Robust Bayesian Vision Transformer for Image Analysis and Classification

Fazlur Rahman Bin Karim
Electrical and Computer Engineering
University of Texas Rio Grande Valley, Edinburg, Texas
fazlurrahmanbin.karim01@utrgv.edu

Dimah Dera
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology, Rochester, NY
dimah.dera@rit.edu

***Abstract*—The tremendous success of the Transformer neural networks in natural language processing (NLP) boosts the interest in integrating and applying Transformer models to computer vision applications. The Vision Transformer (ViT) model adeptly captures extensive inter-dependencies among input sequences by employing the self-attention mechanism, thereby transforming picture data into semantically significant representations. In recent times, ViT has demonstrated superior performance in image classification tasks by implementing the transformer architecture, surpassing the capabilities of convolutional neural networks. Nevertheless, these deterministic architectures cannot evaluate the uncertainty associated with predictions, a crucial aspect in divergent and noisy situations. In order to guarantee the effectiveness and reliability of ViT in critical applications, the Bayesian Inference facilitates the process of making probabilistic predictions. Estimating the Bayesian posterior distribution of the network parameters enables a systematic method for reasoning about predictive uncertainty. The major difficulty in this process lies in propagating the posterior distribution through numerous non-linear layers of ViT architecture, which is mathematically cumbersome. In this paper, we propose a Bayesian Vision Transformer (Bayes-ViT) model, which seeks to make predictions as well as quantify the uncertainty associated with the output decision. The variational optimization approximates the posterior distribution over the unknown model parameters by minimizing the evidence lower bound (ELBO) loss function. The variational moments are propagated through the sequential, non-linear layers of Bayes-ViT by employing the first-order Taylor approximation. The covariance matrix of the predictive distribution effectively manifests the uncertainty associated with the output prediction. Extensive experiments on benchmark datasets (MNIST and Fashion-MNIST) exhibit (1) the superior robustness against noise and adversarial attacks compared to the deterministic ViT and (2) the self-evaluation ability based on the prediction uncertainty that becomes more evident when noise levels increase.**

***Index Terms*—Robust Vision Transformer, Bayesian Inference, Uncertainty Quantification, and evidence lower bound (ELBO).**

I. Introduction

Recent years have witnessed the rise of deep learning (DL) as a leading research field. Convolutional Neural Networks (CNNs) have been the most prevalent DL models for diverse computer vision applications such as classification, segmentation, and object detection [1], [2]. However, due to their localized receptive fields, these models have a limited ability to learn long-range dependencies in images. On the contrary, the attention-based structure of Transformer has excelled in modeling global relationships and can effectively capture such dependencies based on the success garnered in the natural language processing (NLP) domain [3]. The Vision Transformer (ViT) performs at the leading edge on vision tasks with its capability to handle long-range dependencies between input sequences [4]. ViT applies an encoder-based transformer structure to a series of non-overlapping image patches to perform image classification tasks [7].

Recently, researchers have been employing ViT extensively in numerous image classification applications, including medical diagnosis [5], [6]. For example, as part of the MIA-COVID-19 challenge, Gao *et al.* suggested COVID-ViT to differentiate COVID-19 from non-COVID photos of chest radiography as part of solving the binary classification problem [8]. In another medical imaging application, variants of ViT were trained and fine-tuned for meticulously identifying brain tumors while investigating Magnetic Resonance Imaging (MRI) images [9]. Breast cancer detection is another exciting field of research where ViTs have demonstrated impressive performance [10]. The ViT model has also been observed to be effective in other image classification applications, such as the detection and monitoring of deforestation activities [11].

Despite the success of the ViT models in image analysis and classification, ViTs, like most neural networks, produce deterministic predictions and can be sensitive to slight variations in input data (lack of robustness). Deterministic ViTs are prone to overconfident predictions in noisy environments and do not provide a direct measure of uncertainty associated with their predictions. The quantification of uncertainty in model predictions provides justification for the model performance when there is a shift in the distribution of input data, such as in the case of predictions in noisy environments. The consideration of model uncertainty holds significant importance in critical applications closely associated with human life. Failure to identify when models are likely to be erroneous might result in detrimental outcomes and reduce their efficacy in overcritical applications.

This paper proposes a novel Bayesian Vision Transformer

The work was supported by the National Science Foundation CRII-2153413 Award.

