PREDICTING TISSUE STIFFNESS FROM MICROSCOPIC FIBER STRUCTURE IMAGES USING DINOV2

Deep learning course, Electrical engineering faculty, Tel Aviv University

Nofar Sachs, Omri Segev

***Abstract*— Measuring tissue stiffness is vital for assessing tissue health and function and can serve as a diagnostic tool for various pathologies. However, current methods such are costly, time-consuming, invasive, and require specialized equipment. Recent advancements in computer vision, particularly with foundation models trained on large image datasets, show promise in feature extraction that can benefit deep learning solutions. In this paper, we will introduce a novel fine-tuned model based on Meta AI's DINOv2 pre-trained model, specifically designed to predict tissue stiffness. Our model has surpassed DINOv2 in this task and offers a non-invasive, accessible solution.**

I. INTRODUCTION

The stiffness of a tissue, a critical biomechanical property, indicates the tissue's resistance to deformation when a force is applied. This intrinsic characteristic significantly impacts cellular and organ function. The stiffness of a tissue environment affects how cells behave, influencing their movement (migration), adhesion to surfaces, and specialization (differentiation). Changes in tissue stiffness are linked to various diseases. For example, in cancer, stiffer tumor environments may promote tumor growth and spread. Understanding this link could lead to new diagnostic and treatment approaches. However, existing techniques such as AFM, indentation tests, and biopsies are expensive, time-consuming, often invasive, and require specialized equipment. Optical microscopy offers a readily available, non-invasive way to visualize tissue structure, including the direction of stiffness-related fibers, which is crucial for understanding tissue stiffness.

With the rise of deep learning technology, its remarkable performance in the field of image recognition has garnered considerable attention. Muhammad et al. introduced a novel approach for predicting Dh using gene expression programming [1]. This method efficiently extracts features related to images and objects and can undertake tasks such as classification and regression. The stiffness of a tissue may be associated with the distribution and thickness of fibers in the images. The distribution and thickness of fibers vary among tissues with different levels of stiffness. However, quantifying these differences is challenging, and traditional image processing methods are inadequate for analyzing them. Therefore, utilizing deep learning to analyze the distinct fiber structure of tissues and predict their stiffness holds significant practical potential. The implementation of this research could streamline experimental procedures, reduce human effort and expensive processes, and enhance the reliability of measurement results.

The prediction of tissue stiffness heavily relies on deep learning feature extractors, with the most commonly used ones being convolutional neural networks (CNNs) [2-4] and vision transformers (ViTs) [5,6]. CNNs utilize a local perception approach, focusing on specific parts of the input image through the convolution operation to extract features from these areas. This enables the network to effectively capture local structures and patterns within the image [7]. In contrast, ViTs are well-suited for processing large-scale image data. Based on the self-attention mechanism and treating pixels in the image as a sequence, ViTs, introduced in 2020 by Dosovitskiy et al. [8], have demonstrated comparable or even superior performance in image classification and regression tasks compared to CNNs, as they avoid introducing local biases and maintain computational efficiency when handling large-sized images.

II. RELATED WORK

A. ViT for regression

The solution for regression problems using ViTs can be complex since most Vits were originally designed to solve classification problems. Wessels et al. suggest a ViT-based solution to the prediction of survival probability based on histopathological tissue section images, using (DINO-ViT) to extract image features and using the resulting feature vector in a Cox regression model to predict overall and disease-specific survival, resulting in a fair ability of the model to identify high-risk patients using histological images [9]. Another related work by Wang et al. where an image regression model based on a vision transformer (ViT) was introduced to predict the shear strength of transparent soil. The results achieved correlations of 0.93 and 0.94 in the two prediction tasks, thus outperforming existing deep learning models[10].

B. Stiffness prediction

The problem we faced was studied and investigated and a variety of solutions were proposed, both deep learning and machine learning solutions. A deep learning approach introduced by Behera et al. was an adaptive learning rate-based multilayer perceptron technique for determining Young's

modulus of tissues. With results of consistently low mean absolute errors, they managed to prove the reliability of the proposed technique[11]. Another attempt to predict tissue stiffness, precisely myocardial stiffness, was introduced by Babaei et al. where both a neural network as the feature extraction mechanism and a machine learning model as the regression model were trained to predict the stiffness of the myocardial tissue stiffness. With excellent agreements with ML predictions, the trained ML model offers a feasible technology to estimate patient-specific myocardial properties[12].

