**Response to Reviewer 1**

**Comment**:

Basic reporting

The revised manuscript has undergone significant improvements, with a stronger emphasis on scientific evidence to support the study's motivation. The authors addressed all of the major concerns raised about their work, resulting in a more comprehensible version to readers. I suppose that the current version is now suitable for publications with minor modifications:

**Response**:

Thank you for your supportive comments. We have significantly revised our manuscript to correct all the mentioned errors.

**Comment**:

- Line 98: "…. For a feature-target pair (x, y) where x ...": "(x, y)" and "x" need to beitalicised like other mathematical signs, variables, and operators.

**Response**:

attention layer and a position-wise feed-forward layer. For a feature-target pair $(x, y)$ where $x \equiv \{x_{categorical}, x_{continuous}\}$, $x_{categorical}$ and $x_{continuous}$ represent for all categorical features and continuous

**Comment**:

- Line 128: "... In our modeling experiment ..." -> In our modeling "experiments"(plural)

**Response**:

where $y$ is the actual label and $\hat{y}$ is the predicted probability. In our modeling experiments, we designed

**Comment**:

- Line 130-131: "... One epoch required around 1.2 seconds to train and 0.2 secondsfor testing." -> Your sentence should written in a parallel structure. "to train" -> "for training".

**Response**:

Intel i7-12700 CPU with 64GB RAM and an NVIDIA GeForce RTX 3090 Ti GPU. One epoch required around 1.2 seconds for training and 0.2 seconds for testing.

**Comment**:

- Line 147: ... epoch 27. -> "... epoch 27th"

**Response**:

The models' validation loss converged around epoch $27^{th}$. Both training and validation loss continues

**Comment**:

- Line 153-154: "... than the other setup models (b), and (c)." -> "... than the models of setups (b) and (c)."

**Response**:

with three additional epochs. The model of setup (*a*) shows better performance than the models of setups (*b*), and (c). The variations in AUCROC values, however, are not significantly different. Models of

**Comment**:

- Line 175: "... value of 0.38, whereas other methods obtains an AUCPR value" ->value of 0.38, whereas other methods "obtain" (fixed verb) AUCPR "values" (plural)

**Response**:

value of 0.38, whereas other methods obtain AUCPR values of at most 0.37. Figure 3 visualizes the areas under the curves of all the models.

**Comment**:

- Line 178-179: "Table 3 gives information on the performance of all models over multiple trials." -> "Table 3 gives information on the performance of all models over "ten" (concrete number) trials.

**Response**:

to avoid sampling bias. Table 3 gives information on the performance of all models over ten trials. The

**Comment**:

- Line 186: "The p-values" -> The p-values (italicised "p")

**Response**:

each machine learning model (Table 4). The *p*-values of these pairwise comparisons between our model and the other models confirm the statistical significance of these results.

**Comment**:

- Line 190, 192, 197, 204: “Transformers” -> “Transformer-based models”

**Response**:

makes it challenging for Transformer-based models to grasp the semantics and relationships within the data, leading to suboptimal performance. Data sparsity is also a concern, as infrequent or absent words and patterns can impede the learning process. Finally, the high capacity of Transformer-based models may be underutilized with small datasets, limiting their ability to capture complex relationships. Mitigation strategies include transfer learning, data augmentation, regularization techniques, and domain adaptation. These approaches can partially address the limitations, but it is important to acknowledge the inherent challenges of training large-scale models with small datasets. Transfer learning allows leveraging knowledge from related tasks or domains, while data augmentation increases training data diversity. Regularization techniques prevent overfitting and improve generalization, and domain adaptation aligns representations for better adaptation to new domains. These strategies enhance the Transformer-based model’s performance, generalization, and adaptability.

**Limitations**

Despite good outcomes, our model still has limitations that need to be improved in the future. Like other models in the Transformer family, our model requires high computational cost compared to other deep learning architectures. Besides, longer training duration and limited parallelization are also common issues of Transformer-based models. On the other hand, parameter tuning in a Transformer-based model is highly sensitive to create the optimal models.

**Comment**:

**Experimental design**

The experiments were well-designed to achieve the study’s objectives.

**Validity of the findings**

The newly added statistics provide more insights into the model's robustness and applicability.

**Additional comments**

I have no additional comments for this article.

