M. Mavroforakis, H. Georgiou

# Genetic Profiling of Olives for Location of Origin and Variety Discrimination Using Machine Learning

2022 IEEE International Conference on Internet of Things and Intelligence Systems

mmavrof@inttrust.gr
hgeorgiou@unipi.gr

**IoTaIS '22** @ 24-26 Nov. 2022, Bali, Indonesia

# Summary

- Genetic profiling in olives
- Problem statement (DA/ML)
- Methodology highlights
- Pre-processing pipeline
- ML models training
- Results & Discussion
- Enhancements & Future work
- Conclusions

**Main idea:**

- Quantify DNA-specific discriminatory properties via targeted biomarkers
- Extract "fingerprints" from the genome via data-driven feature extractors (pre-processing)
- Exploit discrimination capabilities to address top-level challenges, i.e., location of origin and variety

# Genetic profiling of olives

- **Quality assurance and auditing,** extremely important in food supply chain
- **Polymerase Chain Reaction (PCR)** used for DNA replication from leaf and fruit
- **High Resolution Melt (HRM)** with fluorescence tracking using biomarkers
- Output: HRM transition profiles (curves) of targeted genome regions

Challenge: Can we exploit these HRM profiles for data-driven top-level discrimination tasks, such as location of origin, olive variety, sample purity ?

# Problem Statement

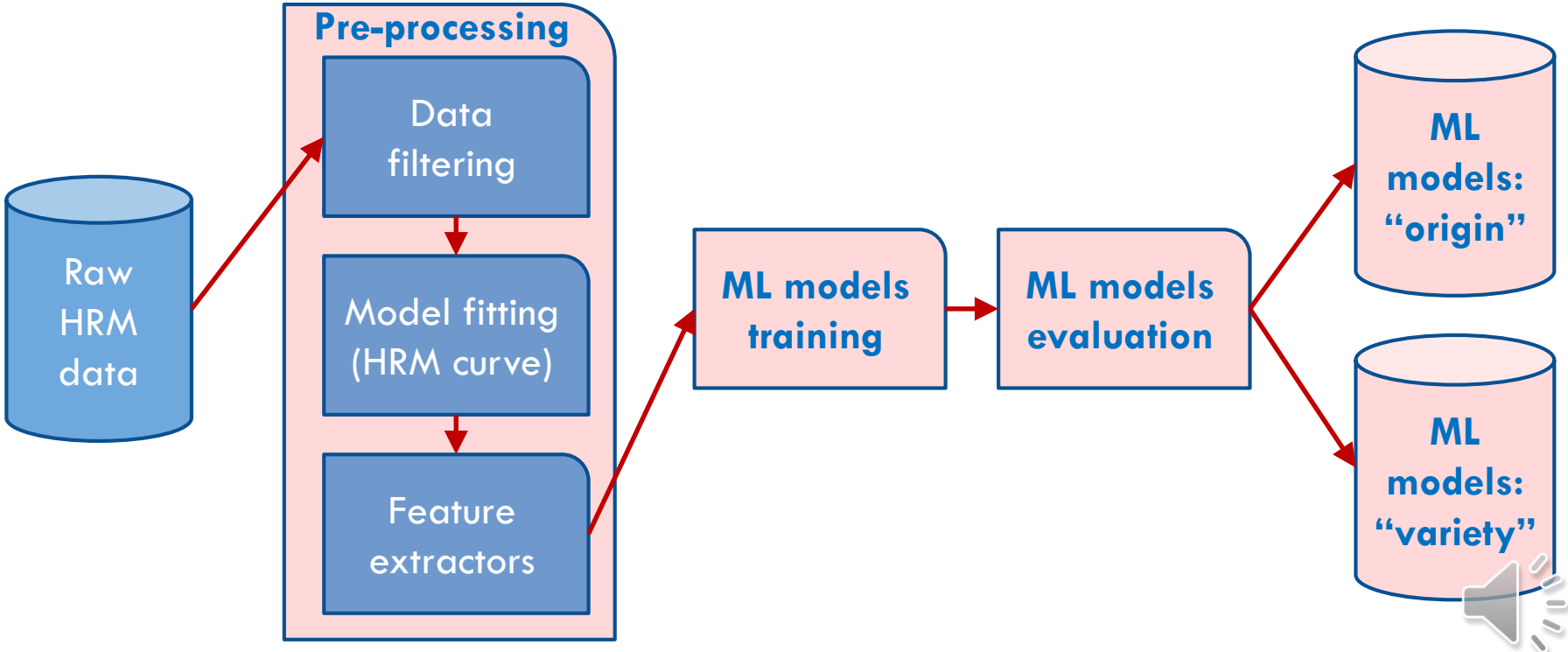
Task to be addressed:

