Predicting Lifespan/Age of *C.elegans* using Multimodal Learning

Jien Li
Brown University
Providence, RI 02903
jien_li@brown.edu

Yiming Yu
Brown University
Providence, RI 02903
yiming_yu@brown.edu

Yifan Lu
Brown University
Providence, RI 02903
lu_yifan@brown.edu

Jingyi Lu
Brown University
Providence, RI 02903
jingyi_lu1@brown.edu

1. Introduction

Studying the biology of aging has caused tremendous enthusiasm in recent years. One of its sub-field aims to develop methods to accurately predict age and lifespan, as means to understand how organisms age as well as tools to expedite drug screening efforts. This field has spent a lot of effort in complex organisms, using DNA methylation profiles to predict age. However, these methods are invasive, damaging larger organisms by drawing blood, and terminating smaller organisms such as flies. This invasiveness can prevent us from accurately measuring the future lifespan of these animals. Therefore, a potent non-invasive method to extract features from animals and accurately predict age/lifespan is a critical puzzle to solve. In this project, we aim to construct age prediction models for *Caenorhabditis elegans* (*C. elegans*) using only images/videos. *C. elegans* is the ideal species here due to its status as a popular model organism and its transparent nature, enabling the detection of both external and internal features.

1.1. Problem Type(do we need this section?)

This project tackles a multi-task learning problem on image and video data, focusing on predicting two related outcomes. The primary task is a regression problem, where the model predicts age and daysRemain, both continuous variables indicating the number of days alive and remaining until dead. To complement this, we transform daysRemain into discrete categories using quantile-based discretization, creating a classification task to predict life stages as categorical labels.

1.2. Related Work

Only a limited amount of previous research was conducted on this subject. However, they either predicted only age (Jiunn-Liang Lin, 2020), required fluorescent imaging of mutant *C. elegans* (Yao Song, 2022), or demanded complex tracking cameras in behavior assays (Celine N. Martineau, 2020), rendering these efforts inaccessible for most researchers. We aim to achieve high performance with simple brightfield imaging on wild type *C. elegans*, leveraging multiple data modalities to increase accuracy.

2. Data

All of our data is gathered from first-hand imaging experiments. Specifically, wild-type N2 strain *C. elegans* were cultured and imaged in facilities provided by the Silva García lab at Brown university.

2.1. Animal Handling

All N2 *C. elegans* were cultured in standard agar plates, seeded with *Escherichia coli* (*E. coli*) strain OP50. Transfering was performed using a thin platinum pick. 195 embryos from day 1 to day 3 adult C. elegans were transferred to 35mm separate plates, and were individually housed at 20°C throughout their lifespan. After they reached the day 1 adult stage, when progeny production started, they were transferred to a new plate every 2 days to remain individually housed, and the progenies were discarded. An animal is determined dead when there is no active motion and no involuntary contraction after being gently stretched with the pick.

2.2. Image Collection

Every other day of the experiment, starting from the embryo stage, each worm would be imaged directly in the plates they were housed in with a Axiocam 705 microscope under brightfield settings. The light source is tweaked to enable maximum contrast. Images and videos were captured with Zen 3.6 (Pro) microscopy software. At 16.2x magnification, a 60s video of the animals' crawling motion was recorded. At 100x magnification, an overview of the worm was imaged, capturing the head and the majority of the body. At 260x magnification, 3 images targeting the head, body, and tail were recorded, focusing on the pharynx, valva, and intestine, respectively. A 10s video was also recorded at 260x to capture pharynx movement. The resulting .czi files were compressed and exported to .png and .avi file formats, for a total of over 3,500 data entries.

2.3. Motion Feature Extraction

The 60s motion videos were quite large (~300GB) and consisted mostly of empty spaces. Therefore, it is beneficial to first condense these data into motion features

using existing software-“tierpsy-tracker”. Detailed instructions were included in their github pages, and the extraction parameters we used were included in our github pages. Post extraction, features that were empty for all animals were removed from our dataset, and features that were empty for a subset of animals were filled with mean imputation. At the end, we were left with around 3,000 features per worm in a numerical tabular format.

2.4. Input Clean-up and Target Variables

The input variables were the images, the 10s head videos, and the tabular motion features. The images were resized to 224x224 pixels and intensities confined to between 0 and 1. The tabular features were also min-max transformed to values between 0 and 1 on a per column basis. The target variables include both “age” (how many days since embryo) and “daysRemain” (how many days until death). Additionally, both of these target variables were classified into 5 bins based on percentile as “age_group” and “daysRemain_gorup”, enabling classification tasks.

