

Appendix

This appendix provides additional details for "TimeSliver: Interpretable Time-Series Classification using Symbolic-Linear Representation". Implementation details for TimeSliver and the backbone models are presented in Section A. Class distribution for the four datasets used in the interpretability study are provided in Section B. Detailed results on interpretability and predictive performance are given in Sections C and D, respectively. The broader impact of this work is discussed in Section E.

A Additional Implementation Details

A.1 Dataset Description

Synthetic Dataset. As described in [42], each sample in the synthetic dataset consists of 6 features and 500 time steps. To simulate realistic temporal dependencies, each feature includes a baseline sine wave with a frequency uniformly sampled from the range $[2, 5]$. Two randomly selected features per sample are injected with discriminative sine waves, each supported over 100 time steps, with frequencies drawn from a discrete uniform distribution in the range $[10, 50]$. In the remaining four features, a square wave is optionally added with 50% probability, also using frequencies sampled from the same range. The classification task is binary: the model must predict whether the sum of the two discriminative frequencies exceeds a predefined threshold, set to $\tau = 60$.

Audio Dataset. We use a manually curated subset of the ESC-50 audio dataset, focusing exclusively on animal sounds. This subset was selected to leverage the temporal localization of animal sounds, which typically occur within short bursts in the observation window, as opposed to environmental sounds that span the entire duration and yield robust results even with randomly sampled segments. This temporal sparsity makes animal sounds particularly useful for evaluating interpretability methods that rely on temporal attribution. For preprocessing, we extract Mel-frequency cepstral coefficients (MFCCs) from the audio using a Mel spectrogram with 40 Mel bands, employing standard settings such as centered windowing and normalization.

EEG Dataset. This dataset comprises single-channel EEG recordings collected from 20 subjects, with the objective of classifying five sleep stages: wake, N1, N2, N3 (non-REM stages), and REM (rapid eye movement). The temporal structure of EEG signals makes this dataset well-suited for tasks requiring time-series modeling and interpretation. We balance all the classes in the dataset before using it for the study.

FordA Dataset. We adopt the data preprocessing and train-test splits for the FordA dataset as defined in the MTS-Bakeoff benchmark [35].

A.2 Dataset Details

Information such as the number of variates (v), maximum sequence length, and dataset splits is provided in Table 5.

Table 5: Summary of the four datasets used in the interpretability study

Dataset	Num. of Variates, v	Max Seq. Length	Train	Valid	Test
Synthetic	6	500	5000	500	500
Audio	40	501	280	60	60
EEG	1	3000	5005	1295	3515
Ford-A	1	500	853	106	119

A.3 Model Details

The complete details of TimeSliver for all four datasets are given in Table 6. Additionally, the details for the other three backbones used in the interpretability study are given in Table 7.

Table 6: Architecture details of TimeSliver used for different datasets

Dataset	Num. of categorical bins, n	Num. of columns in \mathcal{O} , $n \times v$	Latent vector size, q	Segment size, m	Trainable parameters
Synthetic	15	90	36	7	5,858
Audio	10	400	12	1	20,110
EEG	25	25	12	10	6,441
Ford-A	70	70	36	10	8,280

Table 7: Number of trainable parameters for different model architectures across datasets.

Dataset	CNN	COLOR	Transformer
Synthetic	42,378	2,660	46,714
Audio	224,938	8,206	370,498
EEG	74,981	43,309	230,805
Ford-A	42,058	26,536	361,090

A.4 Training and Optimization Details

All experiments are conducted on a server running Ubuntu OS, equipped with NVIDIA RTX A6000 GPUs, using the PyTorch framework. During model training, we employ the Adam optimizer with a learning rate ranging from 3×10^{-4} to 1×10^{-3} . Validation accuracy is used for early stopping and to save the best model checkpoint.

A.5 Predictive Results on different backbone

Table 3 presents the predictive performance of the four deep learning models used as backbones in the interpretability study. The CNN backbone is used for all post-hoc interpretability methods, while the Transformer is employed for attention tracing and the Grad-SAM method. COLOR, originally developed for protein sequence design, is inherently interpretable. The predictive performance of TimeSliver on the four datasets used in the interpretability study is within 3–4% of the best-performing model. All the post-hoc methods are implemented using the Captum library¹ in PyTorch.

