) TESTING MODULE

Fig. 2 represents the block diagram of the test module. Here the trained SVM models generated by the three blocks in the training module are used. Fig. 2a shows the workflow for predicting the class of any sample. The test CXR image is sent to the three blocks (Block A, Block B, and Block C) which outputs the class prediction. As we have three blocks referring to different CNN architectures (i.e., VGG-16, DenseNet121, and VGG-19) as fixed-feature extractors, so each block generates its prediction, these three predictions are used to make a final decision based on the Voting method (i.e., maximum occurring predicted class is taken as final result).

The internal generic working of each block of the test module is same and is shown in Fig. 2b. CNN architecture (related to each block) is used for extracting features from the test CXR image which are then sent to each of the three trained SVM models (referring to each mini-training set, i.e., set1, set2, and set3). Each SVM trained model predicts the class of test samples. These class predictions are then combined together based on the voting method.

Finally, when the output from each block is obtained, then these three outputs are combined together via Voting Method as shown in Fig. 2a.

3) EXPERIMENTAL DETAILS

In all the experiments, we keep the default input size of the images, i.e., 224 × 224. Note that for VGG-16 and VGG-19 we use all the 25088 extracted features in the training and testing process, however, because of the really large number of features in the case of DenseNet121 (i.e., 50176), we have to minimize the number of features else we cannot apply Kernel SVM because of memory issue. Principal Component Analysis (PCA) is applied for feature reduction, the total number of features is reduced to 4758. This number is calculated using the Singular value decomposition (svd) solver method [43]. Moreover, instead of 8296 samples, we keep 7137 samples of Normal class in the training set, i.e., 3 times more than the number of samples in the pneumothorax class. It is done for making class-balanced training subsets.

For authentication, we test our proposed framework on the Random Sample of NIH Chest X-ray (RS-NIH) dataset as well. We select normal and pneumothorax samples from the dataset, thus we have an imbalance dataset with the ratio of 11:1, i.e., for every sample of pneumothorax, there are 11 samples of normal CXRs. It is important to mention here that in the training set we keep 2376 samples of the Normal class instead of 2435 samples, so that each mini-training set has an equal number of normal and pneumothorax samples. The only difference while experimenting with this dataset is that here, we make 11 subsets of normal class samples instead of three subsets. The rest of the implementation is the same as explained above.

The results on the test set for each separate Block (i.e., Block A, Block B, and Block C) along with our proposed VDV model on the SIIM dataset and RS-NIH dataset are reported in Section IV.

IV. DATASETS AND EXPERIMENTAL SETUP

A. DATASETS

1) SIIM ACR PNEUMOTHORAX DATASET

The first dataset selected for our experimentation purpose is available on Kaggle [44], which contains stage-1 training and testing data from "SIIM-ACR Pneumothorax Segmentation competition", in Portable Network Graphics (png) format. There are 12047 chest X-rays (CXR) along with training and testing lists. The training list contains 8296 normal CXR images and 2379 CXRs with pneumothorax, while the testing set contains 1082 normal CXRs and 290 images of the other class. The original size of X-ray images is 1024 × 1024. However, for our experimentation purpose, we resize the images to 224 × 224. The main reason for selecting this dataset is that classification results have never been reported on this dataset before. Also, it provides the same number of RLE (Run Length Encoded) masks which can later be used for segmentation purpose. Table 4 summarizes the details of the dataset, where N represents Normal CXRs, and P represents CXRs with pneumothorax. Some of the images from this dataset are represented in Fig 3.

TABLE 4. Details Of SIIM dataset.

Attribute		Value
Resolution		1024 x 1024
Dataset size		12047
No of classes		2
Training set	N	8296
	P	2379
Testing set	N	1082
	P	290

N: Normal, P: Pneumothorax

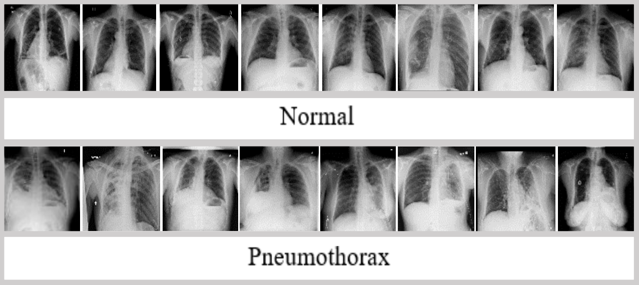

FIGURE 3. Chest X-ray images from the SIIM dataset.

2) RANDOM SAMPLE OF NIH CHEST X-RAY DATASET (RS-NIH)

The second dataset on which we have performed experiment is "Random Sample of NIH Chest X-ray Dataset" (RS-NIH) which is provided by the National Institutes of Health NIH and is available on Kaggle [45]. The full NIH Chest X-ray-14 dataset (NIH-CXR) contains 112,120 images, with 15 classes, covering 14 different thoracic pathologies and 15th being the normal case. The *"RS-NIH"* chosen for our research purpose is a smaller version of the NIH-CXR dataset and contains 5% of the total number of samples, and each pathology is present in the same ratio as is present in the full dataset. Each image has a resolution of 1024 × 1024. It contains 3044 images of No-finding, 967 images of Infiltration, 664 Effusion, 508 Atelectasis, 313 Nodule, 284 Mass, 271 Pneumothorax, 226 Consolidation, 176 Pleural Thickening, 141 Cardiomegaly, 127 Emphysema,118 Edema, 84 Fibrosis, 62 Pneumonia and 13 images of Hernia. For our experiment, we keep the images of No-finding cases (i.e., normal) and pneumothorax samples. Since the NIH-CXR dataset is divided into 80% training and 20% testing set, we also split our RS-NIH data into training and testing set with the same ratio.

