Leveraging AI and IOT for effective explainable prediction of neurodegenerative diseases

Audio and Face Digital Markers based on Machine Learning & Deep Learning Foundation Models for Parkinson's Disease Assessment

Mounim A. EL Yacoubi

mounim.el_yacoubi@telecom-sudparis.eu

Institut Polytechnique de Paris / Telecom SudParis / SAMOVAR Lab

Context : Parkinson's Disease (PD)

- **Second most common neurodegenerative disease**
  - Affects 1% of people over 60 years
- **Impact on Central Nervous System**
  - Destruction of dopaminergic neurons in the substantia nigra
  - Causes Motor deficits
    - Rigidity, bradykinesia, rest tremor
  - Causes non-motor symptoms
    - Depression, anxiety, dysautonomia
- **Delayed Onset of Symptoms**
  - Symptoms occur years after disease onset
  - 60% of dopaminergic neurons already lost by diagnosis
- **Importance of Early-stage Detection**
  - Allows testing of treatments before irreversible brain damage
  - Slows down or halts disease progression

Motivation and Objectives

- **Hypomimia, known as Facial bradykinesia, Masked Face**
  - Common early-stage symptom of Parkinson's Disease
  - Characterized by
    - Decrease in facial movement
    - Loss of emotional expression in the face
- **Dysarthria**
  - Speech disorder when the muscles a person uses to speak become weakened
- **Dysphonia, impairment in the ability to speak normally due to muscle tightenss**
  - harsh, weak or breathy quality of voice
- **Negative Social Consequences**
  - Lack of facial expressions may lead to social rejection by others

Objective

- Parkinson's disease assessment based on hypomimia using face & Audio videos
  - DIGIPD project : Validating DIGItal biomarkers for better personalized treatment of PD

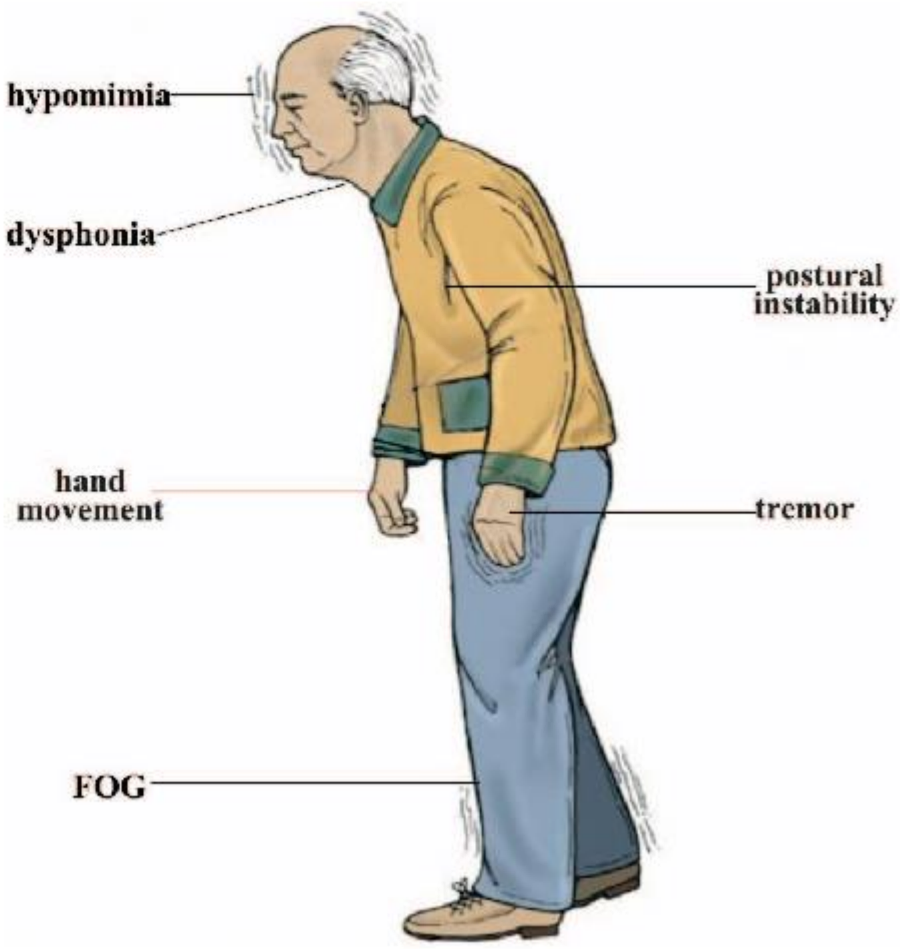


Figure 1. The typical symptoms of PD

IEEE int. Conference on Bioinformatics and Biomedicine 2017
PdAssist: Objective and quantified symptom assessment of Parkinson's disease via smartphone
Yiqiang ChenXiaodong YangBiao ChenC. MiaoHanchao Yu

DIGIPD project

DIGIPD

Validating DIGItal biomarkers for better personalized treatment of Parkinson's Disease

- ERA PerMed : ERA-Net Cofund, supported by 32 partners from 23 countries, cofunded by the EU
  - Joint European Transnational Call for collaborative innovative research projects in Personalised Medicine
- ERA PerMed **DIGIPD**: Validating DIGItal biomarkers for better personalized treatment of Parkinson Disease's (PD)
- Partners

• IP Paris / Telecom SudParis	France
• ICM - Brain Paris institute	France
• Fraunhofer Society	Germany
• University Hospital Erlangen	Germany
• Portabiles Healthcare Technologies	Germany
• University of Luxembourg (UL)	Luxembourg
• Université de Namur	Belgium
• Association Parkinson Madrid	Spain

https://www.digipd.eu/

IP Paris / TSP Contributions to DIGIPD

Audio and Face Digital Markers (DM) based on Machine Learning & Deep Learning (DL) Foundation Models

Research Team

- Institut Polytechnique de Paris, Telecom SudParis
  - *Anas Filali Razzouki* (*Face Digital Markers*), *Quang Dao Vu* (*Voice Digital Markers*), *Dijana Petrovska-Delacrétaz, Mounîm El-Yacoubi*
- ICM - Paris Brain Institute, Sorbonne Université, Inserm, CNRS, APHP, Hôpital Pitié-Salpêtrière, Paris, France
  - *Laetitia Jeancolas, Graziella Mangone, Sara Sambin, Alizé Chalançon, Manon Gomes, Stéphane Lehéricy , Jean-Christophe Corvol, Marie Vidailhet, Isabelle Arnulf*

Outline

- ICEBERG Dataset
- **PD assessment based on facial AUs**
  - Feature extraction-based on facial AUs
  - PD vs. HC classification
  - Interpretability (feature importance)
  - PD sex effect analysis
  - Longitudinal analysis
  - Correlation between AUs with clinical scores and DAT-scan
- **Face-based PD assessment based on Vision Foundation Models**
  - Optical flow extraction
  - Foundation Models based Video Vision Transformers (FM-ViViTs)
  - Classification of PD vs. HC
  - Interpretability
  - Fusion foundation models and AU-based classifiers for PD classification
- **Voice-based PD assessment based on Speech Foundation Models**
- **Perspectives**

ICEBERG Dataset

- **ICEBERG Protocol**
  - Longitudinal study at the Paris Brain Institute (ICM)
  - Aim to identify and validate biomarkers of PD
- **Participants**
  - Early-stage PD patients (disease duration < 4 years)
  - HC subjects had no neurological disorders
  - Participants visited the hospital once a year for 5 years
  - Participants underwent several tests
    - Neurological examination, motor and cognitive tests
    - biological sampling, brain Scans
    - and **<u>audiovisual recordings</u>**

	PD		HC	
Biological sex	Male	Female	Male	Female
No. of videos (294)	**126**	**77**	**58**	**33**
No. of subjects (154)	**70**	**39**	**26**	**19**
Age (years)	64.2 ± 9.4	65.6 ± 8.6	63.4 ± 9.5	63.1 ± 8.5
Hoehn yahr	1.9 ± 0.3	1.86 ± 0.55	-	-
MDS-UPDRS III total	33.9 ± 6.9	28.9 ± 8.3	3.9 ± 2.7	5.5 ± 3.3
MDS-UPDRS III face item	1.1 ± 0.5	0.9 ± 0.4	-	-

ICEBERG Video-Feb2023 dataset recordings

ICEBERG Audio-Visual Database

ICEBERG : initially designed to detect PD from speech

- **Recording Details**
  - The recording session lasts 15 to 20 minutes
  - Participants perform 25 speech tasks
    - Rapid repetitions of syllables :
      - /pa/,/pou/, /kou/, /poupa/, /pakou/, **/pataka/, /bagada/,** /patikou/, /pabikou/,/padikou/
    - Maintain sound /a/ for as long as possible
    - Pronounce sound /a/ like a siren
    - **Monologue**, reading (text, dialogue)
    - Repetitions of short sentences
    - Repeat the syllables /pa/, /kou/, and /pa kou/ slowly
    - Silence
  - Webcam characteristics
    - Frame rate = 24 fps
    - Resolution = 1920 * 1080 pixels

