MULTI-LAYER ATTENTION-BASED EXPLAINABILITY VIA TRANSFORMERS FOR TABULAR DATA: SUPPLEMENTARY MATERIAL

I. DATASETS

The proposed explainability model is tested on three datasets: UCI Forest CoverType [44], KDD'99 Network Intrusion dataset [45], and a real-world proprietary dataset as described below. These where selected due to their relatively large number of features (with a fair mix of numerical and categorical) and samples, allowing adequate conceptual aggregation.

Forest CoverType Dataset (CT)

In CT, the goal is to predict the most common cover type for each 30m by 30m patch of forest. We use the three most represented classes, resulting in approximately 425,000 training and 53,000 validation samples. The dataset consists of 10 quantitative features and two qualitative features, which were organized into the following five concept groups:

- *Generals*: Elevation, aspect, and slope of the patch
- *Distances*: Horizontal and vertical distances to hydrology[4], horizontal distances to roadways and fire points
- *Hillshades*: Shades at 9am, noon, and 3pm
- *Wild areas*: 4 different wilderness areas
- *Soil types*: 40 different types of soil

Network Intrusion Dataset (NI)

In NI, the classification task is to distinguishing between "bad" connections (intrusions or attacks) and "good" connections. Approximately 1,000,000 samples were used for training and almost 75,000 for validation, with each sample consisting of 53 features. The concept groups are defined following [46]:

- *Basic*: 20 features regarding individual TCP connections
- *Content*: 14 features regarding the connection suggested by domain knowledge
- *Traffic*: 9 features computed using a two-second time window
- *Host*: 10 features designed to assess attacks which last for more than two seconds

Real-World Dataset (RW)

The proprietary real-world dataset constitutes a binary classification problem. Tens of thousands of samples were used for training and validation. Each sample has approximately 100 features, which were subsequently arranged into 8 concept groups. Our access to this dataset has been granted for a restricted duration, resulting in its exclusion from certain experiments.

II. IMPLEMENTATION HYPERPARAMETERS

The experiments were implemented in Python and ran using GeForce RTX 2080 Ti GPU and Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz for all datasets except RW, for which Tesla V100 GPU and Intel Xeon CPU E5-2697 v4 @2.30Hz were used.

The same hyperparameters were used for all teacher networks: $N = 2$, $h = 4$, $d = 64$ and 128 neurons in the internal layer. These parameters are standard choices for transformer encoders for TD; on the lower end for N and h, and on the higher end for d and neurons. The student's architecture is identical, but with $M = 4$ and $h = 1$. For training, we chose a dropout rate of 0.1 to prevent overfitting while avoiding a large reduction of network's capacity. Additionally, we used a temperature of 2, which provided a balance between producing reliable soft targets and avoiding to overly flatten the underlying probability distribution. A constant batch size of 128 and the Adam [51] optimizer were employed. Between six and ten lambdas were tested for each dataset's training loss. The lambda corresponding to the highest F1 (for CT and NI) and accuracy (for RW) was selected for the final results, yielding $\lambda_{CT} = 0.005$, $\lambda_{NI} = 0.01$, and $\lambda_{RW} = 0.9$.

[4]The dataset information file states that this refers to the "nearest surface water features."

