Pain-State Classification in Rodent fMRI Using a Vision Transformer

Santiago López Campos
Tecnológico de Monterrey
Av. Gral Ramón Corona No 2514, Colonia Nuevo México, 45201 Zapopan, Jal.

***Abstract*—For this paper, we wanted to see if using less data and more epochs and the vision transformer architecture, we could obtain higher accuracy results compared to the previous study that used 3D VGG16 *from Macías Padilla et al.* [1]. Our vision transformer was the ViT-B/16. In order to complete this study, we decided to only test on the pain timestamps. This included, week one, week7, male vs female tests. In other words, we managed to have 4 test scenarios to compare with the previous study. For two of our tests, we managed to obtain similar results compared to the previous study, and for the other two we got lower scores. We believe that this is due to the fact that transformers rely heavily on data, in those two scenarios we had one rat less. Furthermore, we were able to get a lower standard deviation in three of the four scenarios. Given this, we can conclude that the use of a vision transformer can give us great results, subject to the amount of data.**

I. Introduction

Chronic pain is different to normal pain in which it lasts for a long time, usually more than three months [2]. Over time, adjacent problems can occur such as physical and mental fatigue, sleep problems, anxiety, depression and many other problems. It has no cure, and can lead to many financial burdens due to continuous treatments. Similarly, women are well more affected than men, which is why it's important to understand why and where that pain occurs. An experiment in which this phenomenon was shown, was the one where there was induced orofacial pain, which then stress was applied which resulted in observing visceral pain.

A similar species to the human is the rodent, they function similarly when it comes to brain function due to the fact that both are mammals. Using fMRI, we can conduct the same experiment as mentioned before in which we can obtain fMRI data in order to analyze and comprehend the origins of pain in the brain. Male and Women rodents are used in order to comprehend why one is affected more than the other. In this experiment, two different times were used, week one with pain and week seven with pain. From week one, it's shown the amount of pain women suffer as to men, and from week one to week seven it is shown how pain goes away from men but stays for females.

Neural networks have shown to be able to classify between different timepoints and different sexes when it comes to analyzing fMRI data. Previous works have used architectures such as 3D-VGG16, such as the study by Macías Padilla [1], in which a standard VGG16 was inflated to 3D. However, CNN's are able to understand local patterns but not comprehend patterns happening all around.

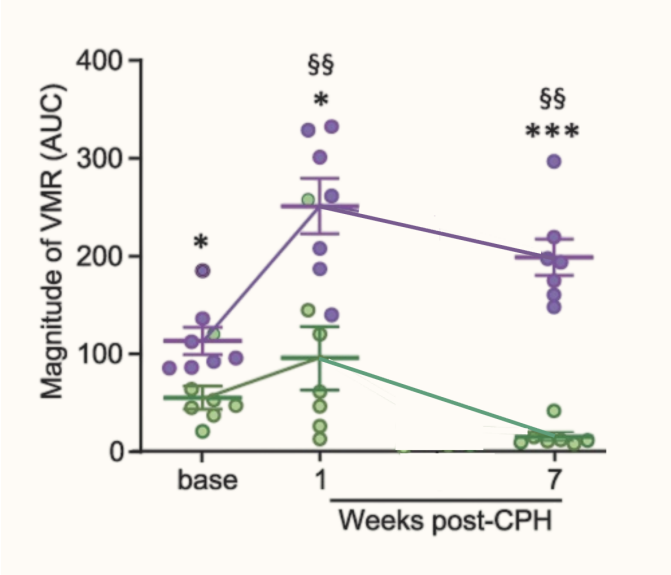


Fig. 1. Magnitude of VMR (AUC) across weeks post-CPH in male and female rats. Reproduced from Thesis Defense: Neuroimaging-Based Pain Detector Using Artificial Intelligence Approaches, Brayhan Alan Macías Padilla (with permission).

In recent times, new architectures have emerged such as the vision transformers which instead of using convolutions to analyze patterns, uses self attention which makes it possible to understand patterns throughout the full space. Due to its recent appearance, there are not too many studies with regards to medical imaging. Unlike the 3D-VGG16, slices of the 3D data are used here as data. In this research, less data is used than the previous study due to the fact that the vision transformer is already pre-trained on about 15 million images. Still, 50 epochs are used due to the fact that overfitting is less likely to occur.

Given this, the purpose of this research is to analyze whether the architecture of ViT-B/16 can give us better values than other architectures which do not use transformers. While there is more work to be done in order to understand pain fully, understanding which architecture is better for the job is a great leap forward.