To prepare the raw video data for training, we implemented a series of preprocessing steps designed to standardize input dimensions and enhance meaningful feature extraction. First, the raw RGB video frames were converted to grayscale to simplify the input data while retaining critical motion and spatial features. Next, frames were resized to a standardized resolution of 112x112 pixels. To address variations in temporal length, all videos were uniformly sampled to consist of 16 frames. Finally, pixel values were normalized to the [0, 1] range to stabilize the learning process.

3. Methodology

To achieve the ultimate goal of multimodal learning, the project employed separate convolutional neural network (CNN) models and transformers for each type of data. The final multimodal learning model was built based on the features and output of the models.

3.1. Image CNN and CvT

The implemented models for image data include a CNN model and a Convolutional Vision Transformer (CvT) model. The CNN model is designed with convolutional layers, pooling, batch normalization, random dropout, data augmentation, and an MLP to produce the output. The CvT model is structured into three stages. Each stage begins with a convolutional layer that generates an embedding, which is then projected into the transformer inputs using a separable convolution structure (`Depth-wise Conv2d → BatchNorm → Point-wise Conv2D`). The transformer processes the inputs, and the output is resized back into a 2D format before proceeding to the next stage. The final output is produced by an MLP. The Convolutional Vision Transformer is specifically designed to capture both spatial features and global relationships across tokens within the image, enhancing its ability to gather comprehensive information for accurate predictions.

3.2. Video CNN

3.2.1 (2+1)D ResNet and ViViT

Video data modeling focuses on extracting meaningful spatial and temporal patterns from video frames. Two prominent architectures, (2+1)D ResNet and ViViT (Video Vision Transformer), were employed to capture these patterns effectively.

The (2+1)D ResNet is an extension of 3D convolutional neural networks that decomposes 3D convolutions into sequential 2D spatial convolutions and 1D temporal convolutions. The architecture outputs a compact spatial-temporal representation that serves as the input to the subsequent layers. This design increases the network’s complexity while maintaining computational efficiency and optimization stability. The detailed architecture consists of:

- Initial Block

A 377 Conv2Plus1D layer with 16 filters to capture spatial and temporal features in early layers. Batch normalization and ReLU activation are applied to stabilize and enhance feature extraction.

- Residual Blocks

Four residual blocks progressively increase the depth of feature maps: 16, 32, 64, and 128 filters with 333 kernel sizes; skip connections incorporated to preserve low-level features and alleviate gradient vanishing issues; spatial downsampling performed using a custom resizing operation after each residual block, reducing spatial resolution by half.

The ViViT block applies self-attention mechanisms across video frames, learning spatiotemporal dependencies beyond the local receptive fields of convolutional layers. The transformer operates on mid-level features extracted from the (2+1)D ResNet, leveraging pre-trained ViViT weights on large-scale video datasets.

3.2.2 Pumping Rate Detection

Based on the goal of multimodal learning, one possible numerical feature to be extracted from the head videos of *C. elegans* was the pumping rate of pharynx. To detect the pumping action, a 1-dimensional attention-based CNN model on optical flow and a 3-dimensional attention-based CNN model on spatial and temporal features were examined and compared.

According to the nature of the videos, data preprocessing for both models involves turning the frames into grayscale, matching the real color of the videos and reducing the unnecessary computation time from loading the videos into RGB frames. The true value of the pharyngeal pumping rate was manually counted to be compared with the output of the two models.

The 1D CNN model employed the functions from the OpenCV library to calculate the flow magnitudes per frame with the following equation:

$$flow\ magnitude = \sqrt{(u^2 + v^2)}$$

where u is the horizontal displacement of pixels and v is the vertical displacement. The array of flow magnitudes was then passed to the function "`find_peaks`" to count the pumping events and 10 seconds was used as the length of videos to calculate the rate. An attempt of adding an attention layer to the flow magnitude was made to improve the focus of the model on the pharynx of C.elegans.

The 3D CNN model processed the video data in the same way as the 1D CNN model, converting the videos into frames of images with a channel value of 1. Since the pharynx was pumping at a roughly constant rate across time, applying temporal attention to the model was redundant and only one spatial attention layer was added between the two 3D convolution layers. The model was directly turned into the evaluation mode to estimate the pumping counts without pre-training the model because the pumping rate was an unknown variable and the collection of such information demanded a significant amount of time and effort.

