A NEOTERIC ENSEMBLE DEEP LEARNING NETWORK FOR MUSCULOSKELETAL DISORDER CLASSIFICATION

S. Nazim[*], *S.S. Hussain*[†], *M. Moinuddin*[‡], *M. Zubair*[§], *J. Ahmad*[¶]

Abstract: The healthcare area is entirely different from other industries. It is of the highly significant area and people supposed to gain the utmost care and facilities irrespective of the cost. Reliable image detection and classification is considered a significant capability in medical image investigation problems. The key challenge is that the whole image has to be searched for a particular event and then classified accordingly but it is necessary to ensure that any important piece of information or instance shouldn't be skipped. With regards to image analysis by radiologists, it is quite restricted because of its partiality, the intricacy of the images, wide variations that happen amongst various analysts and weariness. However, the introduction of deep learning is a promising way to improve this situation by sorting out the issue according to human leaning mechanism consequently it brings high-tech changes in medical image classification problems. In this context, a new ensemble deep learning topology is being proposed in the direction of a more precise classification of musculoskeletal ailments. In this regard, a comparison has been accomplished based on different learning rates, drop-out rates, and optimizers. This comparative research proved to be a baseline to gauge the up-to-the-mark performance of the proposed ensemble deep learning architecture.

Key words: *deep learning, artificial intelligence, musculoskeletal disorder, machine learning*

Received: August 31, 2020 **DOI:** 10.14311/NNW.2021.31.021
Revised and accepted: December 31, 2021

[*]Sadia Nazim; Faculty of Information Technology, Department of Computer Science; Salim Habib University, NC-24, Deh Dih, Korangi Creek, Karachi-74900, Pakistan, E-mail: sadia.nazim@shu.edu.pk

[†]Syed Sajjad Hussain – Corresponding author; Faculty of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, SZABIST, Block 5 Clifton, Karachi-75600, Pakistan, E-mail: sshussainr@gmail.com , dr.sajjad@szabist.edu.pk

[‡]Muhammad Moinuddin; Electrical and Computer Engineering Department; King Abdulaziz University, Al Ehtifalat St, Jeddah-21589, Saudi Arabia, Center of Excellence in Intelligent Engineering Systems (CEIES), King Abdulaziz University, Al Ehtifalat St, Jeddah-21589, Saudi Arabia, E-mail: mmsansari@kau.edu.sa

[§]Muhammad Zubair; Faculty of Engineering, Sciences & Technology; Faculty of Computer Science, IQRA University, North Karachi Campus, Sector 7b North Karachi Twp, Karachi-75850, Pakistan, E-mail: zubair@iqra.edu.pk

[¶]Jawwad Ahmad; Department of Electrical Engineering; Usman Institute of Technology, ST-13, Block-7, Gushan-e-Iqbal, Abu-Hasan Isphahani Road, Karachi-75300, Pakistan, E-mail: jawwad@uit.edu

1. Introduction

Over the past few years, the requirements of medical imaging, such as Magnetic resonance imaging (MRI), Computed tomography (CT) scans and X-rays, have been significantly increased. The requirement and availability of medical imaging for disease diagnosis, is rapidly beating the capability of radiologists especially in low and middle-income countries. Hence this results in wrong decisions and ineffective treatment. Still, this problem is substantially solved by the advancement in artificial intelligence, which involves using the computer-aided system to analyze and investigate medical images so that reliable and timely treatment can be made possible [24]. While considering a roadmap from AI to deep learning, the phrase AI is simply a universal term consisting of problem-solving processes and procedures. Next, under the umbrella of AI, machine learning comes in which machines are to be learned to solve specific problems based on the provided information [42]. Subsequently, such networks have emerged, which are grounded in deep learning. Hence it may be stated that both ML and DL are the architectures of AI. Therefore DL is an explicit sort of ML [17] with the difference of automated feature extraction mechanism instead of the manual feature extraction process as in ML [31, 23, 29]. In the same way, DL networks outperform for massive datasets [39, 6] compared to smaller ones in which ML is the most suitable process. The primary constituent of the deep learning layer is the neuron, whose functionality is the imitation of human neurons, which has weight and bias against each input connection [14, 21].

Although deep learning shows outstanding performance in medical image classification, some of the challenges have also been encountered in medical image classification applications. The first hindering element is the provision of the high-grade annotated dataset. Deep learning algorithm yields excellent results by feeding the huge and wide variety of datasets which is the biggest challenge in such scenarios [18].

Even though several researchers have performed medical image classification with pre-trained deep learning models on their selected medical image datasets and improved the results and accuracies. Some applied ensemble learning but hybridized the pre-trained models with DL standard networks. Nobody works to develop their topology by the fusion of classic deep learning networks and perform training from inception against radiographs datasets. There is no pre-trained network available which has been trained, especially against medical image datasets from scratch. Hence the key objective of this research is to produce the most perfect and efficient mechanism with better-generalized behaviour. This motive is achieved by merging the benefits of classic deep learning topologies in conjunction with ensemble learning and train it against radiographs datasets to generate more competitive results. Hence an innovative ensemble product has been introduced that will turn out to be the more effective and powerful means of performing diagnosis, feature extraction, and classification against the radiograph datasets.

2. Background

2.1 Ensemble learning

The idea of ensemble learning is based on the very old saying, "unity is strength". It is a powerful machine learning technique in which various weak learning algorithms are blended appropriately in such a fashion to generate a powerful model. It considerably boosts the generalization capacity of the ensemble model. Ensemble learning starts with the selection of base models to be assembled. Ensemble models can be categorized as homogeneous and heterogeneous. The ensemble model is termed homogeneous in a case; identical dilute models are trained in distinctive manners [41]. The strong ensemble model is denoted as heterogeneous if divergent topologies are blended to obtain improved performance. There are three significant aggregating techniques in the direction of blending the base models. *Bagging* is one of the benchmark methods based on the parallel but independent learning of homogeneous learning algorithms on different datasets and then combines the multiple predictions generated by them. *Boosting* technique introduced the sequential aggregation of weak homogeneous base topologies and then combined the predictions in a very conclusive manner. *Stacking* methodology deals with the heterogeneous weak models that learn them in parallel and link them by train a meta-model to yield an estimation constructed on the different weak algorithms predictions [16].

