Leveraging Explainable Artificial Intelligence for Understanding the Effect of Model Capacity on Training Dataset Size

Tushar Prakash*
Electronics and Communication Department
Delhi Technological University
Delhi, India
tusharprakash_2k19ec197@dtu.ac.in

Tashvik Dhamija*
Electronics and Communication Department
Delhi Technological University
Delhi, India
tashvikdhamija_2k19ec194@dtu.ac.in

Rohit Kumar
Electronics and Communication Department
Delhi Technological University
Delhi, India
rohit.kumar@dtu.ac.in

Jeebananda Panda
Electronics and Communication Department
Delhi Technological University
Delhi, India
jpanda@dce.ac.in

***Abstract*—Automated Image classification has seen substantial growth in recent times. However, several applications suffer from the limited availability of training data, such as the classification of medical images, where data collection is mostly limited by privacy concerns of human subjects. As a result, to compensate for the limited availability of training data, most of these applications employ custom-made lightweight architectures. While state-of-the-art deep models for computer vision applications usually exploit architectures with huge model capacity. However, the increase in model complexity and size necessarily doesn't guarantee better performance on small medical datasets. We study this phenomenon in the context of medical images, where several existing studies report that most sophisticated deep networks for computer vision trained on large datasets such as Image Net do not generalize on medical image applications, due to huge model capacity, subsequently leading to overfitting on smaller training datasets. In this research, we exploit *explainable artificial intelligence* to analyze the features learned by state-of-the-art deep models for smaller medical image training datasets and contrast them with the features learned for larger medical training datasets. In particular, we exploit Shapley Additive explanations (SHAP) features to perform a qualitative comparison of feature relevance maps and understand how different standard models when trained with different training sizes understand discriminative image patterns to perform classification. Furthermore, we also compare SHAP features on scenarios in which the same model focuses on images belonging to different classes. Experiments on two datasets of different sizes have been presented to understand the dependence of model complexity on the number of samples in the training dataset. Results demonstrate that simpler models learn generalizable SHAP features that allow them to perform better on small datasets, unlike larger models when trained on smaller datasets. Likewise, bigger models when trained on larger datasets learn more distinctive and diverse features that allow them to outperform smaller models.**

* indicates equal contributions from the authors.

979-8-3503-3224-7/22/$31.00 ©2022 IEEE

***Index Terms*—Explainable Artificial Intelligence (XAI), SHAP, medical image classification.**

I. Introduction

Image classification caters to a broad category of problem statements. From robotics to security, image classification plays an important role in a plethora of applications. With the shift to learnable methods, it is one of the prominent areas that has seen a major performance boost with the advancements of deep models in computer vision. Across the last few years, models have progressed from simple machine learning algorithms to convolutional neural networks [1]–[3] to transformer architectures [4]. Multiple variants of these architectures have been proposed and compared on standard in-the-wild datasets like ImageNet [5] which are generally large in size. Based on these comparisons, these models are deemed to be state-of-the-art across multiple subdomains of computer vision and its applications.

Focusing on domain-specific problems, a lot of research targets very specific datasets and environments. There is a massive amount of research that is needed to adapt these proposed models from one subdomain to another, even in the same application field [6]–[8]. This creates a gap between the standard and specific models and there is a lack of understanding of what general model can be used on a new problem statement without domain-specific research.

In order to study how standard networks perform on different sizes of datasets from an application domain, we take medical image classification as a case study. Medical Image classification [9] is a branch of computer vision that caters to equipping models to perform classification tasks on medical datasets which are mostly different from natural image datasets. There exist various different types of classification

