## *Abstract*

***Segmentation techniques can be utilized to help solve everyday life problems. In the medical field itself, semantic segmentation techniques can be applied to help detect the spread of cancer in cancer patients early. In the field of semantic segmentation, Convolutional Neural Network (CNN)-based approaches, such as Fully Convolutional Network (FCN) and DeepLabv3+, have been known to dominate the field. In addition, the success of the Transformer approach in the field of natural language processing has attracted many researchers' attention to utilize the approach in solving semantic segmentation problems. This triggered the development of Vision Transformer (ViT) as a new alternative in solving semantic segmentation. Unlike the convolution approach that uses a shifted kernel to obtain local contextual information, the ViT approach accepts the image as a patch that can later be used to generate local as well as global contextual information. One model inspired by the ViT architecture, SegFormer, combines a Hierarchical Transformer Encoder to generate fine features at low resolution and coarse features at high resolution and a lightweight All-MLP decoder to combine the multi-level features generated from the encoder to create the final segmentation mask. In this study, the SegFormer model was used to segment whole-body bone scan images. As a result, by comparing the SegFormer model with several convolution models namely FCN and DeepLabv3+, the performance of SegFormer successfully beats both convolution models with the highest mIoU value achieved of 77.86% on the test data.***

***Keywords: bone scan image, semantic segmentation, Transformer, convolution, SegFormer, DeepLabv3+, FCN***