A data-driven, computationally efficient processing sequence for 'fingerprinting' and discriminating DNA strands of olives with regard to:
(a) their variety, namely "Task 1"
(b) their location of origin, namely "Task 2"

Constraints for the proposed solution:

- **Robust genetic profiling,** that can be used as a 'fingerprint' of the source product throughout the food supply chain
- **Lightweight models,** readily applicable to Internet-of-Things (IoT), edge computing and microcontroller (MCU) applications
- **Production-grade performance,** i.e., efficient and accurate model

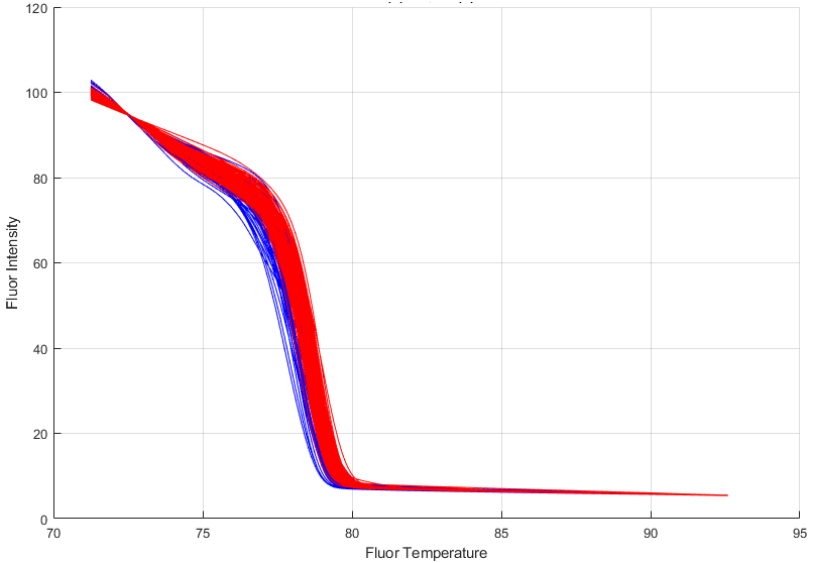
# Methodology

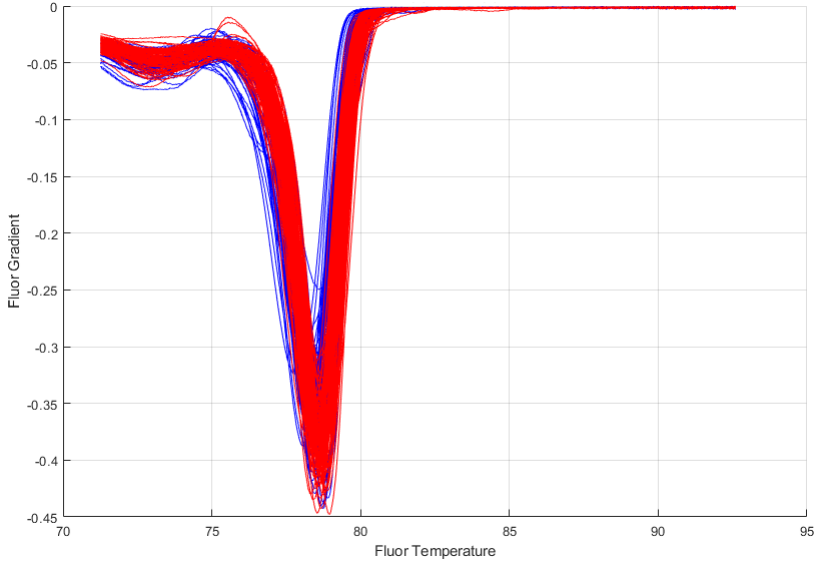


# Pre-processing pipeline

1. Raw data series import (HRM curves)
2. Visual inspection of the data, removal of errors