PD subject performing
3 selected speech tasks

PD Analysis based on Facial Action Units

- Handcrafted features: based action units signal *derivatives*

    ➔Fed as input to XGBoost classifier to detect hypomimia

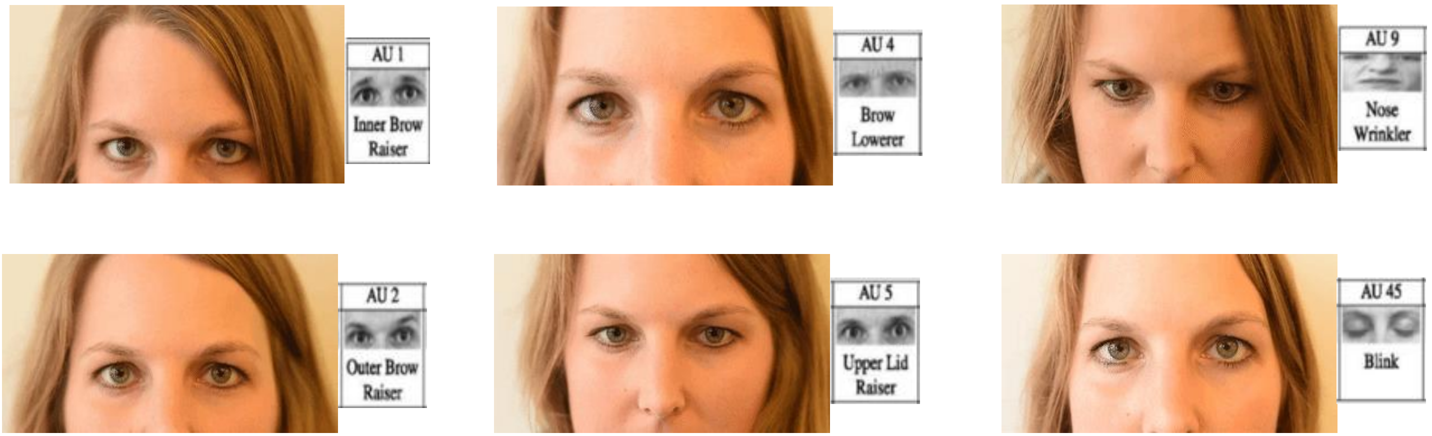
- Interpretability: reveal facial regions linked to hypomimia

- Effect of sex and longitudinal analysis

- Correlation between AUs and Clinical Scores and DatScan

Facial Action Units (AUs)

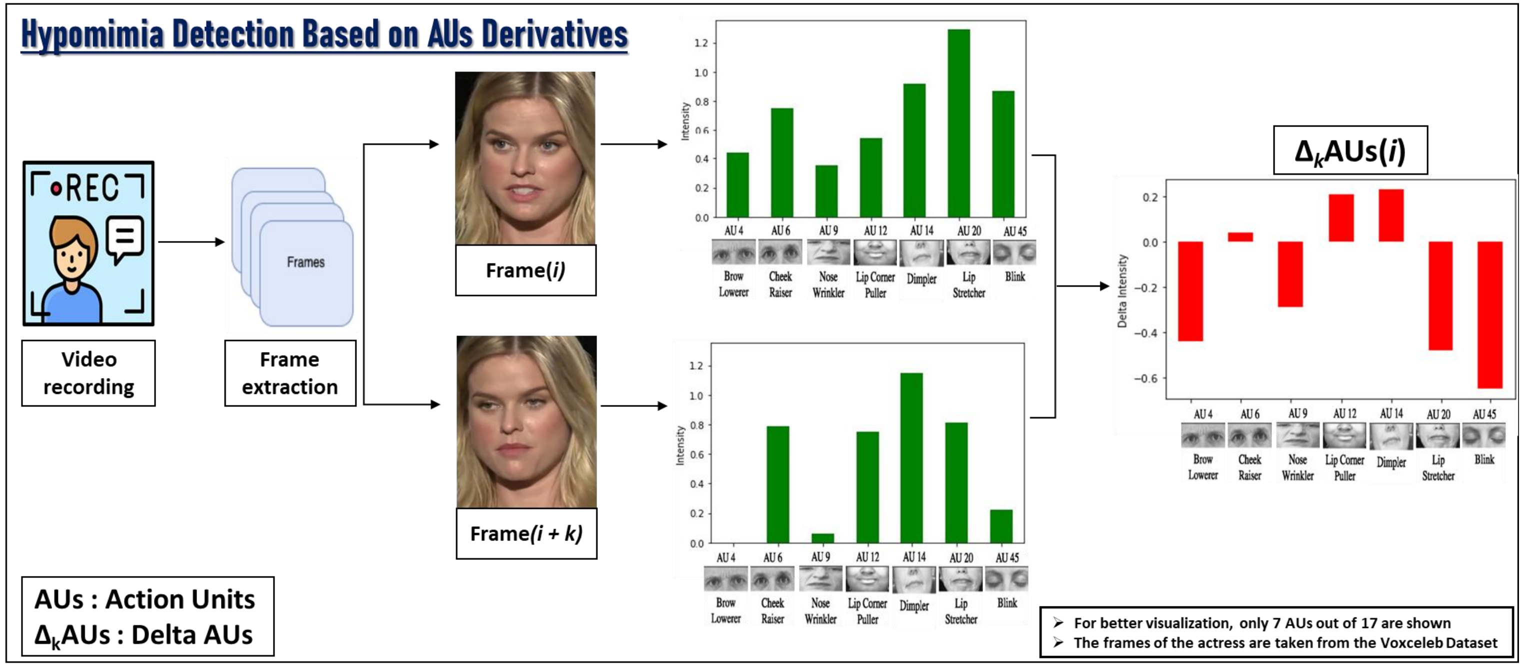
- **Action Units (AUs)**
  - Developed by Carl-Herman Hjortsjö, and later adopted by Paul Ekman and Wallace V. Friesen
  - AUs = basic movements of facial muscles
    - Extracted at each frame with intensity from 0 to 5
    - Compact representation
  - Each AU = specific movement pattern in the face



Visualization Reference: iMotions.com

Feature Extraction

- We use the OpenFace software that extracts 17 AUs out of 44 for each frame
- Movement encoding : Derivative of AUs with step k : $\Delta_k AU(i) = InAU(Frame(i+k)) - InAU(Frame(i))$
  - Step k tuned according video frame rate and speech task



A video is represented as time series of AUs or Δ_kAUs

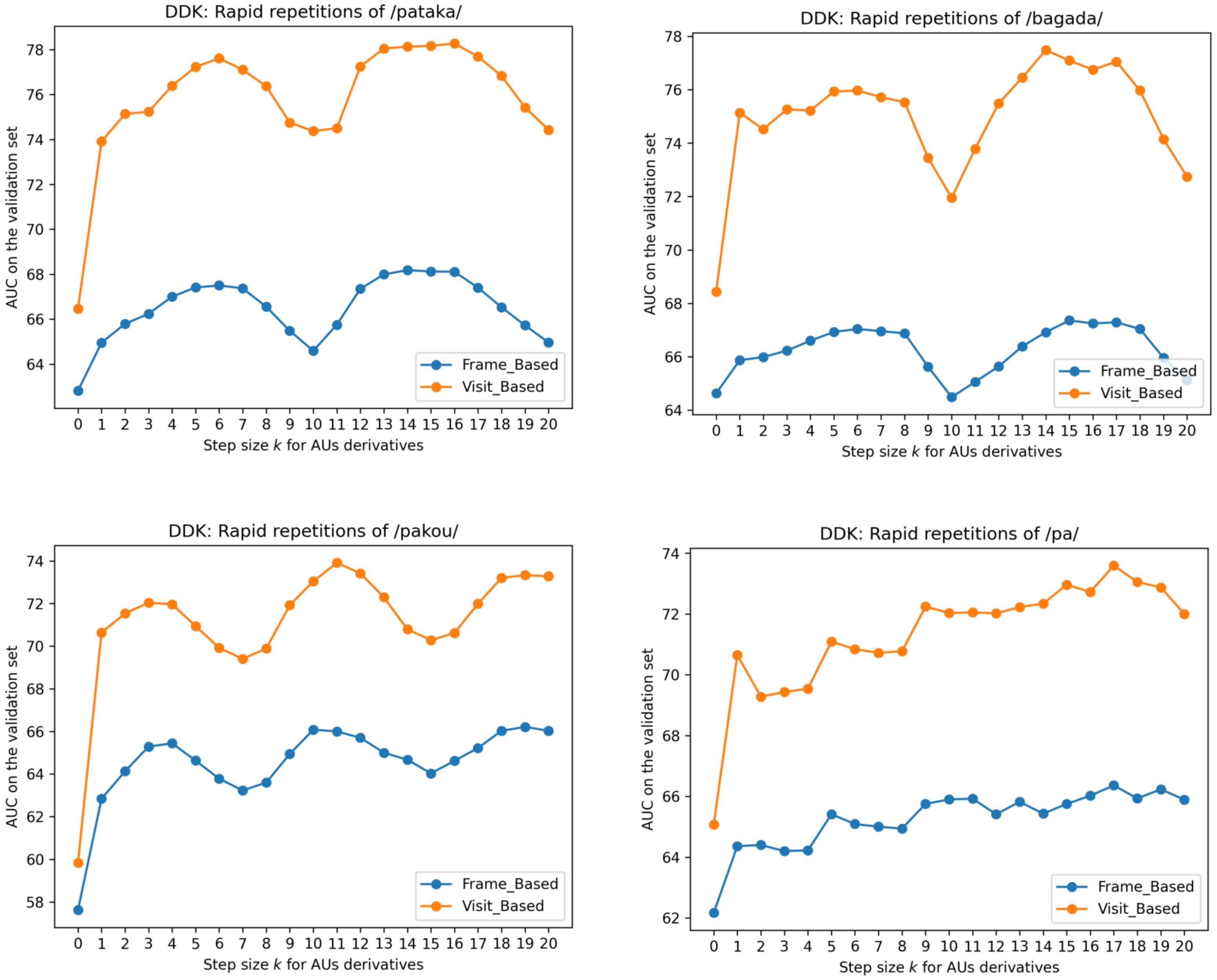
- **Video Global representation:** 28 Statistical measures are calculated across the AUs or Δ_kAUs frames
  - Including **basic descriptive** statistics (e.g., mean, percentiles, etc.), **entropy** measures, **frequency** domain measures
  - Advantage over local representation enhances robustness, captures temporal patterns, allows for simpler explainability and correlation analysis

