CS-550 - Spring 2023 Survey: Ensemble Learning Applications and Techniques for Medical Image Classification

Salih Deniz Uzel (deniz.uzel@bilkent.edu.tr)
22201382

***Abstract*—This survey examines ensemble learning techniques in medical image classification studies, including deep neural networks. Efficiency and applicability of the applications of ensemble learning techniques in Medical Image Classification comparatively analyzed. The latest comprehensive study made on this field is reviewed [1].**

***Index Terms*—ensemble learning, medical AI, machine learning, medical image classification.**

I. Introduction

In the field of medicine, imaging tools have an important place in diagnosis and treatment. Doctors make decisions by evaluating the data obtained from these devices with other auxiliary diagnostic methods. Today, medical experts and computer scientists are working on software improvement and development of post-processing tools in order to obtain better results from imaging devices. With the spread of artificial intelligence techniques and the increase in practitioners, researches are carrying out studies for the use of these techniques in the field of medicine for diagnosis, treatment regulation and risk estimation. Applications made in this field, which is still very new and developing, aim to expand diagnostic services, to alleviate the shortage of experts and financial burden, and to find solutions to existing problems in the medical field. Although much work has been done in the field of Medical AI, very few have been used in real life for safety and regulatory reasons. The reporting guidelines were published in 2019 to check the applicability of the studies, standardize and make them reliable [6], [7]. In 2020, a study called Standards for Reporting of Diagnostic Accuracy Studies (STARD) was carried out for standardization in the field of medical AI [8].

Ensemble learning has become a popular research area in medical image analysis due to its ability to improve the robustness and accuracy of classification models. By combining the predictions of several models, ensemble learning can surpass the single models, produce more accurate predictions [11]. In addition, ensemble learning techniques are used together with feature extraction methods and various statistical methods used in sampling to find solutions to domain-specific problems. The use of ensemble learning techniques has increased in the field of Medical AI, especially in the fields of radiology, ophthalmology, gastroenterology, pathology imaging [2]. Along with the detection of diseases from images, studies have also begun on estimating risk factors for future diseases. For example, in a study published in 2019, the AI model was trained to detect the risk of cancer that may occur in 3 years [3]. These estimates were then used to schedule CT scan dates for patients. On the contrary, a study conducted in 2007 states that CT scans cause 0.4% of cancer cases in the United States and this rate will increase between 1.5% and 2% in the future [4] and Huynh's study group carried out a study in which they could obtain CT images using MRI images with ensemble learning techniques as a solution to CT-induced cancer study [4], [27]. In a study published in 2013, high-performance models were developed in terms of run-time complexity and classification accuracy for the detection of Diabetic retinopathy [28]. The developed method has been used successfully in disease screening in this field, where the practitioners of a profession such as Ophthalmology, which requires intensive training and expertise, are few but relatively large numbers of patients. In the study, time spent on eye screening procedures for the detection of retinopathy shortened considerably and the success rate was increased [9], [10]. Ensemble learning techniques applied in medical image classification not only diagnose the pathology of an organ from the images of that organ, but also make it possible to detect a different disease from a part of the body. In the study by Mitani et al., Anemia was detected from retinal fundus images by ensemble of deep learning models [5].

The purpose of this research is to provide an comparative overview of common ensemble learning techniques and applications in the medical image classification field. Additionally, it aims to provide insights about the research topics and application areas of the usage of ensemble learning techniques in the field of medicine.

II. Preliminary

Ensemble learning techniques aim to increase the prediction accuracy of a machine learning model. The techniques combines multiple models and try to produce final predictions that is more inclusive in terms of generalizing the task compare to single model. They can effectively benefit from the properties of different models and structures. Ensemble learning techniques can be helpful when the data is noisy, unbalanced or model is over-fitting [11]. In the medical image classification domain the amount of data samples is insufficient and their distribution highly unbalanced or unknown. In such cases ensemble models might be a direct solution to these problems or it helps enhancing the accuracy of the models.