III. Dataset and Data Processing

The dataset for the task is a set of microscopic images of tissues and their stiffness measurements. The images and measurements were taken in the lab of Prof. Ayelet Lesman, School of Mechanical Engineering, Tel Aviv University. The images were produced using optical microscopy, and the stiffness measurements and their coordinates were taken using optical tweezers, and their values are in Pascal (Pa).

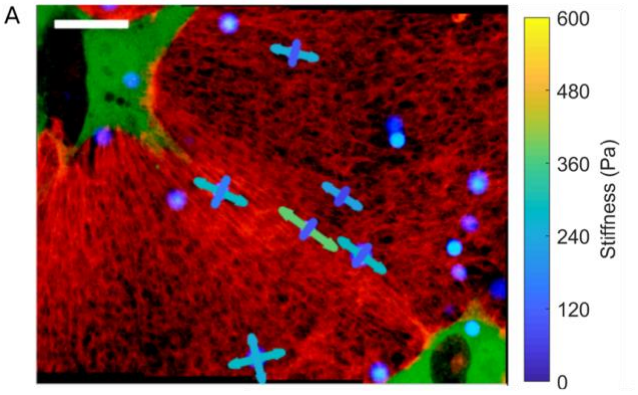


Figure 1: a sample image of a tissue (red) between two cancer cells(green). The stiffness measurements are denoted in colorful dots according to the bar.

The challenge for this task was the dataset size. The dataset consists of 242 samples. When trying to evaluate fine-tuned ViTs, the model performance can be bottlenecked by the dataset size[13]. For the fine-tuning of Dinov2 for a tumor classification task, Huang et al. used a dataset of 10015 images for training and 1512 for testing[14]. For a regression task, which isn't the intuitive use for ViTs, Wang et al. [4] trained a ViT on a dataset containing a total of 20000 pre-processed samples.

A. Image Augmentation and Normalization

A popular approach [15] to increase an image dataset size is by using image augmentation. Image augmentation is the procedure by which an existing dataset is expanded by transforming the original dataset to create new data, and in such a way that the new data are label preserving. The goal is to increase the variance of the dataset while ensuring that new data are meaningful and do not merely add unnecessary volume to the dataset[16]. Another challenge was to choose the relevant augmentations that would preserve valuable features for the regression problem. The chosen augmentation was horizontal and vertical flips. Although it allows a dataset size growth by only a factor of 4 (all possible flip combinations), this augmentation best preserves the complex fiber structure that holds the potential to derive tissue stiffness. The target stiffness measurement coordinates were adjusted accordingly. Additionally, the images were normalized by mean and standard deviation.

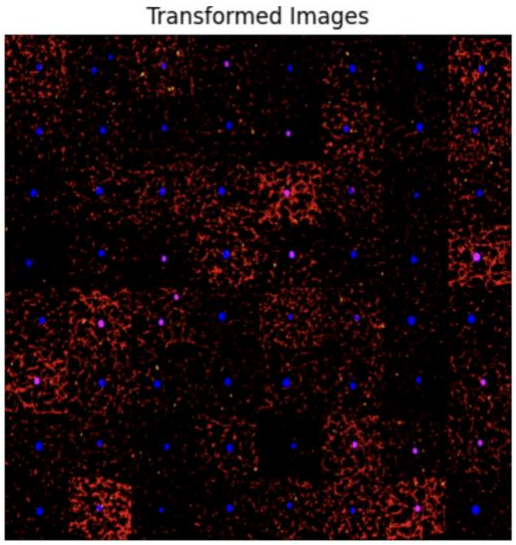


Figure 2: Augmented images after the processing pipeline(horizontal and vertical flip, normalization)

B. Batches and epochs

The batch size was set to 64, with RGB three-channel input of 224x224 pixels. We then ran 10 epochs for each training.

IV. Methods

DINOv2, short for self-DIstillation with NO labels (version 2), [17] represents a state-of-the-art self-supervised learning framework that plays a key role in our approach. The DINOv2 model, pre-trained on the ImageNet [19] dataset comprising 1.3M unlabeled 2D natural images. It employs self-distillation to acquire knowledge from unlabeled images in a self-supervised manner. DINOv2 features a dual-network architecture consisting of a teacher and a student network. Both networks share similar network architectures, typically built on Vision Transformers (ViTs) [8], but undergo different training methods. When presented with an input image, the student and teacher networks are provided with different image augmentations, while being required to extract consistent features. Throughout the training process, the parameters in the student network are optimized using gradient descent algorithms such as stochastic gradient descent (SGD) and Adam[18], while the parameters in the teacher network are updated via the moving average of their counterparts in the student network. The trained teacher network is often utilized as the final product. This general-purpose foundation model serves as a potent feature extractor for downstream tasks such as classification and regression.