II. Methodology

The total number of rodents for this experiment were 25, 12 male and 13 female. For the three timepoints of testing, the

testing sessions resulted in 620 timestamps for fMRI volumes. In the previous study, they used 570 out of the 620 for baseline measurements, 135 for both week one and seven due to the fact that the time in pain was smaller. However, in the previous study, those 135 were turned into 570 with data augmentation. In this case, only 135 data points were used. For week one, due to errors in pre-processing, one rat for the males was not able to be used. Given this, the experiments had to be tweaked in order to accommodate this.

III. Model Architecture

For this research, the architecture of the vision transformers was used. More specifically, the ViT-B/16 model was used. This model used 16 x 16 patches which were obtained by dividing the images used as data into 14 patches. This said, the images that were used had a size of 224 x 224. The number of transformers that the architecture had were 12, the same number of attention heads. Other than the 14 tokens, an additional token is used in order to grab the overall idea, which later on is passed to an mlp head. The size embeddings in this case were of 768. The model was trained on ImagenetNet-21k that consisted of around 14 million images with an additional 1.28 million images from ImageNet-1k to fine tune the model. The original ViT-B/16 was pre-trained with 300 million images. While not being able to use it, using 15 million images still gives us a good starting point.

IV. Training Configuration

To train the model, we changed the 3D input to 2D image slice using the axial dimension. The original input had size 42, 65, and 29. Given this, we had images of 42 by 65, totaling a total of 3915 images for each rat. In order to reach the dimensions of 224 x 224, we had to scale up the image. We used 32 batches in accordance with the previous study. Given that vision transformers need more data, 50 epochs were used unlike the previous study that only used between 10 and 15. Since it is harder or more expensive to train, overfitting did not seem to be a problem. Adam was used for backprop with a learning rate of 3e-5. The loss function that was used was sparse_categorical_crossentropy. Every layer in the architecture was trained on, the last layer was changed from having 1000 categories to two.

V. Evaluation Metrics

For this study, we decided to focus on the accuracy and on the standard deviation of such results. While the original study used full 3D volumes as data, in this study only 2D slices were used. Similarly, we would use three rodents from each category for the test set. The reason for this was to have a balanced test set.

VI. Results

The four scenarios can be observed in Fig. 2, W1 M vs F (W1), W7 M vs F (W7), W7 F vs W1 F and W7 M vs W1M. Other scenarios such as W1 male vs W7 female were not tested since they were not tested in the previous study.

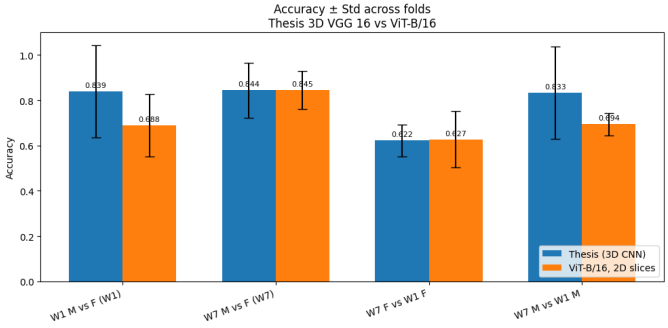


Fig. 2. Test scenario comparison between 3D VGG16 and ViT-B/16.

TABLE I
Accuracy and Standard Deviation for 3D VGG16 and ViT-B/16 Across Test Scenarios

Scenario	3D VGG16	ViT-B/16
W1 M vs F (W1)	0.839 ± 0.204	0.688 ± 0.138
W7 M vs F (W7)	0.844 ± 0.122	0.845 ± 0.085
W7 F vs W1 F	0.622 ± 0.070	0.627 ± 0.124
W7 M vs W1 M	0.833 ± 0.204	0.694 ± 0.049

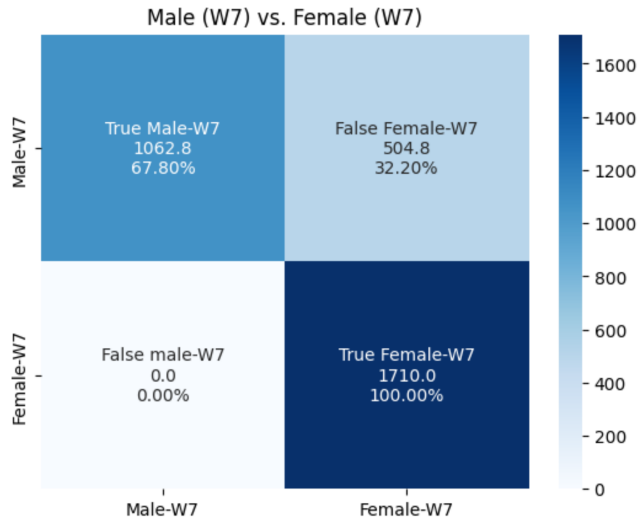
From the graph, we can observe that two scenarios resulted in similar or better scores. Those two scenarios were W7 M vs F (W7) and W7 F vs W1 F. The other two scenarios, W1 M vs F(W1) and W7 M vs W1 M had way lower scores. As seen from Fig. 2 and Table I, we can observe that the standard deviation is smaller for three out of the four scenarios.