2.2 Hyper-parameter selection

Expeditious and effective training of deep neural networks could be an art as far as a science. The truth is that these queries couldn't be answered. In deep learning or machine learning cases, network effectiveness varies deeply based on the selection of hyper-parameter values. Hyper-parameter interrogation aims to traverse different hyper-parameter arrangements to discover the best one, which gives maximum performance. Typically, the hyper-parameter interrogation process is exhaustively manual. It is proved that the exploration area is massive, and the estimation of each interrogation can be costly [14]. This section offers the summarized importance of these hyper-parameter values in estimating any deep learning model's performance.

2.2.1 Learning rate

The learning rate is an essential hyper-parameter, denoted by α (alpha). It is used to fine-tune how accurately a network congregates on a result and can perform classification or prediction. Moreover, the additional time causes raised in cloud GPU costs. In the same way, elevated learning rates might create such a model that could not perform reliable predictions. Therefore, the most authentic learning rate is the one that can converge the network within a reasonable period [36].

2.2.2 Drop-out

Deep Learning networks are getting started even more deepened and broader. In the deep neural network the goal is to attain the best accuracies with these extensive networks. However, in such types of networks, the primary issue is the over-fitting

problem. Hence, in 2012, the idea of drop-out arose, and this conception brought a revolutionary change in deep learning [22]. It completely transformed the idea of understanding the whole weights altogether to discover a small percentage of weights in the network in every iteration process [8, 1]. This technique answered the problem of over-fitting, which often occurs in large networks hence the possibility of developing the larger complex. Still, more precise Deep Learning networks may be increased [3, 38]. In advance of drop-out, regularization was considered a significant research area. These regularizations are not sufficient enough to solve the problem of over-fitting solely because of co-adaptation, which is the major problem of large networks. The traditional L1 and L2 regularization couldn't resolve this problem. Moreover, they also perform predictive ability-based regularization. The critical drawback of this technique would restrain the size and accuracy of the network [38]. Later a new regularization technique evolved, which overcome the problem of co-adaptation [22]. Now it may be possible to denser and broader networks that completely participate in prediction.

2.2.3 Optimization algorithm

Let's suppose the value of a cost function is updated against the performance parameters trained during the training process. The cost function computes the difference between forecasted values and estimated values [11]. This difference is an error, which is the mean of the difference between predicted values and actual values [5]. *Gradient descent* is the first and foremost important algorithm, [15], which is a significant way to calculate the optimum model's performance with minimum loss function values in neural networks [32]. The back-propagation method is used to train a neural network. According to this technique, model propagate in a backward direction along with the calculated gradient of the error function and update weights accordingly [14]. *Root mean square propagation optimizer* (*RMSprop*) is analogous to gradient descent except for methods of calculating the gradient. It constrains the fluctuations along the y-axis; hence the learning rate can be increased, and it may converge faster along the x-axis [19]. RMSprop differs from AdaGrad because AdaGrad vanishes the learning rate vigorously because of the increasing value of the denominator [27]. In this connection, there is an idea of decaying the denominator and avoid its instant growth [40]. *Adaptive gradient algorithm* (*AdaGrad*) permits the adjustment of the learning rate according to the parameters permits by the algorithm. Hence it formulates huge updates for variable settings and minor updates for normal parameters. It results in a learning rate smaller and smaller, and convergence becomes slower and slower, which causes a long time to train a model. Its main weakness is that its learning rate-η is always decreasing and decaying [40, 12]. This issue of diminishing learning rate is resolved in another algorithm called AdaDelta.

AdaDelta is the progression of AdaGrad, which settled the vanishing learning rate issue of AdaGrad. Rather than collecting the entire earlier squared gradients, it restrains to the specific number of previous gradients. It calculates the average at a particular time t, then based on the past average, it generates the current gradient [2]. *Adaptive moment estimation* (*Adam*) refers to Adaptive moment Estimation. It is also one of the methods to calculate the learning rate against each parameter.

Adam is amongst the most efficient algorithm as its convergence and learning speed are faster than others [25]. It also resolves other issues like decaying learning rate, slow convergence, high variance issues [40]. These problems may direct to fluctuation in the error function. Adam is an adjustable learning rate method. It merges the benefits of two SGD enhancements – (RMSProp) and (AdaGrad) and calculates distinct adjustable learning rates for different parameters [12, 26]. Initially, Adam updates the rapidly growing averages of gradients and the squared gradient. This algorithm first revises the exponential progressing averages of the gradient and the squared gradient, which are the estimates of the first and second moments [28].

3. Datasets

Large-scale datasets are considered as the fuel for the high-tech performance of deep learning models. One of the most extensive available musculoskeletal datasets is MURA. It consists of 14,863 upper extremity musculoskeletal images from 12,173 patients; among them, 9,045 are labelled as standard, and the remaining 5,818 are categorized as abnormal. Each finding consists of one or several images that are manually labelled as either normal or abnormal. The upper extremity includes the shoulder, humerus, elbow, forearm, wrist, hand, and finger [30]. Tab. II and Fig. 2 depict the normal and abnormal image distribution against seven classes of the MURA dataset and some of its sample images, respectively.

	Foot	Knee	Hip	Ankle
Normal	312	435	91	285
Abnormal	36	99	3	36
Total	348	534	94	321

Tab. I *Image distribution against classes of LERA dataset.*

	Elbow	Finger	Forearm	Hand	Humerus	Shoulder	Wrist
Normal	3114	3275	1309	4225	818	4481	5981
Abnormal	2221	2196	805	1659	735	4323	4206
Total	5335	5471	2114	5884	1553	8804	10187

Tab. II *Image distribution against classes of MURA dataset.*

The MURA dataset has been collected by the institutional review board, HIPPA compliant, and communication system (PACS) of Stanford University. This dataset has been collected over a long period, i.e., from 2001 to 2012. Therefore, it consists of a wide variety of musculoskeletal images. To cover the musculoskeletal images of both upper and lower extremities, the second dataset, which has been adopted for the training of the proposed DL model, is LERA (Lower extremity radio-graphs).