Note that two different protocols have been followed while splitting the dataset. First being a *random split of data* and the second being a *patient-wise split*, i.e., CXRs from the same

TABLE 5. Details of RS-NIH dataset.

Attribute		Value
Resolution		1024 x 1024
Dataset size (with 14 classes)		5606
No of classes chosen		2
Training set	N	2376
	P	216
Testing set	N	609
	P	55

N: Normal, P: Pneumothorax

patient can only be present in either training or testing set. The details of this dataset are given in Table 5.

B. EXPERIMENTAL SETUP

Keras with Tensorflow backend is used in our research with Python as a programming language. Our research comprises two parts, first one is to compare different existing approaches to tackle the imbalance problem and the second one is to propose a framework (VDV) for automatic diagnosis of pneumothorax which is tested on two different datasets. In both the experiments, the main task is feature extraction and classification of CXR images as normal or pneumothorax.

For the first part, i.e., comparison of CI techniques, we select a pre-trained VGG-16 model (with ImageNet weight) as a fixed-feature extractor based on its structural simplicity [35], and Linear SVM as a classifier.

For the proposed VDV network, three different CNN architectures are selected which are VGG-16, VGG-19, and DenseNet121. The details of input and output sizes, number of parameters, and number of layers in each architecture are summarized in Table 6. These pre-trained models with ImageNet weights are utilized for the extraction of features from the images. The last fully connected (FC) layers of these pre-trained models are removed as those are meant for classification purpose, instead, we have used polynomial kernel SVM as a classifier with gamma value 0.002 and C equal to 100. The values for kernel SVM are selected using the grid search method.

Note that for the proposed framework, we chose poly kernel SVM as it is a proven fact that SVM with poly kernel performs better than Linear SVM [46]. Moreover, instead of using the last fully connected dense layers for classification, we chose SVM as it is found to be more effective [47], [48].

C. PERFORMANCE MEASURES

For the evaluation of a model, the selection of performance metrics is important. As our training as well as testing data is imbalanced, only accuracy is not a good performance measure [34] that is why we select Area under Receiver Operating Characteristic curve (AUC) and Recall as our main performance metrics. In addition to these, we also report the results with other performance metrics which include Accuracy, Specificity, Precision, Geometric mean (G-mean) [49], F1 and F2 score [50]. AUC is calculated by the calculating the area under Receiver Operating curve which is defined in terms of true positive and false positive rate [51]. In all the following expressions, TN, TP, FN, and FP denote True Negative, True Positive, False Negative, and False Positive respectively. The expressions for calculating Accuracy, Recall, Precision, and Specificity are given below:

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

The combination of recall and precision is an important metric known as F-score. It is calculated using F_β, where β is assigned a different value, based on the problem statement. If the aim is to avoid misclassification of negative samples as positive ones, i.e., giving more importance to precision, then β is assigned a value equal to 0.5. However, if it is intended to never miss positive class samples, like in our case, the aim is to make a classifier that should avoid missing pneumothorax samples, i.e., giving more importance to Recall, then value of β is set to 2. If both precision and recall are given equal importance, then β is assigned a value equal to 1. In our experiments, we have calculated F_1 and F_2 score by substituting β as 1 and 2 respectively. The expression for F_β and G-mean are given below:

$$F_{\boldsymbol{\beta}} = \left(1+\boldsymbol{\beta}^2\right) \frac{Recall \times Precision}{(\beta^2.Precision) + Recall} \tag{8}$$

$$G - mean = \sqrt{Recall \times specificity} \tag{9}$$

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. RESULTS

The first part of our work where the comparison of existing class-imbalance approaches is performed utilizes an openly available SIIM Pneumothorax dataset. Accuracy, Recall/Sensitivity, Specificity, and AUC for all the experiments are reported in this section. Table 7 summarizes the results for different existing CI approaches experimented within this research. Here, Column 2 (i.e., No. of Training Samples) refers to the total number of CXR images in each class, used in every approach, separately. Based on the highest AUC value achieved in case of the ensemble model, which is 80.02%, it can be inferred that the under-bagging-based DLE outperforms other existing approaches for CI issue. Moreover, it can be observed that sensitivity value is highest of all in case of the DLE model which shows that maximum correct identification of the pathology is achieved using an ensemble model, compared to any other existing approach.

Based on these results, we propose our framework named as VDV model, the detailed performance of which is summarized in Table 8. As our proposed framework is a model-level-ensemble of three data-level-ensembles using

TABLE 6. Parameters configuration of CNN architectures.

CNN	Shape		Features	Parameters		Layers
	Input	Output		Trainable	Non-Trainable	
VGG-16	224×224×3	7×7×512	25,088	14,714,688	0	19
VGG-19	224×224×3	7×7×512	25,088	20,024,384	0	22
DenseNet-121	224×224×3	7×7×1024	50,176	6,953,856	83,648	427

TABLE 7. Comparison of different existing approaches for class imbalance problem.

