Encoding Visual Attributes in Capsules for Explainable Medical Diagnoses

Rodney LaLonde[1], Drew Torigian[2], and Ulas Bagci[1]

[1] Center for Research in Computer Vision, University of Central Florida
[2] Penn Medicine, University of Pennsylvania

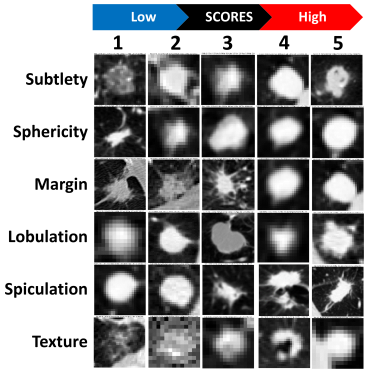
Abstract. Convolutional neural network based systems have largely failed to be adopted in many high-risk application areas, including healthcare, military, security, transportation, finance, and legal, due to their highly uninterpretable "black-box" nature. Towards solving this deficiency, we teach a novel multi-task capsule network to improve the explainability of predictions by embodying the same high-level language used by human-experts. Our explainable capsule network, ***X-Caps***, encodes high-level visual object attributes within the vectors of its capsules, then forms predictions based solely on these human-interpretable features. To encode attributes, *X-Caps* utilizes a new routing sigmoid function to independently route information from child capsules to parents. Further, to provide radiologists with an estimate of model confidence, we train our network on a distribution of expert labels, modeling inter-observer agreement and punishing over/under confidence during training, supervised by human-experts' agreement. *X-Caps* simultaneously learns attribute and malignancy scores from a multi-center dataset of over 1000 CT scans of lung cancer screening patients. We demonstrate a simple 2D capsule network can outperform a state-of-the-art deep dense dual-path 3D CNN at capturing visually-interpretable high-level attributes and malignancy prediction, while providing malignancy prediction scores approaching that of non-explainable 3D CNNs. To the best of our knowledge, this is the first study to investigate capsule networks for making predictions based on radiologist-level interpretable attributes and its applications to medical image diagnosis. Code is publicly available at https://github.com/lalonderodney/X-Caps.

Keywords: Explainable AI · Lung Cancer · Capsule Networks.

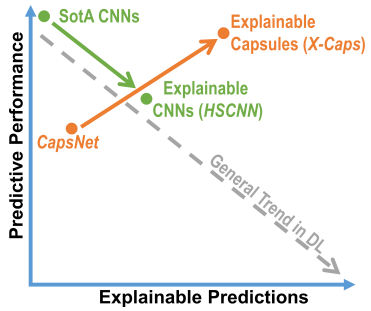
1 Introduction

In machine learning, predictive performance typically comes at the cost of *interpretability* [4,7,14,24]. While deep learning (DL) has impact many fields, there exist several high-risk domains which have yet to be comparably affected: military, security, transportation, finance, legal, and healthcare among others, often citing a lack of interpretability as the main concern [3,18,23]. As features become less *interpretable*, and the functions learned more complex, model predictions become more difficult to explain (Fig. 1b). While some works have begun to press towards this goal of explainable DL, the problem remains largely unsolved.

Interpretable vs. Explainable: There has been a recent push in the community to move away from the *post-hoc interpretations* of deep models and instead create explainable models from the outset [24,27]. Since the terms *interpretable* and *explainable* are often used interchangeably, we want to be explicit about our definitions for the purposes of this study. An *explainable* model is one which provides explanations for its predictions *at the human level* for a *specific task*. An *interpretable* model is one for which some conclusions can be drawn about the internals/predictions of the model; however, they are not explicitly provided by the model and are typically at a lower level. For example, in image classification, when a deep model predicts an image to be of a cat, saliency/gradient or other methods can attempt to *interpret* the model/prediction. However, the model is not *explaining* why the object in the image is a cat in the same way as a human. Humans classify objects based on a taxonomy of characteristics/attributes (*e.g.* cat equals four legs, paws, whiskers, fur, etc.). If our goal is to create *explainable* models, we should design models which explain their decisions using a similar set of "attributes" to humans.



(a) Lung nodules with high-level visual attribute scores as determined by expert radiologists. Scores were given from 1 – 5 for six different visual attributes related to diagnosing lung cancer.