Ensemble learning techniques are widely used in the field of medical image classification to improve the accuracy of predictions such as bagging, boosting, and stacking. The approach for selecting the model depends on on the given problem and with proper hyper-parameters, decision model can easily surpass the single non-ensemble algorithms. However, ensemble models requires great computational power and cherry-picking the best results may cause over-fitting problem. In machine learning, ensemble methods such as bootstrapping, bagging, random forests, boosting are commonly used and shown to be very effective for improving the decision model performance in medical image classification field [1], [23].

A. Stacking

The stacking technique involves training multiple models from the same data. And the prediction of these models which are also called ”base models” are used by main model as input features. The main model called ”meta model” is trained with these features [12]. It learns to use and combine these predictions and make the final prediction. Models that are used for stacking can be any type of machine learning model. One disadvantage is, since the multiple models are capable of learning different representations, the main model is more prone to over-fitting.

B. Bootstrap Aggregation

Bootstrap Agraggation (Bagging) and Random Forest methods are similar to Stacking in terms of training more than one model [13], [14], [16]. However, predictions are not given as inputs to another model to train with. While multiple different type of models can be trained in Bagging, the Random Forests is only created from decision trees. The prediction methodology of the both ensemble method in general is averaging or majority voting depends on whether it is a regression task or classification task. Random Forest can be seen as a bagging method as it uses Bootstrapping and multiple models. They differ in the use of different types of models. The sampling method used for both bagging and random forests is called ”Bootstrapping”.

Bootstrapping [14], in statistic, is a technique that resampling a dataset by randomly selected samples from the main dataset with replacement. With replacement means that the randomly selected data can be selected again. For each model, the main dataset is re-sampled with replacement in order to capture the different patterns. Models are trained with these sub-sampled datas. Using bootstrapping and selecting different features can prevent or reduce over-fitting. For the evaluation, the samples that are not selected during Bootstrapping are used to test that model. This process also called out-of-bag (OOB) [16]. Bootstrapping method can also help sampling in such situation where the real life distribution of the classes are unknown. Due to random sampling with replacement feature, bootstrapping helps the ensemble model to learn under represented class samples to some extend. In addition to that, using class weight might be a solution to that problem but this solution comes with newly introduced variables such as over-fitting, generalization problems, biased performance estimates. These problems can also be encountered when applying the above ensemble methods, but an additional hyper-parameters increases the interpretability of the model. Additionally, there are studies showing that overfitting can be reduced by using the bootstrap method in stacking [15].

C. Boosting

In boosting ensemble method, models are trained sequentially. The models performs slightly better than random guessing however not accurate enough to be used as a classifier. Therefore, they are called weak classifiers [17]. Each models are trained on a sampled data set which it’s sample distribution depends on the previous model’s classification performance [18]. Respectively, the first model is trained on the input data set. After testing this model, new input data set is sampled in a way that miss-classified samples are selected more. Subsequently, the next model is trained on this newly sampled dataset. The performance of the boosting ensemble model is evaluated by getting the weighted voting of the model based on their prediction performance. By looking at the sampling strategy, it can be said that, the boosting is helping the less represented samples to be learned by increasing the weight of the miss-classified classes. In boosting, data sampling depends on the classification performance of the models. Samples are selected in favor of miss-predicted samples and models are tested during the each step in the sequence [18], [19]. On the other hand, for the bagging method, sampling is made with the same probability of being selected for each sample and models are trained and tested in parallel.

D. Hybrid

Hybrid ensemble learning is a technique that used to improve the accuracy and generalization of prediction systems by combining different ensemble learning techniques strengths. However, due to their complex nature, they can be computationally expensive, although they are helpful in sampling, feature selection, training, evaluation, and pruning processes. Different sample selection methods such as bootstrapping methods can be used to select training samples for medical image classification models to handle unknown distributions and unbalanced datasets. In the parallel architecture, multiple ensemble methods are trained independently [23]. Additionally, trained models can be pruned with respect to their contribution to prediction accuracy. For example, multiple deep learning models can be combined, trained, and evaluated using a bagging method. Alternatively, models can be cascaded, which is also known as Cascaded Ensemble Learning. In this approach, one model can be used to learn the discriminant features of the task, while an ensemble learning method can be used to select features for another model's input [20], [21]. Hybrid models can be seen as a combination of solutions produced to the problems encountered and they are being successfully implemented in the domain of medical image classification.