Technique	No of Training Samples		ACC (%)	REC (%)	SPE (%)	AUC (%)
	Normal	Pneumothorax				
Weight balancing	8296	2379	79.08	48.96	87.15	78.8
Under-sampling	2379	2379	72.15	68.62	73.10	77.67
Over-sampling	8296	8296	77.7	50	85.20	77.76
Ensemble	2379 (in each subset)	2379 (in each subset)	75.22	**79.65**	74.09	**80.02**

ACC: ACCURACY, REC: RECALL, SPE: SPECIFICITY

TABLE 8. Performance of proposed VDV framework on both datasets.

	ACC (%)	REC (%)	SPE (%)	PREC (%)	F1 (%)	F2 (%)	G-mean (%)	AUC (%)
				SIIM DATASET				
VGG-16	77.55	83.79	75.87	48.21	61.2	73.01	79.73	86±0.01
VGG-19	77.04	82.06	75.69	47.5	60.17	71.64	78.8	86±0.01
DenseNet-121	76.32	80.68	75.04	46.4	58.94	70.31	77.81	85±0.00
VDV (SIIM)	**78.27**	**85.17**	**76.43**	**49.2**	**62.37**	**74.3**	**80.68**	**86±0.00**
				RS-NIH DATASET (RANDOM SPLIT)				
VDV(RS-NIH)	82.68	**90.9**	81.93	31.25	46.5	65.78	86.3	**95±0.01**
				RS-NIH (WITH PATIENT WISE SPLIT)				
VDV (RS-NIH)	69.12	**85.45**	67.65	19.26	31.43	50.64	76.03	**77±0.06**

PREC: Precision, F1: F_1 score, F2: F_2 score

three different CNN architectures, the first three rows show the performance of each data-level-ensemble (utilizing one of three CNN architectures each). The last three rows show the performance of the proposed VDV model on the SIIM and RS-NIH datasets respectively. Note that the individual model performance is reported for the SIIM dataset only. The AUC value achieved by our framework on the SIIM dataset is 86.0% and the sensitivity value of 85.17%. As our VDV framework gives far better results especially in terms of sensitivity as compared to a single data-level-ensemble (utilizing single CNN architecture), thus it strengthens the idea that such an ensemble enhances the rate of correct classification of the pathology samples.

Moreover, in case of random split of the RS-NIH dataset, our proposed framework achieved 95.0% AUC, 82.68% accuracy, and 90.9% recall value. On the other hand, following a patient-wise data split, the proposed VDV model achieved an AUC of 77.06%, and a recall value of 85.45%.

The model performance in terms of AUC is shown in Fig 4a and 4b. Fig 4a represents the performance of our proposed model on the SIIM dataset, where the ROC curves for individual under-bagging-based data-level-ensemble (utilizing single CNN architecture) along with the ROC curve of the VDV model on SIIM dataset are plotted. The performance of our proposed framework in terms of AUC on the RS-NIH dataset is represented in Fig 4b.

The confusion matrix for the performance of the VDV model on the SIIM and RS-NIH test set (with random split and patient wise split) are shown in Table 9, Table 10 and Table 11 respectively. For the SIIM pneumothorax dataset, our proposed model correctly identifies 247 pneumothorax cases while 43 are misclassified, and the total number of correctly classified Normal CXRs is 827 while 255 are misclassified. For the RS-NIH dataset, with the random split of data, 50 out of 55 samples are correctly classified as pneumothorax while 499 out of 609 samples are correctly identified as Normal CXRs. For patient-wise data split of RS-NIH dataset, 47 out of 55 and 412 out of 609 samples are correctly classified as pneumothorax and Normal x-rays respectively.

1) JUSTIFICATION

The performance comparison of our proposed model for SIIM Pneumothorax dataset is provided in Table 12. We can directly compare our result with [55], in which the same

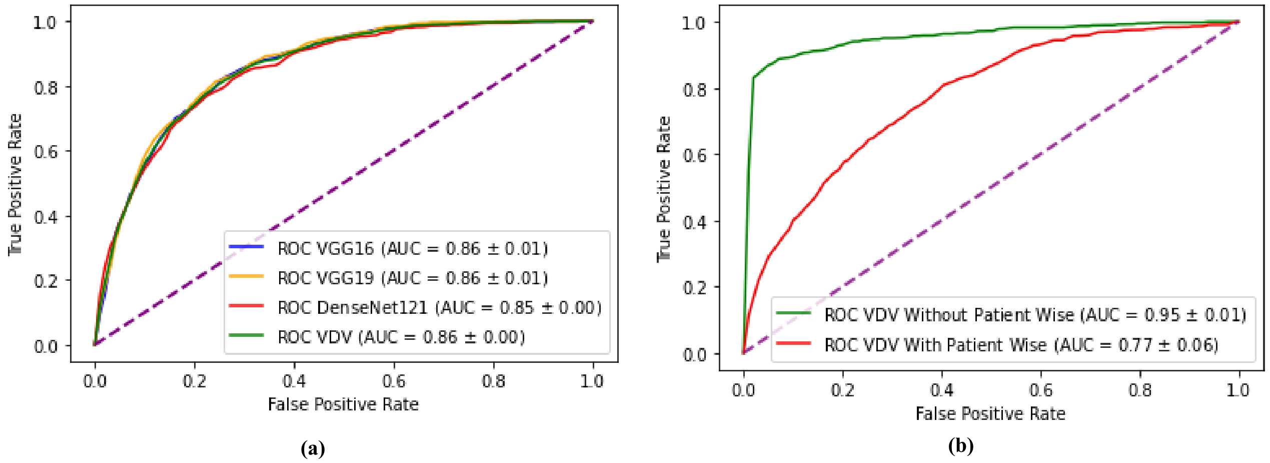

FIGURE 4. AUC plot proving the effectiveness of proposed VDV model. a) On SIIM dataset. b) On RS-NIH dataset.