(b) A symbolic plot showing the general trade-off between explainability and predictive performance in deep learning (DL) [4,7,14,24]. Our proposed *X-Caps* rebuts the trend of decreasing performance from state-of-the-art (SotA) as explainability increases and shows it is possible to create more explainable models *and* increase predictive performance with capsule networks.

Fig. 1: Encoding visual attributes (a) for explainable predictions (b).

Why capsule networks? Capsule networks differ from convolutional neural networks (CNNs) by replacing the scalar feature maps with vectorized representations, responsible for encoding information (*e.g.* pose, scale, color) about each feature. These vectors are then used in a dynamic routing algorithm which seeks to maximize the agreement between lower-level predictions for the instantiation parameters (i.e. capsule vectors) of higher-level features. In their introductory work, a capsule network (*CapsNet*) was shown to produce promising results

on the MNIST data set; but more importantly, was able to encode high-level visually-interpretable features of digits (*e.g.* stroke thickness, skew, localized-parts) within the dimensions of its capsule vectors [25].

Lung cancer diagnosis with a multi-task capsule network: In diagnosing the malignancy of lung nodules, similar to describing why an image of a cat is catlike, radiologists explain their predictions through the language of high-level visual attributes (i.e., radiographical interpretations): subtlety (sub), sphericity (sph), margin (mar), lobulation (lob), spiculation (spi), and texture (tex), shown in Fig. 1a, which are known to be predictive (with inherent uncertainty) of malignancy [8]. To create a DL model with this same level of radiographical interpretation, we propose a novel multi-task capsule architecture, called ***X-Caps***, for learning visually-interpretable feature representations within capsule vectors, then predicting malignancy based solely on these interpretable features. By supervising different capsules to embed specific visually-interpretable features, multiple visual attributes are learned simultaneously, with their weights being updated by both the radiologists visual interpretation scores as well as their contribution to the final malignancy score, regularized by the segmentation reconstruction error. Since these attributes are not mutually-exclusive, we introduce a new routing sigmoid function to independently route child capsules to parents. Further, to provide radiologists with an estimate of model confidence, we train our network on a distribution of expert labels, modeling inter-observer agreement and punishing over/under confidence during training, supervised by human-experts' agreement.

We show even a relatively simple 2D capsule network can better capture high-level visual attribute information than the state-of-the-art deep dual-path dense 3D convolutional neural network (CNN) while also improving diagnostic accuracy, approaching that of even some black-box methods (*e.g.,* [28,29]). *X-Caps* simultaneously learns attribute and malignancy scores from a multi-center dataset of over 1000 CT scans of lung cancer screening patients. **Overall, the contributions of this study are summarized as:**

1. The first study to directly encode high-level visual attributes within the vectors of a capsule network to perform explainable image-based diagnosis *at the radiologist-level.*
2. Create a novel modification to the dynamic routing algorithm to independently route information from child capsules to parents when parent capsules are not mutually-exclusive.
3. Provide a meaningful confidence metric with our predictions at test by learning directly from expert label distributions to punish network over/under confidence. Visual attribute predictions are verified at test via the reconstruction branch of the network.
4. Demonstrate a simple 2D capsule network (*X-Caps*) trained from scratch outperforming a state-of-the-art deep pre-trained dense dual-path 3D CNN at capturing visually-interpretable high-level attributes and malignancy prediction, while providing malignancy prediction scores approaching that of non-explainable 3D CNNs.

2 Related work

The majority of work in explainable deep learning has focused around *post hoc* deconstruction of already trained models (*i.e.* interpretation). These approaches typically rely on human-experts to examine their results and attempt to discover meaningful patterns. Zeiler and Fergus [31] attached a deconvolutional network to network layers to map activations back to pixel space for visualizing individual filters and activation maps, while also running an occlusion-based study of which parts of the input contribute most to the final predictions. *Grad-CAM* [26] is a popular approach which highlights the relative positive activation map of convolutional layers with respect to network outputs. *InfoGAN* [5] separates noise from the "latent code", maximizing the mutual information between the latent representations and the image inputs, encoding concepts such as rotation, width, and digit type for MNIST. In a similar way, capsule networks encode visually-interpretable concepts such as stroke thickness, skew, rotation, and others [25].

A number of recent studies have proposed using *CapsNet* for a variety of medical imaging classification tasks [1,11,12,13,19,22,30]. However, these methods nearly all follow the exact *CapsNet* architecture, or propose minor modifications which present nearly identical predictive performance [16,20]; hence, it is sufficient to compare only with *CapsNet* in reference to these works.