Experiments: PD vs. HC Classification

- For each video task, Δ_kAUs calculated with step k from 1 to 20
- For each k, Δ_kAUs are input to XGBoost
  - XGBoost ➔ better adapted to tabular features & imbalanced class distribution
- Validation : **5-fold nested cross-validation (CV)**
  - Nested CV ➔ splits data into 5 outer folds for testing, with 5 inner folds for training and validation
    - Validation : XGBoost hyperparameters + step k optimization + classification threshold
    - Test: acts as a blind tests (unbiased estimate of performance)
- Evaluation Metrics: Area Under the Curve (AUC), Balanced Accuracy (BA)
  - **AUC** ➔ threshold independence, frequently employed in clinical studies
  - **BA** ➔ proficiency in addressing class imbalance
- Optimal k (k*) $\Rightarrow$ highest average AUC on the validation sets
- **Subject prediction** = Mean classification scores across visits

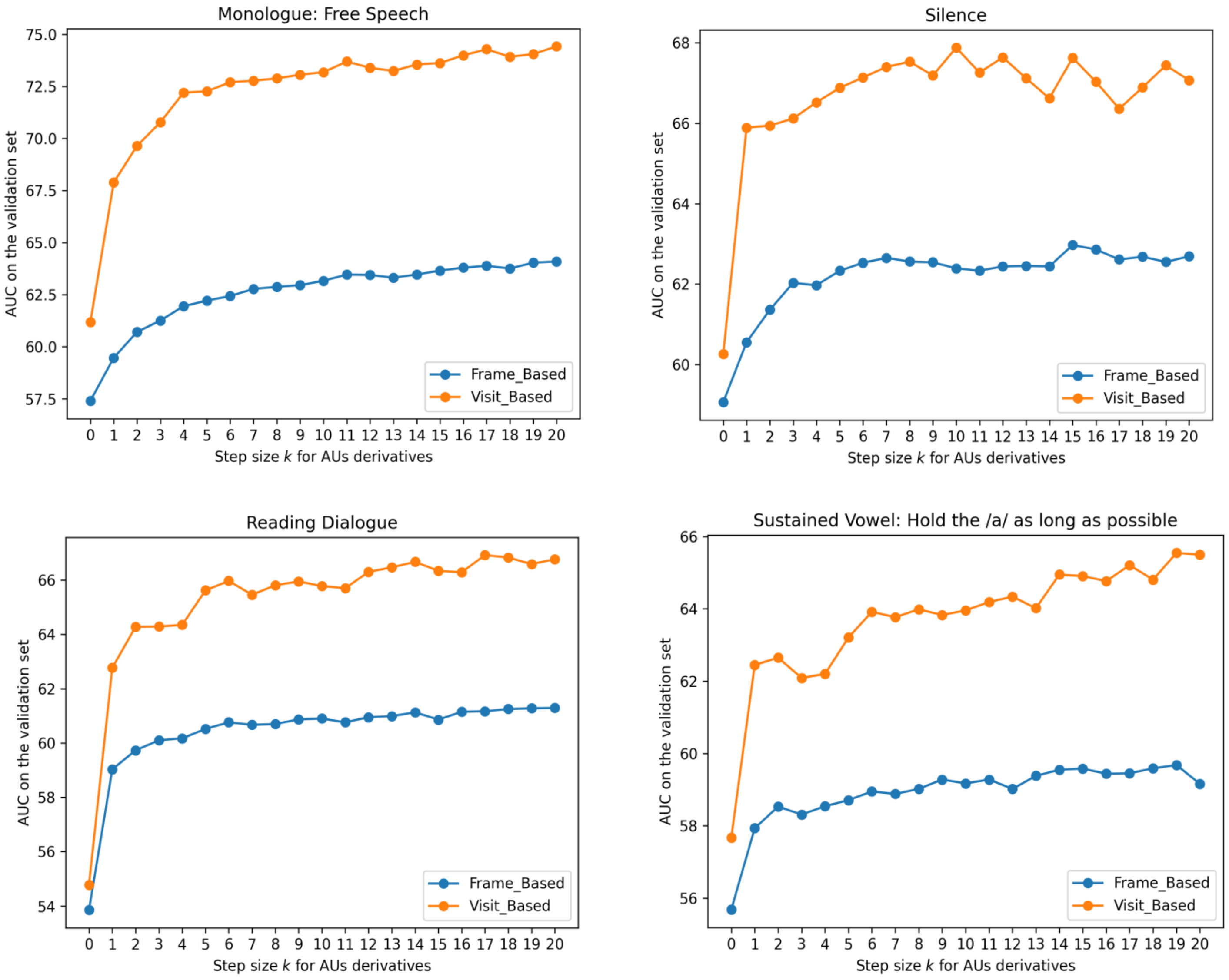
Results: Optimal Step k (k*) for DDK Tasks

- DKK (Diadochokinesi)
  - Rapid repetition of syllables
- The graphs exhibit a periodic pattern
  - characteristic of syllable repetition
- Graph period (P) ≈ average duration (in frames) of an expression
- The more syllables, longer the period:
  - For /pataka/ or /bagada/, P = 10, $k*$ = 6
  - For /pakou/, P = 7, $k*$ = 4
  - For /pa/, P = 4, $k*$ = 1
- AUC of Δ_{k*}AUs >> AUC of AUs ($k = 0$)



Results: Optimal Step k (k^*) for Other Tasks

- The graphs exhibit an aperiodic pattern
- The AUC with Δ_kAUs for k>0 better than k=0
  - $k = 0$, only AU intensities are used
- Silence task (no mouth movement)
  - Movement is captured in the eye region



Results: PD vs. HC Classification

- Best SMs for classification:
  - Basic descriptive statistics
    - e.g., variance, median, maximum, range
  - Signal power : average energy of the signal
  - Total power : overall energy across frequencies
  - Histogram entropy: measuring signal's complexity
- AUC of 91.4% for PD vs. HC
  - ➔Effective hypomimia detection in PD
- AUC : Δ_{k^*} AUs > AUs

Task	Signals	Best Statistical Measure (SM)	AUC (%)		BA (%)	
			VB	SB	VB	SB
/pataka/	$\Delta_{k^*=6}$ AUs	signal power	**79**	82,4 ± 3,3	73,6	76,8 ± 3,8
		absolute variance	78,7	**83,9 ± 3,1**	71,6	74,4 ± 4,0
		absolute histogram entropy	77,9	81,3 ± 3,4	72,2	74,5 ± 3,9
	AUs	total power (spectral density)	76,8	81,4 ± 3,4	69,1	74,6 ± 4,0
		histogram entropy	76,4	82,9 ± 3,3	67,5	75,0 ± 3,9
/bagada/	$\Delta_{k^*=6}$ AUs	absolute median	**80,1**	**86,6 ± 2,9**	69,8	73,9 ± 4,0
		absolute fourth moment	74,5	78,7 ± 3,7	72,8	70,7 ± 4,1
	AUs	75 percentile	74,6	80,6 ± 3,5	68,3	70,6 ± 4,1
		max	71,9	76,6 ± 3,9	69,3	69,4 ± 4,1
Monologue	$\Delta_{k^*=10}$ AUs	absolute histogram entropy	**78,4**	**80,8 ± 3,5**	70,2	71,0 ± 4,1
		absolute 75 percentile	76,5	79,1 ± 3,6	69,5	67,9 ± 4,2
	AUs	range	76,8	79,1 ± 3,6	69,2	70,6 ± 4,1
		max	76,4	75,5 ± 3,9	70,3	69,4 ± 4,1
Tasks fusion	Δ_{k^*} AUs	SMs Combined	**87,4**	**91,4 ± 2,2**	77,2	78,9 ± 3,8
	AUs	SMs Combined	84,7	87,3 ± 2,8	70	76,3 ± 3,9

