Improving Partially Supervised Hidden Markov Models with Soft Labels from Temporal Fuzzy Clustering

Filip Wichrowski, Katarzyna Kaczmarek-Majer

Department of Stochastic Methods
Systems Research Institute, Polish Academy of Sciences
fwichrow@ibspan.waw.pl

Outline

Background: tracking the progression of bipolar disorder through speech

Bipolar disorder is a mental health condition marked by alternating states of depression, mania, mixed state, and stable mood (euthymia).

Goal: track the patient's state over time using speech.

The main challenge is **label sparsity**: a patient's state is evaluated only during infrequent psychiatric visits, which are both costly and time-consuming.

Consequence: highly limited labeled data that renders many methods ineffective.

What is the goal?

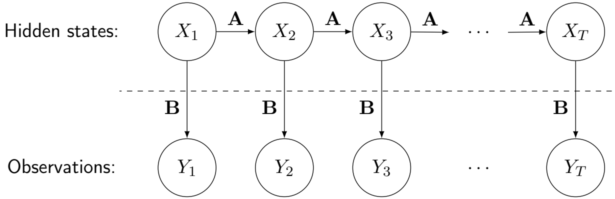
To improve partial supervision (in scarce labels scenarios) in hidden Markov models by applying soft labelling based on fuzzy clustering of observations.

Introduction to hidden Markov models

Let X_t be a Markov process with a space $S = \{s_1, ..., s_N\}$ of hidden states, and Y_t a stochastic process with continuous observations in $V \subseteq \mathbb{R}^k$, for some $N, k \in \mathbb{N}_+$.
A hidden Markov model[1] (HMM) is defined by a triple $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, where:

- $\mathbf{A} = \{a_{ij}\} = \{\mathbb{P}(X_t = s_j \mid X_{t-1} = s_i)\}$ is a transition matrix,
- $\mathbf{B} = \{b_j(Y_t)\} = \{p(Y_t \in V \mid X_t = s_j)\}$ is the emission probability,
- $\pi = \{\pi_i\} = \{\mathbb{P}(X_1 = s_i)\}$ is the initial distribution.

for $i, j = 1, ..., N$ and $t = 1, ..., T$.



[1] L. R. Rabiner, 1989, *A tutorial on hidden Markov models*, in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286.

Introduction to hidden Markov models

By default, HMMs are unsupervised, with learning typically performed using the Baum-Welch algorithm, a variant of the expectation–maximization method. Incorporating partial supervision during training involves constraining the set of possible paths in the HMM's lattice representation (treillis), as shown in the Figure 1.

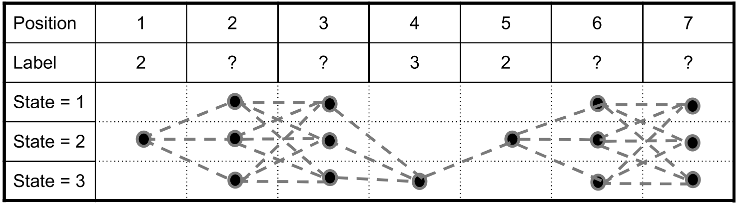


Figure 1: A constrained lattice. Known states $\{X_1, X_4, X_5\}$ restrict the set of possible paths.

However, this approach doesn't support soft or uncertain labels.

Partial supervision with soft labels: weight matrix

To allow for soft partial labels, we introduce a stochastic **weight matrix**

$$\Phi = \{\varphi_{tj}\} \quad t \in \{1, ...T\}, \qquad (1)$$
$$j \in \{1, ..., N\},$$

where $\varphi_{tj} \in [0, 1]$ and $\sum_{j=1}^{N} \varphi_{tj} = 1 \ \forall_t$.

Each row of Φ defines a vector of weights $\varphi_{t\cdot} = (\varphi_{t1}, ..., \varphi_{tN})^\top$ that is used to scale the emission probability

$$\tilde{b}_j(Y_t) = \varphi_{tj} b_j(Y_t) \quad \forall t, j. \qquad (2)$$

For example, if Y_t is strongly believed to originate from state j, it can be assigned a higher weight, with lower weights distributed across the other states. If $\varphi_{tj} \in \{0, 1\} \ \forall t, j$, we recover the original approach.

How to construct weight matrix?

Constrained Temporal fuzzy C-Means (CT-FCM)

We propose to derive weights from the Fuzzy C-Means (FCM) membership matrix.
However, two issues are identified with the vanilla FCM

- Partial label information is not utilized,
- Temporal dependencies within the data are ignored.