3. Data filtering for noise removal
4. 1st derivative data series generation
5. Data corrections / pre-fitting adjustments (prefix)
6. Curve-fitting functions (Gaussian, rational)
7. Curve-fitting quality metrics (RMSE)
8. Visual assessment / validation of curve-fitting
9. Statistical filtering for curve-fitting quality
10. Feature vectors to training dataset

# Pre-processing pipeline (*cont.*)




HRM curves of fluorescence-versus-Temperature for two olive varieties
(left = raw series, right = 1st derivative series)

# Pre-processing pipeline (*cont.*)

Gaussian fit :

$$f_g(x) = \alpha\, e^{\left(\frac{x-\mu}{\sigma}\right)^2}$$

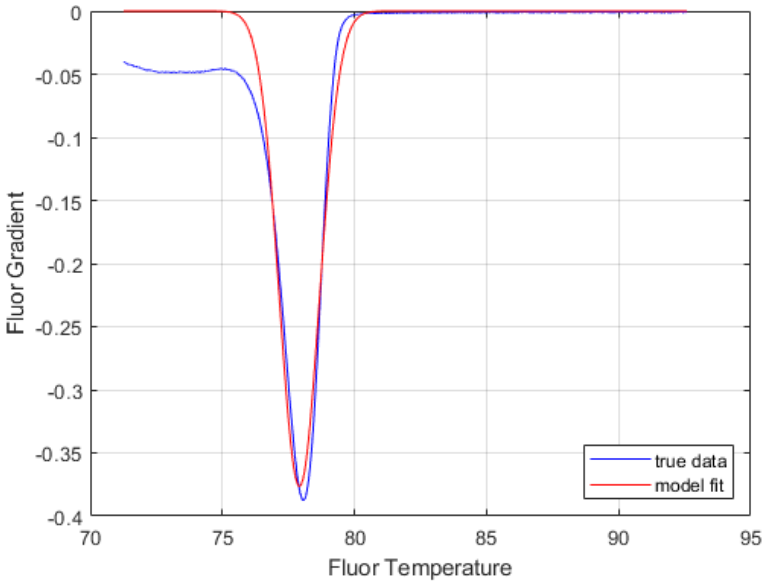
Rational polynomial fit :

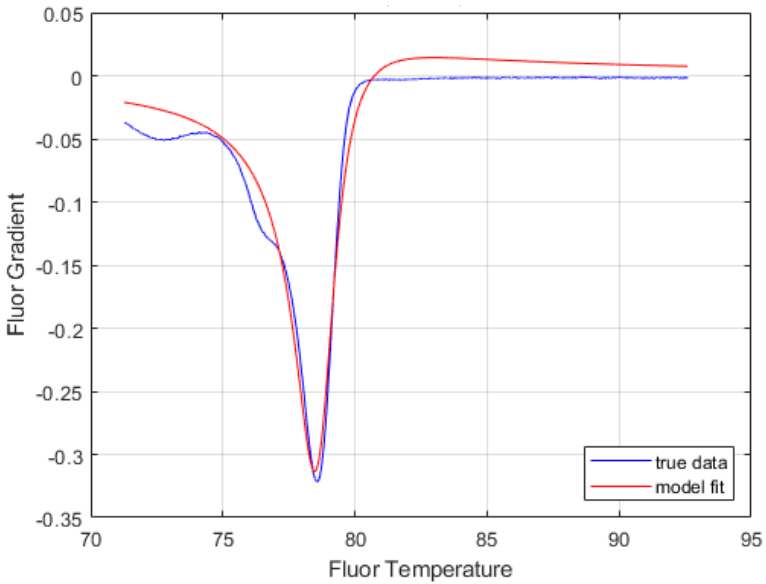
$$f_r(x) = \frac{p_1 x + p_2}{x^2 + q_1 x + q_2}$$