Table 8: Accuracy (mean \pm std) over 3 runs for different predictive backbone and dataset (Supporting results for Section 3 in the main paper).

Dataset	CNN	COLOR	Transformer	TimeSliver
Synthetic	0.93 \pm 0.028	0.93 \pm 0.014	0.95 \pm 0.0071	0.93 \pm 0.014
Audio	0.78 \pm 0.0071	0.80 \pm 0.021	0.80 \pm 0.000	0.81 \pm 0.0071
EEG	0.78 \pm 0.019	0.73 \pm 0.039	0.68 \pm 0.038	0.72 \pm 0.042
FordA	0.92 \pm 0.0071	0.93 \pm 0.000	0.83 \pm 0.0071	0.88 \pm 0.000

B Class Distribution

The class distribution for all four datasets is shown in Figure 5, indicating that there is no class imbalance in any of the datasets used in the interpretability study.

¹<https://github.com/pytorch/captum>

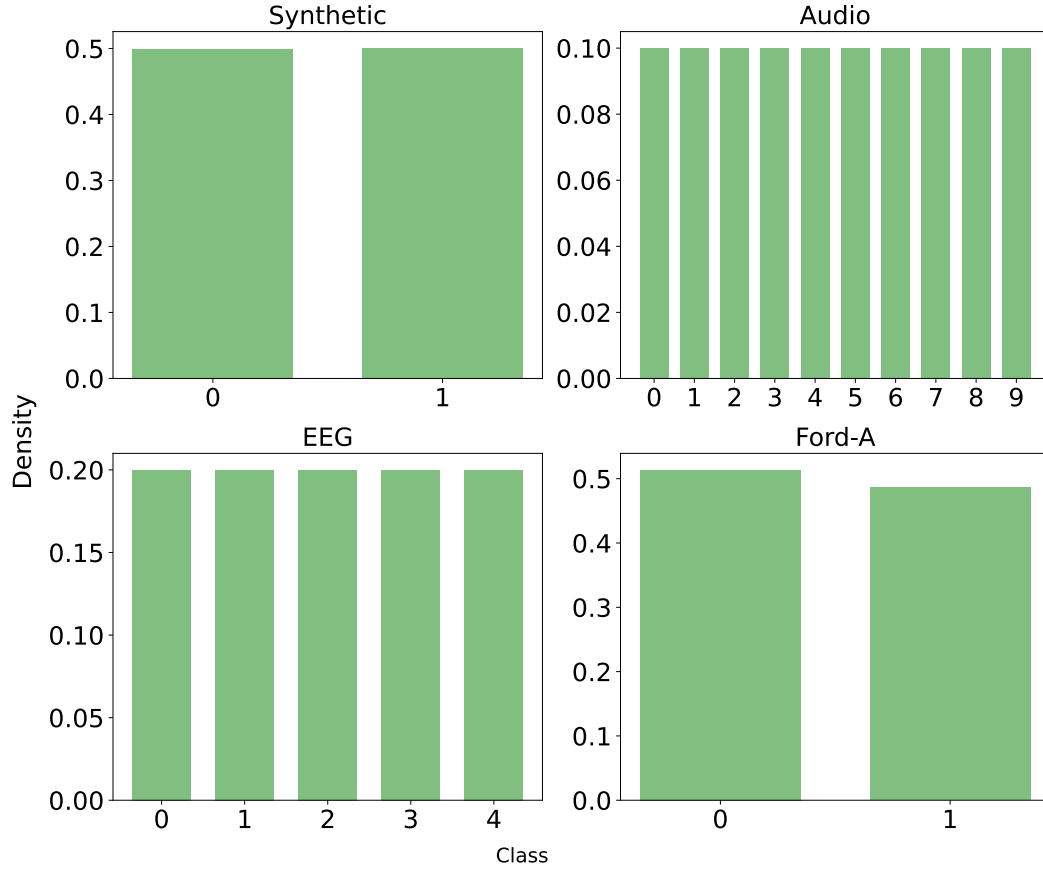


Figure 5: Class distribution among different datasets.