In the area of lung nodule malignancy, many DL-based approaches have been proposed [28,29], with further methods being developed with complicated post-processing techniques [9], curriculum learning methods [21], or gradient-boosting machines [32]. However, adding such techniques is beyond the scope of this study and would lead to an unwieldy enumeration of ablation studies necessary to understand the contributions between our proposed capsule architecture and such techniques. For a fair comparison in this study, we compare our method directly against *CapsNet* and explainable CNN approaches. *HSCNN* [27] creates one of the first explainable methods, by designing a dense 3D CNN which first predicts visual attribute scores then predicts malignancy from those features. This decreased the overall performance as compared to other 3D networks [29] but provided some explanations for the final malignancy predictions.

3 Capsules for encoding visual attributes

Our approach, referred to as *explainable capsules*, or *X-Caps*, was designed to remain as similar as possible to our control network, *CapsNet*, while allowing us to have more control over the visually-interpretable features learned. *CapsNet* already showed great promise when trained on the MNIST data set for its ability to model high-level visually-interpretable features. With this study, we examine the ability of capsules to model *specific* visual attributes within their vectors, rather that simply hoping these are learned successfully in the more challenging lung nodule data. As shown in Figure 2, *X-Caps* shares a similar overall structure as *CapsNet*, with the major differences being the addition of the supervised labels for each of the *X-Caps* vectors, the fully-connected layer for malignancy

prediction, the reconstruction regularization also performing segmentation, and the modifications to the dynamic routing algorithm.

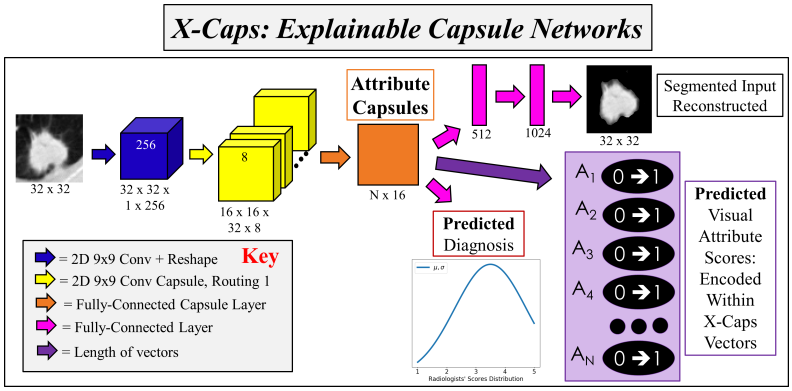


Fig. 2: *X-Caps*: Explainable Capsule Networks. The proposed network (1) predicts N high-level visual attributes of the nodule, (2) segments the nodule and reconstruct the input image, and (3) diagnoses the nodule on a scale of 1 to 5 based on the visually-interpretable high-level features encoded in the X-Caps capsule vectors. The malignancy diagnosis branch is attempting to model the distribution of radiologists' scores in both mean and variance.

The first layer of our proposed network is a 2D convolutional layer which extracts low-level features. Next, we form our primary capsules of 32 capsule types with $8D$ vector capsules. Following this, we form our attribute capsules using a fully-connected capsule layer whose output is N $16D$ capsule types, one for each of the visual-attributes we want to predict. Unlike *CapsNet* where each of the parent capsules were dependant on one another (e.g. if the prediction is the digit 5 it cannot also be a 3), our parent capsules are not mutually-exclusive of each other (i.e. a nodule can score high or low in each of the attribute categories). For this reason, we needed to modify the dynamic routing algorithm presented in *CapsNet* to accommodate this significant difference. The key change is the "routing softmax" employed by *CapsNet* forces the contributions of each child to send their information to parents in a manner which sums to one, which in practice effectively makes them "choose" a parent to send their information to. However, when computing prediction vectors for independent parents, we want a child to be able to contribute to all parent capsules for attributes which are present in the given input. With that motivation, the specific algorithm, which we call "routing sigmoid", is computed as

$$r_{i,j} = \frac{\exp(b_{i,j})}{\exp(b_{i,j}) + 1}, \tag{1}$$

where $r_{i,j}$ are the routing coefficients determined by the dynamic routing algorithm for child capsule i to parent capsule j and the initial logits, $b_{i,j}$ are the prior probabilities that the prediction vector for capsule i should be routed to parent capsule j. Note the prior probabilities are initially set to 1 rather than 0 as in *CapsNet*, otherwise no routing could take place. The rest of the dynamic routing procedure follows the same as in [25].