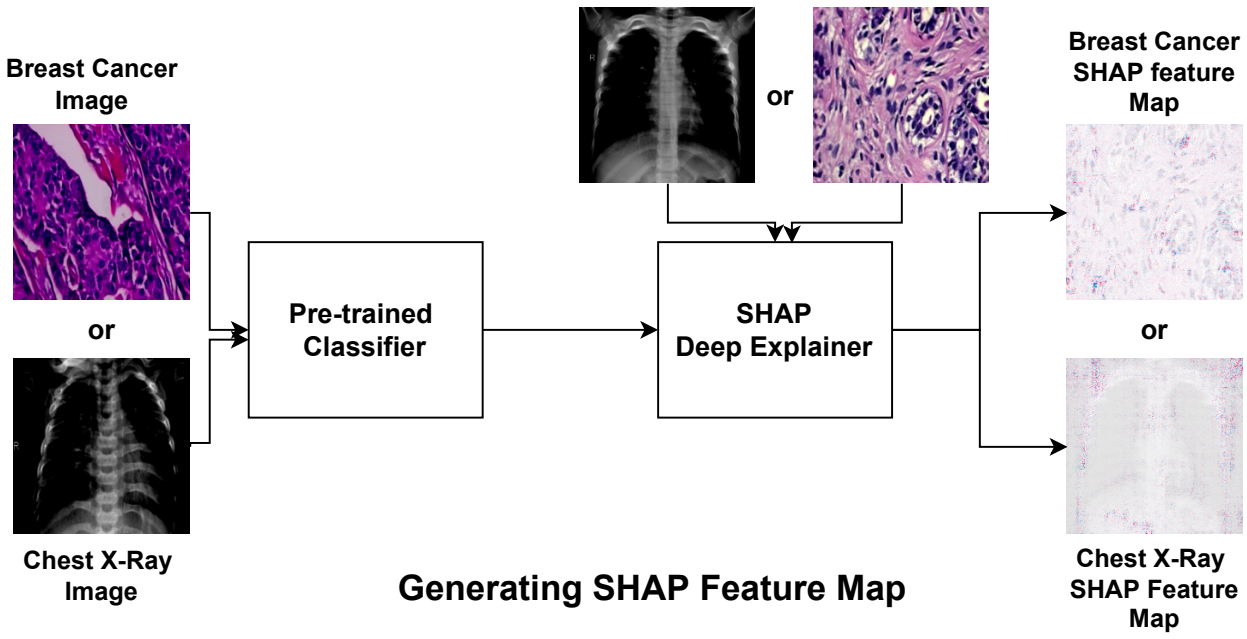

Fig. 1. Flowchart presenting the proposed method. SHAP features are obtained from a pre-trained deep classification model. SHAP deep explainer model is trained on 10% test images in an unsupervised manner. More precise features are obtained for a shallower model when the dataset is small and more precise features are obtained for a deeper model when the dataset is larger.

problems wherein, the data exhibits low intra and inter class variance. This makes adapting domain-specific models even harder and a need for analysis on what standard model works better in what kind of scenarios is much needed.

Research Contribution: On a mission to understand the effect of model capacity on training dataset size, in this paper, we choose three standard and common architectures and perform simple classification on Cancer [10] and Pneumonia[1] datasets. We compare the performance of our models and analyze their SHAP [11] feature maps to understand the focus on the model and how it learns from the images. Then, we compare the SHAP values and determine how these models act differently on datasets of different sizes. Through this experimentation, we develop an understanding of how to choose the best network out of the various standard models for a certain problem statement.

II. Literature Review

A. Medical Image Classification

Image classification using machine learning algorithms like SVM [12], K-NN [13] and more was a forward step that helped improve the performance on many datasets. Due to the simple nature of these algorithms, there was a need to use more complex models like neural networks [14]. With the proposal of Convolutional Neural Networks like AlexNet [15] and VGG [16], a massive improvement in classification was seen as these models were designed to do well with natural images. Further improvements were seen in form of the introduction of residual connections in ResNet [17] and ResNext [18]. Further, using attention in vision models allowed networks to attend to important regions of an image [19], [20].

Medical image classification umbrellas a plethora of problems like COVID detection [21], Cancer classification [22] and more. Various models based on CNNs have also been proposed to solve such problems like [23], [24]. Even models like [25] have been experimented with to improve performance. A common problem with all these networks is that they are proposed for specific datasets under controlled conditions [26] and require revisions when being adapted to new medical problems or datasets.

B. Explainable Artificial Intelligence (XAI)

Deep models are generally black box and therefore, unsafe for real applications. XAI aims at introducing interpretability and transparency in deep models. Various XAI technqiues are proposed in the literature. Among those, Grad-CAM [27]–[30] is a commonly used too to visualize salient image regions by backpropagation of the gradients. Quantification of uncertainty is a important XAI technique to understand model's confidence in prediction and noisy input regions [31]–[33]. LIME [34] suggests that a model can also be explained by perturbing an input and see how the output changes. SHAP [11] is a XAI tool to understand the marginal contribution of a feature in a deep model's performance.