Therefore, we propose a **constrained temporal FCM** (CT-FCM), with an objective:

$$J_{(m,\lambda_{\mathrm{TS}},\lambda_{\mathrm{PS}})}(U,V) = \sum_{j=1}^{c} \left[\underbrace{\sum_{t=1}^{T} u_{tj}^{m} ||x_t - v_j||_2^2}_{\text{Vanilla (FCM)}} + \lambda_{\mathrm{TS}} \underbrace{\sum_{t=1}^{T-1} ||u_{(t+1)j} - u_{tj}||_2^2}_{\text{Temporal smoothing (TS)}} + \frac{\lambda_{\mathrm{PS}}}{|\mathcal{T}_X|} \underbrace{\sum_{t \in \mathcal{T}_X} ||u_{tj} - \mathcal{M}_{tj}||_2^2}_{\text{Partial supervision (PS)}} \right].$$

Here, $\mathcal{T}_X$ is a set of time points with known labels.
For each $t \in \mathcal{T}_X$, the row $(\mathcal{M}_{t1}, \ldots, \mathcal{M}_{tN})$ is a one-hot vector indicating the known label $X_t = j$, i.e., $\mathcal{M}_{tj} = 1$ and $\mathcal{M}_{ti} = 0$ for all $i \neq j$. If $t \notin \mathcal{T}_X$, then $\mathcal{M}_{tj} = 0$ for each j.

Constrained Temporal Fuzzy C-Means (CT-FCM)

Therefore, the goal is to find (U, V) such that for given $(m, \lambda_{\text{TS}}, \lambda_{\text{PS}})$

$$J_{(m,\lambda_{\text{TS}},\lambda_{\text{PS}})}(U, V) \text{ s.t. } \sum_{i=1}^{c} u_{ti} = 1 \tag{3}$$

is minimized. Using Lagrange multipliers yields a

$$\mathcal{L}_{(m,\lambda_1,\lambda_2)}(U, V, \lambda) = J_{(m,\lambda_1,\lambda_2)}(U, V) + \sum_{t=1}^{T} \lambda_t \left(\sum_{i=1}^{c} u_{ti} - 1 \right) \tag{4}$$

$$\frac{\partial \mathcal{L}_{(m,\lambda_1,\lambda_2)}(U, V, \lambda)}{\partial \lambda_t} = \sum_{i=1}^{c} u_{it} - 1 = 0 \tag{5}$$

$$\frac{\partial \mathcal{L}_{(m,\lambda_1,\lambda_2)}(U, V, \lambda)}{\partial v_i} = -2 \sum_{t=1}^{T} u_{ti}^{m} (x_t - v_i) = 0 \tag{6}$$

Constrained Temporal Fuzzy C-Means (CT-FCM)

Further

$$\frac{\partial \mathcal{L}_{(m,\lambda_1,\lambda_2)}(U,V,\lambda)}{\partial u_{it}} = \frac{\partial\,\mathrm{FCM}(U,V)}{\partial u_{it}} + \frac{\partial\,\mathrm{TS}(U)}{\partial u_{it}} + \frac{\partial\,\mathrm{PS}(U,V)}{\partial u_{it}} + \lambda_t \tag{7}$$

Then

$$\frac{\partial\,\mathrm{FCM}(U,V)}{\partial u_{it}} = m u_{it}^{m-1} ||x_t - v_i||^2 + \lambda_t \tag{8}$$

$$\frac{\partial\,\mathrm{TS}(U)}{\partial u_{it}} = 2\lambda_1 \begin{cases} u_{1i} - u_{2i}, & t = 1 \\ 2u_{ti} - u_{(t+1)i} - u_{(t-1)i}, & 1 < t < T \\ u_{Ti} - u_{(T-1)i}, & t = T \end{cases} \tag{9}$$

$$\frac{\partial\,\mathrm{PS}(U)}{\partial u_{it}} = 2\lambda_2 (u_{ti} - \mathcal{M}_{ti}) \mathbb{1}\{t \in \mathcal{T}\} \tag{10}$$

For convenience assume $m = 2$.