This dataset also covers a broad range of bones and joint abnormalities of lower extremity areas of the human body. It is also a diverse-natured dataset because of its collection over a wide range of time from 2003 to 2014. It is also being released HIPAA-compliant Stanford University Medical Center. Tab. I and Fig. 1 show standard and anomalous image dissemination and sample images against the foot, knee, ankle, or hip bones.

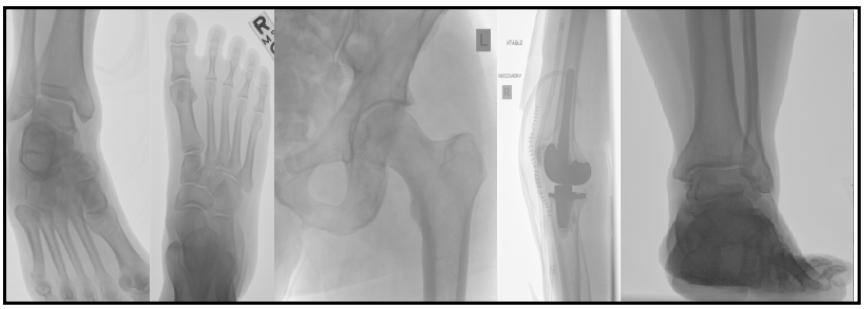

Fig. 1 *Sample images belong to different classes of the LERA dataset [35].*

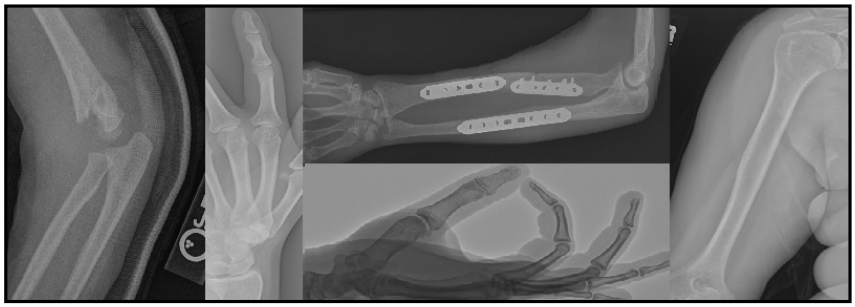

Fig. 2 *Sample images belong to different classes of the MURA dataset [30].*

4. Applications of deep learning in fracture detection

This section portrays the participation of different authors towards the contribution of deep learning models in fracture detection with the help of different types of medical images. For this purpose, an extensive literature review has been conducted and summarized in this section. A lot of research has been done in deep learning, especially for the applications concerned with the interpretation and diagnosis of medical images [10]. These tasks may vary from identifying and detecting diseases to the classification and analysis of abnormalities in medical images [20]. The development of a comprehensive deep learning network for medical images requires huge but labeled datasets to train the network. Numerous researches have revealed deep learning applications in fracture diagnosis. In [9], the author performed abnormality detection through the DL model and trained it against the

CT scans from the CQ500 and Qure25k datasets. The evaluation metrics of this classification are based on sensitivity, specificity with approximately 0.9 values for each. In [33] deep learning-based computer-aided diagnosis (CAD) system has been developed to classify the bone fracture images. The U-Net-based Deep CNN model has been trained to classify the MURA dataset and yielded better accuracies. They claimed that such a model would become a helpful tool for radiologists and doctors to analyze and investigate bone fractures. In [7], a comparative analysis among the pre-trained and state-of-the-art deep learning models have been performed in bone and joints against healthy vs unhealthy classification. As a result, this study provides knowledge about the best-performing deep learning model in this scenario. A large publically available musculoskeletal dataset named MURA is used in this research. In [13], the researchers proposed a deep convolution neural network model for automated proximal femur segmentation with the help of a dataset consisting of MR images and achieved very high scores for accuracies, precision, and recall against manual segmentation scores. In [34], the investigative study explored computer-aided classification and tracking of wrist fractures through the Inception-ResNet Faster R-CNN DL model. The ability of this model has been tested against the evaluation matrices such as sensitivity and specificity and achieved high scores of them, indicating the best performance of the proposed model. In [4], the researchers suggested an ensemble deep CNN approach to classifying the healthy and abnormal interstitial lungs through CT scans. The resulting performance showed the sensitivity, specificity, geometric mean, and accuracy score 88.9 %, 98.4 %, 93.5 %, 93.61 %, respectively. In [37], the study analyzes the recognition and identification of hip fractures with the help of deep learning pre-trained models, Inception-V3 and ImageNet, and performed classification by transfer learning. The performance of this transfer learning approach has been measured based on the AUC-ROC curve with scores of 0.9944, accuracy values of 96.9 %, and confusion matrix. Although several researchers have performed medical image classification with pre-trained deep learning models, the proposed topology excelled in many folds.

5. Implementation details of proposed ensemble network

The proposed model has been trained and tested against the classes for LERA and MURA. 70 % of datasets is used for training in this connection, and the remaining 30 % is utilized for testing. Some preprocessing has been done on both datasets before launching training and testing. The images are labelled properly based on normal and abnormal categories against each class so that the model could be trained for both types of images and perform classification perfectly.

The proposed topology is a 51 layered model developed by aggregating two classic deep learning models: CNN and LSTM but in different arrangements. Tab. III provides a complete picture of the arrangement and specifications of layers. The layer's arrangements have been performed either in series or parallel to attain the optimum performance of the model. Unfortunately, there is no set of rules for defining the best possible configuration of layers in the model. Instead, an experi-

Layers Arrangement	ConvBase Specifications	ConvBase Specifications	ConvBase Specifications
Input	sequence input layer(96*96*1)		
	sequenceFoldingLayer		
Parallel	Convolution2dLayer (3*3*8) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2)	Convolution2dLayer (3*3*16) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2)	Convolution2dLayer (3*3*32) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2)
	concatenationLayer(3)		
Parallel	Convolution2dLayer (3*3*8) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2) dropoutLayer	Convolution2dLayer (3*3*16) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2) dropoutLayer	Convolution2dLayer (3*3*32) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2) dropoutLayer
	concatenationLayer(3)		
Series	Convolution2dLayer (3*3*8) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2) ...	Convolution2dLayer (3*3*16) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2) ...	Convolution2dLayer (3*3*32) batchNormalizationLayer reluLayer maxPooling2dLayer(2,2) dropoutLayer
	sequenceUnfoldingLayer		
	flatten layers		
Series	Two lstmLayers		
	fullyConnectedLayer		
	softmax layer		
Output	outputLayer()		