In our research, we harness the teacher network of a pre-trained DINOv2 model to extract robust and discriminative

features, which serve as input for our subsequent regression neural network. The DINOv2 model is pre-trained on the ImageNet dataset, comprising 1.3 million unlabeled 2D natural images. It excels at processing 2D slices from volumetric computed tomography (CT) or magnetic resonance (MR) images and holds great potential for valuable feature extraction from 2D tissue fiber microscopy images.

For the training, we used combinations of the following training approaches:

- Training the Multi-Layer Perceptron (MLP) head of the model.
- Fine-tuning the entire DinoV2 model using QLorA.

Training the MLP head is a common transfer learning approach where the part of the model that is being trained is the output layer so the properties and expertise of the model in completing other tasks are preserved and applied to the given different task. In our task, the MLP head was modified in a way that the SoftMax activation layer was removed to input regressive output instead of probabilities for classification tasks. The head was then trained to fit our task.

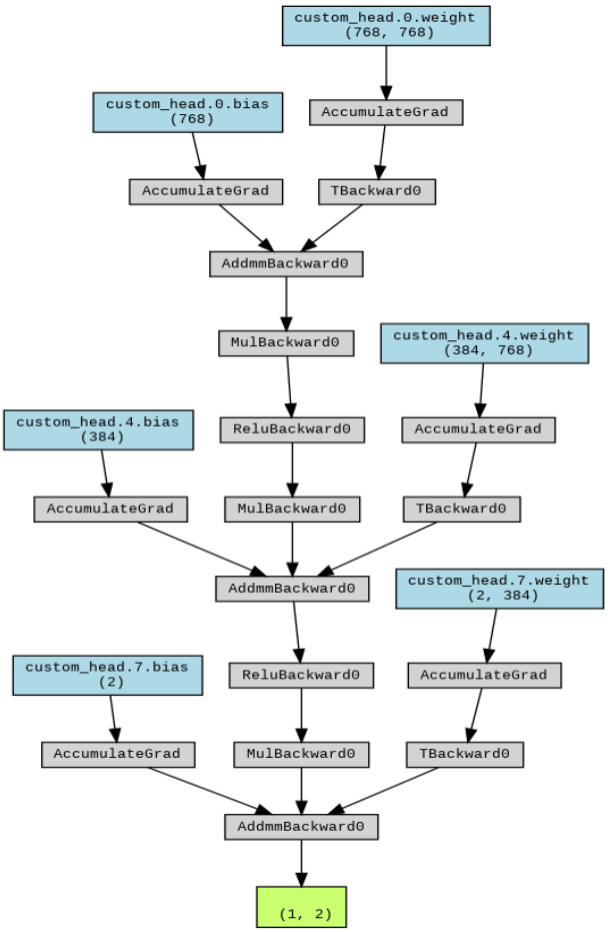


Figure 3: model modified MLP head architecture

LoRA [21] (Low-Rank Adaptation of Large Language Models) is a technique used to fine-tune large language models (LLMs) for specific tasks or domains. This method involves introducing trainable rank decomposition matrices into each layer of the transformer architecture, which reduces the trainable parameters for downstream tasks while keeping the pre-trained weights frozen. The core concept behind LoRA is that while the weight matrices of a machine learning model are of high rank, the weight updates created during fine-tuning are of low intrinsic rank. In other words, we can fine-tune the model using a much smaller matrix than we would need if we were training it from scratch, without experiencing a significant loss of performance.

QLoRA [22] is an extended version of LoRA that operates by quantizing the precision of the weight parameters in the pre-trained LLM to 4-bit precision. Normally, trained model parameters are stored in a 32-bit format, but QLoRA compresses them to a 4-bit format. This reduction in precision significantly decreases the memory footprint of the LLM, making it feasible to fine-tune it on a single GPU. Consequently, this method enables LLM models to run on less powerful hardware, including consumer GPUs.

V. EXPERIMENT

We experimented with four training approach combinations:

a. Training approach a: training only the MLP head of the model.

b. Training approach b: Training the MLP head and finetuning the model simultaneously.

c. Training approach c: Training the MLP head, then fine-tuning the model and training the MLP head.

d. Training approach d: Fine-tuning the model and training the MLP head, then training the MLP head.