III. Studies

Ensemble learning techniques are already techniques that require high computational power. In addition to the classification models used in ensemble learning techniques today, deep learning models, which have been used quite a lot recently, have also been included. In this section, ensemble learning techniques are gathered under two groups, traditional ensemble learning techniques which do not use deep learning models and deep ensemble learning techniques which includes deep neural networks for the ensemble models. We included the studies in both groups on the common points, differences and usage areas of traditional ensemble learning and deep ensemble learning.

A. Traditional Ensemble Learning Techniques

The biggest limitation in the field of Medical Image classification is insufficient data. Since machine learning models are built on statistical distribution, data is essential for the creation of accurate and generalization of the systems. It is very difficult to obtain correctly labeled data in the medical field. For this reason, ensemble learning techniques have been used extensively with sampling and feature selection. Traditional models seem to have data sampling and feature selection as a common point in the studies in the same topic area.

Studies have compared the proposed model performances by using different sampling and feature selection methods, and they have shown the contribution of sampling and feature selection methods to performance [23], [27]. Another common point was image pre-processing [23], [27], [28]. In the study conducted for the detection of diabetic retinopathy called DREAM, in which 28 training and 61 test images were used in the training of the models, various calibrations were made in order to use the features that will benefit the diagnosis in the image extensively. In the retinopathy [28], pre-processing for the detection of bright and red lesions had a direct effect on accuracy. In other studies, pre-process operations such as exposure equalization, removal of variance, and contrast adjustment were carried out for the detection of pathology in line with the knowledge of the domain expert [23], [27], [28], [33].

In the study by Kuncheva et al., the research group tested and compared single classifiers and ensemble learning techniques on brain Functional Magnetic Resonance Imaging (fMRI) data [23]. The researchers used the Random Subspace Sampling (RS) technique [24], which involves generating multiple subsets of the features for training multiple models. For the classification, majority voting and averaging is used. To determine the best RS ensemble size and RS feature size, multiple ensemble models consisting of Support Vector Machines (SVM) were trained. Best configuration determined by creating a kappa-error diagram [25]. RS with SVMs outperformed other ensemble learning models such as AdaBoost, Bagging, Random Forests. The closest scores are Bagging with SVMs and Single SVM, respectively, followed by Multi Layer Perceptron (MLP).

The popularity and success of SVM stand out in the studies reviewed. However, I could not comment on whether this high accuracy is due to the fact that the practitioner has more knowledge and experience about this method or because of the success of the SVM method in this area. Takemura et al. used adaboost for Discrimination of Breast Tumors in Ultrasonic Images and compared the results with SVM [26]. Adaboost achieved 0% error rate, while SVM achieved an average of 0.8% error rate. It is difficult to understand whether the models included in the study have overfit and generalization problems. Since there is no standardization until 2020 for the studies in the field of Medical Image classification [8], it is very difficult to draw a deterministic result from among the studies. An extensive work done in 2021 on the classification of chest radiology results with some standardization by using high computational power will be explained in the deep ensemble learning section. Overall, over-fitting and generalization problems seem to depend on hyper-paramaters of the ensemble models. Each highly cited successful study has extensive feature selecting and sampling strategy along with the data pre-processing. For this reason, I also examined the studies conducted with traditional ensemble learning techniques in the medical field with relatively different motivations. One of them is a study by Hyunh et al [27].

The goal of the study by Hyunh et al is to predict CT scan images from fMRI images based on the 2007 study showing that CT scans cause cancer [4], they conducted this research for the possibility of reducing CT use. The main problem is that areas with bone and air cannot be visualized well in MRI, but can be easily distinguished on CT. Their main strategy is focused on feature extraction. Researches trained a structured random(SR) forests with leave-one-out cross validation method using a different strategy. To train structurel random forest they use extract multiscale features from MRI images and corresponding CT images. Model is evaluated on the train set and then they extract new features from these predicted images. And these features are used to train new structured random forests. They repeat the process until convergence. To evaluate the success of the prediction Mean Absolute Error (MAE) and Peak Signal-to-signal noise ratio is used. The proposed method achieved higher accuracies compared to classic RF, classic SR and Atlas method.