As seen from Fig. 3, there are a couple differences between both architectures, for the 3D VGG16, the model managed to get a 0% when getting false male, while the ViT-B/16 got 7.64% wrong. When it comes to classifying between males, the ViT-B/16 outperformed the 3D VGG16 by obtaining 75.58% compared to 67.80%. However, when classifying females, the 3D VGG 16 surpassed the ViT-B/16 by obtaining a 100% compared to 92.36%. From Fig. 4, we can observe that the ViT-B/16 was better at predicting W1 as compared to W7. On the other hand, the 3D VGG16 was better at predicting W7 compared to W1.

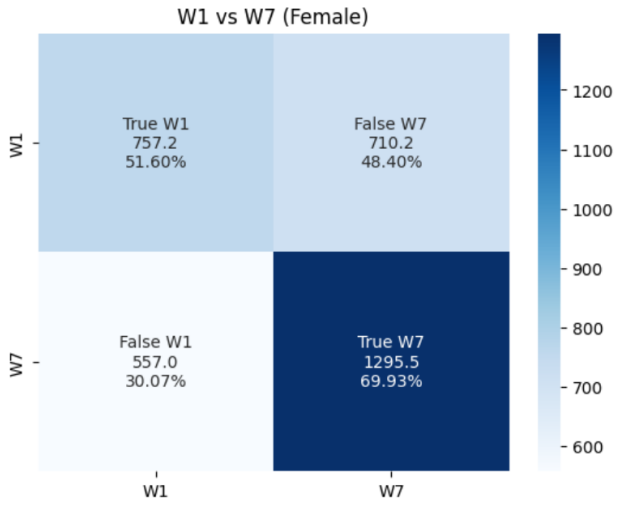
Lastly, as seen from both Fig. 6 and Fig. 7, both were the worst scenarios for the vision transformers. For male W1 vs female W1, the model focused on classifying 35.07% of males as females. In W1 vs W7 male, the model incorrectly categorized 27.27% male as female. The model also incorrectly classified 9.09% female as males. In the other hand, for the 3D VGG16 architecture, it had higher false positives and false negatives which resembles the standard deviations.

VII. Discussion

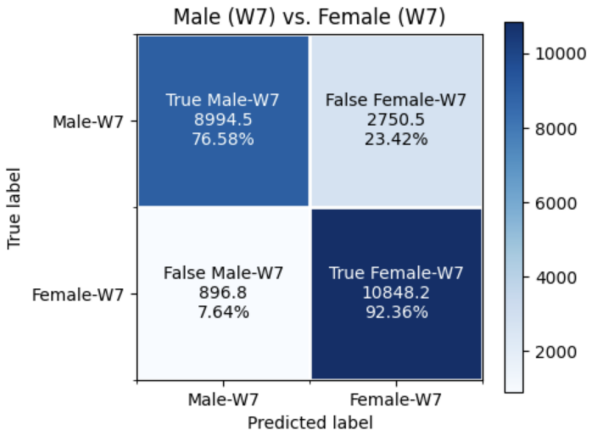
The hypothesis for getting lower scores is that since there was a rat less than the other experiments, due to the hungry nature of transformers, the model performed worse. A single loss in image might not impact a CNN much, but it will impact a vision transformer. Given that W1 F is similar to W7 Female, as shown in Fig. 1, it would be understandable that our transformer would need more data to distinguish better,



(c) Male (W7) vs. Female (W7)



(a) Week 1 vs. Week 7 (Female)



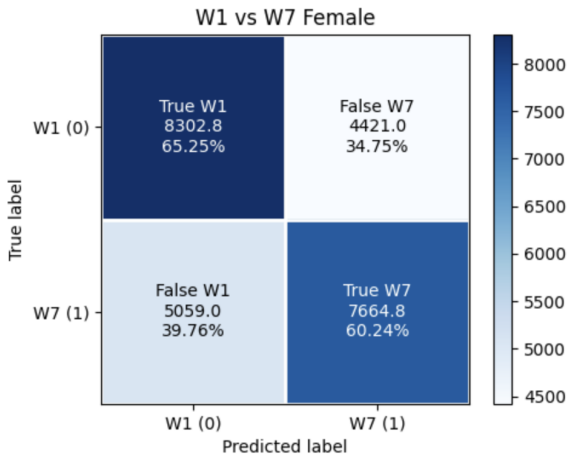


Fig. 3. Confusion matrix comparison between 3D VGG16 and ViT-B/16 for W7 male vs W7 female.