Motivated by the XAI methods and to study how size training dataset affects features learnt by a deep model, we use SHAP feature maps to compare how the model relatively

[1]https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

understands important information from different parts of an input image. This helps us perform a fruitful analysis and support our experimental results.

III. Proposed Method

For this study, we begin with three state-of-the-art deep architectures that are pre-trained on ImageNet dataset [5]: ResNet [17] that uses residual connections; ResNeXt [18] uses cardinality as a new dimension in addition to the dimensions of depth and width and Swin Transformer [4] uses the attention locally then at a global scale, it makes Swin Transformer more computationally fast. Swin Transformer shows the highest accuracy on the ImageNet dataset, followed by ResNeXt and then ResNet.

We trained all three architectures on the two medical datasets (described in Section IV) in a supervised manner. After training these models for the classification task, we compute SHAP values to look for the features identified by the models during classification. However, as presented in Figure 1, to obtain the explainable feature map from SHAP, we provide 10% unannotated test images to train the deep explainer model of SHAP. With the help of SHAP, we can calculate the marginal contribution of each pixel in the given test image, also known as shaply values. By doing so, we get two feature maps belonging to each class and the same size as the input image. The SHAP feature map consists of blue and red pixels distributed throughout the image. Here, red coloured pixeled region signifies the essential or main features that the trained classifier use for the correct classification. At the same time, the blue pixeled region defines the features that push the classifier to make the incorrect classification. These features are the supporting results in our hypothesis that the complex model fails for small medical datasets.

IV. Datasets

To demonstrate our findings, we have used datasets from two medical imaging modalities (i) The Breast Cancer Histopathological Image Classification (BreakHis) [10] and (ii) Chest X-Ray images of pneumonia patients. For both datasets, we have two different classes, malignant and benign for breast cancer and normal and pneumonia for chest X-Ray.

(I) *The Breast Cancer dataset (BreakHis)*: It contains 9,109 microscopic images of 40X, 100X, 200X, and 400X magnifying scales. Breast tumour tissue is collected from 82 different breast cancer patients. To make our study more specific, we only use a 200x magnifying factor for training and testing. Hence the final number of samples used for training and testing are 1,923 from which 608 are benign, and 1,315 are malignant samples.

(II) *The Chest X-Ray images of pneumonia patients*: It consists of 5,863 chest X-Ray images with two different categories, i.e., Pneumonia and Normal.Where the number of Pneumonia images are 4,273 while number of Normal images are 1,583.

V. Results

A. Results on Larger Dataset

In this section, we will be comparing the classification performance of different state-of-the-art deep architectures, i.e., ResNet, ResNeXt and Swin Transformer, on a larger dataset called Chest X-Ray Pneumonia dataset with 5,863 total samples.

From Figure 2, we can see the classification accuracy increase with respect to the model capacity. For ResNet, the classification accuracy is around 86.7%, increasing to 88.5% when we use ResNeXt.With the use of Swin Transformer, classification accuracy is highest at 91%. These results demonstrate that the classification accuracy increases with the parallel increase of model capacity for the larger dataset.

To further support our hypothesis, we use SHAP to generate the feature maps for all the architectures. In Fig. 3(a), We can see the SHAP feature map of a given pneumonia sample generated using Swin Transformer, it is pretty accurate, and the red pixel intensity is also relatively high at a scale of 0.15. This signifies that Swin Transformer can find the best relevant features during the classification. In contrast, the red pixel intensity for ResNeXt(shown in Fig. 3(b)) is lower than Swin Transformer but higher than ResNet, i.e. 0.04, and the features are not as accurate as Swin Transformers. In the case of ResNet((shown in Fig. 3(c))), the red pixel intensity is lowest, which means low confidence, and the feature map is quite distorted and highlights the irrelevant features.

B. Results on Smaller Dataset

In contrast to the above conclusion, the classification accuracy decreases concerning model capacity with a smaller dataset, i.e., The Breast Cancer Histopathological Image, with only 1,923 samples for training and testing.

From Fig. 2, we can see the classification accuracy with respect to model capacity. For ResNet, the classification accuracy is around 95.4%, decreasing to 94.3% when we use ResNeXt.With the use of Swin Transformer, classification accuracy is lowest at 92%. These results demonstrate that the classification accuracy decreases with the parallel increase of model capacity for the larger datasets.