Predicting malignancy from visually-interpretable capsule vectors: In order to predict malignancy scores, we attach a fully-connected layer to our attribute capsules with output size equal to the range of scores. We wish to emphasize here, our final malignancy prediction is coming solely from the vectors whose magnitudes represent *visually-interpretable* feature scores. Every malignancy prediction score has a set of weights connected to the high-level attribute capsule vectors, and the activation from each tells us the exact contribution of the given visual attribute to the final malignancy prediction for that nodule. Unlike previous studies which look at the importance of these attributes on a global level, our method looks at the importance of each visual attribute in relation to a specific nodule being diagnosed. To verify the correctness of our attribute modeling, we reconstruct the nodules while varying the dimensions of the capsule vectors to ensure the desired visual attributes are being modeled. At test, these reconstructions give confidence that the network is properly capturing the attributes, and thus the scores can be trusted. Confidence in the malignancy prediction score, in addition to coming solely from these trusted attributes, is provided via an uncertainty modeling approach.

Previous works in lung nodule classification follow the same strategy of averaging radiologists' scores for visual attributes and malignancy, and then either attempt to regress this average or performing binary classification of the average as below or above 3. To better model the uncertainty inherently present in the labels due to inter-observer variation, we propose a different approach: we attempt to predict the *distribution* of radiologists' scores. Specifically, for a given nodule where we have at minimum three radiologists' score values for each attribute and for malignancy prediction, we compute the mean and variance of those values and fit a Gaussian function to them, which is in turn used as the ground-truth for our classification vector. Nodules with strong inter-observer agreement produce a sharp peak, in which case wrong or unsure (*i.e.,* low confidence score) predictions are severely punished. Likewise, for low inter-observer agreement nodules, we expect our network to output a more spread distribution and it will be punished for strongly predicting a single class label. This proposed approach allows us to model the uncertainty present in radiologists' labels in a way that no previous study has and provide a meaningful confidence metric at test time to radiologists.

Loss and regularization: As in *CapsNet*, we also perform reconstruction of the input as a form of regularization. However, we extend the idea of regularization to perform a pseudo-segmentation, similar in nature to the reconstruction used by [15,17]. Whereas in true segmentation, the goal is to output a binary mask of pixels which belong to the nodule region, in our formulation we attempt

to reconstruct only the pixels which belong to the nodule region, while the rest are mapped to zero. More specifically, we formulate this problem as

$$R^{x,y} = I^{x,y} \times S^{x,y} \mid S^{x,y} \in \{0,1\}, \text{ and} \tag{2}$$

$$\mathcal{L}_R = \frac{\gamma}{X \times Y} \sum_{x}^{X} \sum_{y}^{Y} \|R^{x,y} - O_r^{x,y}\|, \tag{3}$$

where $\mathcal{L}_R$ is the supervised loss for the reconstruction regularization, γ is a weighting coefficient for the reconstruction loss, $R^{x,y}$ is the reconstruction target pixel, $S^{x,y}$ is the ground-truth segmentation mask value, and $O_r^{x,y}$ is the output of the reconstruction network, at pixel location (x, y), respectively, and X and Y are the width and height, respectively, of the input image. This adds another task to our multi-task approach and an additional supervisory signal which can help our network distinguish visual characteristics from background noise. The malignancy prediction score, as well as each of the visual attribute scores also provide a supervisory signal in the form of

$$\mathcal{L}_a = \sum_{n}^{N} \alpha^n \|A^n - O_a^n\| \text{ and } \mathcal{L}_m = \beta \|M - O_m\|, \tag{4}$$

where $\mathcal{L}_a$ is the combined loss for the visual attributes, A^n is the average of the attribute scores given by at minimum three radiologists for attribute n, N is the total number of attributes, α^n is the weighting coefficient placed on the n^{th} attribute, O_a^n is the network prediction for the score of the n^{th} attribute, $\mathcal{L}_m$ is the loss for the malignancy score, M is a Gaussian distribution over malignancy scores with mean and variance computed from scores given by at minimum three radiologists, O_m is the network prediction for the malignancy score distribution, and β is the weighting coefficient for the malignancy score. In this way, the overall loss for *X-Caps* is simply $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_a + \mathcal{L}_R$. For simplicity, the values of each α^n and β are set to 1, and γ is set to $0.005 \times 32 \times 32 = 0.512$.