Tab. III *Specifications of the proposed ensemble model.*

mental approach has been adopted to attain the best one. This model starts with the sequence input layer to ensure that the provided input must be in sequence form. Later, after passing this sequenced input through sequencedFoldingLayer, the input is ready to feed to 3 sets of convolution bases, which are arranged in a parallel manner. Each convolution base comprises convolution2d, batch normalization, relu, and maxpool2d layers. These initial convolution bases are used to extract the low-level features, then the extracted feature matrix is concatenated and handed over to 3 other parallel arranged convolution bases. Next to the parallel arrangement of convolution bases, there exist three serially arranged convolution bases. Finally, the convolution bases have identified enough feature details, so the resultant map is submitted to the sequence unfolding layer, which is later connected to the serially arranged two LSTM layers. After-ward, the LSTM layer delivered the results to a classifier that performs the classification of musculoskeletal images. This classifier consists of fully connected, softmax, and output layers.

6. Proposed ensemble network training parameters & performance analysis

There is no technique available to inquire about the best suitable values against the following hyper-parameters. The only method is the hit and trial method. The best parametric values for one dataset may work well for other datasets. However, there are some rules of thumb available for finding suitable values. Hence the trial and error approach is used in determining the best possible set of hyper-parameter values, as described in Tab. IV, for the optimum performance of the proposed model.

Datasets	No of Epochs	nLearning Rates	MiniBatchSize	Learning Algorithms	Drop-out Rates
MURA	40	0.002 0.001	30	Adam Sgdm Rmsprop	0.1 0.3 0.5 0.8
LERA	150	0.002 0.001	30	Adam Sgdm Rmsprop	0.1 0.3 0.5 0.8

Tab. IV *Proposed ensemble network training parameters.*

The performance analysis of the proposed model has been performed by considering the best two learning rates, 0.001 and 0.002, against three optimization algorithms. By taking into account four drop-out rates, training is accomplished. Still, the mini-batch size remains kept constant, i.e., 30, and the number of epochs for MURA and LERA datasets are set to be 40 and 150, respectively. Moreover, it is also compared with the previously available scores.

Tab. V effectively summarizes the performance analysis of the proposed model against all classes of LERA datasets at different values of hyper-parameters. Here the AUC-ROC values are taken into consideration because of the comparison with existing AUC-ROC values. The Foot class depicts that the AUC-ROC probability values of the proposed network at drop-out rate 0.5 with learning rate 0.002 and for Adam optimizer show better performance compared to the existing value. At the same time, the ankle class reveals its better performance compared to the existing value at drop-out rate 0.5 with learning rate 0.001 and for Adam optimizer. The comparison chart for hip class represents that at drop-out rate 0.3 with learning rate 0.002 and for rmsprop optimizer the proposed model's scores are the best. Whether the knee class reveals that the AUC-ROC probability values of the proposed network at drop-out rate 0.1 with learning rate 0.001 and for rmsprop optimizer supports better performance compared to the existing value.

Tab. VI summarizes that for forearm and hand classes, the proposed model exhibits its best performance at a drop-out rate of 0.5 with a learning rate of 0.002 and Adam optimizer. While in the case of the shoulder, wrist, and finger classes, the proposed network's accuracy at drop-out rate 0.5 with a learning rate of 0.001 and for Adam optimizer displays the maximum obtained scores. As far as humerus class is concerned, it summaries that the best accuracies of the proposed model

	Dropout rate 0.1		Dropout rate 0.3		Dropout rate 0.5		Dropout rate 0.8	
FOOT	LR0.001	LR0.002	LR0.001	LR0.002	LR0.001	LR0.002	LR0.001	LR0.002
Adam	88.5714	90.1687	96.3500	97.4138	87.7843	**97.0103**	79.0169	89.7102
Sgdm	93.8188	87.6376	72.6889	91.4894	50	64.8936	50	50
Rmsprop	91.4343	89.2517	92.6999	94.8826	93.5620	92.9017	91.2326	92.2414
Highest Existing Scores [35]				88.7				
HIP								
Adam	80	80	80	82.2222	89.4444	84.4444	58.3333	42.7778
Sgdm	73.8889	81.6667	68.8889	76.6667	50	70.5556	50	50
Rmsprop	74.4444	92.2222	89.4444	**100**	76.6667	89.4444	46.1111	68.3333
Highest Existing Scores [35]				93.4				
KNEE								
Adam	96.3158	97.3684	97.3684	98.9474	98.7045	96.3968	53.0769	65.5061
Sgdm	84.6154	96.3158	67.6923	93.8057	60.7692	74.0486	50	50
Rmsprop	**99.4737**	94.8583	97.4089	95.2632	97.8947	98.9474	62.5506	85.3441
Highest Existing Scores [35]				90.5				
ANKLE								
Adam	92.8758	88.3007	92.4837	92.0261	**98.0392**	95.9477	89.5425	77.0588
Sgdm	82.3529	91.8954	57.7778	87.7124	81.7647	75.4902	50	50
Rmsprop	93.0065	93.0065	95.6863	95.817	97.9085	91.1111	58.6275	86.6013
Highest Existing Scores [35]				81.3				

Tab. V *Proposed ensemble network training results against all classes of LERA dataset.*

occur at drop-out rate 0.3 with learning rate 0.002 and for Adam optimizer. The elbow class reviews that the proposed model accuracies at drop-out rate 0.5 with learning rate 0.001 and for rmsprop optimizer validates the best results.