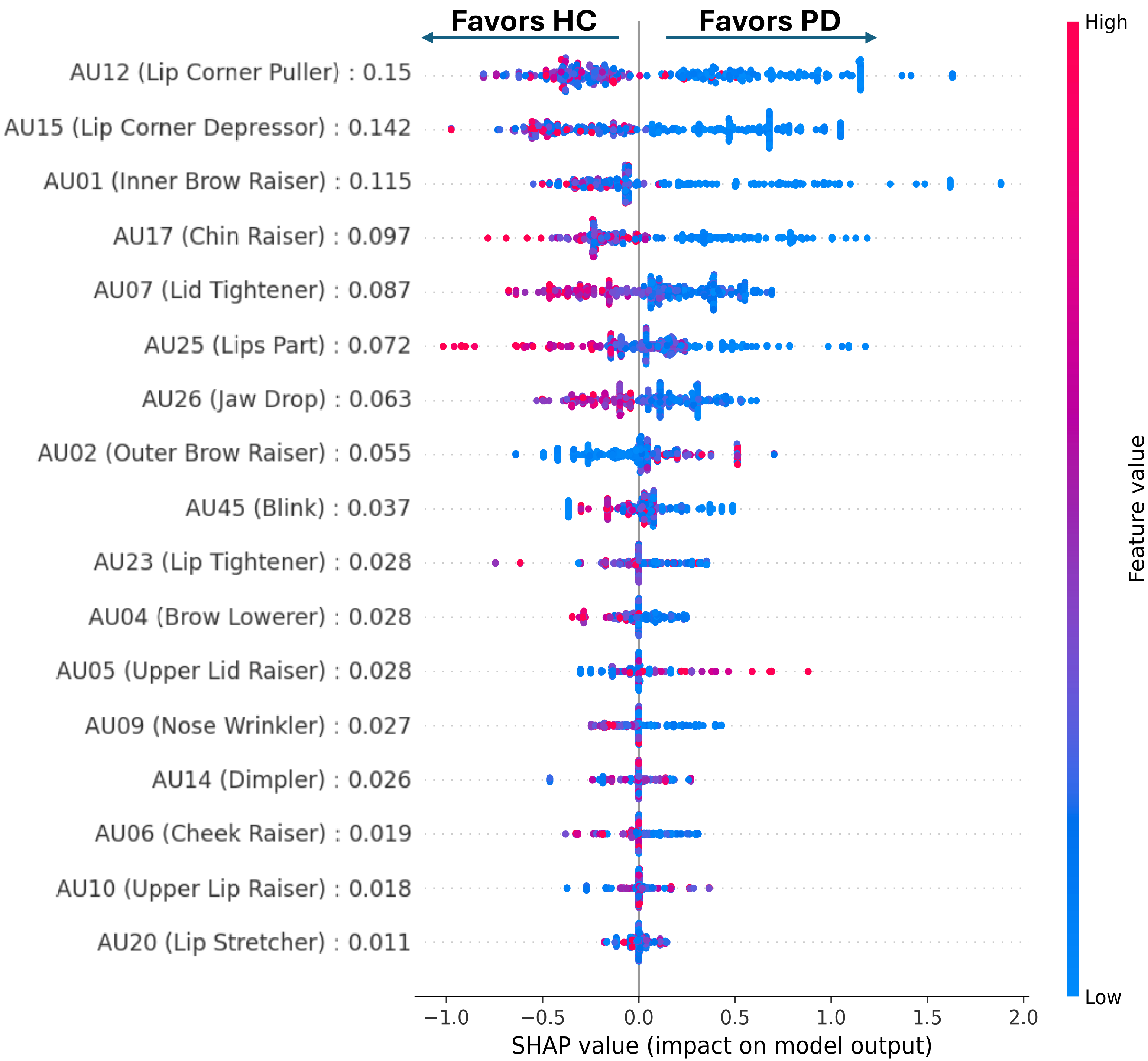
[] Anas Filali Razzouki, M.A. El-Yacoubi, et al. (2024) "Leveraging Action Unit Derivatives for Early-Stage Parkinson's Disease Detection" **Innovation and Research in BioMedical engineering (IRBM).**

Interpretability with $AbsVar(\Delta_k AUs)$ Based Model SHAP Technique

- Task = \pataka\
- Each dot = one AU feature for a video
- Lower feature values favor PD prediction
- Higher feature values favor HC prediction

✓ Consistent with hypomimia definition

- Top 7 AUs:
    - AU12, AU15, AU1, AU17, AU7, AU25, AU26
        - Importance score $> (1/17) = 0.058$ = random score
    - Discriminate more PD vs. HC
    - Mostly located in the mouth region
        - Except AU1 (inner brow raiser), AU7 (lid tightener)
    - ✓ Consistent with speech task scenario



Visualization of Important AUs on the ICEBERG Database

Important AUs found by SHAP
- AU12 : Lip Corner Puller
- AU01 : Inner brow raiser
- AU17 : Chin raiser

- Dynamic automatic encoding of AUs across frames
- Arrow length = Intensity of AU
  - Higher Intensity of AU ~
  - Lower Intensity of AU ~

Lower AU Intensities

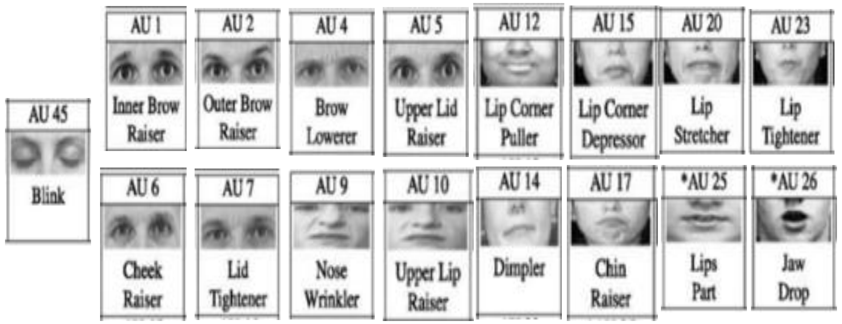
PD Person

Healthy Person

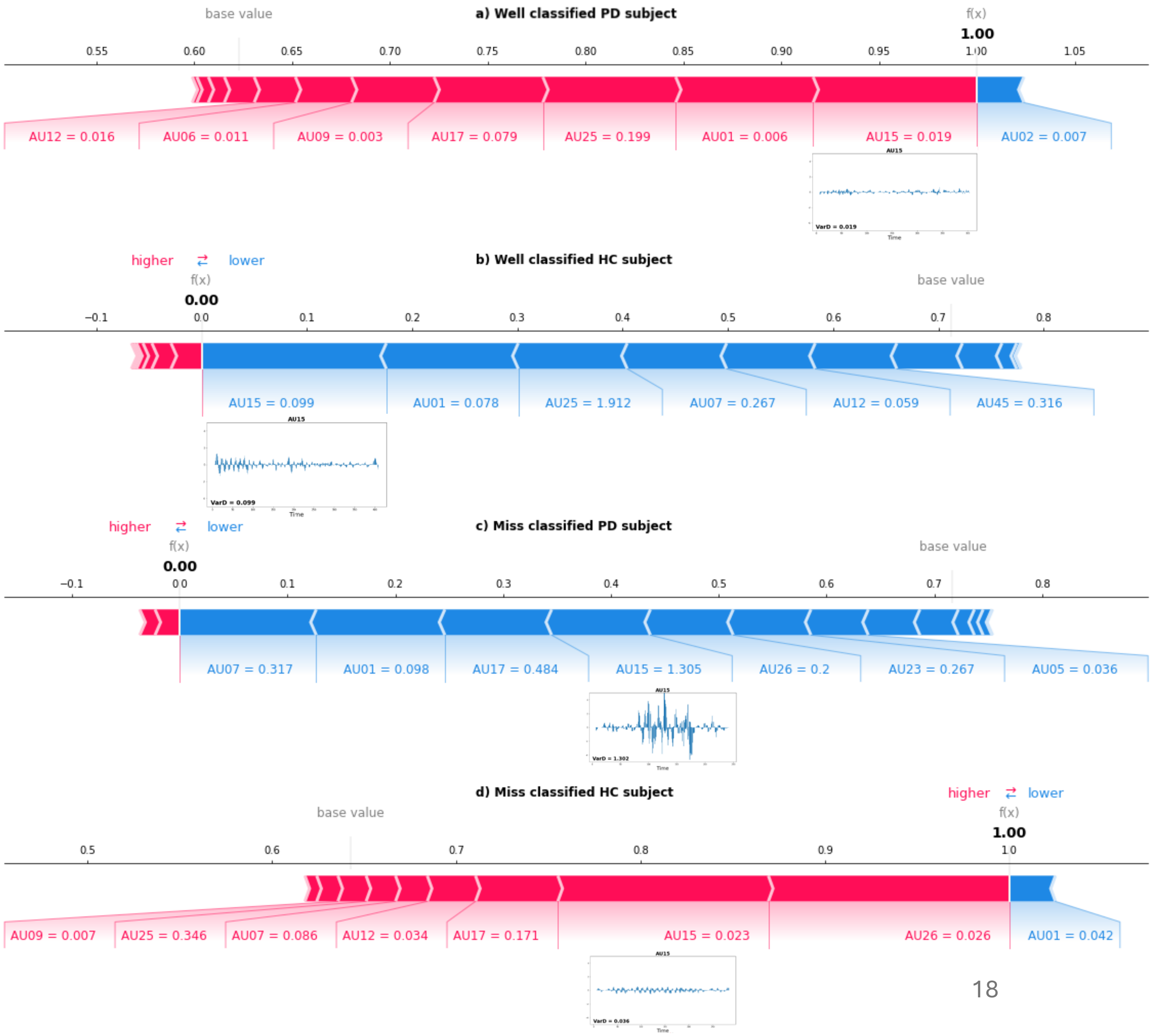
Higher AU Intensities

- Healthy person has higher intensity activation of important AUs compared to PD person