904 C Detailed Interpretability Results

905 C.1 Evaluating Positive Temporal Attribution

906 Figure 6 shows the mean $e(u)$ versus unmasking percentage ($u\%$) curves obtained using different
 907 interpretability methods, along with their standard deviations. The trend of the curves clearly
 908 demonstrates that TimeSliver outperforms the baseline methods in the lower unmasking range
 909 (5–20%), highlighting its effectiveness in identifying the most critical time points.

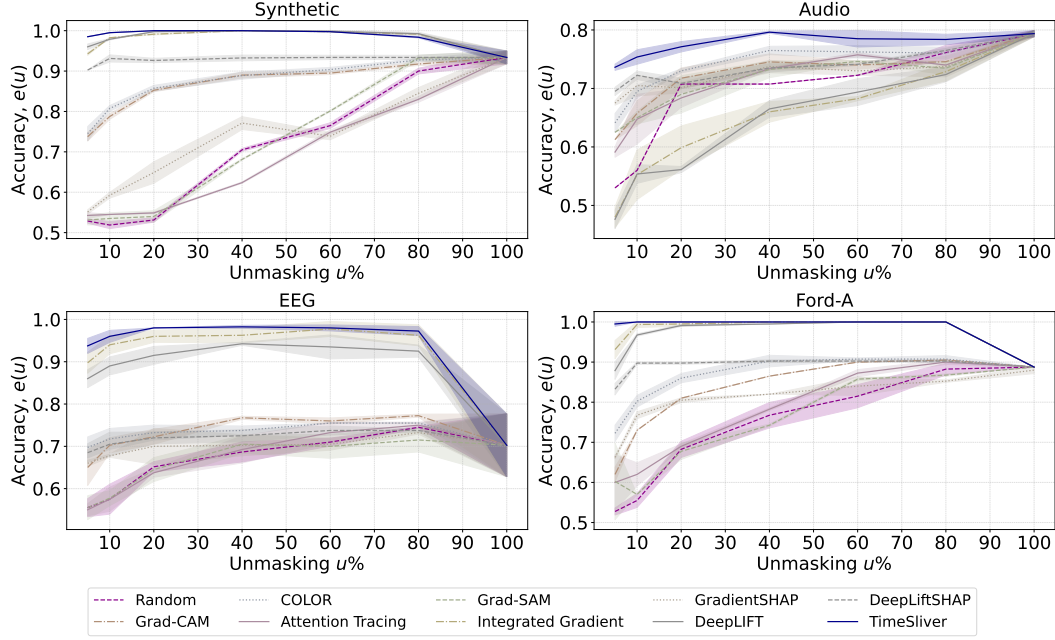


Figure 6: Positive attribution study. Accuracy curves $e(u)$ plotted against the unmasking percentage $u\%$ for various methods on three datasets: a) Audio, b) EEG SSC, and c) Ford-A. Each curve represents the mean accuracy over three runs (Supporting results for Section 3.1 in the main paper).

The areas under the curves shown in Figure 6, $\mathcal{I}(100)$ and $\mathcal{I}(20)$, are used to quantitatively compare the different interpretability methods. Table 9 reports the mean and standard deviation of $\mathcal{I}(100)$ and $\mathcal{I}(20)$ for each method.

Table 9: Comparison of positive attribution results, with the mean $\pm std$ $\mathcal{I}(100)$ and $\mathcal{I}(20)$ values. The best results are **bolded**, and the second-best are underlined. \uparrow indicates that larger values are preferred. In the main paper, only the mean values are presented in Table 1.