4 Experiments, results, limitations, and ablations

Experiments we performed on the LIDC-IDRI data set [2]. Five-fold stratified cross-validation was performed, with 10% of each training set used for validation and early stopping. *X-Caps* was trained with a batch size of 16 using Adam with an initial learning rate of 0.02 reduced by a factor of 0.1 after validation loss plateau. Consistent with the literature, nodules of mean radiologists' score 3 were removed (leaving 646 benign and 503 malignant nodules) and predictions were considered correct if within ± 1 of the radiologists' classification [9,10]. The results summarized in Table 1 illustrate the prediction of visual attributes with the proposed *X-Caps* in comparison with an adapted version of *CapsNet*, a deep dense dual-path 3D explainable CNN (*HSCNN* [27]), and two state-of-the-art non-explainable methods which do not have extra post-processing or learning strategies. Compared methods results are from the original reported works.

Table 1: Prediction accuracy of visual attributes with capsule networks. Dashes (-) represent values which the given method could not produce. *X-Caps* outperforms the state-of-the-art explainable method (*HSCNN*) at attribute modeling (the main goal of both studies), while also producing higher malignancy prediction scores, approaching state-of-the-art non-explainable methods performance.

	Attribute Prediction Accuracy %						**Malig-**
	sub	sph	mar	lob	spi	tex	**nancy**
Non-Explainable Methods							
3D Multi-Scale + RF [28]	-	-	-	-	-	-	86.84
3D Multi-Crop [29]	-	-	-	-	-	-	87.14
CapsNet [25]	-	-	-	-	-	-	77.04
Explainable Methods							
3D Dual-Path *HSCNN* [27]	71.9	55.2	72.5	-	-	83.4	84.20
Proposed *X-Caps*	**90.39**	**85.44**	**84.14**	**70.69**	**75.23**	**93.10**	**86.39**

Our results show that a *X-Caps* has the ability to model visual attributes far better than *HSCNN* while also achieving better malignancy prediction. Further, we wish to emphasize the significance of *X-Caps* providing increased predictive performance *and* explainability over *CapsNet*. This goes against the assumed trend in DL, illustrated with a symbolic plot in Figure 1b, that explainability comes at the cost of predictive performance, a trend we observe with *HSCNN* being outperformed by less powerful (*i.e.* not dense or dual-path) but non-explainable 3D CNNs [28,29]. While *X-Caps* slightly under-performs the best non-explainable models, it is reasonable to suspect that future research into more powerful 3D capsule networks would allow explainable capsules to surpass these methods; we hope this study will promote such future works.

As two limitations of our work, we did not tune the weight balancing terms between the different tasks and further investigation could lead to superior performance. Also, we found capsule networks can be somewhat fragile; often random initializations failed to converge to good performance. However, this might be due to the small/shallow network size and its relation to the Lottery Ticket Hypothesis [6] rather than anything specific to capsules.

Ablation studies: To analyze the impact of each component of our proposed approach, we performed ablation studies for: (1) learning the distribution of radiologists' scores rather than attempting to regress the mean value of these scores, (2) removing the reconstruction regularization from the network, and (3) performing our proposed "routing sigmoid" over the original "routing softmax" proposed in [25]. The malignancy prediction accuracy for each of these ablations is (1) 83.09%, (2) 80.30%, and (3) 80.69%, respectively, as compared to the proposed model's accuracy of 86.39%. This shows retaining the agreement/disagreement information among radiologists proved useful, the reconstruction played a role in improving the network performance, and our proposed modifications to the dynamic routing algorithm were necessary for passing information from children to parents when the parent capsule types are independent.

5 Discussions and concluding remarks

Available studies for explaining DL models, typically focus on *post hoc* interpretations of trained networks, rather than attempting to build-in explainability. This is the first study for directly learning an interpretable feature space by encoding high-level visual attributes within the vectors of a capsule network to perform explainable image-based diagnosis. We approximate visually-interpretable attributes through individual capsule types, then predict malignancy scores directly based only on these high-level attribute capsule vectors, in order to provide malignancy predictions with explanations *at the human-level*, in the same language used by radiologists. Our proposed multi-task explainable capsule network, *X-Caps*, successfully approximated visual attribute scores better than the previous state-of-the-art explainable diagnosis system, while also achieving higher diagnostic accuracy. We hope our work can provide radiologists with malignancy predictions which are explained via the same high-level visual attributes they currently use, while also providing a meaningful confidence metric to advise when the results can be more trusted, thus allowing radiologists to quickly interpret and verify our predictions. Lastly, we believe our approach should be applicable to any image-based classification task where high-level attribute information is available to provide explanations about the final prediction.