To further support our hypothesis, we use SHAP to generate the feature maps for all the architectures. In Fig. 4(a), We can see the SHAP feature map of a given malignant sample for ResNet50 is quite crisp, and features have more confidence in their prediction, specified by the dark red pixels with a maximum value of 0.2. whereas the red pixel intensity for ResNeXt(shown in Fig. 4(b)) is lower, i.e., 0.10, and the features are not uniform, the model cannot find the optimal features during the classification. In Swin Transformer's case((shown in Fig. 4(c))), despite high red pixel intensity, the feature map is distorted and highlights irrelevant features or backgrounds.

C. Class-wise Features

To gain further insight into our proposed hypothesis, We generated the SHAP feature maps for classes belonging to

Normal images in the Chest X-Ray dataset and Benign in the case of the Breast cancer dataset.

In Fig. 5, we can see the more complex architectures can also generate a better SHAP feature map for the Normal X-Ray sample. The uniformity and precision of the features generated by Swin Transformer (shown in Fig. 5(a)) are a little bit better than ResNeXt's (shown in Fig. 5(b)) feature map and a lot more accurate than ResNet(shown in Fig. 5(c)) one.

Similarly, In Fig. 6, we can see that the more complex architectures can not generate a better SHAP feature map for Benign cancer Samples. the uniformity and precision of ResNet(shown in Fig. 6(a)) is a little bit better than ResNeXt(shown in Fig. 6(b)) feature map and a lot accurate than Swin Transformer(shown in Fig. 6(c)).

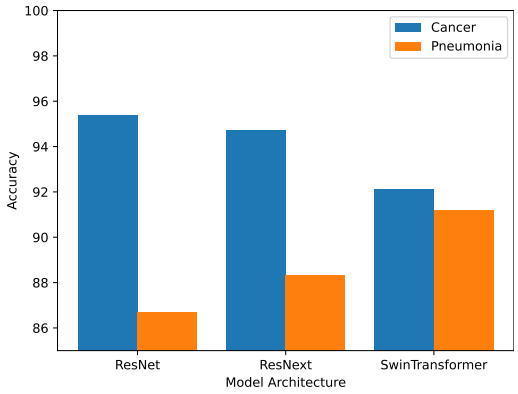

Fig. 2. Performance of models on Cancer and Pneumonia datasets

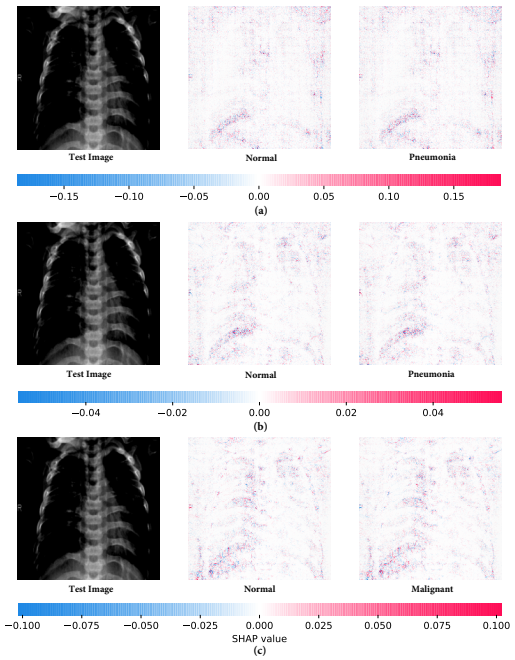

Fig. 3. SHAP feature map for pneumonia x-ray samples for different models (a)Swin Transformer, (b)ResNeXt and (c)ResNet.