Type	Method	Synthetic		Audio		EEG		FORD-A	
		$\mathcal{I}(100)\uparrow$	$\mathcal{I}(20)\uparrow$	$\mathcal{I}(100)\uparrow$	$\mathcal{I}(20)\uparrow$	$\mathcal{I}(100)\uparrow$	$\mathcal{I}(20)\uparrow$	$\mathcal{I}(100)\uparrow$	$\mathcal{I}(20)\uparrow$
Post-hoc	Random	69.92 \pm 0.04	7.87 \pm 0.11	67.90 \pm 0.30	9.06 \pm 0.02	62.66 \pm 1.85	9.33 \pm 0.38	73.89 \pm 1.96	8.89 \pm 0.25
	Grad-CAM	83.93 \pm 0.42	12.01 \pm 0.06	69.79 \pm 0.38	10.05 \pm 0.07	67.23 \pm 0.96	10.70 \pm 0.33	81.43 \pm 0.04	11.07 \pm 0.04
	Integrated Gradient	93.71 \pm 0.14	14.68 \pm 0.00	63.69 \pm 0.04	8.34 \pm 0.46	83.19 \pm 0.89	14.24 \pm 0.29	93.65 \pm 0.19	14.76 \pm 0.14
	GradSHAP	71.99 \pm 0.58	9.07 \pm 0.09	69.53 \pm 0.31	10.48 \pm 0.17	63.76 \pm 2.27	10.41 \pm 0.26	78.54 \pm 0.64	11.44 \pm 0.03
	DeepLift	93.87 \pm 0.23	14.73 \pm 0.03	63.35 \pm 0.39	8.15 \pm 0.11	80.43 \pm 0.70	13.63 \pm 0.32	93.11 \pm 0.08	14.41 \pm 0.07
	DeepLiftShap	88.47 \pm 0.66	13.87 \pm 0.09	70.55 \pm 0.40	10.70 \pm 0.08	65.34 \pm 1.78	10.66 \pm 0.34	85.30 \pm 0.27	13.30 \pm 0.07
Inherently Interpretable	Attention Tracing	67.06 \pm 0.27	8.19 \pm 0.03	69.15 \pm 0.49	9.75 \pm 0.42	63.16 \pm 2.42	9.28 \pm 0.30	76.47 \pm 0.36	9.60 \pm 0.42
	Grad-SAM	71.08 \pm 0.72	8.04 \pm 0.21	69.00 \pm 0.12	9.89 \pm 0.24	62.11 \pm 2.92	9.38 \pm 0.34	74.17 \pm 0.01	9.17 \pm 0.14
	COLOR	84.52 \pm 0.42	12.21 \pm 0.03	71.46 \pm 0.67	10.47 \pm 0.14	66.48 \pm 1.47	10.85 \pm 0.34	83.95 \pm 0.85	12.12 \pm 0.22
Ours	TimeSLIVER	93.89 \pm 0.19	14.93 \pm 0.00	74.30 \pm 0.68	11.35 \pm 0.15	83.99 \pm 0.61	14.52 \pm 0.15	93.87 \pm 0.01	14.99 \pm 0.01

D Detailed Predictability Results

The predictive performance of TimeSliver on all 26 UEA datasets, along with the results of the baseline methods, is presented in Table 10.

E Broader Impact

Given TimeSliver’s superior interpretability, practitioners can make decisions with greater trust, transparency, and accountability in high-stakes applications, while also advancing the state-of-the-art in fundamentally interpretable architectures for time-series classification. For instance, in a medical setting, TimeSliver can highlight specific segments of EEG time-series data that influence its prediction. If these segments align with a clinician’s own assessment, it can enhance confidence in