References

1. Afshar, P., Mohammadi, A., Plataniotis, K.N.: Brain tumor type classification via capsule networks. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3129–3133. IEEE (2018)
2. Armato III, S., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical Physics **38**(2), 915–931 (2011)
3. Bloomberg, J.: Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box'. http://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/#6ceaf3793b4a (11162018), forbes Magazine
4. Bologna, G.: A model for single and multiple knowledge based networks. Artificial Intelligence in Medicine **28**(2), 141–163 (2003)
5. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems. pp. 2172–2180 (2016)
6. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635 (2018)
7. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)

8. Hancock, M., Magnan, J.: Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms. Journal of Medical Imaging **3**(4), 044504 (2016)
9. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3d cnn-based multi-task learning. In: International Conference on Information Processing in Medical Imaging. pp. 249–260. Springer (2017)
10. Hussein, S., Gillies, R., Cao, K., Song, Q., Bagci, U.: Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process. In: 14th International Symposium on Biomedical Imaging (ISBI). pp. 1007–1010. IEEE (2017)
11. Iesmantas, T., Alzbutas, R.: Convolutional capsule network for classification of breast cancer histology images. In: International Conference Image Analysis and Recognition. pp. 853–860. Springer (2018)
12. Jiménez-Sánchez, A., Albarqouni, S., Mateus, D.: Capsule networks against medical imaging data challenges. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 150–160. Springer (2018)
13. Kandel, P., LaLonde, R., Ciofoaia, V., Wallace, M.B., Bagci, U.: Su1741 colorectal polyp diagnosis with contemporary artificial intelligence. Gastrointestinal Endoscopy **89**(6), AB403 (2019)
14. Kuhn, M., Johnson, K.: Applied predictive modeling, vol. 26. Springer (2013)
15. LaLonde, R., Bagci, U.: Capsules for object segmentation. arXiv preprint arXiv:1804.04241 (2018)
16. LaLonde, R., Kandel, P., Spampinato, C., Wallace, M.B., Bagci, U.: Diagnosing colorectal polyps in the wild with capsule networks. In: 17th International Symposium on Biomedical Imaging (ISBI). IEEE (2020)
17. LaLonde, R., Xu, Z., Jain, S., Bagci, U.: Capsules for biomedical image segmentation. arXiv preprint arXiv:2004.04736 (2020)
18. Lehnis, M.: Can We Trust AI If We Don't Know How It Works? http://www.bbc.com/news/business-44466213 (15062018), bBC News
19. Mobiny, A., Lu, H., Nguyen, H.V., Roysam, B., Varadarajan, N.: Automated classification of apoptosis in phase contrast microscopy using capsule network. IEEE transactions on medical imaging **39**(1), 1–10 (2019)
20. Mobiny, A., Van Nguyen, H.: Fast capsnet for lung cancer screening. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 741–749. Springer (2018)
21. Nibali, A., He, Z., Wollersheim, D.: Pulmonary nodule classification with deep residual networks. International journal of computer assisted radiology and surgery **12**(10), 1799–1808 (2017)
22. Pal, A., Chaturvedi, A., Garain, U., Chandra, A., Chatterjee, R., Senapati, S.: Capsdemm: capsule network for detection of munros microabscess in skin biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 389–397. Springer (2018)
23. Polonski, V.: People Don't Trust AI–Here's How We Can Change That. http://www.scientificamerican.com/article/people-dont-trust-ai-heres-how-we-can-change-that/ (10012018), scientific American
24. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
25. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems. pp. 3856–3866 (2017)

26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
27. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Systems with Applications (2019)
28. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: International Conference on Information Processing in Medical Imaging. pp. 588–599. Springer (2015)
29. Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J.: Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition **61**, 663–673 (2017)
30. Shen, Y., Gao, M.: Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization. In: International Workshop on Machine Learning in Medical Imaging. pp. 389–397. Springer (2018)
31. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
32. Zhu, W., Liu, C., Fan, W., Xie, X.: Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 673–681. IEEE (2018)