TABLE 9. Confusion matrix for SIIM dataset classification.

	Predicted Class	
Actual Class	Normal	Pneumothorax
Normal	827	255
Pneumothorax	43	247

TABLE 10. Confusion matrix for RS-NIH dataset classification with random split.

	Predicted Class	
Actual Class	Normal	Pneumothorax
Normal	499	110
Pneumothorax	05	50

TABLE 11. Confusion matrix for RS-NIH dataset classification with patient-wise split.

	Predicted Class	
Actual Class	Normal	Pneumothorax
Normal	412	197
Pneumothorax	08	47

dataset and same testing set was used. It can be clearly seen that although their reported AUC is slightly greater than the AUC value achieved by our proposed model, however the sensitivity/ recall achieved by our proposed model is far greater than the value reported in [55]. Since our aim is to increase the rate of correct classification of positive class samples, so we can safely claim that our proposed model surpasses the existing techniques. Additionally, few other papers are also presented in Table 12 (in which pneumothorax classification was the main concern) and we have tried to prove the superiority of our proposed model based on the CI ratio and the size of the dataset in those paper. The total number of normal and pneumothorax CXR used by each researcher in the training and testing set is also given in Table 12. The sub-column referred as B in the table shows if the dataset is class-balanced or imbalanced in nature. The last column depicts if the dataset used is publicly available or not. It can be seen that although better results are obtained in terms of AUC value, however the datasets utilized are either completely balanced or have minimal imbalance ratio. Moreover the sample size of those datasets is also small. While the dataset utilized in this research is not only highly imbalance, but also has large number of samples in both training and testing set, and achieved much better results especially in terms of Recall. Furthermore, they have used private datasets while the dataset that we have used is publicly available, so other researchers can add to this work.

For the RS-NIH dataset, we can directly compare our results with [52] in which the RS-NIH dataset has been used for classification purpose, as presented in Table 13. Note that in [52], multi-label classification was performed considering 14 different chest diseases, so we have reported their achieved AUC value for pneumothorax classification. It is important to mention here that we have referred and compared our results with the paper in which pneumothorax was considered as a separate class in the classification problem, while we have not considered the papers in which the same dataset was used to differentiate between normal and abnormal CXRs without considering any specific pathology. Additionally, the papers using NIH Chest X-ray-14 datasets are not considered for comparison purpose as RS-NIH contains only 5% part of NIH Chest X-ray-14. Thus, these are *two different datasets* and hence their results can't be directly compared.

B. DISCUSSION

Our work in this paper comprises two parts, so we will discuss the result of each part separately. The fact that under-bagging based data-level-ensemble outperforms other existing approaches for CI problem is because it makes use of the whole dataset in such a way that training is done piecewise using class-balanced subsets of whole data. It is generally observed that a balanced dataset performs better than CI

TABLE 12. Comparing performance of VDV model with other classification models using SIIM dataset.

Year	Author	Training Set				Testing Set				Handling Class Imbalance	Public	Results (%)
		N	P	IR	B	N	P	IR	B			
2018	Chan [9]	36	22	1.6:1	×	16	10	1.6:1	×	Ignored the minimal imbalance	×	ACC= 82.20
2018	Yoon [10]	24	24	1:1	✓	15	15	1:1	✓	Completely balanced	×	ACC=96.60 REC=100 SPEC=93.8
2019	S. Park [11]	10887	1343	8.1:1	×	250	253	0.9:1	✓	Under-sampling for class balance	×	REC=89.7 SPEC=96.4
2018	Taylor [12]	7095	2214	3.2:1	×	1553	437	3.5:1	×	Under-sampling for class balance	×	REC=55 PREC=90 AUC=82
2020	Wang [55]	7250	2079	3.4:1	×	1082	290	3.7:1	×	Loss function and deep learning network named ChexLocNet	✓	REC= 78 PREC= 78 AUC= 87
2021	**VDV**	7137	2379	3.4:1	×	1082	290	3.7:1	×	Comparison of multiple approaches and selection of ensemble model	✓	**ACC=78.27** **REC=85.17** **PREC=76.4** **AUC=86.0**

N: No. of Normal CXRs, P: No. of CXRs with Pneumothorax, B: Balance or Imbalance Dataset, IR: Imbalance Ratio, Public: Dataset Public or Private

TABLE 13. Comparing performance of VDV model with other classification models using RS-NIH dataset.