Constrained Temporal Fuzzy C-Means (CT-FCM)

Let's put

$$d_i = [||x_1 - v_i||^2, \ldots, ||x_T - v_i||^2]^\top, \qquad \lambda = [\lambda_1, \ldots, \lambda_T]^\top,$$

$$u_i = [u_{i1}, \ldots, u_{iT}]^\top, \qquad \tau = [\tau_1, \ldots, \tau_T]^\top, \ \tau_t = 1 \text{ iff } t \in \mathcal{T}_X, \text{ else } 0.$$

For a given $i = 1, \ldots, c$, we have

$$2 \underbrace{\left(\operatorname{diag}(d_i) + \lambda_1 L + \lambda_2 \operatorname{diag}(\tau)\right)}_{A} u_i + \underbrace{\lambda - 2\lambda_2 \operatorname{diag}(\mathcal{M}_i)\tau}_{-B} = 0 \tag{11}$$

where

$$L = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 1 \end{bmatrix} \qquad \mathcal{M}_i = \begin{bmatrix} \mathcal{M}_{1i} & & \\ & \ddots & \\ & & \mathcal{M}_{Ti} \end{bmatrix} \tag{12}$$

Constrained Temporal Fuzzy C-Means (CT-FCM)

For each centroid $i = 1, ..., c$, we have:

$$2 \underbrace{\left(\mathrm{diag}(d_i) + \lambda_1 L + \lambda_2 \,\mathrm{diag}(\tau)\right)}_{A} u_i + \underbrace{\lambda - 2\lambda_2 \,\mathrm{diag}(\mathcal{M}_i)\tau}_{-B} = 0 \tag{13}$$

$$Au_i = B \tag{14}$$

Since A is tridiagonal (banded and positive-semidefinite), it is invertible.

Obviously, a solution to (13) must satisfy, for each t, a stochastic constraint. Therefore, it is either projected onto a probability simplex or a dual-elimination is used to satisfy KKT conditions for constrained optimization.

Let's pause and ponder

To summarise: the membership matrix U derived from CT-FCM is used as a weight matrix Φ (that encodes partial soft labels) to train a hidden Markov model.

$$\text{Partial labels} + \text{observations} \rightarrow \text{CT-FCM} \rightarrow \text{U} \rightarrow \Phi \rightarrow \text{HMM}$$

Simulation example

(1) **Generate** a sequence of hidden labels $\{X_t\}_{t=1}^T$, where each $X_t \in \{\text{green}, \text{red}\}$, and conditionally generate observations $\{Y_t\}_{t=1}^T$, according to

$$Y_t \mid X_t = \text{red} \sim \mathcal{N}(-\mu, 0.75),$$
$$Y_t \mid X_t = \text{green} \sim \mathcal{N}(+\mu, 0.75).$$

with $\mu \in \{0.25, 0.35, 0.45\}$.

(2) **Remove** a fixed proportion of labels from X uniformly at random to get partial labels X^*.

(3) **Define a grid** $\mathcal{G}$ of parameters $(\lambda_{\text{TS}}, \lambda_{\text{PS}})$; here, $\mathcal{G} = \{0, 1, 5, 10\} \times \{0, 1, 10, 10^2, 10^3\}$.

(4) **Choose an optimal tuple** $(\lambda_{\text{TS}}, \lambda_{\text{PS}})$ leveraging a walk-forward validation, as random CV breaks the temporal structure of the data:

(a) using first t_i observations, $i = 1, ..., n$, fit $\text{CT-FCM}(\lambda_{\text{TS}}, \lambda_{\text{PS}})$ and then fit HMM,

(b) calculate log-likelihood ℓ_i on the last $T - t_i$ observations,

(c) calculate weighted log-likelihood $(\sum_{i=1}^t t_i \ell_i)/(\sum_{i=1}^t t_i)$,

(d) proceed with $(\lambda_{\text{TS}}^*, \lambda_{\text{PS}}^*)$ that maximizes (c).

Simulation example

(5) **Fit the model**: using the whole training data, get the membership matrix from $\text{CT-FCM}(\lambda^*_{\text{TS}}, \lambda^*_{\text{PS}})$ and use it as a weight matrix to train a hidden Markov model $\text{CT-FCM} + \text{HMM}$.

(6) **Test the model**: generate data (X^t, Y^t), predict labels (Viterbi decoding), calculate Adjusted Rand Index (ARI) for $\text{CT-FCM} + \text{HMM}$. Repeat 50 times.

(7) **Repeat** steps $(1) - (6)$ 50 times to get the distribution of ARI values.