3.3. Motion Feature EN/RF/MLP

The tabular motion features were fitted through multiple models to predict age and daysRemain. Random Forest (RF) and Elastic Net (EN) are two popular methods to predict age from tabular datasets, such as DNA methylation profiles. Additionally, we want to test the power of a simple Multilayer Perceptron (MLP) in this task for its flexibility and ease of integration into a multimodal learning architecture. Standard RF was utilized with 100 estimators and 42 random states. Similarly the EN was used with 1.0 alpha and 0.5 l1 ratio. The MLP contained 3 layers of size 512, 256, and 128, respectively, with relu activation, batch normalization, and 0.3 dropout at each layer.

3.4. Multimodal Learning

The multimodal learning architecture combined the last layer outputs from the image models (the resnet50 model in this case), as well as the motion feature model by concatenation, and fed it through 2 dense layers of size 64 and 32, respectively.

4. Results

4.1. Image CNN and CvT

The goal of the project is to achieve accurate predictions of the C.elegans' age based on the input image. The performance metric used is the Mean Absolute Error, which quantifies how closely the predicted values align with the actual data. This metric is appropriate as it effectively captures the deviation between predictions and truth labels, providing a clear measure of model accuracy. Experiments involved feeding the data into both the CNN and CvT models. While neither model yielded highly accurate predictions, the CNN consistently demonstrated superior performance but produced lower Mean Absolute Error compared to the CvT model. This was further supported by the plotted results, which showed the CNN model converging toward an ideal fit, whereas the CvT model failed to do so.

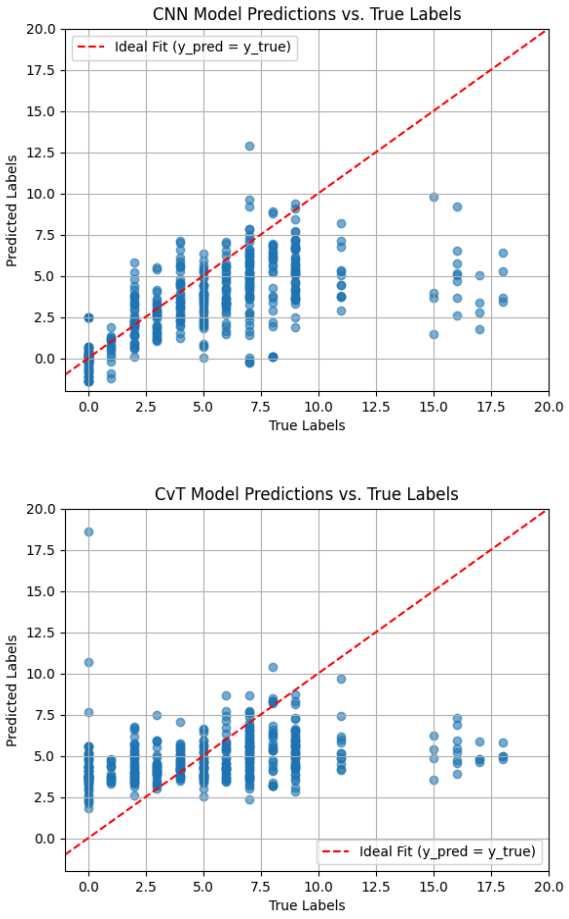

Figure 1: CNN and CvT Model Predictions vs. True Labels

4.2. Video CNN

4.2.1 (2+1)D ResNet and ViViT

We present the classification task results here, as the regression task for video data yielded similarly suboptimal outcomes that need further improvement. The plots reveal that while the model's training and validation loss consistently decrease over 30 epochs, indicating optimization, validation accuracy remains stagnant at approximately 0.58, and training accuracy shows pronounced oscillations. This suggests overfitting, where the model learns the training set effectively but struggles to generalize to unseen data. The flat validation accuracy points to potential issues such as dataset limitations, model complexity, or insufficient overfitting mitigation. Improvements could involve applying stronger regularization, augmenting the dataset, or simplifying the model architecture, as elaborated in Section 6.