7. Simulation results

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Any classification algorithm can be analyzed based on a summary generated by a confusion matrix. For example, if the dataset comprises several classes with unequal images, then the classification algorithm misrepresenting the accuracies; hence, the confusion matrix provides the understanding of errors and the types of errors of the classifier. So this section comprises of simulation results generated by the Deep Network Designer tool of MATLAB 2019. Confusion matrices and simulation graphs have been generated against all the classes of both datasets. Fig. 4 visualize the model performance through the graph and illustrate the confusion matrix for a single run against the foot class of the LERA dataset, respectively. The same is the case for Fig. 5 against the finger class of the MURA dataset.

The Figs. 6 and 7 demonstrate the aggregate values of confusion matrices for randomly selected 21 runs against the foot and finger class of LERA and MURA datasets.

	Dropout rate 0.1		Dropout rate 0.3		Dropout rate 0.5		Dropout rate 0.8	
FOREARM	LR0.001	LR0.002	LR0.001	LR0.002	LR0.001	LR0.002	LR0.001	LR0.002
Adam	71.4511	68.6120	68.1038	70.6625	65.2997	**74.0900**	72.0820	67.3502
Sgdm	71.4511	62.7760	65.6151	61.9874	63.5647	63.7224	61.8297	54.5741
Rmsprop	72.8707	71.9243	70.1893	69.0852	70.9779	71.2934	64.6688	70.5047
Highest Existing Scores [30]	73.7							
HAND								
Adam	75.9207	73.5977	76.2040	75.2408	75.5241	**76.1473**	69.5751	72.6346
Sgdm	71.5581	74.2210	72.2946	71.9547	71.6714	69.2351	71.7847	71.7847
Rmsprop	73.8244	71.7847	75.7507	71.9547	72.5779	73.5977	69.4051	74.1643
Highest Existing Scores [30]	85.1							
WRIST								
Adam	75.3927	73.7893	75.8508	68.8482	**78.1086**	77.8796	70.6152	74.7055
Sgdm	74.9673	73.6911	71.0079	73.5929	63.3181	67.0484	62.8272	42.2448
Rmsprop	75.2618	74.8691	76.7343	75.8181	77.6505	77.6505	67.1466	69.1427
Highest Existing Scores [30]	93.1							
HUMERUS								
Adam	69.4624	70.1075	72.6882	**74.1935**	72.6882	72.4731	69.8925	68.8172
Sgdm	69.0323	67.3469	61.9355	61.2903	52.9032	56.3441	52.6882	52.6882
Rmsprop	69.6774	69.8900	70.3226	72.9032	72.6882	72.4731	66.4378	71.6129
Highest Existing Scores [30]	60.0							
SHOULDER								
Adam	68.0016	69.4834	71.1254	70.2843	**71.871**	69.2831	66.0793	65.8791
Sgdm	70.1241	69.0429	67.3608	70.4846	61.674	65.839	52.0224	56.2275
Rmsprop	66.8002	69.0028	69.964	70.4045	68.9227	71.7261	68.1618	67.4409
Highest Existing Scores [30]	72.9							
FINGER								
Adam	71.1152	69.7136	70.2620	71.7855	**72.0293**	69.3100	64.2291	68.3729
Sgdm	68.6167	68.0683	68.4948	70.2011	64.8995	66.7276	49.1164	40.8288
Rmsprop	71.7855	70.9324	71.1152	69.2261	71.7855	69.4089	69.4089	67.0323
Highest Existing Scores [30]	38.9							
ELBOW								
Adam	68.5300	71.7391	75.0518	73.4990	72.4638	71.8427	70.2899	69.6687
Sgdm	72.3602	70.7039	69.6687	70.2899	68.9441	68.9441	68.9441	68.9441
Rmsprop	70.7039	70.1863	74.0166	74.6377	**75.7764**	73.6025	70.2899	70.2899
Highest Existing Scores [30]	71.0							

Tab. VI *Proposed ensemble network training results against all classes of MURA dataset.*

8. Conclusion & future work

This research has been inspired by the diffusion of deep neural networks in the medical imaging industry. Although several researchers have performed medical image classification, they mostly use pre-trained models for this purpose and try to achieve good accuracies. So in this study, a neoteric ensemble deep learning approach is developed. Its much-improved performance is recorded because of the best arrangements of layers and the combination of two different deep learning models. Unfortunately, there is no particular pattern available for the arrangement of layers and the combinations of classic models. Instead, the trial-and-error methodology has been adopted to achieve the best medical image classification. After discussing its detailed structure, a comparative analysis is being made based on the better of the two learning rates and four drop-out rates for three optimizers and concluded the best set of learning and drop-out rates and optimizer for each class. The ensemble model represents much improved results for all classes in the case of the LERA dataset.

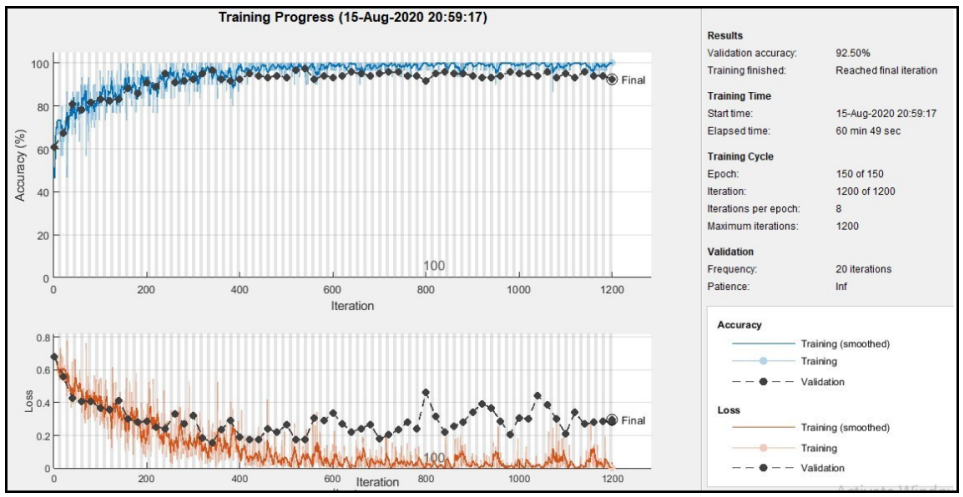

Fig. 3 *Simulation result for foot class of the LERA dataset, generated by Deep Network Designer tool of MATLAB 2019.*