Ensemble learning techniques are widely used in the field of ophthalmology. One of the highly cited example is DREAM study [28]. The main purpose in of the paper is detecting the diabetic retinopathy with the highest sensitivity AUC evaluation method. Undetected retinopathy 90% of the time ends with sight loss and requires great expertise to detect.

Most of the popular ensemble learning techniques are used in this study, which also includes intensive image pre-processing. The research group tries to achieve sensitivity rate 1. They preprocess 28 training image and extract features by using adaboost method and rank them. For the feature selection they used a method that is proposed by Talavera et al. similar to decision tree prunning process [29]. Classification made cascade, in two hierarchical step. In the first step segmentation is made and in the second step Lesion are classified. Its is observed that the cascade classfication decreases the time complexity around 18-24 percent compare to parallel methods. Researcher used Gaussian Mixture Models (GMM), SVM, AdaBoost, SVM and GMM in parallel and SVM and k-nearest neighbors (kNN) in parallel as classifier in both stages. They indicate that SVMs' are sensitive to data imbalances. One feature that distinguishes this study from other studies is the measurement of time complexities of classifiers and ensemble models. The research team preferred models are, in terms of both time complexity and accuracy, kNN for the classifying red lesions and GNN for the bright lesions.

B. Deep Ensemble Learning Techniques

Deep Neural Networks (DNN) have the ability to learn complex patterns. However, a very large amount of data is needed for deep models to be generalizable without overfitting. If the data is scarce, the parameters of the a model trained on a large number of labeled data for different task, can also be used in another classification task. This process is called transfer learning [30]. Due to the scarcity of data, Transfer Learning is widely used for medical image classification field such as classification, segmentation, and data generating. Segmentation networks have various applications. In the study by G. Wang et al., it is utilized as a solution to noisy data [31]. In this study a self ensembling network is used [32]. At the each training run, a small random noise added to training data of the adapter teacher and adapter student network. After iteration predictions combined and help the models ability generalize with limited amount of data. The model was developed as a solution to noisy data. They used Dice-Loss and MAE for the noisy data loss. In this way, it is claimed that the segmentation accuracy of Covid19-induced pneumonia surpassed other noise-robust models.

Many studies are carried out in the field of machine learning for the detection and prediction of eye diseases. In the highly cited study by De Fauw. et al., researches try to predict 4 different age related eye degenaration with an ensemble of deep learning models [33]. Model combines 5 segmentation and 5 classification models. Manually segmented 3d images and 3d Digital Optical Coherence Tomography (OCT) Scans feeded into ensemble of 3D U-Net segmentation network [34], [35]. This work is the first example where three-dimensional (3d) diagnostic scans used for classification task in this domain. Ensemble segmentation networks outputs Tissue-segmentation map. These output are combined with the tissue maps that have confirmed diagnosis referral decision and evaluated with weighted predictions. Ensemble model outputs for different predictions which are Urgent, Semi-Urgent, Routine, Obsevartion. Proposed method achieves more than 96% area under the ROC curve and outperformed State of the Art (SotA) models. Model achieved 94.5% and 96.6% accuracy on the images of two different devices. Ensemble approach improved the accuray compare to single model. Researchers claimed that the two-stage classification helped the prediction model to achieve better results compare to previous works.

One of the largest and most systematic studies in this area was conducted by the Stanford ML group. Stanford ML group collected 224,316 chest radiographs (X-ray and CT images) from 65,240 patients for 14 different pulmonory diseases [38]. They used a single pretrained 121-layer DenseNet architecture trained on ImageNet dataset [36], [37]. Model outperformed the radiologist scores. In 2019, they made a comptetion on classifying 14 pulmonary diseases. Also authors repeated the study with multiple deep Convolutional Neural Networks (CNN) trained on ImageNet data. However, 121-layer DenseNet outperformed radiology scores [39] again. In 2020, they published the radiograph dataset and opened a classification competition called CheXpedition [40]. Differently this time, Deep CNNs consisting of convolutional neural networks were used as ensembles in all of the top 10 scores in the competition. In 2021, research repeated with immense amount of computation power and most of the SotA Deep CNN Networks are trained and tested [1]. As a result, no correlation found between ImageNet performance and Radiology Image performance of pre-trained and without pre-trained DNN models. It has been found that model architecture is more effective than the depth in the performance of non-pre-trained models.They found that the models trained with ImageNet gave a boost to accuracy and the model depth was not very effective. Additionaly, the models were truncated and gave almost the same results with 3.25x fewer parameters. These models outperformed the top 10 models of the CheXpedition competition, which were ensemble models. Additionally, all models trained by the research group were used Bootstrapping for sampling technique like in the most of the other studies given.