Shap values calculated for 4 individual examples



	sex	age at V0	Statut	wc/mc	pr(PD)	UPDRS 3	Hahn Yahr	mds face
a : 238_V0	M	71	PD	wc	0.98	35	2	1
b : 182_V1	M	41	HC	wc	0.02	4	0	0
c : 247_V0	M	73	PD	mc	0.02	28	2	1
d: 218_V1	M	59	HC	mc	0.98	0	0	0



Sex Effect and Longitudinal Analysis

• Sex Effect

- **Multivariate Analysis (XGBoost)**
  - Males and females have similar AUCs of ~ 84%
  - Males have 9% higher **recall** than females
    - Avg. MDS-UPDRS3 is 5 points higher for males
- **Univariate Analysis**
  - Linear mixed model (factors: PD and Sex)
  - 7 AUs significantly linked to PD
    - Mouth: AU14 (dimpler), AU25 (lips part), AU26 (jaw drop)
    - Eye: AU45 (blink), AU7 (lid tightener), AU4 (brow lowerer)
  - AU45 (blink), AU14 (dimpler) significantly linked to sex
    - Well known in the literature
  - No significant interaction between disease & sex
    - PD effect on AUs is sex-neutral
  - PD Left-onset women blink less than right-onset
    - side-onset = Side of first motor symptoms appear

• Longitudinal Analysis

- **Multivariate Analysis (XGBoost)**
  - AUC at V_0 and V_f are similar AUCs of ~ 78%
  - Recall at V_0 is 7% higher than at V_f
    - Avg. MDS-UPDRS3 decrease from V_0 to V_f (3 points)
- **Univariate Analysis**
  - ANOVA at V_0 and V_f (separately)
  - 4 AUs significantly linked to PD at V_0, not V_f

➔ PD detection harder at V_f than V_0 (not significantly)

  - Patients' Levodopa dosage rose by 60% from V_0 to V_f
  - Patients were recorded while ON medication state

V_0 : Initial Visit, V_f : Final Visit

[] Anas Filali Razzouki, M.A. El-Yacoubi, et al. (2025) "Clinical Interpretability of Parkinson disease's detection based on Facial Action Units" **npj Parkinson's Disease Journal**.

Correlation between AUs and Clinical & DATScan scores

- AU features = $AbsVar(\Delta_k AUs)$ from /pataka/
- The clinical score are measured in both the OFF and ON state

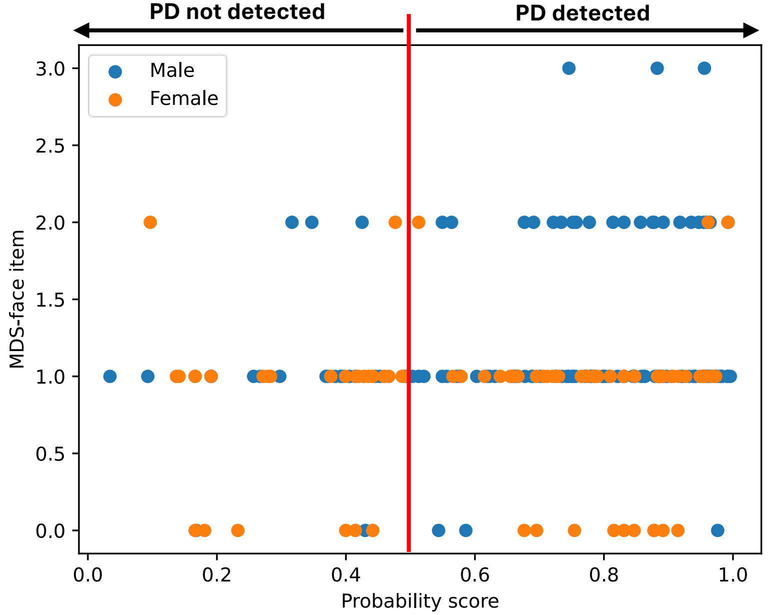
Motor Clinical Items		**Agility**		**Rigidity**				**Bradykinesia**	**Total MDS-UPDRS3**
Limb Side		Left	Right	Upper Left	Upper Right	Lower Right	Neck	All	All
AUs		AU17 (Chin raiser) 	AU15 (Lip corner) 	AU17 (Chin raiser) 	AU07 (Lid tightener) 	AU07 (Lid tightener) 	AU25 (Lips part) 	AU01 (Inner brow raiser) 	AU01 (Inner brow raiser) 
Spearman (r)	OFF Med state	-0.34	**-0.42**	-0.34	**-0.38**	-0.32	-0.3	-0.37	-0.31
	ON Med state	-	-	-	-0.37	-	-	-0.31	-

- Negative significant correlations ($p < 0.05$) between key AUs and some clinical scores
- AUs strongly significant with clinical scores include important AUs found by SHAP for PD vs. HC
- AUs show stronger and more correlations OFF state compared to ON state
  - Patients recorded within 12 hours of morning medication intake

  ➔This may reduce the consistency of ON-state measures, while offering OFF-state measures a stable baseline
- No statistical significance between the 17 AUs and DATScan or MDS-face
  - For Dat-scan, possibly due to only 18 PD patients considered
  - For MDS-face, may be due to dominant class being 1

Relationship Between MDS-Face Scores and PD Predictions

- MDS-UPDRS3 Face item (mds_3_2) used by clinicians to assess the degree of hypomimia
  - 0 : Normal (no hypomimia)
  - 1 : Minimal (only decreased blinking frequency)
  - 2 : Mild: decreased blinking frequency + mask-like facies in the face lower part
  - 3 : Moderate: Mask-like facies with sometimes separated lips when rested mouth



MDS-UPDRS3 face item	No. Videos	Detection Rate	Detection probability
3 (moderate hypomimia)	3	100%	> 0.75
2 (mild hypomimia)	30	83%	Most detections > 0.65
1 (minimal hypomimia)	147	76%	Mixed range of values
0 (no hypomimia)	21	62%	Mixed range of values

Observations

- Trend observed: detection rates increases as MDS-face scores increase
- MDS-Face Score = 0 (no hypomimia): detection rate of 62%
  - $AbsV\_\Delta_{k*}AUs$ **encode subtle muscles movements at very early stage**
- Ability to detect hypomimia even for MDS-UPDRS3 face item = 0

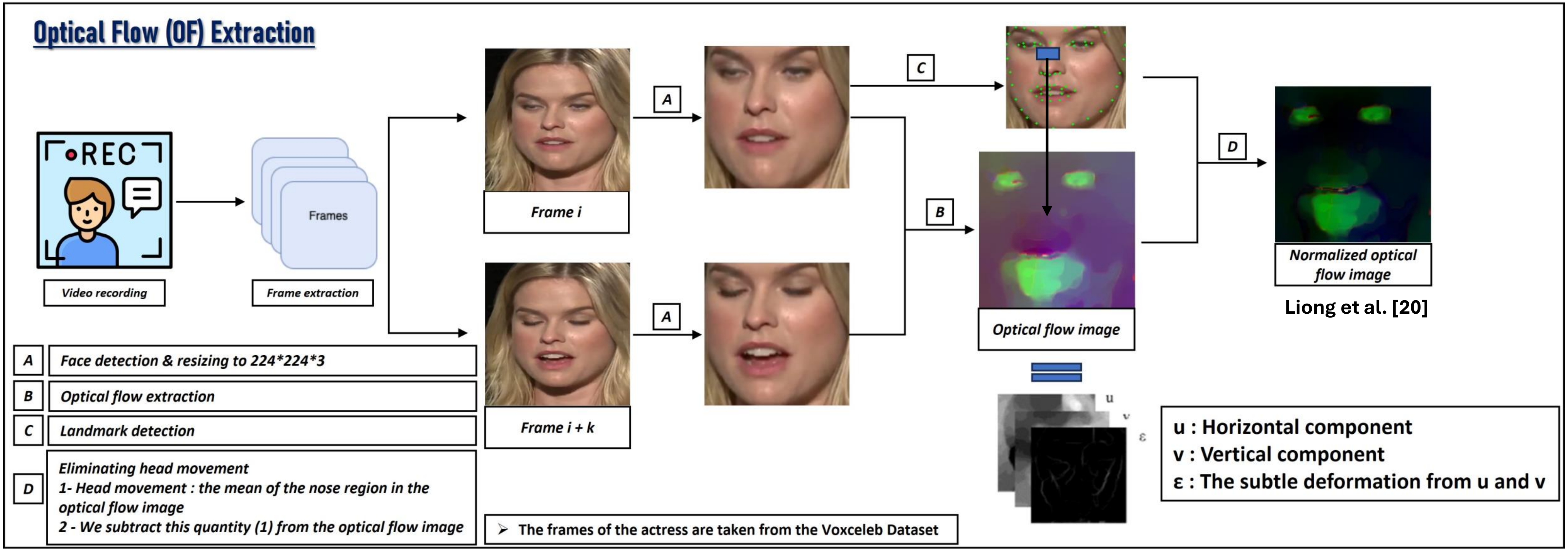
➔ *Potential of our scheme to support clinicians for early-stage detection of hypomimia*