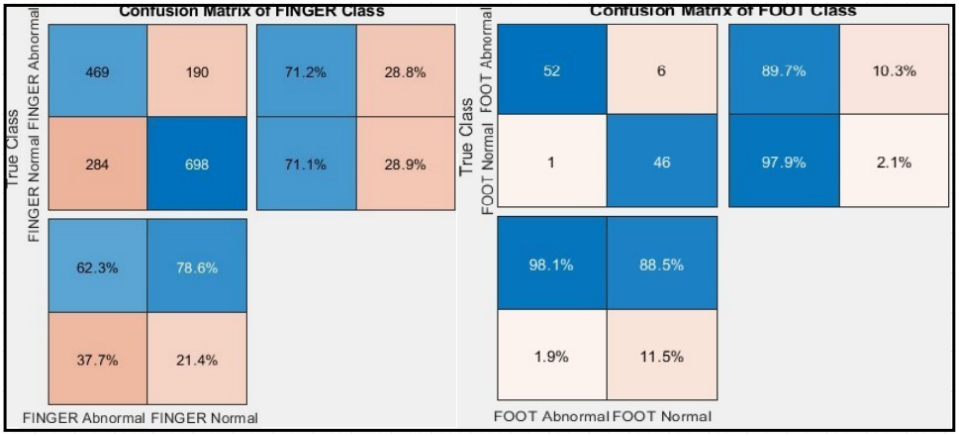

Fig. 4 *Confusion matrices for foot class of the LERA and finger class of the MURA datasets against the single run.*

In contrast, the MURA dataset depicts the improvement for four classes (forearm, finger, humerus, elbow), and scores for shoulder class are almost close to existing values. In cases of wrist and hand classes, the results are satisfactory but not improved. This comparative analysis proves that it is not recommended that there should be a single optimum learning and drop-out rates, but it depends upon varying factors. Hence it is concluded that for deep learning applications, the best values for learning and drop-out rates should be nominated against the best-selected optimizer for every class of a particular dataset.

For future work, we have planned to employ our proposed model against a wide array of datasets. However, the number of hyper-parameters such as number

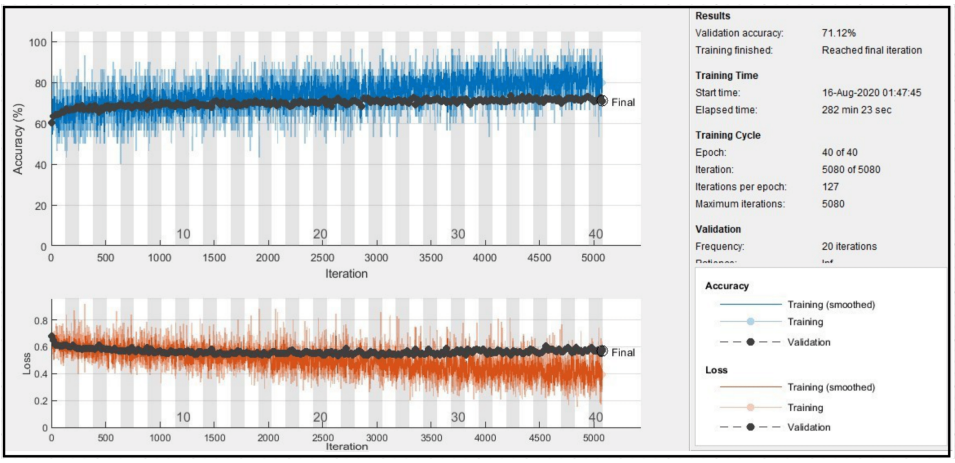

Fig. 5 *Simulation result for finger class of the MURA dataset, generated by Deep Network Designer tool of MATLAB 2019.*

$\begin{bmatrix}50 & 8\\4 & 43\end{bmatrix}$ Run 1	$\begin{bmatrix}55 & 3\\1 & 46\end{bmatrix}$ Run 2	$\begin{bmatrix}50 & 8\\5 & 42\end{bmatrix}$ Run 3	$\begin{bmatrix}54 & 4\\6 & 41\end{bmatrix}$ Run 4	$\begin{bmatrix}55 & 3\\0 & 47\end{bmatrix}$ Run 5	$\begin{bmatrix}57 & 1\\2 & 45\end{bmatrix}$ Run 6	$\begin{bmatrix}51 & 7\\4 & 43\end{bmatrix}$ Run 7
$\begin{bmatrix}53 & 5\\4 & 43\end{bmatrix}$ Run 8	$\begin{bmatrix}52 & 6\\2 & 45\end{bmatrix}$ Run 9	$\begin{bmatrix}50 & 8\\5 & 42\end{bmatrix}$ Run 10	$\begin{bmatrix}54 & 4\\6 & 41\end{bmatrix}$ Run11	$\begin{bmatrix}55 & 3\\0 & 47\end{bmatrix}$ Run 12	$\begin{bmatrix}57 & 1\\2 & 45\end{bmatrix}$ Run 13	$\begin{bmatrix}57 & 1\\4 & 43\end{bmatrix}$ Run14
$\begin{bmatrix}57 & 1\\5 & 42\end{bmatrix}$ Run 15	$\begin{bmatrix}51 & 7\\20 & 27\end{bmatrix}$ Run 16	$\begin{bmatrix}50 & 8\\5 & 42\end{bmatrix}$ Run 17	$\begin{bmatrix}55 & 3\\0 & 47\end{bmatrix}$ Run 18	$\begin{bmatrix}58 & 0\\8 & 39\end{bmatrix}$ Run 19	$\begin{bmatrix}58 & 0\\33 & 14\end{bmatrix}$ Run 20	$\begin{bmatrix}58 & 0\\47 & 0\end{bmatrix}$ Run 21

Fig. 6 *Confusion matrices for foot class of LERA dataset against the individual run.*