As usually done with regression models, the computed loss is Mean Square Error (MSE):

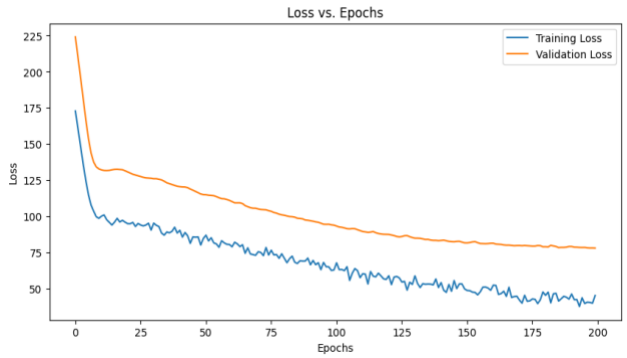
$$MSE = \frac{1}{N}\Sigma(y_i - \hat{y}_i)^2$$

Equation1 – MSE, where y_i is the correct target, $\hat{y}_i$ is the model prediction, N is the number of samples

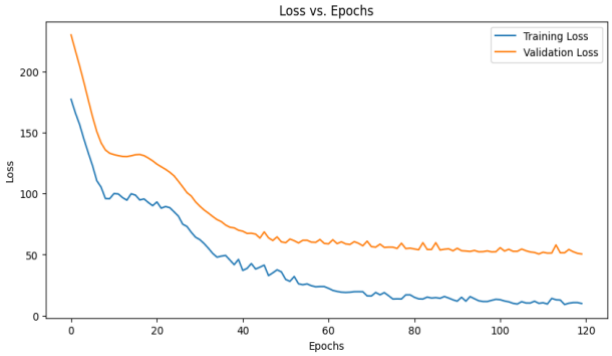
We also calculated the relative error:

$$Relative\ Error = \frac{\sqrt{MSE}}{||stiffness||}$$

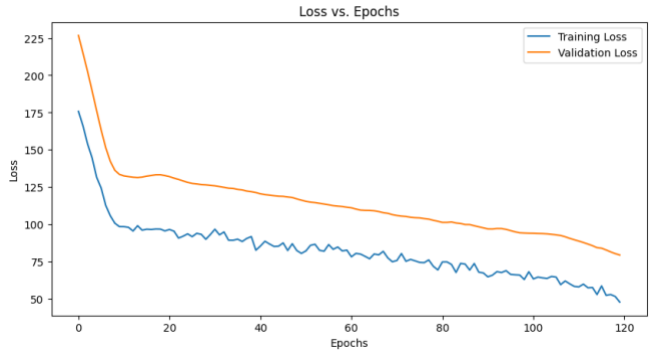
The graphs below represent the training process of each of the different approaches. For each of the graphs, the calculated loss was MSE, for both training and validation.



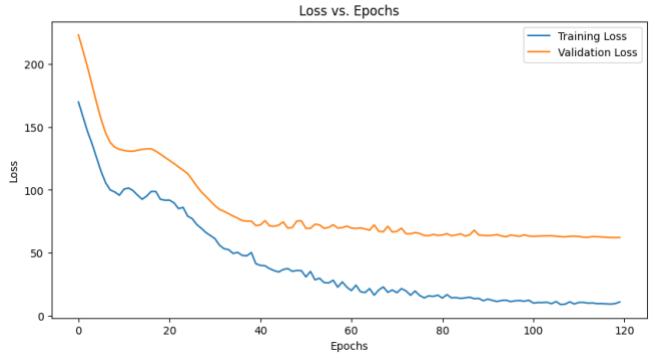
Graph 1: Loss curve of training approach a.



Graph 2: Loss curve of training approach b



Graph 3: Loss curve of training approach c



Graph 4: Loss curve of training approach d

TABLE I. FINAL TRAINING RESULTS

Training approach	*N epochs*	*Training MSE*	*Validation MSE*	*Relative Error*
Approach a[a]	200	45.3172	78.0882	231%
Approach b[a]	120	9.9369	50.4879	185.77%
Approach c[a]	100+20	47.88	79.4	233%
Approach d[a]	100+20	10.8176	62.1288	206%

a. The approach details can be found under the Experiments section

As can be seen from the graphs above, The DINOv2 model struggled with our limited data. As the model loss did drop along the training epochs, there was a consistent delta between the training and validation loss, indicating an overfit of the model. The best performance is of training approach b, with a slightly smaller delta between the validation and training loss. This approach seems to achieve the best generalization of the training data and holds a promise to the feasibility of the training approach for this task, yet the dataset size remains a challenge.