Vision Foundation Models for PD Analysis

- Automatic features based on transformers and optical flow (OF)
- Combined RGB and OF modalities for robust PD analysis
- Explainability: link auto-extracted features to AUs
- Fusion of AUs, OF-based and RGB-based transformer classifiers

Optical Flow Extraction

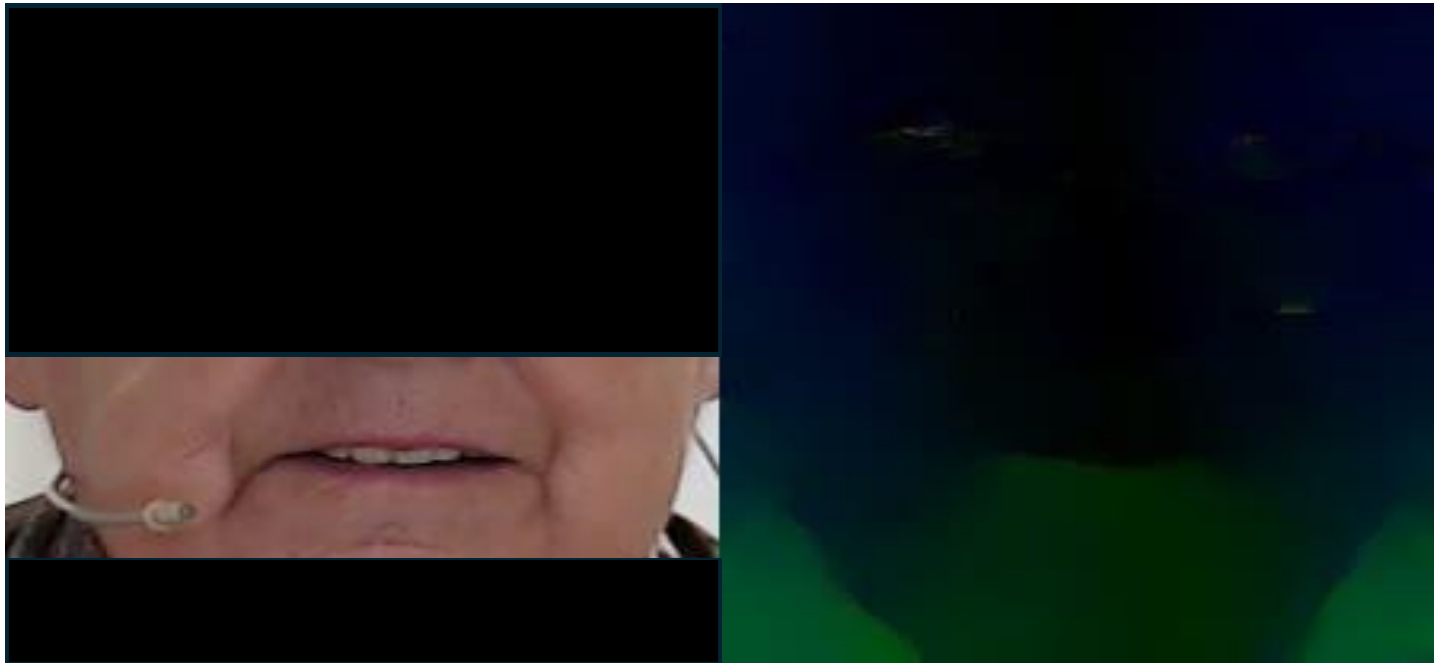
- Movement Encoding: Optical flow between $Frame(i)$ and $Frame(i + k)$
  - The movement is encoded at the pixel levels rather than within specific regions, as seen with ($\Delta_k AU(i)$)
  - Optimal step k^* found with the previous experiments based $\Delta_k AU(i)$



Visualization of Optical Flow Extraction for the ICEBERG data

- Video of PD patient performing /pataka/ speech task
- The optical flow is calculated with step $k = k^* = 6$
  - $k^* = 6$ was the optimal step found with /pataka/ speech task with $\Delta_k AU$

RGB Optical Flow



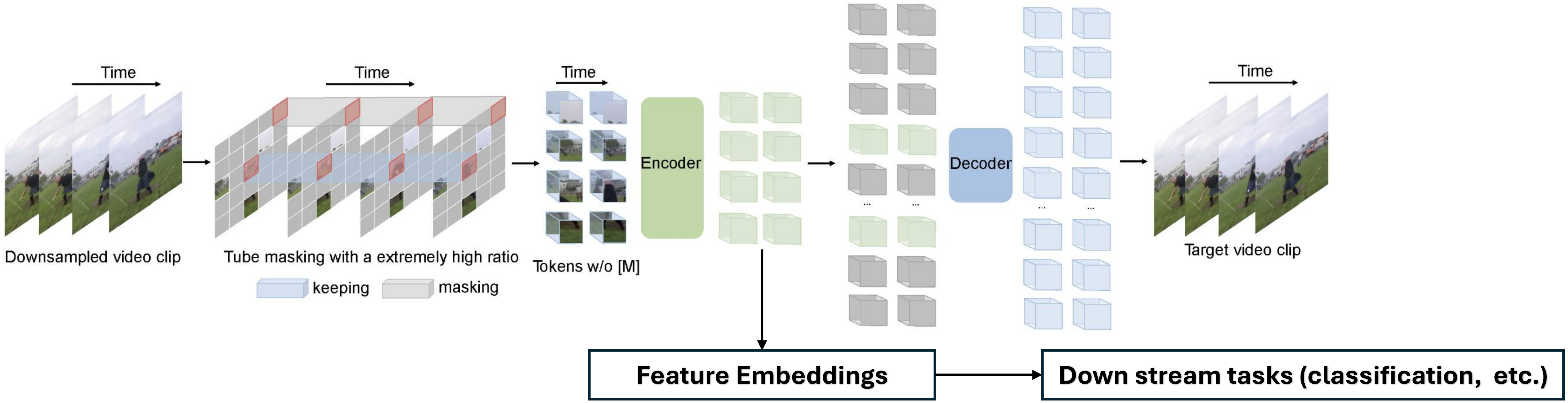
- Optical flow components :
  - Vertical component (v) : green channel
  - Horizontal component (u) : blue channel
  - Subtle deformation from u and v : red channel

Self-Supervised Video Pre-training (SSVP)

- **Masked Autoencoding**: A technique to reconstruct masked or corrupted inputs
  - Applications:
    - NLP: Predicts masked words (e.g., BERT)
    - Computer Vision
      - Image Pre-Training: Learns spatial patterns by reconstructing masked image regions
      - Video Pre-Training: Self-Supervised Video Pre-Training (SSVP)
- **Self-Supervised Video Pre-Training (SSVP)**
  - Key Idea: Captures spatial and temporal patterns by masking and reconstructing video cubes
  - Advantages:
    - Uses unlabeled video data
    - Captures temporal & spatial patterns
- **Prominent Foundation Models-based SSVP Method**
  - **VideoMAE: Video Masked Auto Encoder**
  - MARLIN: Masked Autoencoder for facial video Representation LearnINg
  - V-JEPA: Video-based Joint-Embedding Predictive Architecture

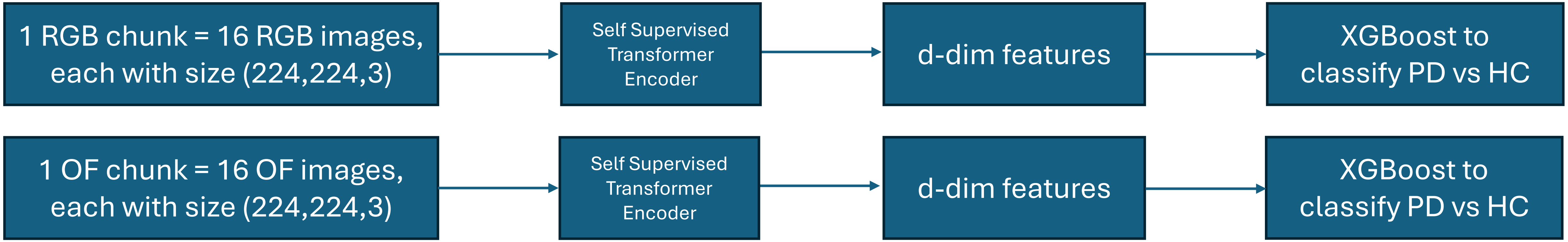
Video Masked Auto Encoder (VideoMAE) Model

- Foundation Model-based Video Vision Transformer (Encoder-Decoder Architecture)
- Masks and reconstructs tubes within videos to learn temporal and spatial dynamics
- VideoMAE is pre-trained on 409,000 videos from two datasets:
  - Kinetics-400 dataset: Contains 400 action classes
  - Something-Something V2 dataset: Contains 174 motion-centric action classes
- VideoMAE ranked among the top 5 state-of-the-art models for action recognition
  - Action recognition datasets: HMDB-51, Something-Something V2, UCF101, and AVA v2.2 datasets