Author	Dataset	Description	Results (%)
Modal [52]	RS-NIH	Multi-label classification of 14 thoracic diseases	AUC= 54.0
		Random split of data	
Proposed model	RS-NIH	Binary classification (normal and pneumothorax CXRs)	**ACC=82.68** **AUC=95.0**
		Patient-wise split of data	
Proposed model	RS-NIH	Binary classification (normal and pneumothorax CXRs)	**ACC=69.12** **AUC=77.06**

data, hence training the classifier on class-balanced subsets of the dataset and then combining their results gives better performance as compared to other techniques. Now as the single DLE (i.e., data-level ensemble with a single CNN architecture as feature extractor) gives quite good results, we have designed a MLE of three data-level-ensembles, with each DLE using different CNN architecture as a fixed feature extractor. Utilizing three different CNN architectures in three data-level-ensembles separately allows our model to give better performance and utilization as opposed to the previously used single data-level-ensemble models.

So far, the work done in order to solve the CI problem proves that single architecture-based DLE outperforms other existing approaches, the same is evident from this paper as well. However, in literature mostly researchers have reported their results using MNIST or CIFAR dataset. Moreover, instead of proposing a new approach, most of the researchers using imbalanced medical images dataset have used a single approach, either oversampling or undersampling. We have not only made a comparison of different existing approaches using our real-life medical images dataset but have also presented a novel framework that finds its roots from the concepts of data-level and model-level ensemble. To our knowledge, this type of ensemble model has never been proposed which not only solves the issue of class imbalance while using the whole imbalanced dataset but also takes advantage of the performance of different CNN architectures.

The comparison of our proposed model with existing literature is provided in Table 12 and Table 13. In Table 12, the comparison of the performance of our proposed VDV model on SIIM datasets with existing work is presented. Since, SIIM dataset was originally presented for the purpose of localization, so mostly researchers have put their effort in localization of disease instead of classification. Hence, we can directly compare our result with [55] in which the authors' focus was on both classification and localization. It can be seen that although their achieved AUC is 87%, while ours is 86%, however our VDV model obtained 85.17% Recall value which is much higher than that achieved in [55], i.e. 78%. As mentioned earlier, the main focus of this study is to increase the rate of correct classification of positive sample, hence the obtained results strongly support the superiority of proposed model.

We have made an indirect comparison with related work based on the imbalance ratio and size of the datasets utilized. In [9]–[11] it can be seen that although better results in terms of accuracy and AUC are achieved however they have used comparatively smaller test datasets, and mostly datasets especially the test sets are balanced or have a minimal ratio of imbalance, while our dataset is imbalance in terms of both training and testing set. Additionally, their datasets are not publicly available. Furthermore, referring to [12], in which the dataset size and the imbalance ratio in both training and testing set is almost the same as in the SIIM dataset, the under-sampling technique was used to solve the CI problem. It can be seen that although their AUC is 82% however the recall value is only 55%, i.e., only 55% positive samples were correctly identified. While the result obtained by our proposed ensemble model is much better with an AUC of 86% and recall of 85.17%. Hence it can be safely said that the ensemble model created by a stack of under-bagging-based ensemble models provides much better results especially in terms of sensitivity.

In addition, we have also tested our proposed framework on the openly available RS-NIH dataset and the comparison with existing work is provided in Table 13. In Table 13, it can be seen that the results obtained by the proposed VDV model using both the data-split protocols (i.e. random split and patient-wise data split) are far better as compared to those achieved in [52], where the random split of data was considered. Note that the higher performance of VDV model in case of the random split of data as compared to patient-wise split is because in case of random split of the dataset, there are chances that Chest X-rays from same patient might be present in both training and testing set, whereas in patient-wise split there is no such overlap. Moreover, to our knowledge, mostly researchers have utilized NIH-Chest X-ray-14 dataset, and very few have used RS-NIH dataset for classification purpose [50], [52], [56]. So we have directly compared our results with [52] in which pneumothorax was considered as a separate class in pathology detection problem while other two papers are deliberately ignored since they have not considered any particular pathology, instead the classification problem was treated as a binary classification problem in which all pathology samples in the dataset were considered as positive while non-pathology samples were considered as negative class.

In the end, we can say that our work surpasses previous works performed to date in the field of pneumothorax, as we have used publicly available datasets intending to allow other researchers to study, comprehend and offer their input. Our results prove that the proposed VDV framework (i.e., MLE of DLEs) performs better than the existing approaches and can be used for any class-imbalance dataset.

VI. LIMITATION OF THIS STUDY

The limitation of this proposed framework is that it cannot be experimented with K-fold cross-validation, because it requires a large number of subsets to be created based on the imbalance ratio. So, in case of bigger datasets with a large number of samples or highly class imbalance datasets, it would be computationally expensive and complex to perform K-fold cross-validation using the VDV model.

VII. CONCLUSION

Pneumothorax can be a deadly disease if not treated in time, thus there is a need to correctly identify it in time. With the advancement in deep learning technology, and its ability to make unsupervised wise decisions, an efficient automatic diagnostic system can be proposed for the detection of pneumothorax. For proposing such a framework for automatic detection of pneumothorax using highly imbalanced data, we have first analyzed different techniques for class imbalance (CI) problem using a medical image dataset. After finding out that data-level-ensemble (i.e., Under-bagging based ensemble) performs best of all, we have presented a model by combining the ideas of the ensemble of models and ensemble of data. Our results have shown that the VDV model outperforms single data-level-under-bagging based ensemble with a single CNN architecture as fixed feature extractor. Our proposed VDV framework achieved 85.17% Recall with 86.0% AUC for the SIIM pneumothorax dataset. For the RS-NIH dataset, 90.9% Recall with 95.0% AUC is achieved for the random split of data. For patient-wise split of data 85.45% recall with 77.06% AUC is obtained. Our achieved results on the both the datasets, i.e. SIIM Pneumothorax and RS-NIH, are higher which also validates the performance of our proposed framework. So, one can use our proposed framework for any imbalanced dataset with a little modification in terms of using different CNN architecture for feature extraction and different resolution of the input image. In the future, we can propose the utilization of this framework for bigger datasets, for example, full NIH Chest X-ray-14 dataset. Also, a segmentation model using SIIM dataset can be developed which will be more helpful to the radiologists for correctly identifying the disease.