VI. CONCLUSION

While the training results show promise for the feasibility of training a DINOv2 ViT-based model as a useful tool for the prediction of tissue stiffness, the dataset size was indeed a bottleneck for the models' performance. There is a need to address this issue to improve the overall performance and create a reliable solution. There are a few possible approaches:

- Expand Data Augmentation: Trying more techniques (cropping, resizing, and elastic transformations) might vary more on the limited dataset.
- Replace DinoV2 with a smaller architecture: A smaller model has fewer parameters to adjust during training, so it will reduce the risk of overfitting.
- Use Unlabeled Data with Self-Supervised Learning: Enhancing feature extraction ability from unlabeled tissue images(For example, training the model to distinguish between slightly rotated versions of the same image).

We find the ViT-based model to be a feasible solution for the discussed task, yet to promise a reliable solution further research is required.

REFERENCES

[1] Raja, M.N.A.; Abdoun, T.; El-Sekelly, W. Smart prediction of liquefaction-induced lateral spreading. J. Rock Mech. Geotech. Eng. 2023.

[2] Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278-2324.

[3] Gurin, J.; Thiery, S.; Nyiri, E.; Gibaru, O.; Boots, B. Combining pretrained CNN feature extractors to enhance clustering of complex natural images. Neurocomputing 2020, 423, 551-571.

[4] Varshni, D.; Thakral, K.; Agarwal, L.; Nijhawan, R.; Mittal, A. Pneumonia detection using CNN based feature extraction. In Proceedings of the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 20-22 February 2019.

[5] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11-17 October 2021.

[6] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, ≈Å.; Polosukhin, I. Attention is all you need. Adv. Neural Inf. Process. Syst. 2017, 30, 6000-6010.

[7] Liu, Y.; Pu, H.; Sun, D.-W. Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. Trends Food Sci. Technol. 2021, 113, 193–204.

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

[9] Wessels F, Schmitt M, Krieghoff-Henning E, Nientiedt M, Waldbillig F, Neuberger M, Kriegmair MC, Kowalewski KF, Worst TS, Steeg M, Popovic ZV, Gaiser T, von Kalle C, Utikal JS, Fröhling S, Michel MS, Nuhn P, Brinker TJ. A self-supervised vision transformer to predict survival from histopathology in renal cell carcinoma. World J Urol. 2023 Aug.

[10] Wang, Z.; Jia, J.; Zhang, L.; Li, Z. ViT-Based Image Regression Model for Shear-Strength Prediction of Transparent Soil. Buildings 2024, 14, 959. https://doi.org/10.3390/buildings14040959 M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.

[11] B. Behera, M. F. Orlando and R. S. Anand, "Prognosis of Tissue Stiffness Through Multilayer Perceptron Technique With Adaptive Learning Rate in Minimal Invasive Surgical Procedures," in IEEE Transactions on Medical Robotics and Bionics, vol. 6, no. 2, pp. 769-781, May 2024.

[12] Babaei, H., Mendiola, E.A., Neelakantan, S. et al. A machine learning model to estimate myocardial stiffness from EDPVR. Sci Rep 12, 5433 (2022).

[13] Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2021). Scaling Vision Transformers. ArXiv. /abs/2106.04560

[14] Huang, Y., Zou, J., Meng, L., Yue, X., Zhao, Q., Li, J., Song, C., Jimenez, G., Li, S., & Fu, G. (2024). Comparative Analysis of ImageNet Pre-Trained Deep Learning Models and DINOv2 in Medical Imaging Classification. ArXiv. /abs/2402.07595

[15] Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2022). A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. ArXiv. https://doi.org/10.1016/j.patcog.2023.109347

[16] Marcus D Bloice, Peter M Roth, Andreas Holzinger, Biomedical image augmentation using Augmentor, Bioinformatics, Volume 35, Issue 21, November 2019, Pages 4522–4524, presented at the IEEE Summer power Meeting, Dallas, TX, June 22–27, 1990, Paper 90 SM 690-0 PWRS.

[17] Caron, M., Touvron, H., Misra, I., J´egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)

[18] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[19] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A largescale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

[20] Song, X., Xu, X., & Yan, P. (2024). General Purpose Image Encoder DINOv2 for Medical Image Registration. ArXiv. /abs/2402.15687.

[21] Hu, E. J., Shen, Y., Wallis, P., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv. /abs/2106.09685

[22] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv*. /abs/2305.14314

APPENDIX

GitHub
https://github.com/NofarSachs/tissue_stiffness_dinov2

ACKNOWLEDGMENT

Prop. Raja Giryes, School of Electrical Engineering, Tel Aviv University.

Prof. Ayelet Lesman, School of Mechanical Engineering, Tel Aviv University.