Simulation example

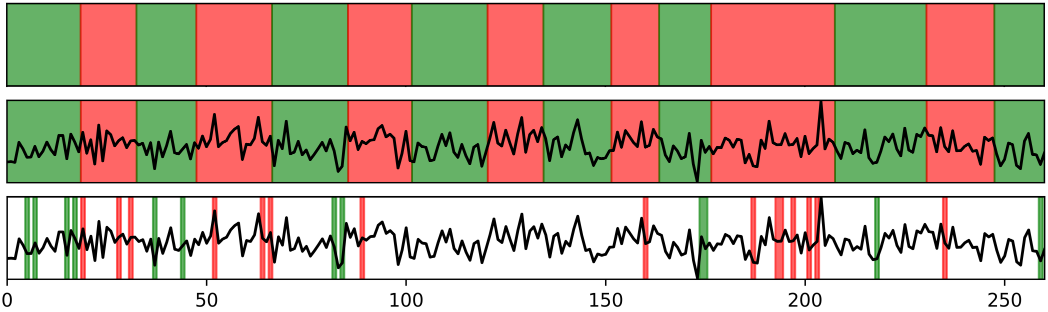


Figure 2: Visualization of hidden labels X and observations Y. **Top panel**: true labels X. The duration of each block is drawn from $\mathrm{Poiss}(20)$. **Middle panel**: true labels X with corresponding observations Y. Here, Y | red $\sim \mathcal{N}(0.3, 1)$ and Y | green $\sim \mathcal{N}(-0.3, 1)$. **Bottom panel**: partial labels after randomly removing 90% of the original labels.

Simulation example

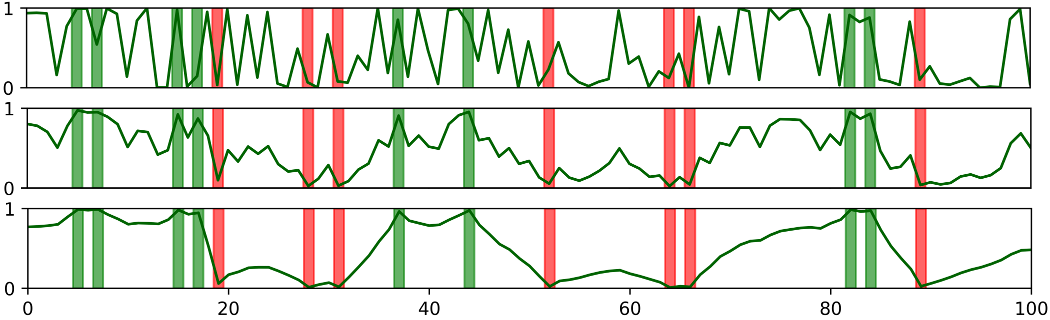
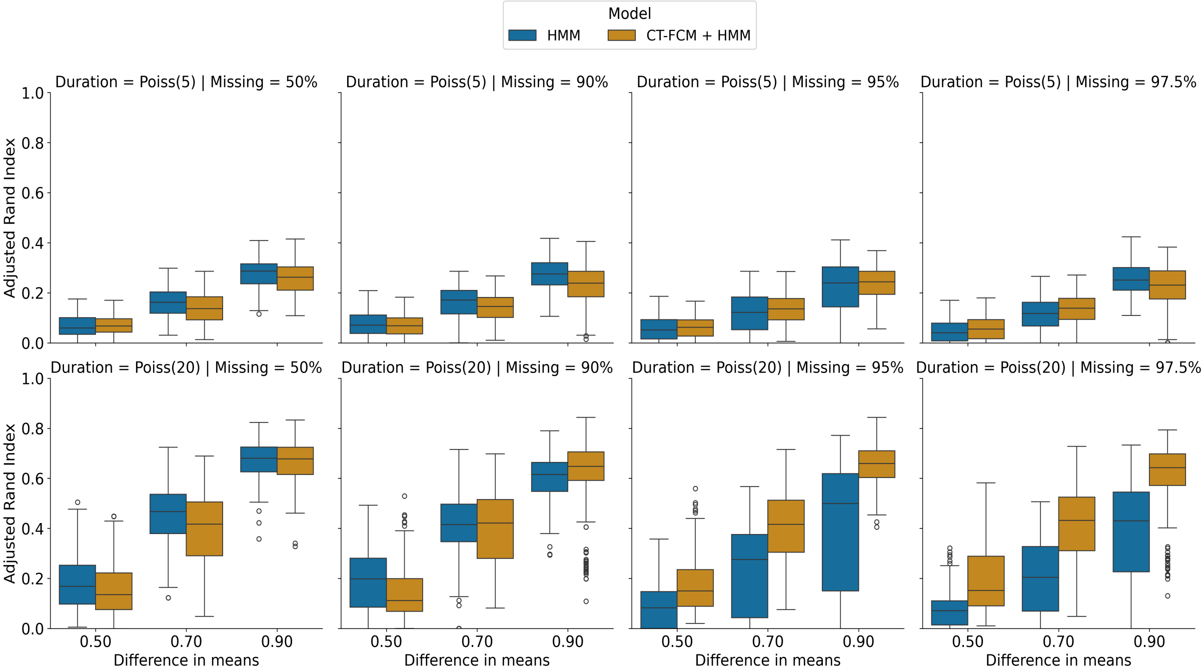


