Data-driven model for fingerprinting plastic waste material using low-cost spectroscopy

Marcin Pietrasik, PhD

Department of Advanced Computing Sciences

Who am I?

- Postdoctoral researcher in data fusion in the Department of Advanced Computing Sciences at Maastricht University
- PhD in Electrical and Computer Engineering from the University of Alberta, Canada

Prof. Anna Wilbik

Prof. Paul Grefen

Dr. Sin Yong Teng

Prof. Kim Ragaert

Agenda

- Three works:
  - The enabling technologies for digitalization in the chemical process industry **(published)**
  - Capturing variability in material property predictions for plastics recycling via machine learning **(published)**
  - Data-driven model for fingerprinting plastic waste material using low-cost spectroscopy **(in progress)**

The enabling technologies for digitalization in the chemical process industry

Motiation

- **Chemical process industry lags** behind other sectors in digitalization
- Barriers: **high costs**, **lack of digital culture/training**, **unclear vision**
- **Objective:** Provide a **structured overview** of enabling technologies for practitioners with little/no digitalization background
- Goals:
  - Identify and categorize enabling technologies
  - Link industry **problem domains** with digital **development aspects**
  - Showcase **case studies** of underutilized but high-potential technologies

Problem Domains

Maintenance

Production

Safety

Supply Chain

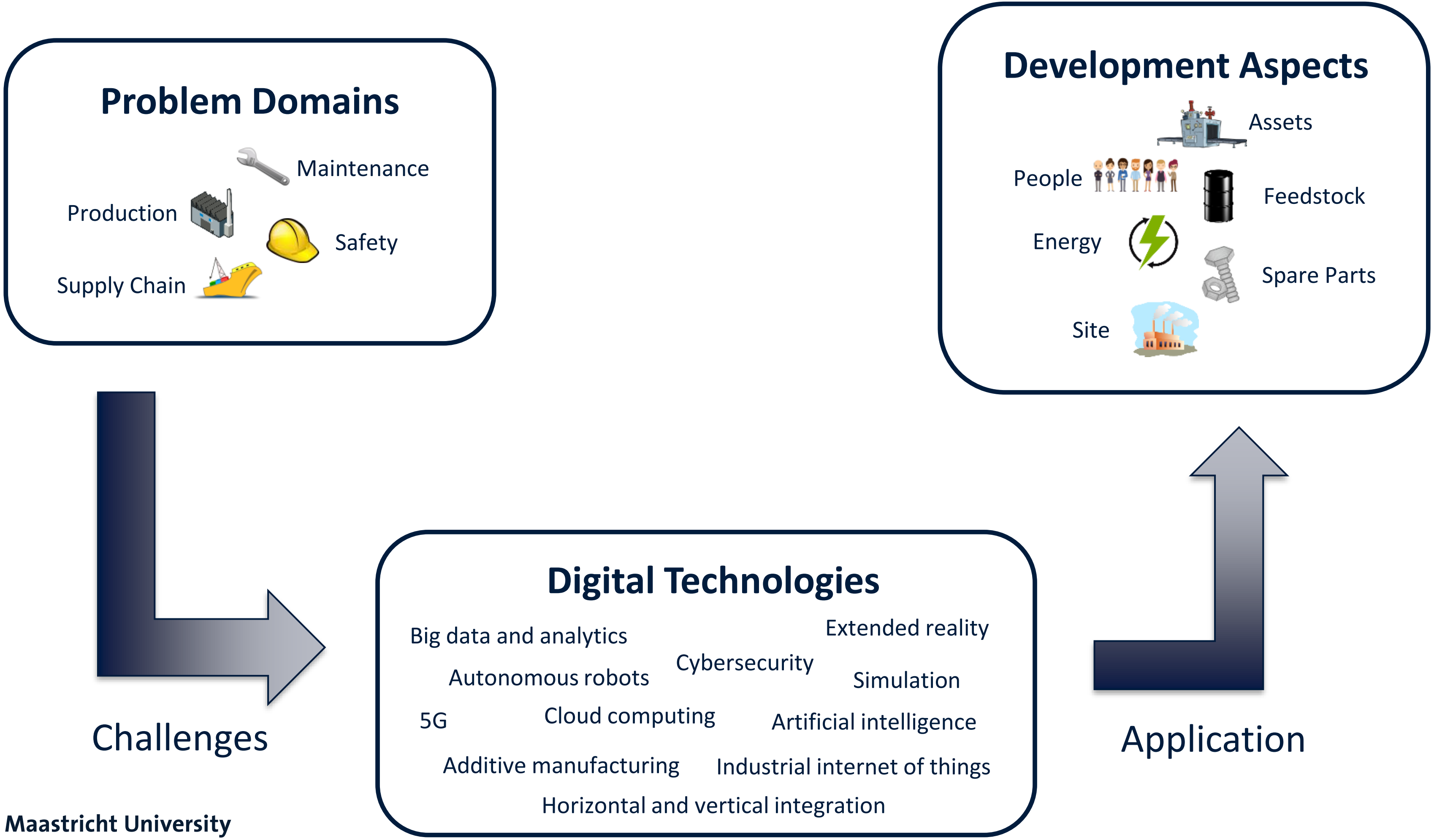
Problem Domains

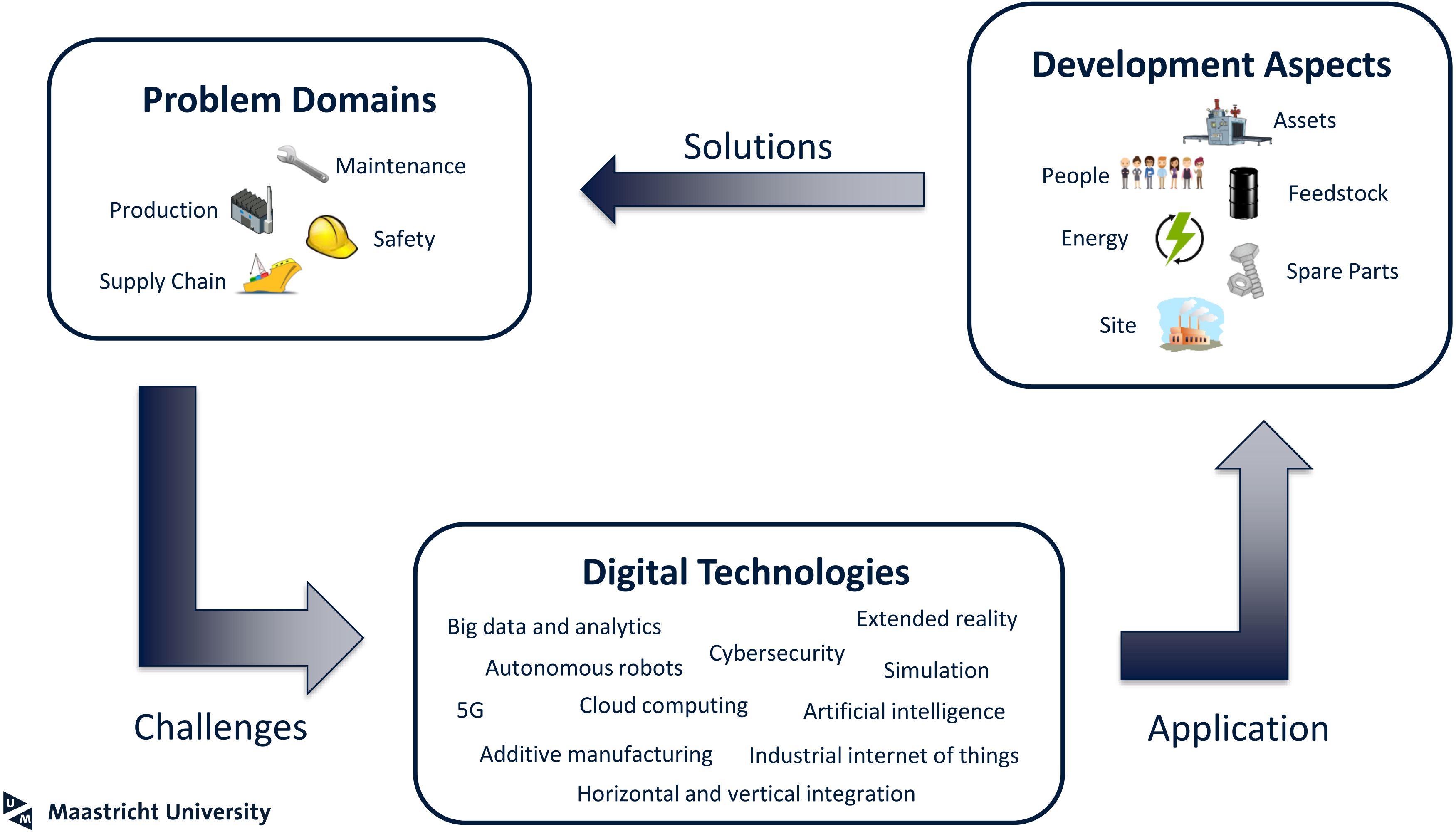
- Maintenance
- Production
- Safety
- Supply Chain