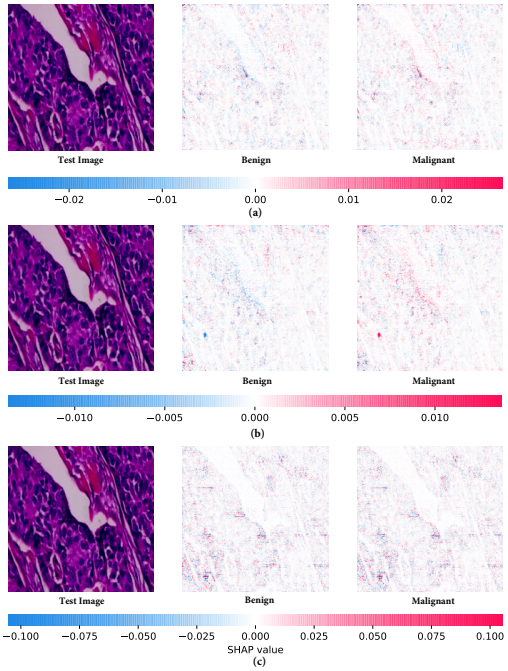

Fig. 4. SHAP feature map for malignant breast cancer samples for different models (a)ResNet, (b)ResNeXt, and (c)Swin Transformer.

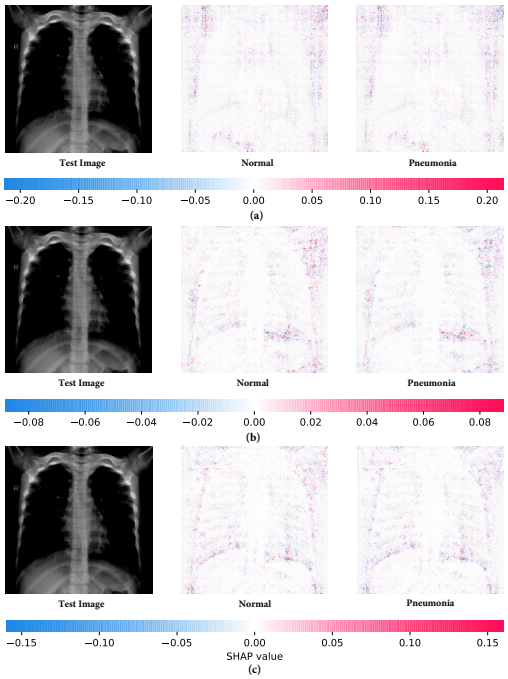

Fig. 5. SHAP feature map for normal x-ray samples for different models (a)Swin Transformer, (b)ResNeXt and (c)ResNet.

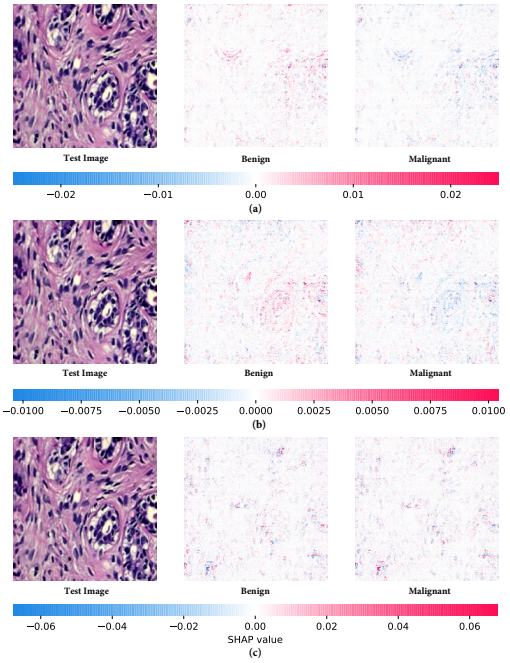

Fig. 6. SHAP feature map for benign breast cancer samples for different models (a)ResNet, (b)ResNeXt and (c)Swin Transformer.

VI. Conclusion

From the study performed, we can conclude that the standard models cannot be generally stated to be state-of-the-art and the most appropriate model that should be used is dependent on the nature of a problem statement. Whilst, complex models successfully beat their predecessors on datasets with enough data samples, simpler models perform relatively better on datasets of smaller size. This is by the virtue of their inability to focus on regions of important information in a precise manner as seen in the SHAP feature maps obtained from these models. Due to the increased complexity and extra parameters, the advanced models suffer on smaller datasets. This proves that when choosing a model for a new problem, there is a need to identify which model shall perform the best irrespective of its ability on large in-the-wild datasets. Further, this reduces the need to perform domain-specific research until absolutely necessary.

References

[1] I. Joshi, A. Anand, M. Vatsa, R. Singh, S. Dutta Roy, and P. Kalra, "Latent Fingerprint Enhancement using Generative Adversarial Networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 895 – 903.