979-8-3503-2969-8/23/$31.00 ©2023 IEEE

(Bayes-ViT) neural network that addresses the task of image classification and quantifies the level of uncertainty associated with the class prediction. We adopt the variational inference and approximate the posterior distribution of the model's parameters by minimizing the evidence lower bound (ELBO) loss function. We propagate the first two moments of the variational posterior distribution over the model's parameters through all layers and non-linear activations of Bayes-ViT using the first-order Taylor approximation (an extension of the work in [19]). The propagated variational moments, i.e., the mean and covariance matrix, help simultaneously learn the mean and covariance of the probabilistic classification output. The mean vector at the output refers to the classification decision, while the covariance matrix conveys the level of uncertainty associated with that prediction. The experimental results demonstrate that the proposed Bayes-ViT outperforms the deterministic ViT in both no-noise and noisy environments. Particularly, the proposed model reveals enhanced robustness when subjected to noise and adversarial attacks during the test time using the benchmark datasets (MNIST and Fashion-MNIST). Moreover, we observe that the output uncertainty (measured by the predictive covariance matrix) increases with the increasing level of natural or adversarial noise. This behavior serves as a warning for human users to identify the failure mode of the model, especially important in mission-critical domain applications.

II. Bayesian Vision Transformer

A. Bayes-ViT Structure

The proposed Bayes-ViT model takes a sequence of non-overlapping image patches as input and linearly projects these patches into vectors, i.e., $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}^p$, where p is the size of each patch. The positional embedding is then applied to provide spatial information about the location of image patches within an input image. The output of the embedding is fed to the encoder structure of the Bayes-ViT model. The encoder structure consists of several layers, including the self-attention function, the multi-layer perceptron (MLP), and the layer normalization. The self-attention mechanism is the core of the Bayes-ViT model because it ascertains the correlation between the image patches within the input sequence. Given a set of n query vectors $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n \in \mathbb{R}^{d_k}$, n key vectors $\mathbf{k}_1, \mathbf{k}_2, \cdots, \mathbf{k}_n \in \mathbb{R}^{d_k}$, and n value vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n \in \mathbb{R}^n$, the attention mechanism maps the query vector, $\mathbf{q}_i$, the key vector, $\mathbf{k}_j$, and the value vector, $\mathbf{v}_j$ and computes a set of output vectors $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n \in \mathbb{R}^q$, such that

$$a_{ij} = \frac{\mathbf{k}_j^T \mathbf{q}_i}{\sqrt{d_k}}, \text{ and } \tilde{\mathbf{a}}_i = \varphi(\mathbf{a}_i) = \frac{\exp(\mathbf{a}_i)}{\sum_{j=1}^{n} \exp(a_{ij})}, \quad (1)$$

$$\mathbf{z}_i = \sum_{j=1}^{n} \tilde{\mathbf{a}}_i \odot \mathbf{v}_j, \text{ for } i, j = 1, \cdots, n. \quad (2)$$

where $\mathbf{q}_i = \mathbf{W}^{(\mathbf{q})}\, \mathbf{x}_i$, $\mathbf{k}_j = \mathbf{W}^{(\mathbf{k})}\, \mathbf{x}_j$, and $\mathbf{v}_j = \mathbf{W}^{(\mathbf{v})}\, \mathbf{x}_j$, and $\mathbf{W}^{(\mathbf{q})}$, $\mathbf{W}^{(\mathbf{k})}$, and $\mathbf{W}^{(\mathbf{v})}$ are the weight matrices. The query, key, and value vectors represent the linear projection of the input sequence, d_k denotes the dimension of the key vector, φ is the softmax function, and $\odot$ is the Hadamard product. The MLP consists of two fully connected layers and a Gaussian error linear unit (GeLU) activation function.

B. Bayesian Inference in the Bayes-ViT Model

We define a prior probability distribution over the model parameters, $p(\mathcal{W})$, where $\mathcal{W} = \{\mathbf{W}^{(\mathbf{q})}, \mathbf{W}^{(\mathbf{k})}, \mathbf{W}^{(\mathbf{v})}, \mathbf{W}^{\text{MLP}}\}$. We impose the independence assumption between the parameters across layers to (1) extract uncorrelated features across layers and (2) develop a feasible optimization problem, as estimating the joint distribution of all layers is mathematically intractable in large models. Using the training samples $\mathcal{D} = \{\mathbf{X}_i, \mathbf{y}_i\}_{i=1}^N$ and the prior distribution $p(\mathcal{W})$, we approximate the true unknown posterior distribution, $p(\mathcal{W}|\mathcal{D})$, with a simpler parametric variational distribution $q_\phi(\mathcal{W})$. The optimal parameters ϕ^* of this variational approximation are estimated by minimizing the Kullback-Leibler (KL) divergence between the approximate and the true posterior distributions, $\text{KL}\left[q_\phi(\mathcal{W}) \| p(\mathcal{W}|\mathcal{D})\right]$, which is known as the evidence lower bound (ELBO), $\mathcal{L}(\phi; \mathcal{D})$.