$\begin{bmatrix}359 & 300\\174 & 808\end{bmatrix}$ Run 1	$\begin{bmatrix}460 & 199\\289 & 693\end{bmatrix}$ Run 2	$\begin{bmatrix}460 & 199\\260 & 722\end{bmatrix}$ Run 3	$\begin{bmatrix}528 & 131\\456 & 526\end{bmatrix}$ Run 4	$\begin{bmatrix}368 & 291\\206 & 776\end{bmatrix}$ Run 5	$\begin{bmatrix}443 & 216\\247 & 735\end{bmatrix}$ Run 6	$\begin{bmatrix}443 & 216\\303 & 679\end{bmatrix}$ Run 7
$\begin{bmatrix}437 & 222\\241 & 741\end{bmatrix}$ Run 8	$\begin{bmatrix}393 & 266\\208 & 774\end{bmatrix}$ Run 9	$\begin{bmatrix}458 & 201\\262 & 720\end{bmatrix}$ Run 10	$\begin{bmatrix}464 & 195\\307 & 675\end{bmatrix}$ Run11	$\begin{bmatrix}343 & 316\\161 & 821\end{bmatrix}$ Run 12	$\begin{bmatrix}410 & 249\\256 & 726\end{bmatrix}$ Run 13	$\begin{bmatrix}402 & 257\\245 & 737\end{bmatrix}$ Run14
$\begin{bmatrix}310 & 349\\166 & 816\end{bmatrix}$ Run 15	$\begin{bmatrix}377 & 282\\235 & 747\end{bmatrix}$ Run 16	$\begin{bmatrix}525 & 134\\442 & 540\end{bmatrix}$ Run 17	$\begin{bmatrix}464 & 195\\307 & 675\end{bmatrix}$ Run 18	$\begin{bmatrix}364 & 295\\229 & 753\end{bmatrix}$ Run 19	$\begin{bmatrix}462 & 197\\292 & 690\end{bmatrix}$ Run 20	$\begin{bmatrix}467 & 192\\354 & 628\end{bmatrix}$ Run 21

Fig. 7 *Confusion matrices for finger class of MURA dataset against the individual run.*

and size of filters, pooling window, and the stride is being set on a trial-and-error basis and prior experience; hence, tuning the appropriate range of values is a tremendous task an active research area. Furthermore, these hyper-parameters values vary with respect to datasets and the deepness of DL model, so we have intended to develop such a generic technique that will help investigate the scale of values that are best suited against a particular dataset and for the variable depth of deep learning models. Apart from the learning and drop-out rates, the number of hidden nodes/layers and the size of the feature map, pooling window, and stride also contribute significantly during training and affect accuracy.

References

[1] ALEXANDER PAULS J.A.Y. Determining Optimum Drop-out Rate for Neural Networks. *The Bridge, The Magazine of IEEE-Eta Kappa Nu*. 2018, 115, pp. 10 –17.

[2] AULT R. Optimization Study of an Image Classification Deep Neural Network. *Honors Projects*. 2020.

[3] BALDI P., SADOWSKI P.J. Understanding dropout. *Advances in neural information processing systems*. 2013, 26, pp. 2814–2822, doi: 10.17744/mehc.25.2.xhyreggxdcd0q4ny.

[4] BERMEJO-PELÁEZ D., ASH S.Y., WASHKO G.R., ESTÉPAR R.S.J., LEDESMA-CARBAYO M.J. Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks. *Scientific reports*. 2020, 10(1), pp. 1–15, doi: 10.1038/s41598-019-56989-5.

[5] CABADA R.Z., RANGEL H.R., ESTRADA M.L.B., LOPEZ H.M.C. Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems. *Soft Computing*. 2020, 24(10), pp. 7593–7602, doi: 10.1007/s00500-019-04387-4.

[6] CAMERON W.D., BENNETT A.M., BUI C.V., CHANG H.H., ROCHELEAU J.V. Cell segmentation using deep learning: comparing label and label-free approaches using hyper-labeled image stacks. *bioRxiv*. 2020, doi: 0.1101/2020.01.09.900605.

[7] CHAKRAVARTY A., GHOSH N., SHEET D., SARKAR T., SETHURAMAN R. Radiologist Validated Systematic Search over Deep Neural Networks for Screening Musculoskeletal Radiographs. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2020.

[8] CHEN L., GAUTIER P., AYDORE S. DropCluster: A structured dropout for convolutional networks. *arXiv preprint arXiv:2002.02997*. 2020.

[9] CHILAMKURTHY S., GHOSH R., TANAMALA S., BIVIJI M., CAMPEAU N.G., VENUGOPAL V.K., MAHAJAN V., RAO P., WARIER P. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*. 2018, 392(10162), pp. 2388–2396, doi: 10.1016/S0140-6736(18)31645-3.

[10] D SOUZA R.N., HUANG P.-Y., YEH F.-C. Structural analysis and optimization of convolutional neural networks with a small sample size. *Scientific reports*. 2020, 10(1), pp. 1–13, doi: 10.1038/s41598-020-57866-2.

[11] DARWISH A., EZZAT D., HASSANIEN A.E. An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis. *Swarm and Evolutionary Computation*. 2020, 52, pp. 100616, doi: 0.1016/j.swevo.2019.100616.

[12] DÉFOSSEZ A., BOTTOU L., BACH F., USUNIER N. A Simple Convergence Proof of Adam and Adagrad. *arXiv preprint arXiv:2003.02395*. 2020.

[13] DENIZ C.M., XIANG S., HALLYBURTON R.S., WELBECK A., BABB J.S., HONIG S., CHO K., CHANG G. Segmentation of the proximal femur from MR images using deep convolutional neural networks. *Scientific reports*. 2018, 8(1), pp. 1–14, doi: 10.1038/s41598-018-34817-6.

[14] DO S., SONG K.D., CHUNG J.W. Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning. *Korean journal of radiology*. 2020, 21(1), pp. 33, doi: 10.3348/kjr.2019.0312.

[15] FARAJTABAR M., AZIZAN N., MOTT A., LI A. Orthogonal gradient descent for continual learning. In: *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 3762–3773.

[16] GANAIE M., HU M., et al. Ensemble deep learning: A review. *arXiv preprint arXiv:2104.02395*. 2021.

[17] HUSSAIN F., HASSAN S.A., HUSSAIN R., HOSSAIN E. Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges. *IEEE Communications Surveys & Tutorials*. 2020, 22(2), pp. 1251–1275, doi: 10.1109/comst.2020.2964534.