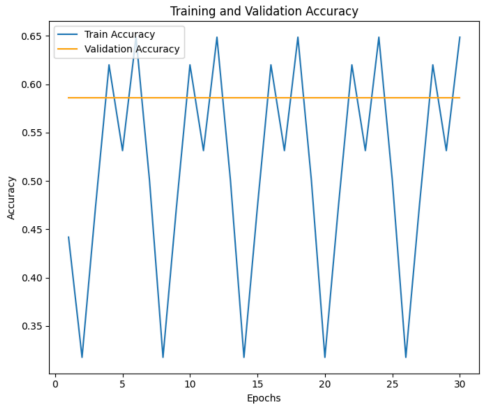

Figure 2: Training and validation accuracy of (2+1)D ResNet model with ViViT for predicting life stages

Figure 3: Training and validation loss of (2+1)D ResNet model with ViViT for predicting life stages

4.2.2 Pumping Rate Detection

For the pumping rate of pharynx extracted from the head videos, the performance of the 1D and 3D CNN models were evaluated and compared with substantially differing results.

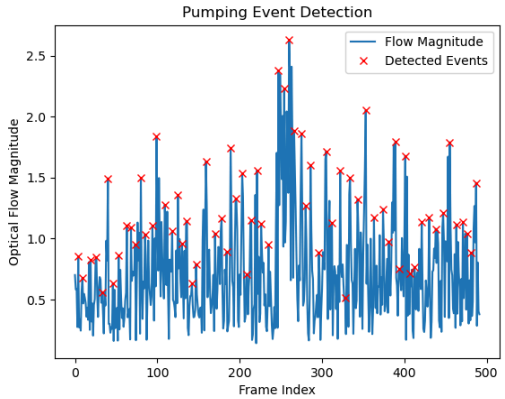

Figure 4: Output of the 1D CNN model on pumping event detection using flow magnitudes.

The pharyngeal pumping rate estimated from the 1D CNN model was 6.7 events per second, much higher than the true rate of 2.2 events/s. Moreover, when the attention was applied, no events were detected by the model. This revealed that adding an attention layer to a sequence of 1D values was not logically rational and made no improvement in the model performance.

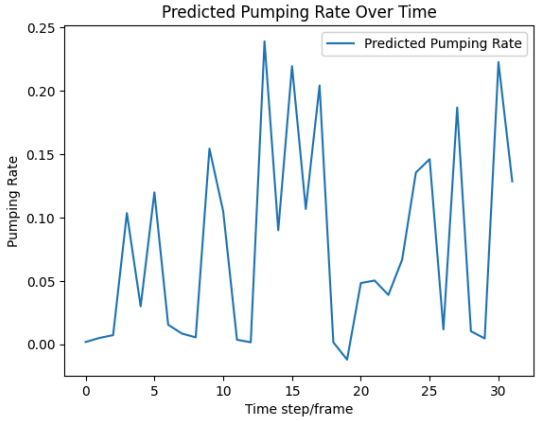

Figure 5: Pumping rate over time estimated from the 3D CNN model on spatial and temporal features.

The 3D CNN model predicted a pharyngeal pumping rate of 0.25 events/s for the same video examined by the 1D CNN model, which also diverted significantly from the true rate of 2.2 events/s. The reliability of the result was further reduced by the randomness of the output, as the model exhibited substantial changes across different random seeds.

Referring back to the head videos, the accuracies and reliabilities of both models were extremely low because the pumping action was not distinct enough compared to the movement of the whole body. The pharynx was a small and mainly transparent organ of C. elegans, preventing the models from distinguishing the motion between the body and the pharynx. Considering the difficulty of manually counting the pumping rate due to the large dataset and incomplete motion capture of the

pharynx in the head videos, excluding the numerical feature of pumping rate from the feature space was deemed a more appropriate choice for multimodal learning.

4.3. Motion Feature EN/RF/MLP

Looking at the visualization of the extracted moving parts from the tierypsy-tracker, the extraction process seemed more challenging and chaotic than we expected. This is potentially due to the high-contrast nature of our recordings making the background more complex than the software was designed to handle. Therefore, it is incredibly surprising to see our model predicting the correct age based on these features. As seen in figure, MLP and RF performed better than EN, potentially due to the high-nonlinearity of our dataset, contrary to most other age-predicting datasets. MLP was the model of choice due to its ease of integration into our multimodal learning.

4.4. Multimodal Learning

Despite both the image-based and motion features-based models had predicted age and/or lifespan, the multimodal learning architecture combining these two datasets failed to make meaningful predictions (fig._). This can be explained by the increase in size of the model, now utilizing 4 images and 3,000 features to predict one output, essentially using 5 times the input. If this is the case, then increasing the sample size of our dataset could prove useful in improving the performance of this model. Alternatively, it may also be possible to borrow weights from individual modalities using transfer learning to overcome this issue.