$$\mathcal{L}(\phi; \mathcal{D}) = -\mathbb{E}_{q_\phi(\mathcal{W})} \left\{\log p(\mathcal{D}|\mathcal{W})\right\} + \text{KL}\left[q_\phi(\mathcal{W}) \| p(\mathcal{W})\right]. \quad (3)$$

The first term in the ELBO loss is the log-likelihood expectation on the given set of training datasets and weight matrices. The second term is the KL divergence between two multivariate Gaussian distributions, i.e., the variational posterior distribution and the prior distribution, defined over the network parameters.

C. Uncertainty Propagation in the Bayes-ViT Model

We propagate the moments of the variational distributions, $q_\phi(\mathcal{W})$, i.e., the mean and covariance matrix, through all layers of the Bayes-ViT model. In our proposed model, all the learnable parameters are random variables. In the self-attention function, we have inner products between two random vectors, Hadamard products between random vectors and non-linear functions applied to random vectors. We will formulate the moment propagation for the self-attention function, and the mathematical relations can then be generalized to all layers.

Let, $\mathbf{w}_h^{(\mathbf{q})}$ be the h^{th} row vector of the weight matrix $\mathbf{W}^{(\mathbf{q})}$ where $h = 1, 2, \cdots, H$ and H is number of hidden nodes. The variational distribution is $\mathbf{w}_h^{(\mathbf{q})} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}_h^{(\mathbf{q})}}, \boldsymbol{\Sigma}_{\mathbf{w}_h^{(\mathbf{q})}})$. We assume the weight vectors are independent of each other and of the input vector $\mathbf{x}_i$. Each row of the matrix $\mathbf{W}^{(\mathbf{q})}$ multiplies the vector $\mathbf{x}_i$ in the matrix-vector multiplication $\mathbf{q}_i = \mathbf{W}^{(\mathbf{q})}\mathbf{x}_i$. Thus, the inner product between each pair of independent random vectors, $\boldsymbol{\mu}_{\mathbf{w}_h^{(\mathbf{q})}}$ and $\mathbf{x}_i$ can be written as $q_i = (\mathbf{w}_h^{(\mathbf{q})})^T\ \mathbf{x}_i$. The mean and covariance of $\mathbf{q}_i$ can be derived as the following,

$$\boldsymbol{\mu}_{\mathbf{q}_i} = \mathbf{M}^{(\mathbf{q})} \boldsymbol{\mu}_{\mathbf{x}_i}, \text{ where } \mathbf{M}^{(\mathbf{q})} = [\boldsymbol{\mu}^T_{\mathbf{w}_h^{(\mathbf{q})}}] \quad (4)$$

$$\boldsymbol{\Sigma}_{\mathbf{q}_i} = \begin{cases} \text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{w}_h^{(\mathbf{q})}} \boldsymbol{\Sigma}_{\mathbf{x}_i}\right) + \boldsymbol{\mu}^T_{\mathbf{w}_h^{(\mathbf{q})}}\ \boldsymbol{\Sigma}_{\mathbf{x}_i}\ \boldsymbol{\mu}_{\mathbf{w}_l^{(\mathbf{q})}} + \boldsymbol{\mu}^T_{\mathbf{x}_i} \boldsymbol{\Sigma}_{\mathbf{w}_h^{\mathbf{q}}}\ \boldsymbol{\mu}_{\mathbf{x}_i}, & h = l \\ \boldsymbol{\mu}^T_{\mathbf{w}_h^{(\mathbf{q})}}\ \boldsymbol{\Sigma}_{\mathbf{x}_i}\ \boldsymbol{\mu}^T_{\mathbf{w}_l^{(\mathbf{q})}}. & h \neq l \end{cases}$$