REFERENCES

[1] Harvard Health. (Jan. 2, 2019). *Pneumothorax*. Accessed: Jun. 12, 2020. [Online]. Available: https://www.health.harvard.edu/a_to_z/pneumothorax-a-to-z

[2] C. Qin, D. Yao, Y. Shi, and Z. Song, "Computer-aided detection in chest radiography based on artificial intelligence: A survey," *BioMed. Eng. OnLine*, vol. 17, p. 113, Dec. 2018, doi: 10.1186/s12938-018-0544-y.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[4] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," Jul. 2017, *arXiv:1707.01836*.

[5] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016, doi: 10.1001/jama.2016.17216.

[6] X. Huang, J. Shan, and V. Vaidya, ''Lung nodule detection in CT using 3D convolutional neural networks,'' in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Melbourne, VIC, Australia, Apr. 2017, pp. 379–383, doi: 10.1109/ISBI.2017.7950542.

[7] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, ''CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,'' Dec. 2017, *arXiv:1711.05225*.

[8] H. Vasudevan, A. Michalas, N. Shekokar, and M. Narvekar, Eds., ''Advanced computing technologies and applications,'' in *Proc. 2nd Int. Conf. Adv. Comput. Technol. Appl. (ICACTA)*. Singapore: Springer, 2020, p. 300, doi: 10.1007/978-981-15-3242-9.

[9] Y.-H. Chan, Y.-Z. Zeng, H.-C. Wu, M.-C. Wu, and H.-M. Sun, ''Effective pneumothorax detection for chest X-ray images using local binary pattern and support vector machine,'' *J. Healthcare Eng.*, vol. 2018, pp. 1–11, Apr. 2018, doi: 10.1155/2018/2908517.

[10] Y. Yoon, T. Hwang, and H. Lee, ''Prediction of radiographic abnormalities by the use of bag-of-features and convolutional neural networks,'' *Veterinary J.*, vol. 237, pp. 43–48, Jul. 2018, doi: 10.1016/j.tvjl.2018.05.009.

[11] S. Park, ''Performance of a deep-learning system for detecting pneumothorax on chest radiograph after percutaneous transthoracic needle biopsy,'' in *Proc. Eur. Congr. Radiol.*, 2019, p. 2053, doi: 10.26044/ECR2019/C-0334.

[12] A. G. Taylor, C. Mielke, and J. Mongan, ''Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study,'' *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002697, doi: 10.1371/journal.pmed.1002697.

[13] X. Li, J. H. Thrall, S. R. Digumarthy, M. K. Kalra, P. V. Pandharipande, B. Zhang, C. Nitiwarangkul, R. Singh, R. D. Khera, and Q. Li, ''Deep learning-enabled system for rapid pneumothorax screening on chest CT,'' *Eur. J. Radiol.*, vol. 120, Nov. 2019, Art. no. 108692, doi: 10.1016/j.ejrad.2019.108692.

[14] T. Lindsey, R. Lee, R. Grisell, S. Vega, and S. Veazey, ''Automated pneumothorax diagnosis using deep neural networks,'' in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, vol. 11401. Cham, Switzerland: Springer, 2019, pp. 723–731, doi: 10.1007/978-3-030-13469-3_84.

[15] A. Blumenfeld, H. Greenspan, and E. Konen, ''Pneumothorax detection in chest radiographs using convolutional neural networks,'' *Proc. SPIE*, vol. 10575, p. 3, Feb. 2018, doi: 10.1117/12.2292540.

[16] O. Geva, G. Zimmerman-Moreno, S. Lieberman, E. Konen, and H. Greenspan, ''Pneumothorax detection in chest radiographs using local and global texture signatures,'' *Proc. SPIE*, vol. 2015, Mar. 2015, Art. no. 94141P, doi: 10.1117/12.2083128.

[17] K. Jakhar, A. Kaur, and D. M. Gupta, ''Pneumothorax segmentation: Deep learning image segmentation to predict pneumothorax,'' Apr. 2021, *arXiv:1912.07329*.

[18] T. J. Jun, D. Kim, and D. Kim, ''Automated diagnosis of pneumothorax using an ensemble of convolutional neural networks with multi-sized chest radiography images,'' Apr. 2018, *arXiv:1804.06821*.

[19] B. S. Raghuwanshi and S. Shukla, ''Class imbalance learning using UnderBagging based kernelized extreme learning machine,'' *Neurocomputing*, vol. 329, pp. 172–187, Feb. 2019, doi: 10.1016/j.neucom.2018.10.056.

[20] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, ''Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 990–994, doi: 10.1109/ICASSP.2018.8461430.