Problem	DTW_D	DTW_I	DTW_A	MUSE	gRSF	CIF	MSeQL	ROCKET	CBoss	STC	RISE	TSF	HC	TapNet	ResNet	Inception	TimeSliver
ArticulatoryWordRecognition	98.87	94.31	98.94	98.87	98.21	97.89	98.98	99.56	97.56	97.51	95.73	94.82	97.99	97.13	98.26	99.10	99.33
AtrialFibrillation	23.56	34.67	22.44	74.00	27.56	25.11	36.89	24.89	30.44	31.78	24.44	29.78	29.33	30.22	36.22	22.00	73.00
BasicMotions	95.25	97.17	99.92	100.00	99.83	99.75	94.83	99.00	98.75	97.92	100.00	98.78	100.00	99.08	100.00	100.00	100.00
Cricket	100.00	95.74	100.00	99.77	97.41	98.38	99.21	100.00	97.55	98.94	97.78	93.15	99.26	97.50	99.40	99.44	98.61
DuckDuckGeese	49.20	29.27	56.67	56.00	44.47	56.00	39.27	46.13	43.07	43.47	50.80	38.87	47.60	58.27	63.20	63.47	56.10
EigenWorms	64.58	44.20	97.85	99.33	83.00	90.33	72.16	86.28	62.80	74.68	81.93	76.62	78.17	83.00	45.45	98.68	89.31
Epilepsy	96.30	67.03	97.37	99.64	97.34	98.38	99.93	99.08	99.83	98.74	99.86	99.83	100.00	96.09	98.16	98.65	98.55
EthanolConcentration	30.15	30.68	29.87	48.64	34.06	72.89	60.18	44.68	39.62	82.36	49.16	45.42	80.68	28.99	28.62	27.92	43.73
ERing	92.91	91.42	92.89	96.89	91.98	95.65	93.19	98.05	84.48	84.28	82.44	89.84	94.26	89.46	87.19	92.10	84.10
FaceDetection	53.28	51.53	—	68.89	55.06	69.17	62.97	69.42	52.32	69.76	51.17	68.95	69.17	52.87	53.13	77.24	69.97
FingerMovements	54.17	55.50	54.93	54.77	54.43	53.90	55.53	55.27	51.03	53.40	52.10	53.17	53.77	51.33	54.70	56.13	65.00
HandMovementDirection	30.32	26.67	30.72	38.02	32.07	52.21	35.23	44.59	28.87	34.95	28.24	48.51	37.79	32.34	35.32	42.39	46.00
Handwriting	61.21	34.33	60.55	51.85	36.06	35.13	54.04	56.67	49.09	28.77	18.27	36.42	50.41	32.95	59.78	65.74	57.10
Heartbeat	68.88	63.80	69.87	73.59	78.49	76.52	72.52	71.76	72.15	72.15	73.22	72.28	72.18	73.97	63.89	80.49	80.49
Libras	88.04	78.63	87.85	90.30	75.56	91.67	86.57	90.61	85.26	84.46	81.67	79.72	90.28	83.63	94.11	88.72	83.33
LSST	54.76	49.57	56.96	63.62	58.05	56.17	60.28	63.15	43.62	57.82	50.58	34.31	53.84	46.33	42.94	33.97	68.53
MotorImagery	56.10	49.63	50.37	52.17	53.80	51.80	53.00	53.13	52.37	50.83	49.83	53.80	52.17	45.37	49.77	51.17	61.90
NATOPS	82.04	76.07	81.48	87.13	82.37	84.41	86.43	88.54	82.48	84.35	80.59	77.72	82.85	90.30	97.11	96.63	98.89
PenDigits	99.28	99.22	99.27	98.68	91.27	98.97	97.14	99.56	95.61	97.70	87.47	94.11	97.19	93.65	99.64	99.68	98.10
PEMS-SF	77.05	80.23	78.73	99.85	91.27	99.86	97.15	85.63	96.57	98.40	98.98	96.76	97.98	79.21	81.95	82.83	94.80
PhonemeSpectra	15.39	10.18	—	25.86	15.27	32.87	30.86	28.35	19.43	30.62	26.78	14.52	32.87	22.17	15.39	36.74	29.00
RacketSports	85.64	81.69	85.79	89.56	87.79	89.30	88.73	92.79	88.90	88.09	84.17	88.29	90.64	85.81	91.23	91.69	92.10
SelfRegulationSCP1	81.81	80.63	81.34	73.58	79.74	85.94	82.86	86.55	81.33	84.73	73.17	84.73	86.02	95.68	76.11	84.69	90.00
SelfRegulationSCP2	54.09	48.48	52.43	49.52	50.62	48.87	49.61	51.35	50.62	51.63	50.28	50.62	51.67	56.05	50.24	52.04	62.00
StandWalkJump	22.00	35.78	25.56	34.67	38.44	45.11	42.00	45.56	36.89	44.00	34.00	33.33	40.67	35.11	30.89	42.00	67.00
UWaveGestureLibrary	92.28	87.58	91.51	90.39	89.59	92.42	91.32	94.43	86.13	87.03	71.11	85.05	91.31	89.59	88.35	91.23	91.00

Table 10: Detailed predictive performance comparison across 26 datasets in UEA (Supporting results for Section 3.2 in the main paper).

922 the model's output. In another example, when used to inform decisions such as adjusting a credit
923 limit based on transaction history, TimeSliver enables inspection of the decision rationale, helping
924 to ensure that it is not influenced by gender-specific transactions—thereby promoting fairness and
925 responsible AI practices.