Feature Extraction based on Foundation Models with Vision

- **Video Decomposition:** each video is decomposed into non-overlapping chunks
  - Each chunk with dimension 16×224×224×3, consisting of optical flow (OF) or RGB images (224×224×3)
- **Feature Extraction:**



Embedding dimension = 768 for VideoMAE and MARLIN
Embedding dimension = 1024 for V-JEPA

- **Evaluation metrics:** Area under the curve (AUC), balanced accuracy (BA)
- **Validation technique:** 5-fold nested cross-validation

Results: PD vs. HC Classification Based on Self Supervised Transformer Encoder-Based Optical Flow or RGB

- Classifier: XGBoost
- Features
  - RGB local embedding features from FM-ViViTs
  - OF local embedding features from FM-ViViTs

Results

- /pataka/ & /bagada/: OF achieved 10% higher AUC than RGB
- Monologue: OF and RGB achieved a similar AUC of 82%
- RGB: Monologue achieved 10% higher AUC than DDK tasks
- Monologue provides:
  - 2.5x more training data than /pataka/
  - 4.5x more training data than /bagada/

  ➔advantageous given the high dimensionality of the training data
- VideoMAE outperforms V-JEPA and MARLIN

  ➔Continue working only with VideoMAE

Task	FM-ViViT	Type	Optical Flow (OF)		RGB	
			AUC (%)	BA (%)	AUC (%)	BA (%)
/pataka/	V-JEPA	VB	73,4	70	62	61,7
		SB	78.6 ± 3.7	73.2 ± 4.0	65.6 ± 4.6	62.6 ± 4.3
	MARLIN	VB	73	66,5	65,5	60,4
		SB	75.2 ± 4.0	68.7 ± 4.1	65.5 ± 4.6	60.5 ± 4.3
	VideoMAE	VB	74,6	69,3	65,2	59,5
		SB	**79.1 ± 3.6**	75.1 ± 3.8	**69.6 ± 4.4**	62.0 ± 4.2
/bagada/	V-JEPA	VB	73	66,4	65,2	61,5
		SB	**79.2 ± 3.6**	70.8 ± 3.9	68.8 ± 4.4	62.4 ± 4.1
	MARLIN	VB	68,3	62	61,9	56,6
		SB	70.5 ± 4.3	64.5 ± 4.2	66.1 ± 4.6	63.5 ± 4.1
	VideoMAE	VB	70,4	64	66,9	61,9
		SB	74.2 ± 4.0	67.3 ± 4.1	**69.4 ± 4.4**	59.3 ± 4.3
Monologue	V-JEPA	VB	74,6	65,8	75,7	69,3
		SB	78.6 ± 3.7	73.8 ± 3.9	79.2 ± 3.6	72.0 ± 4.0
	MARLIN	VB	78,7	71,1	68,7	65,7
		SB	81.8 ± 3.4	75.5 ± 3.9	74.7 ± 4.0	69.4 ± 3.9
	VideoMAE	VB	78,8	72,4	78	75,5
		SB	**82.2 ± 3.3**	76.1 ± 3.8	**81.8 ± 3.4**	78.4 ± 3.2

Interpretability with Embedding Features with SHAP Technique for PD vs. HC

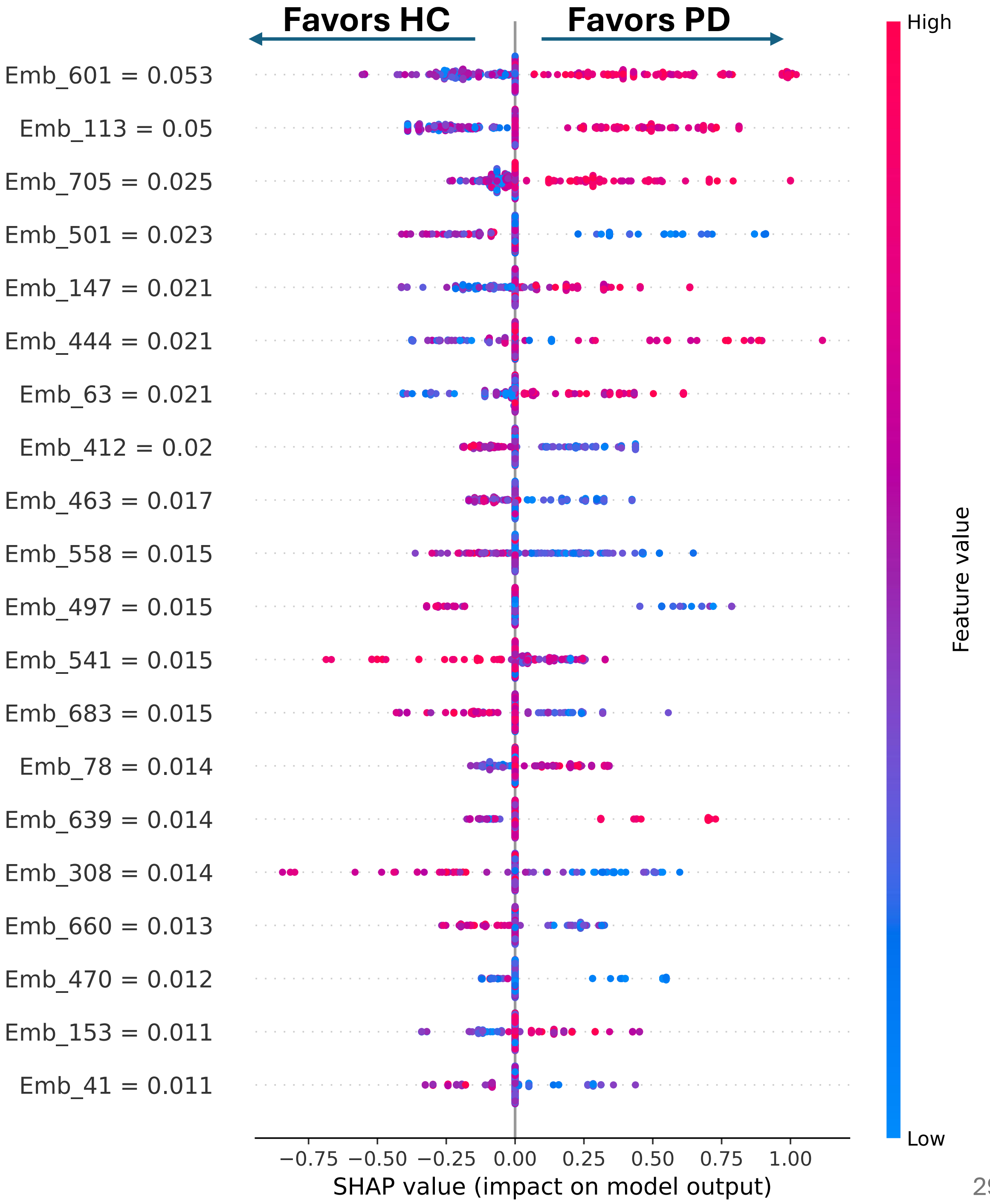
- Task = Monologue
- Features: Global embedding-based VideoMAE from OF
- Classifier: XGBoost
- Each dot = one embedding value of video

Results

- Top 2 features: Emb_601 and Emb_113

	Sperman Correlation (p<0.05)		
$AbsVar(\Delta_k AUs)$	**Emb_601**	**Emb_113**	**Emb_147**
AU26 (jaw drop)	**-0,64**	**-0,6**	**--**
AU23 (lip tightener)	-0,62	-0,53	--
AU06 (nose wrinkler)	-0,55	-0,48	-0.22

- Emb_601 and Emb_113 correlate negatively with AUs
  - Higher Feature Values (lower movement) favors PD prediction
  - Lower Feature Values (higher movement) favors HC prediction
  - ✓ Consistent with hypomimia definition



Fine-tuning the Foundation Model

- Task: **Monologue**, Modality: Optical Flow
- VideoMAE Encoder Fine-Tuning:
  - Add classification layer with 2 classes to VideoMAE encoder
  - 86M total parameters; we finetune only half (42M)
    - Preliminary Findings: Fine-tuning half of the layers performs better than other fine-tuning strategies
  - Validation technique: cross-validation with 100 epochs, without any optimization
    - Time Constraints: Training 100 epochs takes 17 hours per model on the cluster; nested CV not feasible

Method	Modality	AUC (%)	BA (%)	Rec (%)	Spe (%)
Finetune half the layers	VB	79	74,2	68,3	80
	SB	84.6 ± 3.1	79.4 ± 3.4	74.3 ± 4.2	84.4 ± 5.4
Extracted Features + XGBoost	VB	78,8	72,4	74,8	70
	SB	82.2 ± 3.3	76.1 ± 3.8	78.9 ± 3.9	73.3 ± 6.6

- Fine-tuning improved AUC by 2.5% over XGBoost with extracted embeddings
- Potential Improvement: Use nested CV with 5 validation folds for better optimization