[21] M. Buda, A. Maki, and M. A. Mazurowski, ''A systematic study of the class imbalance problem in convolutional neural networks,'' *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[22] G. M. Weiss, ''Mining with rarity: A unifying framework,'' *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 7–19, Jun. 2004, doi: 10.1145/1007730.1007734.

[23] B. Raskutti and A. Kowalczyk, ''Extreme re-balancing for SVMs: A case study,'' *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 60–69, Jun. 2004, doi: 10.1145/1007730.1007739.

[24] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, ''Learning from class-imbalanced data: Review of methods and applications,'' *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.

[25] C. Drummond and R. C. Holte, ''C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling,'' in *Proc. 2nd Workshop Learn. From Imbalanced Datasets*, Washington, DC, USA, vol. 11, Aug. 2003, pp. 1–8.

[26] G. Levi and T. Hassncer, ''Age and gender classification using convolutional neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 34–42, doi: 10.1109/CVPRW.2015.7301352.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[28] T. Jo and N. Japkowicz, ''Class imbalances versus small disjuncts,'' *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 40–49, Jun. 2004, doi: 10.1145/1007730.1007737.

[29] H. Guo and H. L. Viktor, ''Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach,'' *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 30–39, Jun. 2004, doi: 10.1145/1007730.1007736.

[30] F. Chollet. (Jun. 2016). Building powerful image classification models using very little data. The Keras Blog. Accessed: Jul. 1, 2020. [Online]. Available: https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html

[31] X.-Y. Liu, J. Wu, and Z.-H. Zhou, ''Exploratory undersampling for class-imbalance learning,'' *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009, doi: 10.1109/TSMCB.2008.2007853.

[32] S. Sapp, M. J. van der Laan, and J. Canny, ''Subsemble: An ensemble method for combining subset-specific algorithm fits,'' *J. Appl. Statist.*, vol. 41, no. 6, pp. 1247–1259, Jun. 2014, doi: 10.1080/02664763.2013.864263.

[33] R. Barandela, R. M. Valdovinos, and J. S. Sanchez, ''New applications of ensembles of classifiers,'' *Pattern Anal. Appl.*, vol. 6, no. 3, pp. 245–256, Dec. 2003, doi: 10.1007/s10044-003-0192-z.

[34] U. R. Salunkhe and S. N. Mali, ''Classifier ensemble design for imbalanced data classification: A hybrid approach,'' *Proc. Comput. Sci.*, vol. 85, pp. 725–732, Jan. 2016, doi: 10.1016/j.procs.2016.05.259.

[35] M. Marouf, R. Siddiqi, F. Bashir, and B. Vohra, ''Automated hand X-ray based gender classification and bone age assessment using convolutional neural network,'' in *Proc. 3rd Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Sukkur, Pakistan, Jan. 2020, pp. 1–5, doi: 10.1109/iCoMET48670.2020.9073878.

[36] *Transfer Learning, CS231n Convolutional Neural Networks for Visual Recognition*. Accessed: Sep. 8, 2020. [Online]. Available: https://cs231n.github.io/transfer-learning/

[37] S. Bunrit, N. Kerdprasop, and K. Kerdprasop, ''Evaluating on the transfer learning of CNN architectures to a construction material image classification task,'' *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 201–207, Apr. 2019, doi: 10.18178/ijmlc.2019.9.2.787.

[38] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' Apr. 2015, *arXiv:1409.1556*.

[39] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, ''Fundus image classification using VGG-19 architecture with PCA and SVD,'' *Symmetry*, vol. 11, no. 1, p. 1, Dec. 2018, doi: 10.3390/sym11010001.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[41] S. Patel. (May 4, 2017). *Chapter 2: SVM (Support Vector Machine)—Theory*. Accessed: Jun. 7, 2020. [Online]. Available: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[42] R. Rashid, S. G. Khawaja, M. U. Akram, and A. M. Khan, ''Hybrid RID network for efficient diagnosis of tuberculosis from chest X-rays,'' in *Proc. 9th Cairo Int. Biomed. Eng. Conf. (CIBEC)*, Cairo, Egypt, Dec. 2018, pp. 167–170, doi: 10.1109/CIBEC.2018.8641816.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.SVM.SVC.html

[44] Kaggle. (2020). *Chest X-Ray Images With Pneumothorax Masks*. Accessed: Mar. 13, 2020. [Online]. Available: https://kaggle.com/vbookshelf/pneumothorax-chest-xray-images-and-masks

[45] Kaggle. (2017). *Random Sample of NIH Chest X-Ray Dataset*. Accessed: Sep. 6, 2020. [Online]. Available: https://kaggle.com/nih-chest-xrays/sample

[46] M. O. Faruqe and M. A. M. Hasan, ''Face recognition using PCA and SVM,'' in *Proc. 3rd Int. Conf. Anti-Counterfeiting, Secur., Identificat. Commun.*, Aug. 2009, pp. 97–101, doi: 10.1109/ICASID.2009.5276938.

[47] J.-D. Wu and C.-T. Liu, "Finger-vein pattern identification using SVM and neural network technique," *Expert Syst. Appl.*, vol. 38, pp. 14284–14289, Jun. 2011, doi: 10.1016/j.eswa.2011.05.086.

[48] R. P. R. Priya and P. Aruna, "SVM and neural network based diagnosis of diabetic retinopathy," *Int. J. Comput. Appl.*, vol. 41, no. 1, pp. 6–12, Mar. 2012, doi: 10.5120/5503-7503.

