

Response to Reviewer 2 Comments

Based on the first round review of the manuscript entitled A novel fault diagnosis method of rolling bearing based on Integrated Vision Transformer model, the reviewer has the following comments:

Point 1: Please explain about Figure 1 in more details. Please explain all blocks deeply here.

Response 1: The explanation of all blocks is revised in the manuscript.

Point 2: In Figure 1, if ViT is included sub-blocks, please explain here.

Response 2: The sub-blocks of ViT is explained in the manuscript.

Point 3: What is the advantages of ViT model compare to other AI or classical signal/system modeling techniques?

Response 3: The ViT is the improved version of the transformer model, which removes the decoder block of the transformer and has higher computation efficiency and less memory usages. In addition it not only inherit the multi head attention mechanism and relative positional embedding method of transformer, but also adopt the parallel structure, it is prone to capture the global and sensitive fault related information from the original vibration signal. So it has higher diagnosis accuracy and efficiency compared with CNN and RNN

Point 4:How you can validate the stability and robustness of the proposed diagnosis algorithm?

Response 4: Two test methods are used to validate the stability and robustness of the proposed integrated ViT method, which are described as follows: 1) The five diagnosis test are implemented on the same dataset, the mean and minimum and maximum value of diagnosis accuracy of the integrated ViT model are 99.87%, 99.47% and 100% respectively, but their difference among them is very small. This can demonstrate that the proposed method is stable. 2) the fault diagnosis analysis of the integrated ViT diagnosis model are conducted on three different datasets under three different working conditions (0hp load and 1797rpm, 1 HP load and 1772 rpm and 2 HP load and 1750 rpm), the diagnosis accuracy of the integrated ViT on the three datasets is 100%, 99.67% and 99.83% respectively, the maximum difference among these three diagnosis accuracy is 0.33%, and the minimum difference among them is 0.17%.This can validate the robustness of the integrated ViT.

Point 5: In Figure 8, do the DWT and TFR for training and testing have the same design (coefficients and)

Response 5: No, they are different. The DWT has only 4 level, the CWT scale sequence take the following values $c/\text{totalscal}, \dots, c/(\text{totalscal}-1), c/4, c/2, c$. And the basis wavelet functions of DWT and TFR are also different.

Point 6: .Can the proposed method work for acoustic emission (AE) signals as well? (In this manuscript the vibration signal is used to validate the algorithm)

Response 6: It is well known that the acoustic is produced by vibration, so the proposed integrated ViT can work for AE signal as well.

Point 7:What is the limitation of the proposed method? please mention in the conclusion and based on that please mention about the future work.

Response 7: This has been revised in the manuscript.