[18] IRMAKCI I., ANWAR S.M., TORIGIAN D.A., BAGCI U. Deep Learning for Musculoskeletal Image Analysis. In: *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1481–1485.

[19] JIANG X., HU B., CHANDRA SATAPATHY S., WANG S.-H., ZHANG Y.-D. Fingerspelling identification for Chinese sign language via AlexNet-based transfer learning and Adam optimizer. *Scientific Programming*. 2020, 2020, doi: 10.1155/2020/3291426.

[20] KALMET P.H., SANDULEANU S., PRIMAKOV S., WU G., JOCHEMS A., REFAEE T., IBRAHIM A., HULST L.v., LAMBIN P., POEZE M. Deep learning in fracture detection: a narrative review. *Acta orthopaedica*. 2020, 91(2), pp. 215–220, doi: 10.1080/17453674.2019.1711323.

[21] KIRAN R, KUMAR P., BHASKER B. DNNRec: A novel deep learning based hybrid recommender system. *Expert Systems with Applications*. 2020, 144, pp. 113054, doi: 10.1016/j.eswa.2019.113054.

[22] KRAUS M., FEUERRIEGEL S., OZTEKIN A. Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*. 2020, 281(3), pp. 628–641, doi: 10.1016/j.ejor.2019.09.018.

[23] LIU L., OUYANG W., WANG X., FIEGUTH P., CHEN J., LIU X., PIETIKÄINEN M. Deep learning for generic object detection: A survey. *International journal of computer vision*. 2020, 128(2), pp. 261–318, doi: 10.1007/s11263-019-01247-4.

[24] LIU X., FAES L., KALE A.U., WAGNER S.K., FU D.J., BRUYNSEELS A., MAHENDIRAN T., MORAES G., SHAMDAS M., KERN C., et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*. 2019, 1(6), pp. e271–e297, doi: 10.1016/S2589-7500(19)30123-2.

[25] LU Z., LU M., LIANG Y. A Distributed Neural Network Training Method Based on Hybrid Gradient Computing. *Scalable Computing: Practice and Experience*. 2020, 21(2), pp. 323–336.

[26] NAMAN AGARWAL R.A., HAZAN E., KOREN T., ZHANG C. Disentangling adaptive gradient methods from learning rates. *arXiv preprint arXiv:2002.11803*. 2020.

[27] NGUYEN L.C., NGUYEN-XUAN H. Deep learning for computational structural optimization. *ISA transactions*. 2020, 103, pp. 177–191, doi: 10.1016/j.isatra.2020.03.033.

[28] NHU V.-H., HOANG N.-D., NGUYEN H., NGO P.T.T., BUI T.T., HOA P.V., SAMUI P., BUI D.T. Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area. *Catena*. 2020, 188, pp. 104458, doi: 10.1016/j.catena.2020.104458.

[29] NWANKPA C., IJOMAH W., GACHAGAN A., MARSHALL S. Activation functions: Comparison of trends in practice and research for deep learning. 2018, pp. 1–20.

[30] RAJPURKAR P., IRVIN J., BAGUL A., DING D., DUAN T., MEHTA H., YANG B., ZHU K., LAIRD D., BALL R.L., et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*. 2017.

[31] SENGUPTA S., BASAK S., SAIKIA P., PAUL S., TSALAVOUTIS V., ATIAH F., RAVI V., PETERS A. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*. 2020, 194, pp. 105596, doi: 10.1016/j.knosys.2020.105596.

[32] SUN J., NIU Z., INNANEN K.A., LI J., TRAD D.O. A theory-guided deep-learning formulation and optimization of seismic waveform inversion. *Geophysics*. 2020, 85(2), pp. R87–R99, doi: 10.1190/geo2019-0138.1.

[33] TANZI L., VEZZETTI E., MORENO R., MOOS S. X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach. *Applied Sciences*. 2020, 10(4), pp. 1507, doi: 10.3390/app10041507.

[34] THIAN Y.L., LI Y., JAGMOHAN P., SIA D., CHAN V.E.Y., TAN R.T. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*. 2019, 1(1), pp. e180001, doi: 10.1148/ryai.2019180001.

[35] VARMA M., LU M., GARDNER R., DUNNMON J., KHANDWALA N., RAJPURKAR P., LONG J., BEAULIEU C., SHPANSKAYA K., FEI-FEI L., et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nature Machine Intelligence*. 2019, 1(12), pp. 578–583, doi: 10.1038/s42256-019-0126-0.

[36] YU C., QI X., MA H., HE X., WANG C., ZHAO Y. LLR: Learning learning rates by LSTM for training neural networks. *Neurocomputing*. 2020, 394, pp. 41–50, doi: 10.1016/j.neucom.2020.01.106.

[37] YU J., YU S., ERDAL B., DEMIRER M, GUPTA V, BIGELOW M, SALVADOR A, RINK T, LENOBEL S., PREVEDELLO L., et al. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. *Clinical radiology*. 2020, 75(3), pp. 237–e1, doi: 10.1016/j.crad.2019.10.022.

[38] YU T., ZHU H. Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*. 2020, pp. 1 –56.

[39] YUAN Q., SHEN H., LI T., LI Z., LI S., JIANG Y., XU H., TAN W., YANG Q., WANG J., et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*. 2020, 241, pp. 111716, doi: 10.1016/j.rse.2020.111716.

[40] ZHONG H., CHEN Z., QIN C., HUANG Z., ZHENG V.W., XU T., CHEN E. Adam revisited: a weighted past gradients perspective. *Frontiers of Computer Science*. 2020, 14(5), pp. 1–16, doi: 10.1007/s11704-019-8457-x.

[41] ZHOU T., LU H., YANG Z., QIU S., HUO B., DONG Y. The ensemble deep learning model for novel COVID-19 on CT images. *Applied Soft Computing*. 2021, 98, pp. 106885, doi: 10.1016/j.asoc.2020.106885.

[42] ZINCHUK V., GROSSENBACHER-ZINCHUK O. Machine learning for analysis of microscopy images: A practical guide. *Current Protocols in Cell Biology*. 2020, 86(1), pp. e101, doi: 10.1002/cpcb.101.