Figure 3: CT-FCM membership values for green state. **Top panel**: $(\lambda_{\mathrm{TS}}, \lambda_{\mathrm{PS}}) = (0, 0)$. **Middle panel**: $(\lambda_{\mathrm{TS}}, \lambda_{\mathrm{PS}}) = (1, 10)$. **Bottom panel**: $(\lambda_{\mathrm{TS}}, \lambda_{\mathrm{PS}}) = (10, 100)$. Only the first 100 observations are shown for clarity.



Model
HMM
CT-FCM + HMM
Duration = Poiss(5) | Missing = 50%
Duration = Poiss(5) | Missing = 90%
Duration = Poiss(5) | Missing = 95%
Duration = Poiss(5) | Missing = 97.5%
Duration = Poiss(20) | Missing = 50%
Duration = Poiss(20) | Missing = 90%
Duration = Poiss(20) | Missing = 95%
Duration = Poiss(20) | Missing = 97.5%
Adjusted Rand Index
0.0
0.2
0.4
0.6
0.8
1.0
0.50
0.70
0.90
Difference in means

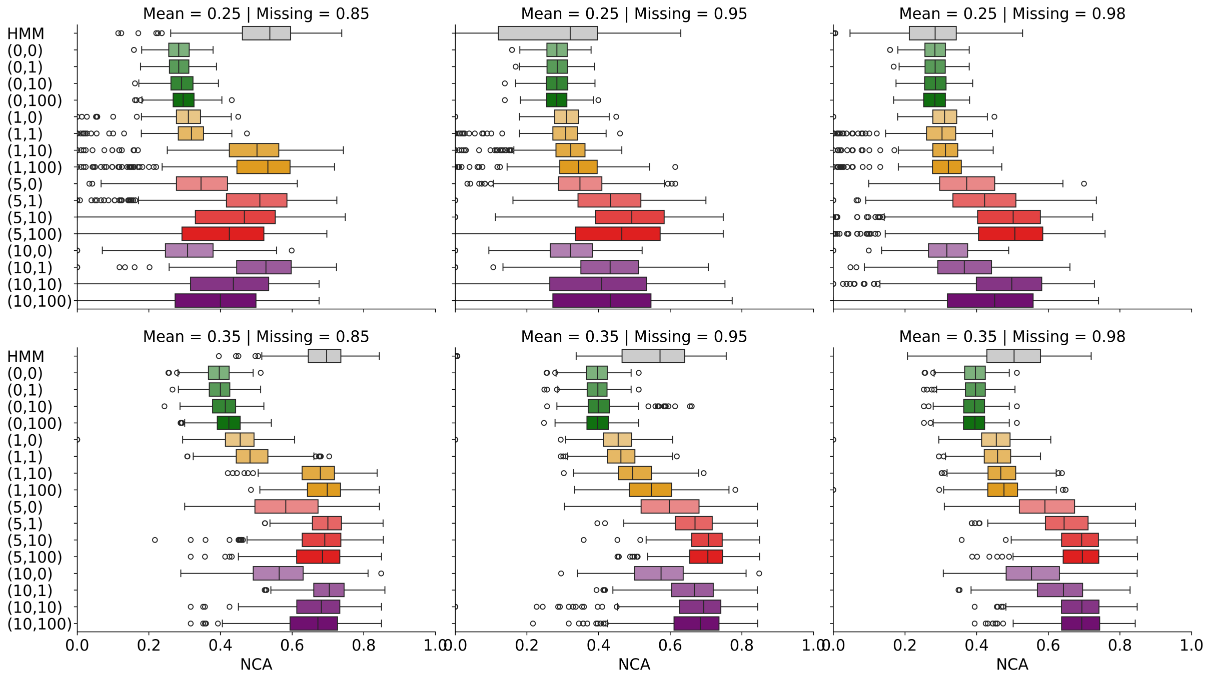
Conclusions

More experiments are necessary!