5. Ethics

Addressing the challenge of predicting the lifespan of *C. elegans* holds significant implications for advancing research in aging and lifespan studies, which are relevant to broader societal concerns, including human health, longevity, and age-related diseases. Deep learning offers highly efficient solutions for processing large datasets, particularly those in the form of images and videos. The

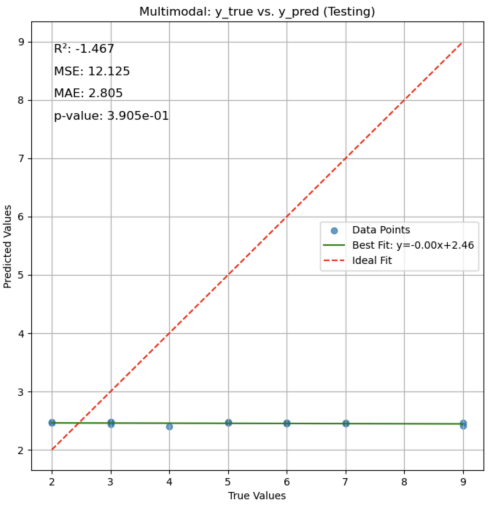

Figure 7: Multimodal training to predict age.

CNN-based models selected by the project are highly effective at automatically learning complex patterns from large image and video datasets, making them ideal for tasks that involve large-scale, unstructured data. Furthermore, using multimodal learning is a logical choice as the primary approach for the project, given that the input dataset includes images, videos, and numerical features derived from the videos.

Focusing on *C. elegans* research, which aligns with the primary goal of the project, the dataset generated for this purpose demonstrates strong representativeness. With frequently recorded growth through images and videos, the dataset is able to capture key biological features and variability relevant to *C. elegans* research, such as different developmental stages and genetic variations, ensuring that the dataset reflects the natural diversity observed in *C. elegans* populations. Additionally, the dataset is structured to mirror real-world experimental settings, making it suitable for developing models that generalize well to other *C. elegans* studies or applications.

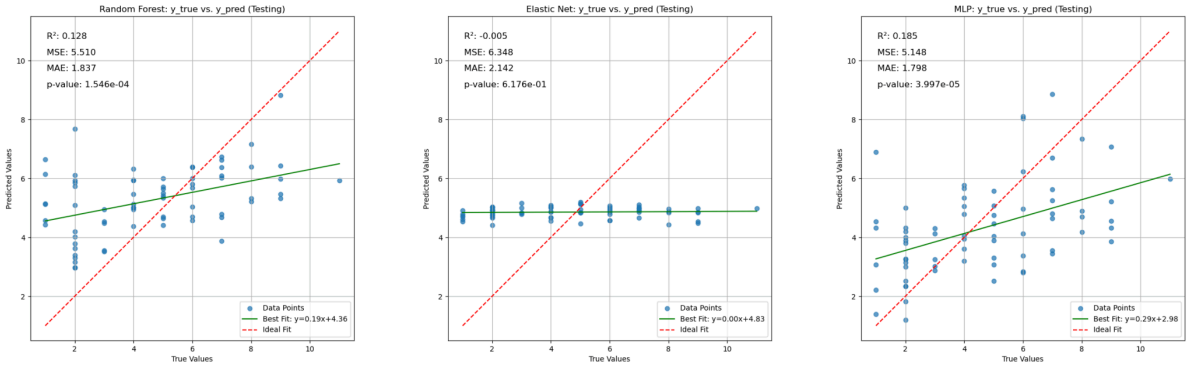

Figure 6: from left to right, RF, EN, and MLP prediction results for age.

6. Reflection

6.1. Image CNN and CvT

In the time scope of this project, the CNN produced better results than the CvT model. However, several factors could have influenced these results.

First, there is a limitation in the dataset. The total dataset contains around 2,000 images, separated into training and testing sets. While this is a reasonable amount of data, it can certainly be improved. This limitation is evident from experiments using a pre-trained ResNet50 model with ImageNet weights, which yielded excellent predictions. However, when the model was not pre-trained, the performance was significantly reduced. This highlights the importance of larger and more diverse datasets for improving model accuracy.

Second, hardware limitations affected the training process. Due to memory constraints, the image size had to be reduced to (128,128) instead of the (224,224) used in the original CvT implementation. Additionally, the batch size was reduced to prevent memory exhaustion. The original CvT documentation recommends training the model for 200 epochs, but this was infeasible due to the extended time required on the available hardware.