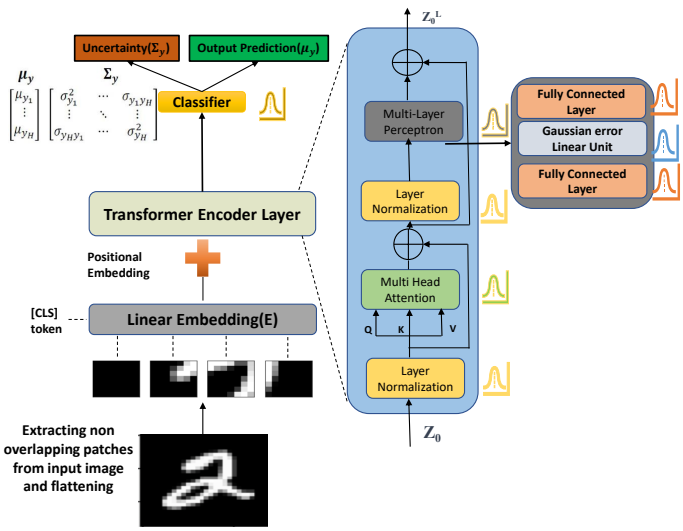

Fig. 1. A schematic diagram of the proposed Bayes-ViT model illustrating the uncertainty propagation through the various non-linear layers of the model.

Similarly, the mean and covariance matrix of the key vector $\mathbf{k}_i$, the value vector $\mathbf{v}_i$ and the vector $\mathbf{a}_i$ in Equation 1 follow the derivation in Equation 4.

The first-order Taylor approximation estimates the mean and covariance matrix after the non-linear activation functions in the model, including the softmax function. Thus, the mean and covariance of $\tilde{\mathbf{a}}_i$ in Equation 1 are derived as follows.

$$\boldsymbol{\mu}_{\tilde{\mathbf{a}}_i} \approx \varphi(\boldsymbol{\mu}_{\mathbf{a}_i}), \quad \boldsymbol{\Sigma}_{\tilde{\mathbf{a}}_i} \approx \mathbf{J}_\varphi \boldsymbol{\Sigma}_{\mathbf{a}_i} \mathbf{J}_\varphi^T, \tag{5}$$

where $\mathbf{J}_\varphi$ denotes the Jacobian matrix of $\tilde{\mathbf{a}}_i$ with respect to $\mathbf{a}_i$ evaluated at $\boldsymbol{\mu}_{\mathbf{a}_i}$. The results presented in Equation 5 hold true for any non-linear activation function, including hyperbolic tangent (Tanh), sigmoid, or rectified linear unit (ReLU). The mean and covariance matrix of the element-wise multiplication in Equation 2, i.e., $\tilde{\mathbf{z}}_i = \tilde{\mathbf{a}}_i \odot \mathbf{v}_j$, are derived as follows,

$$\begin{aligned} \boldsymbol{\mu}_{\tilde{\mathbf{z}}_i} &= \boldsymbol{\mu}_{\tilde{\mathbf{a}}_i} \odot \boldsymbol{\mu}_{\mathbf{v}_j}, \qquad (6) \\ \boldsymbol{\Sigma}_{\tilde{\mathbf{z}}_i} &= \boldsymbol{\Sigma}_{\tilde{\mathbf{a}}_i} \odot \boldsymbol{\Sigma}_{\mathbf{v}_j} + D(\boldsymbol{\mu}_{\mathbf{v}_j}) \boldsymbol{\Sigma}_{\tilde{\mathbf{a}}_i} D(\boldsymbol{\mu}_{\mathbf{v}_j}) + D(\boldsymbol{\mu}_{\tilde{\mathbf{a}}_i}) \boldsymbol{\Sigma}_{\mathbf{v}_j} D(\boldsymbol{\mu}_{\tilde{\mathbf{a}}_i}), \end{aligned}$$

where $D(\boldsymbol{\mu}_{\mathbf{v}_j})$ represents the diagonal matrix whose entries are given by the column vector $\boldsymbol{\mu}_{\mathbf{v}_j}$.

By propagating the variational moments through all layers, we obtain the moments of the predictive distribution, $p(\mathbf{y}|\mathbf{X}, \mathcal{D})$. The mean of $p(\mathbf{y}|\mathbf{X}, \mathcal{D})$, i.e., $\boldsymbol{\mu}_\mathbf{y}$ represents the network's prediction, while the covariance matrix, $\boldsymbol{\Sigma}_\mathbf{y}$, reflects the uncertainty associated with the output classification. Figure 1 illustrates the proposed Bayes-ViT model.