IV. Conclusion

Ensemble learning techniques are widely applied in medical image classification field to improve accuracy and robustness. The studies reviewed in this survey have shown that ensemble learning techniques can be used with different types of medical images such as MRI, CT. Additionally, the ensemble techniques can be combined with the latest machine learning models. In cases where there is not enough data, transfer learning technique gives good results. Considering that the differences between the models vary according to the needs and competencies of the study groups, the studies reviewed show the high performance achieved by using ensemble learning techniques and some studies have benefited people in real life applications. The studies shows a promising direction for future research in the field of medical AI.

REFERENCES

[1] Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021, April). CheXtransfer. Proceedings of the Conference on Health, Inference, and Learning. doi:10.1145/3450439.3451867

[2] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

[3] Huang, P. et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. Lancet Digit. Health 1, e353–e362 (2019).

[4] D. J. Brenner and E. J. Hall, ”Computed tomography—An increasing source of radiation exposure”, N. Eng. J. Med., vol. 357, no. 22, pp. 2277-2284, 2007.

[5] Mitani, A., Huang, A., Venugopalan, S. et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng 4, 18–27 (2020). https://doi.org/10.1038/s41551-019-0487-z

[6] The CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nat Med 25, 1467–1468 (2019). https://doi.org/10.1038/s41591-019-0603-3

[7] Hopewell, S., Boutron, I., Chan, AW. et al. An update to SPIRIT and CONSORT reporting guidelines to enhance transparency in randomized trials. Nat Med 28, 1740–1743 (2022). https://doi.org/10.1038/s41591-022-01989-8

[8] Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol Sounderajah, V., Ashrafian, H., Golub, R. M., Shetty, S., De Fauw, J., Hooft, L., Moons, K., Collins, G., Moher, D., Bossuyt, P. M., Darzi, A., Karthikesalingam, A., Denniston, A. K., Mateen, B. A., Ting, D., Treanor, D., King, D., Greaves, F., Godwin, J., Pearson-Stuttard, J., … STARD-AI Steering Committee (2021). Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ open, 11(6), e047709.

[9] Wolf, R. M., Channa, R., Abramoff, M. D. Lehmann, H. P. Cost-effectiveness of autonomous point-of-care diabetic retinopathy screening for pediatric patients with diabetes. JAMA Ophthalmol. 138, 1063–1069 (2020).

[10] Xie, Y. et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit. Health 2, e240–e249 (2020).

[11] Dong, X., Yu, Z., Cao, W., Shi, Y., &amp; Ma, Q. (2019, August 30). A survey on Ensemble Learning - Frontiers of Computer Science. SpringerLink. Retrieved April 3, 2023, from https://link.springer.com/article/10.1007/s11704-019-8208-z

[12] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. doi:10.1016/S0893-6080(05)80023-1

[13] Breiman, L. (1996) Bagging Predictors. Machine Learning, 24, 123-140. https://link.springer.com/content/pdf/10.1007/BF00058655.pdf https://doi.org/10.1007/BF00058655

[14] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1), 1–26. http://www.jstor.org/stable/2958830

[15] Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., & Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. Applied Soft Computing, 77, 188–204. doi:10.1016/j.asoc.2019.01.015

[16] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[17] Schapire, R. E. (1990). The strength of weak learnability. In Machine Learning (Vol. 5, Issue 2, pp. 197–227). Springer Science and Business Media LLC. https://doi.org/10.1007/bf00116037

[18] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

[19] Breiman, L. (1998). Arcing Classifiers. The Annals of Statistics, 26(3), 801–824. http://www.jstor.org/stable/120055

[20] Xiao, Y., Wang, X., Li, Q., Fan, R., Chen, R., Shao, Y., Chen, Y., Gao, Y., Liu, A., Chen, L., & Liu, S. (2021). A cascade and heterogeneous neural network for CT pulmonary nodule detection and its evaluation on both phantom and patient data. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society, 90, 101889. https://doi.org/10.1016/j.compmedimag.2021.101889