III. Best Context Group Distributions by Sample Classification Output per Dataset

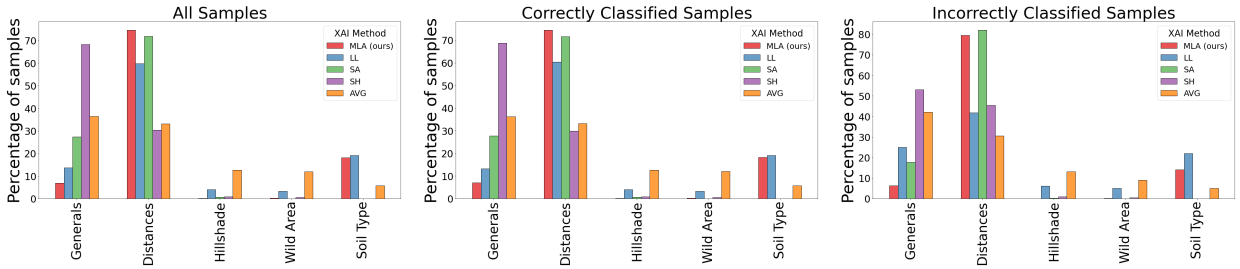


Fig. 1: CT

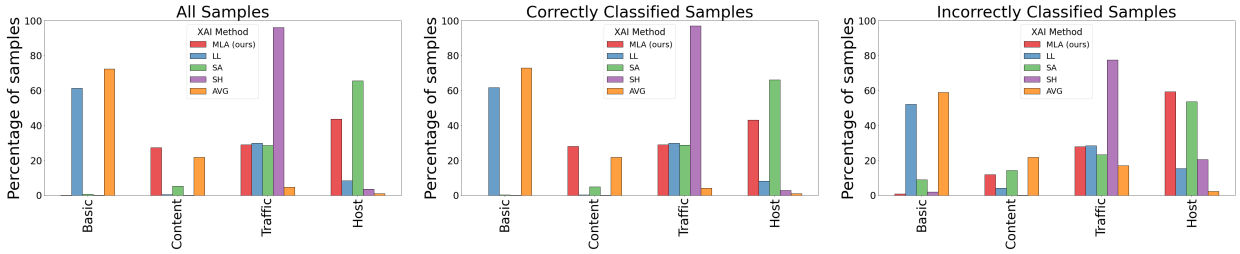
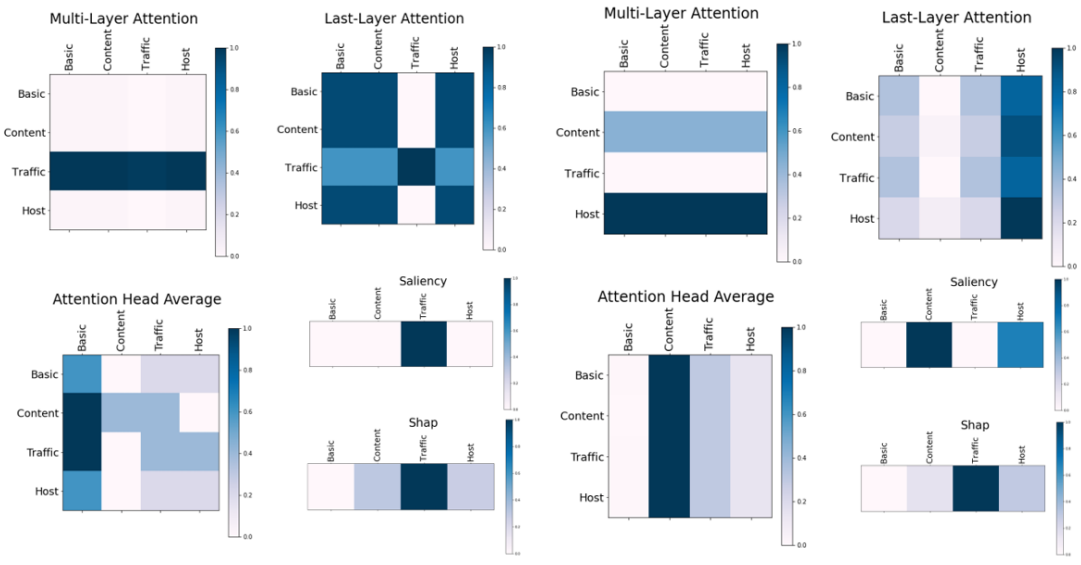


Fig. 2: NI

IV. Network Intrusion - Concept groups explainability coefficients

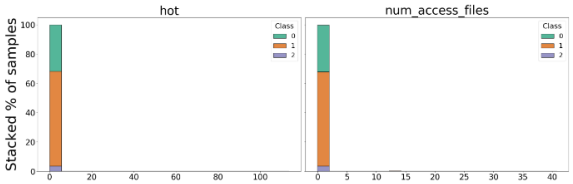
In Fig. 3a, we observe that features related to *Traffic* were given larger values by all methods. In contrast, in Fig. 3b there is a lack of unanimous agreement across methods, but with alignment shown by the attention-based methods.