III. Experimental Results

This section evaluates the performance of the proposed Bayes-ViT model compared to the deterministic ViT using the benchmark MNIST and Fashion-MNIST datasets. Both models are trained on clean datasets and then evaluated during the test time in two separate cases: (1) using clean test samples and (2) after adding various levels of natural noise and adversarial attacks. We use random noise, the fast gradient sign method (FGSM) [20], and the projected gradient descent (PGD) adversarial attacks to evaluate the models in noisy situations [21]. The hyperparameters selected for training the models are outlined in Table I. Table II presents the classification accuracy of the proposed Bayes-ViT model compared to the deterministic counterpart using MNIST and Fashion-MNIST datasets in clean and various noisy situations.

We plot the uncertainty measured by the output variance (diagonal element of the output covariance matrix that corresponds to the predicted class) versus signal-to-noise ratio (SNR) for the various noisy conditions. Figure 2 shows the output variance of the proposed model versus SNR for the MNIST and Fashion-MNIST datasets.

A. Discussion and Robustness Analysis

We observe from Table II that the proposed Bayes-ViT model attains higher accuracy for both MNIST and Fashion-MNIST datasets when tested on clean as well as noisy test samples. Particularly, the proposed model maintains its performance with high accuracy under high levels of random noise and adversarial attacks. We highlight the highest accuracy for the two models for the highest level of noise. For example, the proposed Bayes-ViT obtains 87.01% and 88.59% for the highest level of FGSM and PGD adversarial noise compared to 56.52% and 80.92% for the deterministic ViT on the MNIST dataset. Similarly, the proposed model produces 51.5% and 68% accuracy for the highest level of FGSM and PGD adversarial noise as compared to 28.9% and 45.9% for the deterministic model on the Fashion-MNIST dataset.

The robust behavior of the proposed Bayes-ViT model can be justified by the uncertainty (measured by the covariance matrix) propagated through the model's layers and associated with the output classification. The uncertainty associated with the model's parameters passes the important information from the data and filters out the irrelevant information, which helps improve the model's performance. Furthermore, we observe from Figure 2 that the output uncertainty (variance) gets higher values in the case of noise, especially when there is an adversarial attack. When the noise level increases (SNR decreases), the variance values increase significantly, especially in the adversarial noise cases. The increasing value of the uncertainty adjusts the learning process of the Bayes-ViT model and promotes robustness in its performance. The uncertainty plays a vital role in preserving important features of the data and eliminating weak and redundant features that might be severely affected by the attacks.

TABLE I
Hyperparameter setting in our simulation.

Dataset	Patch Size	No. Encoder layers	No. Hidden Units	Batch Size	No. Epoch	Initial Learning Rate	Final Learning Rate	KL Weight Factor
MNIST	4	5	64	20	300	0.001	0.00001	0.00001
F-MNIST	8	7	64	50	500	0.001	0.00001	0.001

TABLE II

CLASSIFICATION ACCURACY OF THE PROPOSED BAYES-VIT AND DETERMINISTIC VIT MODELS USING MNIST AND FASHION-MNIST DATASETS. BOTH MODELS ARE TESTED FOR VARIOUS LEVELS OF RANDOM NOISE AND FGSM AND PGD ADVERSARIAL ATTACKS.

(a) MNIST Dataset			
Noise Type	Noise level	Bayes-ViT	Deterministic ViT
No Noise		90.08	88.1
Gaussian	0.05	90.01	85.57
	0.1	89.53	84.57
	0.2	**86.50**	**78.22**
FGSM	0.001	89.43	85.7
	0.005	89.43	84.88
	0.01	89.35	83.51
	0.05	**87.01**	**56.52**
PGD	0.001	89.59	86.07
	0.005	89.53	85.12
	0.01	89.34	85
	0.05	**88.59**	**80.92**
(b) Fashion-MNIST Dataset			
Noise Type	Noise level	Bayes-ViT	Deterministic ViT
No Noise		82.44	79.9
Gaussian	0.05	81.60	79.2
	0.1	75.40	72.4
	0.2	**52.20**	**47.7**
FGSM	0.001	81.10	79.23
	0.005	77.97	76.05
	0.01	70.50	69.4
	0.05	**51.50**	**28.9**
PGD	0.001	82.00	79.01
	0.005	80.30	77.2
	0.01	78.60	73.2
	0.05	**68.00**	**45.9**

IV. CONCLUSION

This work presents a novel Bayes-ViT model for image analysis and classification that delineates both robustness and uncertainty awareness. We adopt the Bayesian inference and propagate the mean and covariance matrix of the variational posterior across the model's layers. The mean of the predictive distribution refers to the predicted class, whilst the covariance matrix provides insights about the uncertainty associated with the prediction. The proposed model exhibits robust behavior against random noise and adversarial attacks compared to the deterministic model. Moreover, the learned uncertainty increases significantly with the level of noise, which defines the failure mode of the model under noisy situations. The excellent robustness and self-assessment properties of the proposed model make it highly suitable for critical applications.