prefix weighting :

$$\widehat{y'_t} = y'_t\,(t/\tau)^k\,,\ \forall\, 0 \le t \le \tau$$

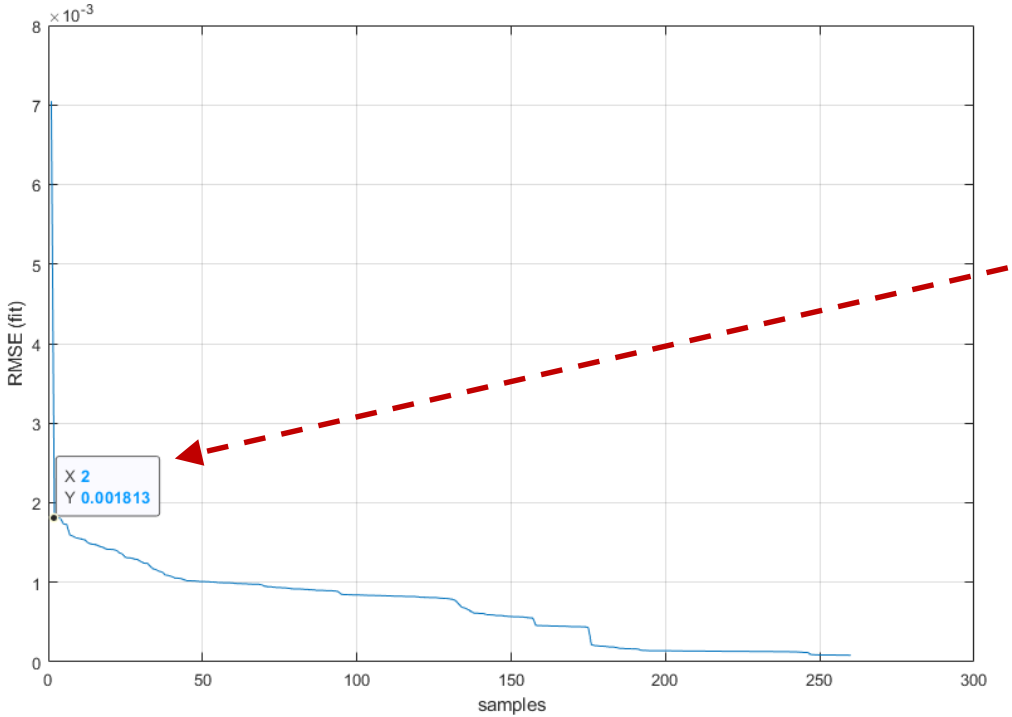
# Pre-processing pipeline (*cont.*)




Examples of HRM curve fitting on 1st derivative series with weighted Gaussian (left) and rational polynomial (right) models for one olive variety biomarker

# Pre-processing pipeline (*cont.*)




RMSE criterion for both Gaussian and Rational model fitting is used to flag HRM-curve outliers, i.e., typically bad samples or biomarker response.