[49] M. da Silva Santos, M. Ladeira, G. C. G. Van Erven, and G. L. da Silva, "Machine learning models to identify the risk of modern slavery in Brazilian cities," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Boca Raton, FL, USA, Dec. 2019, pp. 740–746, doi: 10.1109/ICMLA.2019.00132.

[50] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Informat. Med. Unlocked*, vol. 20, 2020, Art. no. 100391, doi: 10.1016/j.imu.2020.100391.

[51] M. Degiorgis, G. Gnecco, S. Gorni, G. Roth, M. Sanguineti, and A. C. Taramasso, "Classifiers for the detection of flood-prone areas using remote sensed elevation data," *J. Hydrol.*, vols. 470–471, pp. 302–315, Nov. 2012, doi: 10.1016/j.jhydrol.2012.09.006.

[52] S. Mondal, K. Agarwal, and M. Rashid, "Deep learning approach for automatic classification of X-ray images using convolutional neural network," in *Proc. 5th Int. Conf. Image Inf. Process. (ICIIP)*, Shimla, India, Nov. 2019, pp. 326–331, doi: 10.1109/ICIIP47207.2019.8985687.

[53] B. Sun, H. Chen, J. Wang, and H. Xie, "Evolutionary under-sampling based bagging ensemble method for imbalanced data classification," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 331–350, Apr. 2018, doi: 10.1007/s11704-016-5306-z.

[54] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, May 2015, doi: 10.1016/j.patcog.2014.11.014.

[55] H. Wang, H. Gu, P. Qin, and J. Wang, "CheXLocNet: Automatic localization of pneumothorax in chest radiographs using deep convolutional neural networks," *PLoS ONE*, vol. 15, no. 11, Nov. 2020, Art. no. e0242013, doi: 10.1371/journal.pone.0242013.

[56] P. Choudhary and A. Hazra, "Chest disease radiography in twofold: Using convolutional neural networks and transfer learning," *Evolving Syst.*, vol. 12, no. 2, pp. 567–579, Jun. 2021, doi: 10.1007/s12530-019-09316-2.

[57] F. A. Noor, I. Munzerin, A. M. A. Iqbal, T. Islam, and E. Hossain, "An ensemble learning based approach to autonomous COVID19 detection using transfer learning with the help of pre-trained deep neural network models," in *Proc. 24th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2021, pp. 1–6, doi: 10.1109/ICCIT54785.2021.9689825.

TAHIRA IQBAL received the B.S. degree in computer engineering from the Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (CEME NUST), Pakistan, in 2018, and the M.S. degree in computer engineering from CEME NUST. Her research interests include deep learning, machine learning, and medical image analysis.

ARSLAN SHAUKAT received the B.S. and M.S. degrees in computer engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2003 and 2005, respectively, and the Ph.D. degree in computer science from The University of Manchester, U.K., in 2010. He is currently an Associate Professor with the Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, NUST (CEME NUST). He has published various research papers in refereed journals and conference proceedings. His research interests include machine learning, pattern recognition, digital image, and speech processing. He has been a member of technical program committees of numerous international conferences and a reviewer of international journals. He was a recipient of academic awards, including the Best Teacher Award in 2018 and the Best Research Paper Award in 2019.

MUHAMMAD USMAN AKRAM received the Ph.D. degree in computer engineering (specializing in medical imaging analysis) from the National University of Sciences and Technology, Pakistan. He is currently an Associate Professor with the Department of Computer and Software Engineering, CEME NUST. He is one of the youngest Ph.D. recipient in Pakistan and has published more than 115 research papers in peer-reviewed journals and conferences. He has received academic awards, including the Best University Teacher Award in 2016 and the Best Researcher Award in 2019. In addition, he has also worked as a Reviewer for many journals, including IEEE, Elsevier, and Springer.

ABDUL WAHAB MUZAFFAR received the Ph.D. degree in software engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2017. He is currently working as an Assistant Professor with Saudi Electronic University, Saudi Arabia. He is the author of 27 conference papers and journal articles, and has participated in several conferences held in United Arab Emirates, USA, and Thailand. His research interests include data and text mining, software engineering, machine learning, and bioinformatics. He is also an active reviewer of various scientific journals.

ZARTASHA MUSTANSAR received the Ph.D. degree from The University of Manchester, U.K. She was selected by Microsoft Research Cambridge (MSR) to pursue research in physical sciences and engineering in Manchester. She is currently employed as an Assistant Professor with the Research Center for Modeling and Simulation (RCMS), NUST. She has published 32 research papers in various peer-reviewed journals and conferences. Her research interest includes biomechanical engineering, especially associated with health care.

YUNG-CHEOL BYUN studied at the University of Florida as a Visiting Professor, from 2012 to 2014. He currently directs the Machine Learning Laboratory, Department of Computer Science, Jeju National University. Before joining Jeju National University, he worked as a Special Lecturer with Samsung Electronics Company Ltd., in 2000 and 2001. From 2001 to 2003, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI). He was promoted to join Jeju National University as an Assistant Professor, in 2003. He is also serving as the Director for the Information Science Technology Laboratory and other academic societies. He has been hosting the International Conference on Computers, Networks, Systems, and Industrial Engineering (CNSI); and serving as the program chair, the workshop chair, and the session chair for various international conferences and workshops.

• • •