**Response**:

Thank you for your supportive comments.

**Response to Reviewer 2**

**Comment**:

**Basic reporting**

I appreciate the authors' efforts in adding more details and conducting additional experiments to improve the quality of the manuscript. The manuscript is now well structured and understandable to readers. I recommend that this work be considered for publication once the authors have completed correcting several minor points.

**Response**:

Thank you for your supportive comments. We have significantly revised our manuscript to correct all the mentioned errors.

**Comment**:

(1) Figure 2. It is recommended to move the legend of Figure B to the top-right position to avoid overlaying text.

(2) Figure 2. The axis names in Figure B are wrong. It should be "Precision" and"Recall" instead of "TPR" and "FPR".

**Response**:

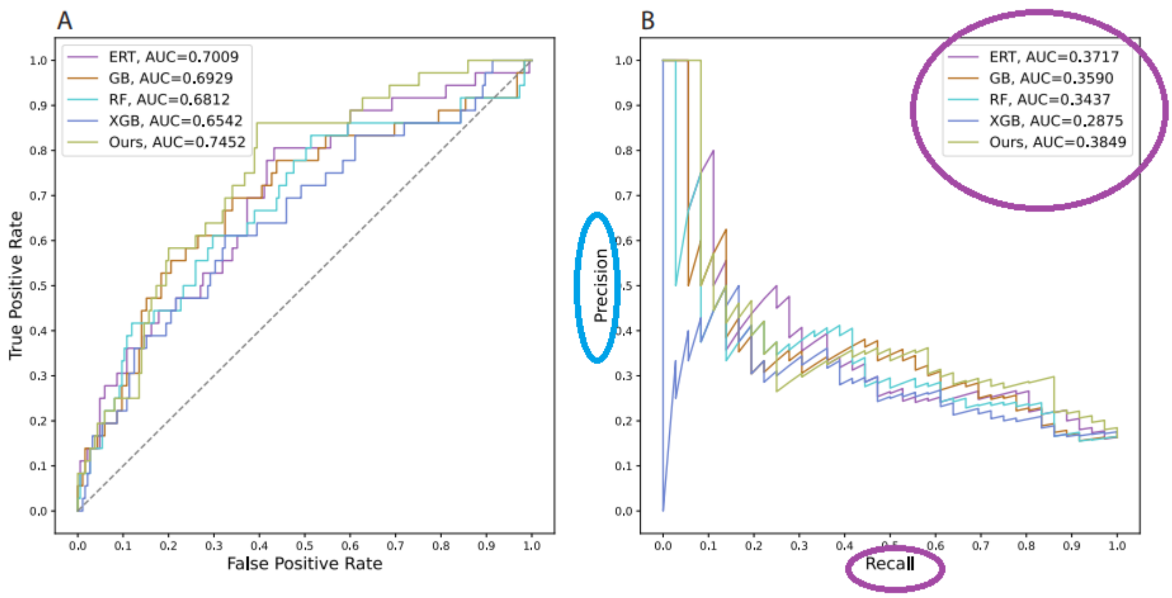


**Figure 3.** Areas under the curves of all the models (A. Receiver operating characteristic curves, B. Precision-recall curves).

**Comment**:

(3) The limitations of the method should be discussed.

**Response**:

### Limitations

Despite good outcomes, our model still has limitations that need to be improved in the future. Like other models in the Transformer family, our model requires high computational cost compared to other deep learning architectures. Besides, longer training duration and limited parallelization are also common issues of Transformer-based models. On the other hand, parameter tuning in a Transformer-based model is highly sensitive to create the optimal models.

7/9

**Comment**:

(4) In the statistical analysis section, which threshold did you choose? (0.05, 0.01,etc.). Is it the one-tail or two-tail test?

**Response**:

the GB and XGB models. Also, to assess the statistical significance of the results, we used two-tailed independent $t$-tests with a confidence interval of 0.95 to compare the performance of our model to that of each machine learning model (Table 4). The $p$-values of these pairwise comparisons between our model

**Comment**:

**Additional comments**

- Line 75: "All selected algorithms for" should be read "All algorithms selected for".

- Line 185: "compare the performance of our model to each machine learning model" should be read "compare the performance of our model to that of each machine learning model".

**Response**:

We have fixed those errors.