# Pre-processing pipeline (*cont.*)

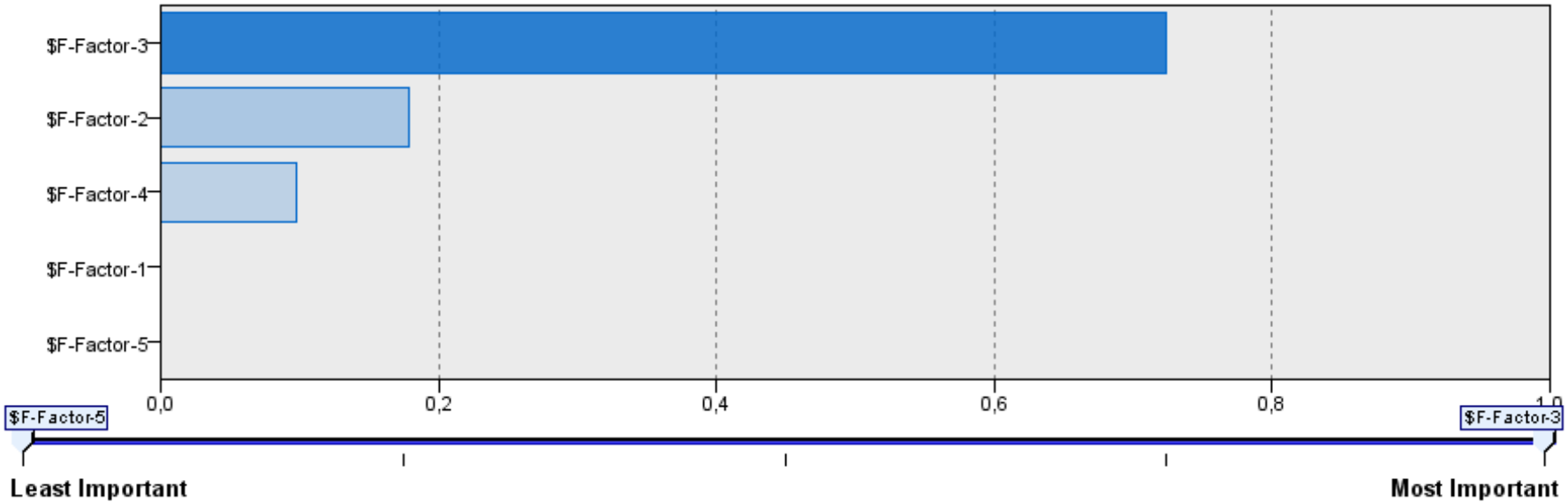
| Symbol | Data | Description |
|---|---|---|
| MK | X | Biomarker used {1,2,3} |
| MIN | S | Minimum value |
| MAX | S | Maximum value |
| ENT | S | Entropy (Eq.4) |
| FVg1 | D | Gaussian fit, scale ($\alpha$) |
| FVg2 | D | Gaussian fit, mean ($\mu$) |
| FVg3 | D | Gaussian fit, sigma ($\sigma$) |
| FVg4 | D | Gaussian fit, error ($RMSE_g$) |
| FVq1 | D | Rational poly fit, numerator 1 ($p_1$) |
| FVq2 | D | Rational poly fit, numerator 2 ($p_2$) |
| FVq3 | D | Rational poly fit, denominator 1 ($q_1$) |
| FVq4 | D | Rational poly fit, denominator 2 ($q_1$) |
| FVq5 | D | Rational poly fit, error ($RMSE_r$) |

*ML model training*

| Feature | Task 1 | Task 2($vK$) | Task 2($vT$) |
|---|---|---|---|
| MK | | | x |
| MIN | | x | x |
| MAX | | x | x |
| ENT | | | x |
| FVg1 | | | x |
| FVg2 | | | x |
| FVg3 | x | x | |
| FVg4 | | x | x |
| FVq1 | x | x | x |
| FVq2 | | | |
| FVq3 | | | |
| FVq4 | | | x |
| FVq5 | x | | |