Third, the task of age prediction for microscopic images of worms might rely primarily on spatial features rather than global relationships. The self-attention mechanism in CvT, while powerful for many tasks, could introduce irrelevant or misleading information for this specific application, negatively affecting the model's accuracy.

With more time and access to better computational resources, improvements could include pre-training the model on a larger library of similar images and experimenting with parameters such as image size, batch size, number of training epochs, and the embedding size of the model. These enhancements could potentially yield more accurate and reliable results.

6.2. Video CNN

The video-based model encountered challenges that impacted its performance. While the integration of ViViT and (2+1)D ResNet layers showcased promising spatial and temporal feature extraction capabilities, limitations in the dataset size and diversity hindered robust generalization. The training process was constrained by hardware limitations, necessitating the reduction of video resolution and batch sizes, which likely impacted feature accuracy. Additionally, the complexity of ViViT's self-attention mechanism may have introduced unnecessary overhead for the regression task, leading to underperformance relative to baseline models. With access to a larger video dataset, coupled with optimized hyperparameters and computational resources, future iterations of the model could potentially achieve better results. Exploring more lightweight architectures such as I3D to adapt pre-trained vision transformers from 2D images to 3D through weight inflation (Zhang et al., 2022) may also improve performance.

For the 3D CNN model, one possible strategy to improve the performance to predict the pumping rate is to incorporate the training portion into the model. With future effort devoted into obtaining the true values of the pumping rate from the head videos, the model would be able to learn the correlation between the pumping action of pharynx and the labels, improving the accuracy of the model significantly and enriching the feature space of multimodal learning with the features and weights learned from the 3D CNN model.

6.3. Motion Feature EN/RF/MLP

This data modality could be the easiest to train, due to its numerical nature and the existence of many previous numerical data-based age predictors, and also the hardest to train, as a result of the chaotic feature extraction process. We are happy to see both RF and MLP successfully predicted age to certain extent but were surprised that EN failed at this task. EN is arguably the most widely used model to predict age in the age-predicting field. To explain this, we think that the primary data mode in this field, DNA methylation profiles, could be a lot more linear than our motion-based dataset. This could result in the elimination of most features by the EN model and an ineffective prediction. On the other hand, RF and MLP provide a lot more flexibility for non-linear datasets.

6.4. Multimodal Learning

Although the multimodal training did yield the most optimal results, it can guide us to make proper adjustments. Compared to training with individual modalities, this multimodal architecture includes 5 times as much training inputs, but the same amount of output labels as the motion features MLP, or 4 times less the amount of output labels as the image-based model. This would make this model incredibly difficult to train. One solution to circumvent this issue would be to use transfer learning to prime the model with individual modality training. Another method could be to augment our input datasets and virtually increase the sample size, although coordinating the augmentation of different data modality could be a tricky task. Finally, the safest way to improve our model is to collect more data, which is definitely feasible should interests align, but not within the time-frame of this project.

7. Division of Labor

Jien Li: Data collection, motion feature-based models, and multimodal learning.

Yiming Yu: Image CNN and Convolutional Vision Transformer (CvT)

Yifan Lu: (2+1)D ResNet and ViViT on head videos

Jingyi Lu: 3D CNN on head videos for detecting pumping rate

8. Code Availability

https://github.com/jienli/Wooooooooorm.git

References

[1] Song, Y., Liu, J., Yin, Y., & Tang, J. (2022). Estimation of Caenorhabditis Elegans Lifespan Stages Using a Dual-Path Network Combining Biomarkers and Physiological Changes. Bioengineering (Basel, Switzerland), 9(11), 689.

[2] J. -L. Lin, W. -L. Kuo, Y. -H. Huang, T. -L. Jong, A. -L. Hsu and W. -H. Hsu, "Using Convolutional Neural Networks to Measure the Physiological Age of Caenorhabditis elegans," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2724-2732, 1 Nov.-Dec. 2021

[3] Martineau, line N., Brown, E. X., & Laurent, P. (2020). Multidimensional phenotyping predicts lifespan and quantifies health in Caenorhabditis elegans. ProQuest, e1008002.

[4] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). *CvT: Introducing Convolutions to Vision Transformers.*

[5] Zhang, Y., Huang, S.-C., Zhou, Z., Lungren, M. P., & Yeung, S. (2022). Adapting Pre-trained Vision Transformers from 2D to 3D through Weight Inflation Improves Medical Image Segmentation. *ML4H 2022*. https://doi.org/10.48550/arXiv.2302.04303