[2] I. Joshi, T. Dhamija, R. Kumar, A. Dantcheva, S. D. Roy, and P. K. Kalra, "Cross-Domain Consistent Fingerprint Denoising," *IEEE Sensors Letters*, 2022.

[3] I. Joshi, T. Prakash, R. Kumar, A. Dantcheva, S. D. Roy, and P. K. Kalra, "Context-Aware Restoration of Noisy Fingerprints," *IEEE Sensors Letters*, 2022.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[6] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, and A. Dantcheva, "Synthetic data in human analysis: A survey," *arXiv preprint arXiv:2208.09191*, 2022.

[7] I. Joshi, "Advanced Deep Learning Techniques for Fingerprint Preprocessing," Ph.D. dissertation, IIT Delhi, 2021.

[8] I. Joshi, A. Anand, S. Dutta Roy, and P. K. Kalra, "On Training Generative Adversarial Network for Enhancement of Latent Fingerprints," in *AI and Deep Learning in Biometric Security*, 2021, pp. 51 – 79.

[9] I. Joshi, S. Kumar, and I. N. Figueiredo, "Bag of visual words approach for bleeding detection in wireless capsule endoscopy images," in *International Conference on Image Analysis and Recognition*, 2016, pp. 575–582.

[10] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[11] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

[12] X. Sun, L. Liu, H. Wang, W. Song, and J. Lu, "Image classification via support vector machine," in *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, vol. 01, 2015, pp. 485–489.

[13] J. Kim, B.-S. Kim, and S. Savarese, "Comparing image classification methods: K-nearest-neighbor and support-vector-machines," in *Proceedings of the 6th WSEAS International Conference on Computer Engineering and Applications, and Proceedings of the 2012 American Conference on Applied Mathematics*, ser. AMERICAN-MATH'12/CEA'12. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2012, p. 133–138.

[14] W. H. Ibrahim, A. A. A. Osman, and Y. I. Mohamed, "Mri brain image classification using neural networks," in *2013 international conference on computing, electrical and electronic engineering (ICCEEE)*. IEEE, 2013, pp. 253–258.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[21] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. Prathosh, "xvitcos: explainable vision transformer based covid-19 screening using radiography," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–10, 2021.

[22] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, p. 105051, 2022.

[23] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, 2014, pp. 844–848.

[24] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classi-

fication," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3478–3488.

[25] B. Gheflati and H. Rivaz, "Vision transformers for classification of breast ultrasound images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2022, pp. 480–483.

[26] M. Margaryan, M. Seibold, I. Joshi, M. Farshad, P. Fürnstahl, and N. Navab, "Improved techniques for the conditional generative augmentation of clinical audio data," *arXiv preprint arXiv:2211.02874*, 2022.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[28] R. Roy, I. Joshi, A. Das, and A. Dantcheva, "3d cnn architectures and attention mechanisms for deepfake detection," in *Handbook of Digital Face Manipulation and Detection*, 2022, pp. 213–234.

[29] I. Joshi, A. Utkarsh, P. Singh, A. Dantcheva, S. Dutta Roy, and P. K. Kalra, "On Restoration of Degraded Fingerprints," *Multimedia Tools and Applications*, pp. 1–29, 2022.

[30] I. Joshi, A. Utkarsh, R. Kothari, V. K. Kurmi, A. Dantcheva, S. Dutta Roy, and P. K. Kalra, "Sensor-Invariant Fingerprint ROI Segmentation using Recurrent Adversarial Learning," in *International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1 – 8.

[31] I. Joshi, A. Utkarsh, R. Kothari, V. K. Kurmi, , A. Dantcheva, S. Dutta Roy, and P. K. Kalra, "On Estimating Uncertainty of Fingerprint Enhancement Models," in *Digital Image Enhancement and Reconstruction*, 2022.

[32] I. Joshi, R. Kothari, A. Utkarsh, V. K. Kurmi, A. Dantcheva, S. Dutta Roy, and P. K. Kalra, "Explainable Fingerprint ROI Segmentation using Monte Carlo Dropout," in *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2021, pp. 60 – 69.

[33] I. Joshi, A. Utkarsh, R. Kothari, V. K. Kurmi, A. Dantcheva, S. Dutta Roy, and P. K. Kalra, "Data Uncertainty Guided Noise-Aware Preprocessing of Fingerprints," in *International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1 – 8.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938