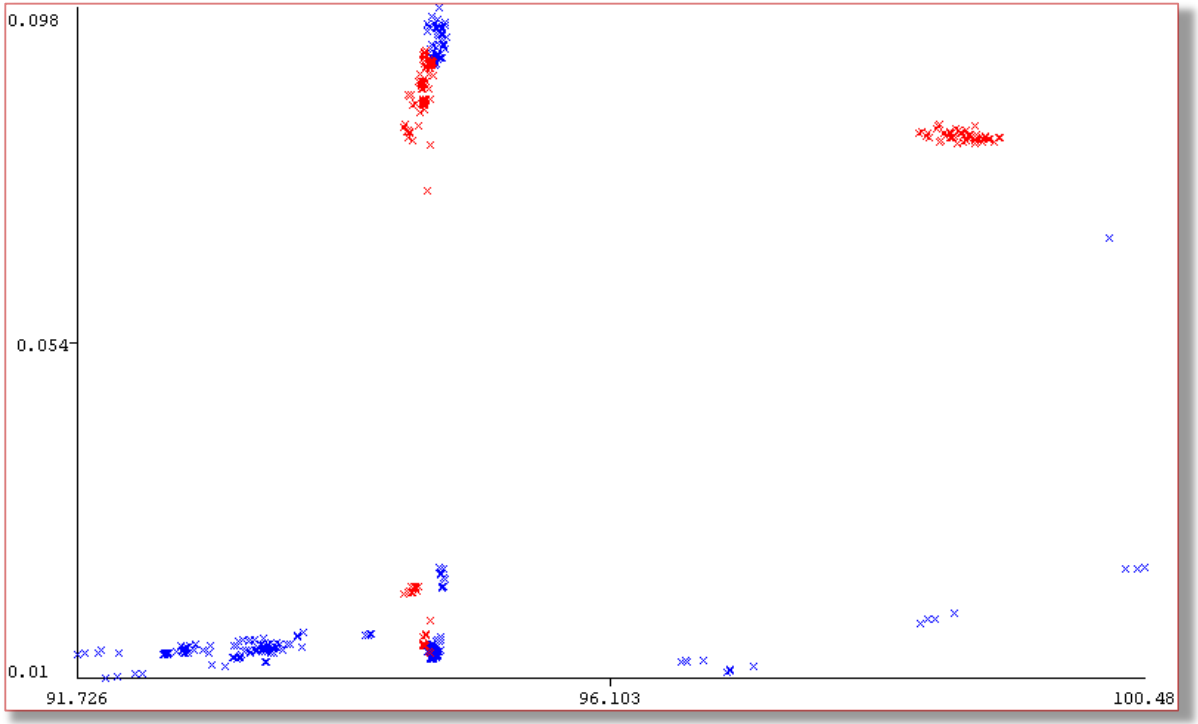
Variety labels: vK = “Koroneiki”, vT = “Tsounati”

# ML training




Feature set compactness after PCA transformation (robust, low-dimensionality)

# ML training (*cont.*)




Example: Geographical discrimination (Task 2) of the same variety (vK) from the regions of Chania (vC=blue) and Rethymnon (vR=red); the plot illustrates the two of the most statistically significant features (X=MAX, Y=FVg4).

# Results

TASK 1: OLIVES VARIETY

| *Classifier models* | *Acc%* |
|---|---|
| **Kstar** | **96.48** |
| **IBkLG** | 94.13 |
| **AdaboostM1** | 94.72 |
| **MultiboostAB** | 95.01 |
| **Random Forest** | 94.72 |
| **Rotation Forest** | 94.43 |

Note: Acc% deviation in all cases is $\leq \pm 1\%$.

# Results *(cont.)*

TASK 2: LOCATION OF ORIGIN

| *Classifier models* | *Acc% (vK)* | *Acc% (vT)* |
|---|---|---|
| **Kstar** | **99.25** | **99.61** |
| **IBkLG** | **99.25** | 98.83 |
| **AdaboostM1** | 98.13 | 98.05 |
| **MultiboostAB** | 97.76 | 97.66 |
| **Random Forest** | 98.88 | 99.22 |
| **Rotation Forest** | 98.51 | 98.44 |

Note: Acc% deviation in all cases is $\leq \pm 1\%$.

# Conclusions

- ✓ Efficient HRM data series filtering and quantification of DNA-specific discriminatory properties via targeted biomarkers
- ✓ Robust “fingerprint” creation from the genome via data-driven HRM-based feature extractors (pre-processing)
- ✓ Highly accurate discrimination capabilities addressing top-level challenges, i.e., location of origin and variety of olives
- ✓ Lightweight design, validated feasibility study, proper for production-level evolution and integrated solutions, e.g. IoT-enabled.

# Enhancements – Future Work

- Inclusion of more biomarkers for targeted discrimination in olive genome
- Experiment with alternative HRM curve fitting models, e.g. GMM
- Experiment with other HRM modalities, e.g. DNA helicity-versus-temperature
- Test weaker / simpler ML models for IoT implementations

# Thank you
# Questions?

IoTaIS '22 @ 24-26 Nov. 2022, Bali, Indonesia

Michael Mavroforakis (MSc,PhD) – Email: mmavrof@inttrust.gr

Harris Georgiou (MSc,PhD) – Email: hgeorgiou@unipi.gr