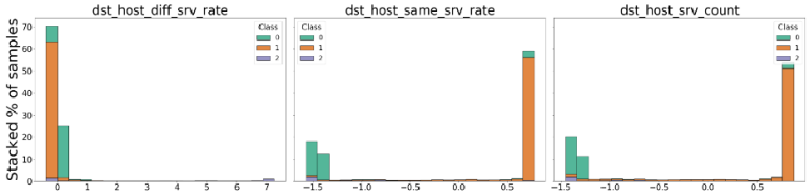
(a) Sample 21259, Class 0

(b) Sample 13927, Class 1

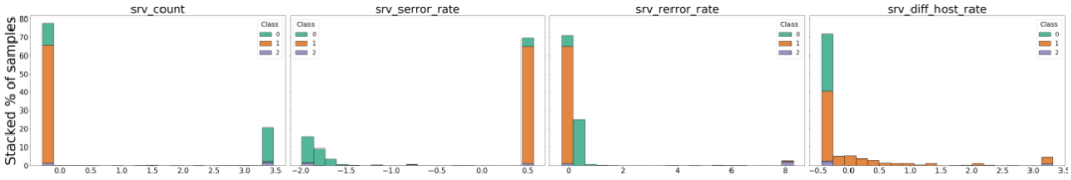
Fig. 3: NI Concept groups explainability coefficients.



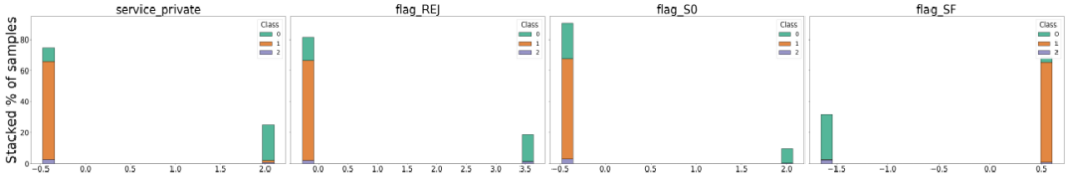
(a) Density of Content features



(b) Density of Host features



(c) Density of Traffic features

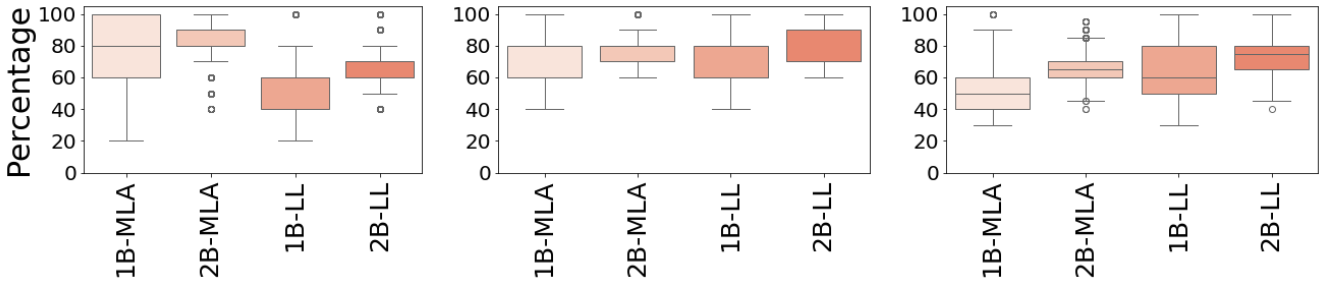


(d) Density of Basic features

Fig. 4: NI Exploratory Data Analysis

V. Stability Analysis

The stability of the explanations is analyzed by quantifying the percentage of distinct runs[5] that agree on the same explanation for each sample. Given the previously discussed observation that SA and SH tend to steadily choose the same groups even across different samples, we focus on methods MLA and LL for this analysis. Fig. 5 shows the boxplots for the best (1B) and two best (2B) *concept groups* per dataset.



(a) CT (b) NI (c) AS

Fig. 5: Percentage of runs that agree on the best (1B) and two best (2B) *concept groups* per method

For these datasets, the 1B *concept groups* comparison of MLA and LL appears inconclusive. For CT, we observe a better performance of MLA but larger variability than LL. On the other hand, the exact opposite can be said for RW, whereas both distributions seem to be identical for NI. It is important to note that correlation between concept groups could have a major impact in the 1B results. In the extreme case in which the data has perfectly correlated groups, the methods are free to choose one group over the other at random. Identifying the *two best concept groups* helps to mitigate this issue. Our experiments using the 2B *concept groups* show that both models are quite stable with averages of over 60% of agreement across runs. Again, the average model-to-model comparison seems to be dataset-dependant, however, MLA consistently shows lower variability than LL, making it more reliable and prone to provide robust and reliable explanations.

[5]Note that one teacher network was trained for each dataset, and its corresponding student networks underwent multiple runs to accommodate randomization. Since AVG is derived from the teacher network rather than the student networks, it is excluded from this analysis.