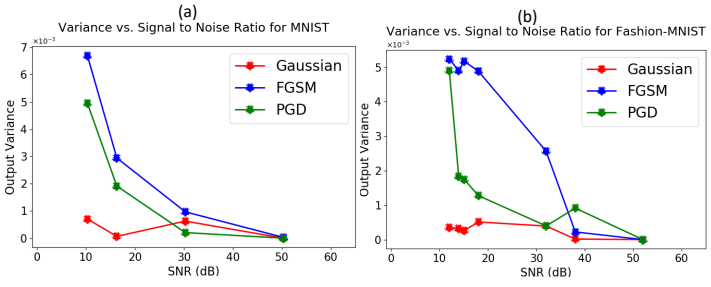

Fig. 2. The predictive variance versus SNR for Bayes-ViT for MNIST and Fashion-MNIST under random noise, FGSM and PGD adversarial attacks.

REFERENCES

[1] N. Jmour, S. Zayen and A. Abdelkrim, "Convolutional neural networks for image classification," in *International Conference on Advanced Systems and Electric Technologies (IC_ASET)*, 2018, pp. 397-402.

[2] F. Milletari, N. Navab and S. -A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *International Conference on 3D Vision (3DV)*, 2016, pp. 565-571.

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[4] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, 'Visformer: The Vision-friendly Transformer," in Proc 2021 IEEE/CVF Int. Conf. Comput. Vision (ICCV), 2021, pp. 569-578.

[5]] Sharma, A. K., & Sharma, N. K. (2023). A Novel Vision Transformer with Residual in Self-attention for Biomedical Image Classification. *arXiv preprint arXiv:2306.01594*

[6] Dai, Y., Gao, Y., & Liu, F. (2021). TransMed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8), 1384.

[7] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. 9th Int. Conf. Learn. Representations (ICLR 2021), Austria, May 3-7, 2021.

[8] X. Gao *et al.*, "COVID-VIT: Classification of Covid-19 from 3D CT chest images based on vision transformer model," 2022 3rd International Conference on Next Generation Computing Applications (NextComp), 2022, pp. 1-4.

[9] Tummala S, Kadry S, Bukhari SAC, Rauf HT. Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling. Curr Oncol. 2022 Oct 7;29(10):7498-7511. doi: 10.3390/curroncol29100590. PMID: 36290867; PMCID: PMC9600395.

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

[11] Kaselimi, M., Voulodimos, A., Daskalopoulos, I., Doulamis, N., & Doulamis, A. (2022). A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring. *IEEE Transactions on Neural Networks and Learning Systems*.

[12] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, 'Visformer: The Vision-friendly Transformer," in Proc 2021 IEEE/CVF Int. Conf. Comput. Vision (ICCV), 2021, pp. 569-578.

[13] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the Big Data Paradigm with Compact Transformers," 2021. [Online]. Available: https://arxiv.org/abs/2104.05704

[14] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in Proc 38th Int. Conf. Mach. Learn., M. Meila, and T. Zhang, Eds., in Proc. of Mach. Learn. Res. (PMLR), vol. 139, 2021, pp. 10347-10357.

[15] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., & Douze, M. (2021). Levit: a vision transformer in convnet's clothing for faster inference. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 12259-12269).

[16] Iwana, B. K., & Kusuda, A. (2023). Vision Conformer: Incorporating Convolutions into Vision Transformer Layers. arXiv preprint arXiv:2304.13991.

[17] Shao, R., & Bi, X. J. (2022). Transformers meet small datasets. IEEE Access, 10, 118454-118464.

[18] Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., & Nadai, M. (2021). Efficient training of visual transformers with small datasets. Advances in Neural Information Processing Systems, 34, 23818-23830.

[19] D. Dera, N. C. Bouaynaya, G. Rasool, R. Shterenberg and H. M. Fathallah-Shaykh, "PremiUm-CNN: Propagating Uncertainty Towards Robust Convolutional Neural Networks," IEEE Trans. Signal Process., vol. 69, pp. 4669-4684, 2021

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proceedings of 3rd International Conference on Learning Representations, (ICLR), 2015.

[21] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.