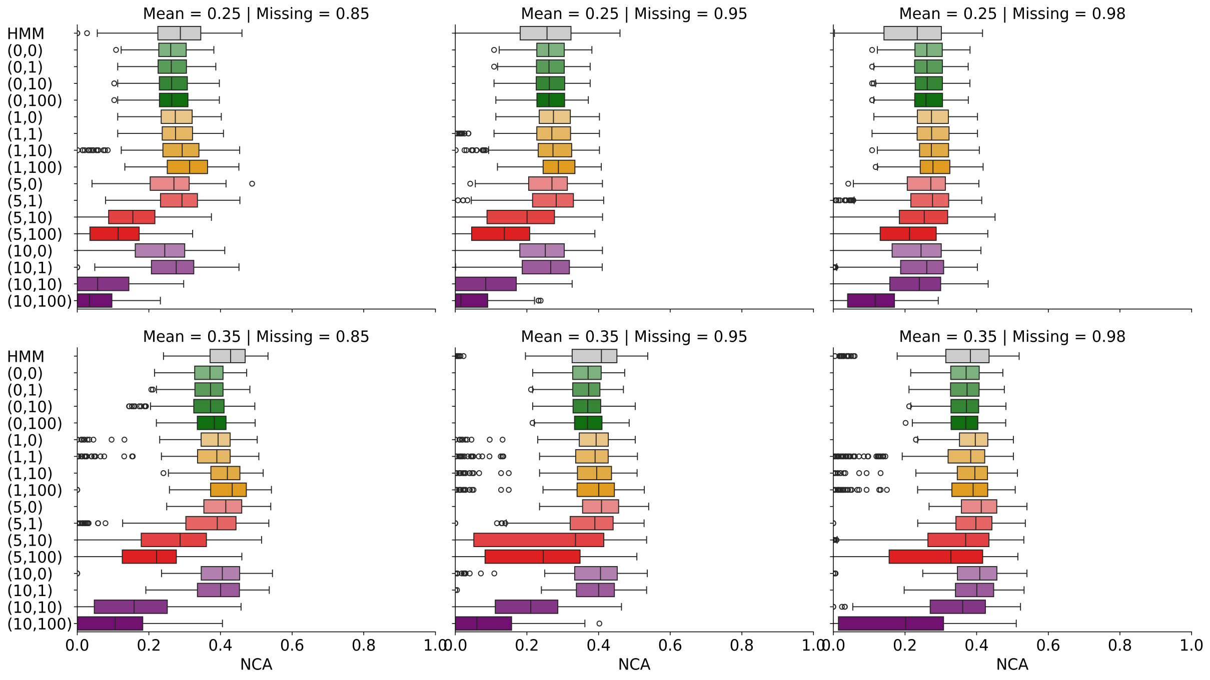
- Fuzzy pre-clustering might enhance the performance of a hidden Markov model, potentially not only in the sparse labels setting.
- A systematic and structured approach for selecting optimal values for λ_{TS} and λ_{PS} is crucial (e.g. walk-forward validation or blocked corss-validation).

Thank you!

The project *ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech* (FENG.02.02-IP.05-0302/23) is carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG).



Mean = 0.25 | Missing = 0.85
Mean = 0.25 | Missing = 0.95
Mean = 0.25 | Missing = 0.98
Mean = 0.35 | Missing = 0.85
Mean = 0.35 | Missing = 0.95
Mean = 0.35 | Missing = 0.98
HMM
(0,0)
(0,1)
(0,10)
(0,100)
(1,0)
(1,1)
(1,10)
(1,100)
(5,0)
(5,1)
(5,10)
(5,100)
(10,0)
(10,1)
(10,10)
(10,100)
0.0
0.2
0.4
0.6
0.8
1.0
NCA



Mean = 0.25 | Missing = 0.85
Mean = 0.25 | Missing = 0.95
Mean = 0.25 | Missing = 0.98
Mean = 0.35 | Missing = 0.85
Mean = 0.35 | Missing = 0.95
Mean = 0.35 | Missing = 0.98
HMM
(0,0)
(0,1)
(0,10)
(0,100)
(1,0)
(1,1)
(1,10)
(1,100)
(5,0)
(5,1)
(5,10)
(5,100)
(10,0)
(10,1)
(10,10)
(10,100)
0.0
0.2
0.4
0.6
0.8
1.0
NCA