Fig. 4. Confusion matrix comparison between 3D VGG16 and ViT-B/16 for W7 female vs W1 female.

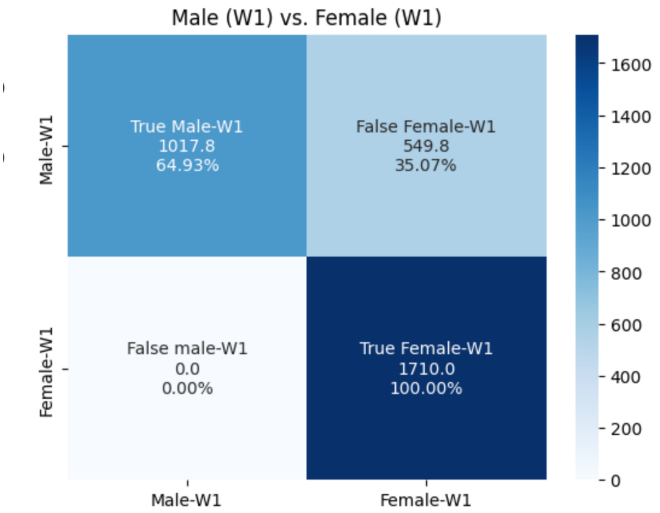
as those two categories are the most alike. However, when looking at model vs model, with just that data, the CNN captures a better response when it comes to similar inputs.The results from W1 male vs W1 female and W7 Male vs W1 Male reflect the lower standard deviation values. In these scenarios, the ViT showed more imbalanced predictions, while the 3D VGG16 showed higher false positives and false negatives.

VIII. Conclusion

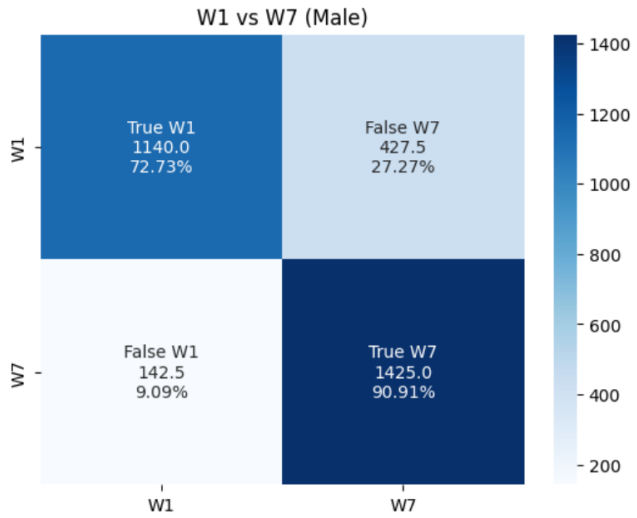
Different architectures can bring in different outcomes when it comes to analyzing a problem. While CNN's have been the state of the art, the use of new architectures can bring insight into new findings. Whether values are higher or lower, advancement in the understanding of pain is crucial for humanity in the l ong run. Of course, different architectures have different advantages and disadvantages. In this case, the vision transformer architecture, or more specific, the ViT-B/16 was used to see if it has any advantage over the architecture 3D VGG16 when it comes to detecting pain in different rodent experiment scenarios. As mentioned before, while the model

	Vit-B16	3D VGG-16
W1 M vs F W1	0.688 +- 0.138	0.839 +- 0.204
W7 M vs F W7	0.845 +- 0.085	0.844 +- 0.122
W7 F vs F W1	0.627 +- 0.124	0.622 +- 0.070
W7 M vs M W1	0.694 +- 0.049	0.833 +- 0.204

Fig. 5. Accuracy and standard deviation comparison between ViT-B/16 and 3D VGG16.