Fusion of AUs, OF-Based & RGB-Based VideoMAE Classifiers

- Task : **Monologue**
- AUs-based experiment
  - Fusion of 2 global features-based classifiers: $Hist\_entropy$(Abs($\boldsymbol{\Delta_{k^*} AUs}$)) and 75_percen(Abs($\boldsymbol{\Delta_{k^*} AUs}$))
- OF-Based and RGB-Based VideoMAE experiment
  - Fusion of 2 global features-based classifiers : RGB-Based VideoMAE and OF-Based VideoMAE
- Fusion Experiment (based on averaged probabilities across both approaches)

Approach	Type	AUC (%)	BA (%)	Recall (%)	Specificity (%)
1) Δ_{k^*} AUs	VB	80.7	73.7	80.7	66.7
	SB	82.9 ± 3.3	76.4 ± 3.9	81.7 ± 3.7	71.1 ± 6.8
2) VideoMAE (OF & RGB)	VB	79.8	69.1	78.2	60
	SB	84.6 ± 3.1	75.7 ± 3.9	82.6 ± 3.6	68.9 ± 6.9
Fusion of (1) and (2)	VB	82.6	71.3	79.2	63.3
	SB	**85.9 ± 2.9**	78.4 ± 3.7	83.5 ± 3.6	73.3 ± 6.6

- AUs Experiment
  - Achieved AUC: 82.9%
- VideoMAE (OF & RGB) Experiment
  - Achieved AUC: 84.6%
- Fusion Experiment:
  - Achieved AUC : 85.9 %

Audio Digital Markers (DM) for PD Detection based on Deep Learning (DL) Foundation Models

ICEBERG subset for DM based on DL

- With high quality recordings in hospital
- ICEBERG subset until **july-2022** (355 subjects)
- 2 labels: healthy and Parkinson (281 subjects)
- Keep only subjects who have at least 1 session with enough speech tasks ➔ 267 subjects
- Use all good quality sessions for training
- Data:
  - 156 males, 111 females
  - 156 Parkinson, 111 healthy
  - Split subjects into 5 folds for cross validation

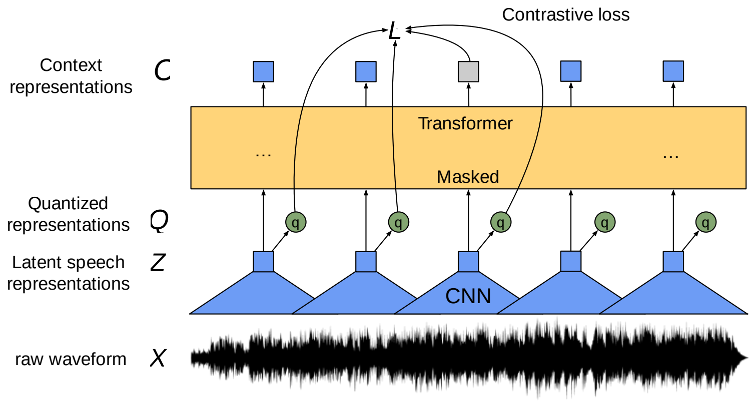
Deep-learning based acoustic features: SOTA Speech Foundation Models

- Speech Foundation models
- Harnessed in 2 ways
  - To extract deep feature representations, taken then as input to Machine Learning classifiers
  - Finetuned on PD/HC Datasets to act as standalone classifiers
- Foundation models considered in our work
  - Wav2vec2.0 (Meta)
  - Whisper (OpenAI)
  - SeamlessM4T (Meta)

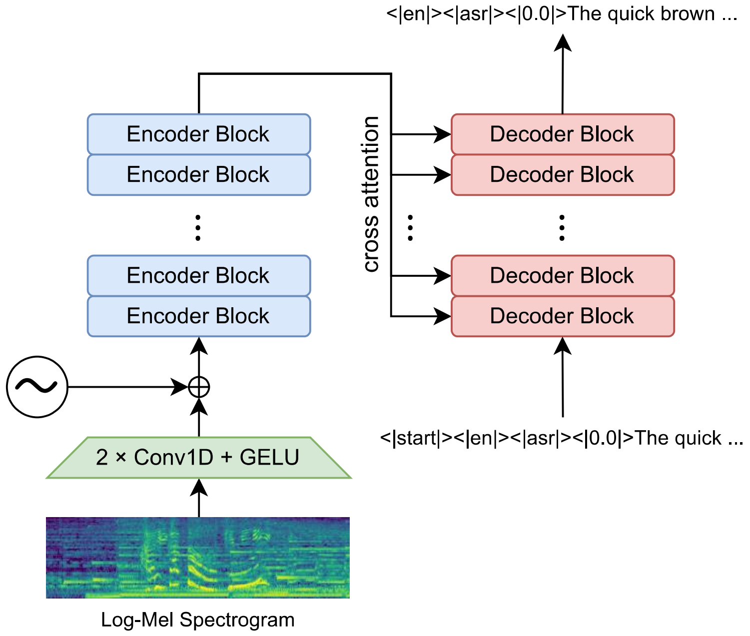
- [1] Baevski A, Zhou Y, Mohamed A, Auli M (2020), « wav2vec 2.0: A framework for self-supervised learning of speech representations. » Adv Neural Inf Process Syst (NeurIPS) 33:12449–12460
- [2] A. Radford, J. W. Kim, et al. (2022) , “Robust speech recognition via large-scale weak supervision,” Tech. Rep., OpenAI.
- [3] Barrault, Loïc, et al. (2023) "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation." arXiv preprint arXiv:2308.11596

Speech Foundation Models

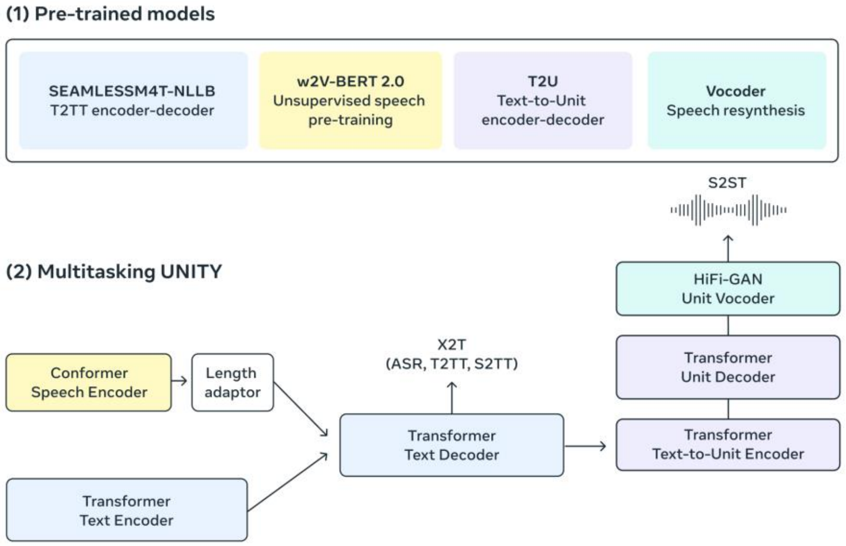
■ Wav2vec2.0
- Pretrained on 53k hours of unlabeled data
- Use raw waveform as input
- Consists of 1DCNN feature encoder and Transformer context network



■ Whisper [2]
- Pretrained on 680k hours of labeled data
- Use Log-Mel spectrogram as input
- Encoder-Decoder Transformer architecture



■ SeamlessM4T
- First foundation model for speech.
- Pretrained on 1 millions hours of unlabeled speech data
- Finetuned to do multiple speech-related tasks



Finetuning foundation models on the Iceberg dataset for PD classification

Model	Train data	Validation data	Gender	AUC	Precision	Recall	F1
wav2vec2.0 finetuned	**speech tasks**	**speech tasks**	**all**	**89.41**	**83.66**	**86.49**	**85.05**
wav2vec2.0 finetuned	speech tasks	speech tasks	male	91.76	85.15	92.47	88.66
wav2vec2.0 finetuned	speech tasks	speech tasks	female	85.38	80.77	76.36	78.5
wav2vec2.0 pretrained + SVM	speech tasks	speech tasks	all	88.53	75.86	89.19	81.99

- With a large enough dataset such as ICEBERG, it is possible to finetune a foundation DL model for better results on PD classification
- Classification performance for males is generally better than for females

[] Dao, Q., & El-Yacoubi, M. A. , et al. (2025). Detection of Early Parkinson's Disease by Leveraging Speech Foundation Models. **IEEE Journal of Biomedical and Health Informatics (JBHI)** (pp. 1–10). https://doi.org/10.1109/jbhi.2025.3548917

Future Directions

- Short-Term:
  - Develop a dual-stream FM-ViViT architecture combining OF and RGB
  - Integrate facial AUs into FM-ViViT for enhanced performance
  - Facial AUs to analyze hypomimia in iRBD patients (at risk of developing PD)
  - Fusion Facial and Audio Digital Markers
- Mid-Term:
  - Incorporate medication timing and dosage into models
    - Allows of finer modeling of PD severity
- Long-Term:
  - Stratify patients based on disease progression using AUs
    - Not feasible with current database (PD patients, on average had only 2 videos visits over 5 years)
      - Necessity to have more participants' longitudinal video recordings
- Vision for Clinical Impact:
  - Develop clinician-friendly tools and software for early diagnosis and monitoring
  - Integrate technology into telemedicine platforms to enhance remote PD management

Thank you

Any questions ?