[21] Yang, Z., Ran, L., Zhang, S., Xia, Y., & Zhang, Y. (2019). EMS-Net: Ensemble of Multiscale Convolutional Neural Networks for Classification of Breast Cancer Histology Images. Neurocomputing, 366, 46–53. doi:10.1016/j.neucom.2019.07.080

[22] Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021, April). CheXtransfer. Proceedings of the Conference on Health, Inference, and Learning. doi:10.1145/3450439.3451867

[23] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden and S. J. Johnston, ”Random Subspace Ensembles for fMRI Classification,” in IEEE Transactions on Medical Imaging, vol. 29, no. 2, pp. 531-542, Feb. 2010, doi: 10.1109/TMI.2009.2037756

[24] A. Bertoni, R. Folgieri, and G. Valentini, “Feature selection combined with random subspace ensemble for gene expression based diagnosis of malignancies,” in Biological and Artificial Intelligence Environments, B. Apolloni, M. Marinaro, and R. Tagliaferri, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 29–36.

[25] D. D. Margineantu and T. G. Dietterich, “Pruning adaptive boosting,” Proc. 14th Int. Conf. Mach. Learning Morgan Kaufmann. San Francisco, CA, 1997, pp. 378–387.

[26] Takemura, A., Shimizu, A., & Hamamoto, K. (2010). Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the AdaBoost algorithm with feature selection. IEEE transactions on medical imaging, 29(3), 598–609. https://doi.org/10.1109/TMI.2009.2022630

[27] T. Huynh et al., ”Estimating CT Image From MRI Data Using Structured Random Forest and Auto-Context Model,” in IEEE Transactions on Medical Imaging, vol. 35, no. 1, pp. 174-183, Jan. 2016, doi: 10.1109/TMI.2015.2461533.

[28] S. Roychowdhury, D. D. Koozekanani and K. K. Parhi, ”DREAM: Diabetic Retinopathy Analysis Using Machine Learning,” in IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 5, pp. 1717-1728, Sept. 2014, doi:10.1109/JBHI.2013.2294635.

[29] Luis Talavera, 1999, L. Talavera, Feature selection as retrospective pruning in hierarchical clustering”, Adv. Intell. Data Anal., vol. 1642, pp. 75-86, 1999.

[30] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C. (2018). A Survey on Deep Transfer Learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018. Lecture Notes in Computer Science(), vol 11141. Springer, Cham.

[31] G. Wang et al., ”A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images,” in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2653-2663, Aug. 2020, doi: 10.1109/TMI.2020.3000314.

[32] Ostyakov, P., Logacheva, E., Suvorov, R., Aliev, V., Sterkin, G., Khomenko, O., & Nikolenko, S. I. (2018, September). Label Denoising with Large Ensembles of Heterogeneous Neural Networks. Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

[33] De Fauw, J., Ledsam, J.R., Romera-Paredes, B. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24, 1342–1350 (2018). https://doi.org/10.1038/s41591-018-0107-6

[34] Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. in Navab N., Hornegger J., Wells W., FrangiA. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol. 9351 (Springer, Cham, Switzerland, 2015).

[35] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science, vol. 9901 (Springer, Cham, Switzerland; 2016)

[36] Huang, Gao, Liu, Zhuang, Weinberger, Kilian Q, and van der Maaten, Laurens. Densely connected convo- lutional networks. arXiv preprint arXiv:1608.06993, 2016.

[37] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hier- archical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Con- ference on, pp. 248–255. IEEE, 2009.

[38] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. CoRR, abs/1711.05225. Retrieved from http://arxiv.org/abs/1711.05225

[39] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., … Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. ArXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1901.07031

[40] Rajpurkar, P., Joshi, A., Pareek, A., Chen, P., Kiani, A., Irvin, J., … Lungren, M. P. (2020). CheXpedition: Investigating Generalization Challenges for Translation of Chest X-Ray Algorithms to the Clinical Setting. ArXiv [Eess.IV]. Retrieved from http://arxiv.org/abs/2002.11379