(b) Male (W1) vs. Female (W1)



(b) Week 1 vs. Week 7 (Male)

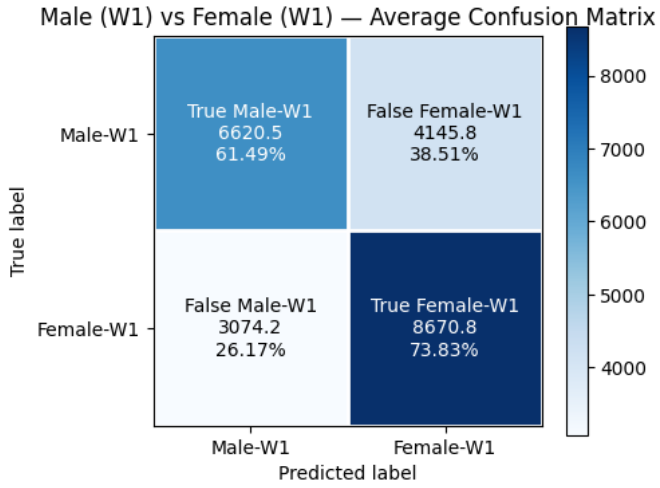


Fig. 6. Confusion matrix comparison between 3D VGG16 and ViT-B/16 for W1 male vs W1 female.

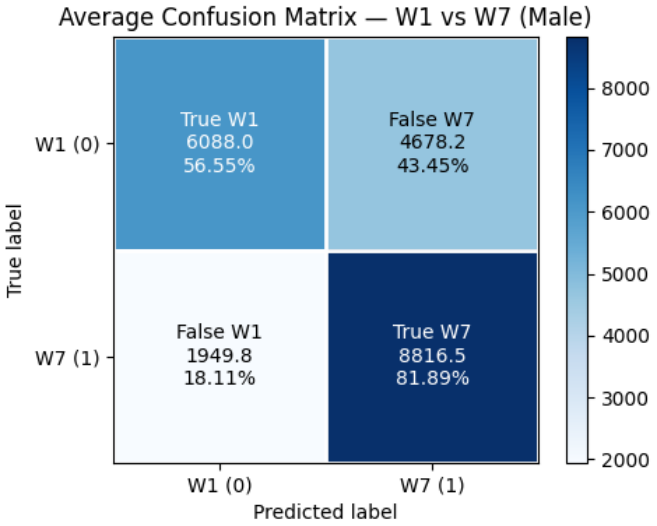


Fig. 7. Confusion matrix comparison between 3D VGG16 and ViT-B/16 for W1 male vs W7 male.

did not use the famous 300 million images for training, it did use 15 million images for pre-training which set us up in a well spot for fine-tuning.

Four scenarios were used to compare against 3D-VGG16. Out of those four scenarios, two achieved a similar result, and two achieved a lower score compared to the 3D VGG16. For the two that got a lower score, there was one less subject compared to the other scenarios. While it might not seem much, for a vision transformer it can be a gigantic hit. Since CNN's learn early on to detect patterns, having one less rodent means that a vision transformer will lose patterns that it could have learned. For this same reason, the original or talked about vision transformer is trained on 300 million images. Furthermore, out of the four scenarios, three got a lower standard deviation. If more data were to be gathered, then for those three results we have a better idea of what values to expect. For the scenario in which a higher standard deviation was obtained, it involved female in week one and female in week seven. Those two weeks had similar patterns which with the data available, it made the vision transformer struggle more.

Vision transformers can definitely be used to analyze different scenarios such as pain. The results obtained give us insight into what is possible if same conditions are applied as other experiments with other architectures. One important thing to note, is that while having a different type of input, some context was lost due to having one dimension less, so comparisons between the 3D VGG16 and the vision transformer are not one to one. Still, the values obtained give us information on our type of experiment, in this case analyzing images in the axial space.

For the future, different changes can be applied and should. One of those changes is the use of data augmentation, as already mentioned, only 135 volumes were used per rat to train the model. In order to get a closer comparison with the 3D VGG16, the volumes should be augmented to 570. Different perspectives should be taken into account, for example the use of coronal slices or sagittal. Another option would be to take all three as one and test on all of them. In addition, the use of the 3D input should be taken into considerations in which an adaptation of the vision transformers is used.

REFERENCES

[1] B. A. Macías Padilla, *3D VGG16-Based fMRI Classification for Pain Detection: A Deep Learning Approach with Explainable AI*, Master's thesis, Tecnológico de Monterrey, 2024.

[2] Harvard Health Publishing, "Stopping pain before it turns chronic," Harvard Medical School, 2021. [Online]. Available: https://www.health.harvard.edu/pain/stopping-pain-before-it-turns-chronic