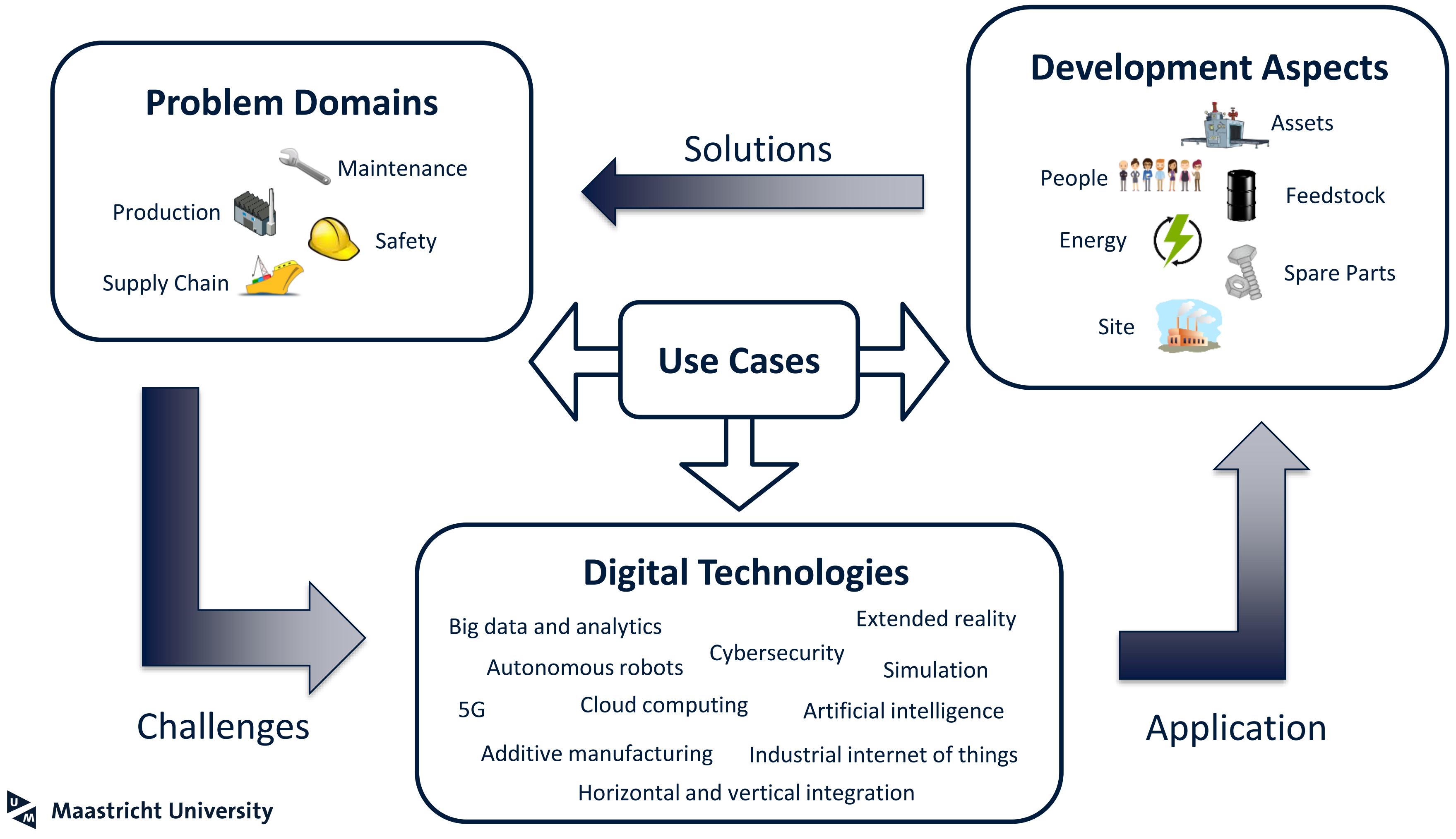
Challenges

Digital Technologies

- Big data and analytics
- Extended reality
- Cybersecurity
- Autonomous robots
- Simulation
- 5G
- Cloud computing
- Artificial intelligence
- Additive manufacturing
- Industrial internet of things
- Horizontal and vertical integration


Problem Domains
Maintenance
Production
Safety
Supply Chain
Challenges
Digital Technologies
Big data and analytics
Extended reality
Cybersecurity
Autonomous robots
Simulation
5G
Cloud computing
Artificial intelligence
Additive manufacturing
Industrial internet of things
Horizontal and vertical integration
Application
Development Aspects
Assets
People
Feedstock
Energy
Spare Parts
Site


Problem Domains
Maintenance
Production
Safety
Supply Chain
Solutions
Development Aspects
Assets
People
Feedstock
Energy
Spare Parts
Site
Challenges
Digital Technologies
Big data and analytics
Extended reality
Cybersecurity
Autonomous robots
Simulation
5G
Cloud computing
Artificial intelligence
Additive manufacturing
Industrial internet of things
Horizontal and vertical integration
Application
Maastricht University

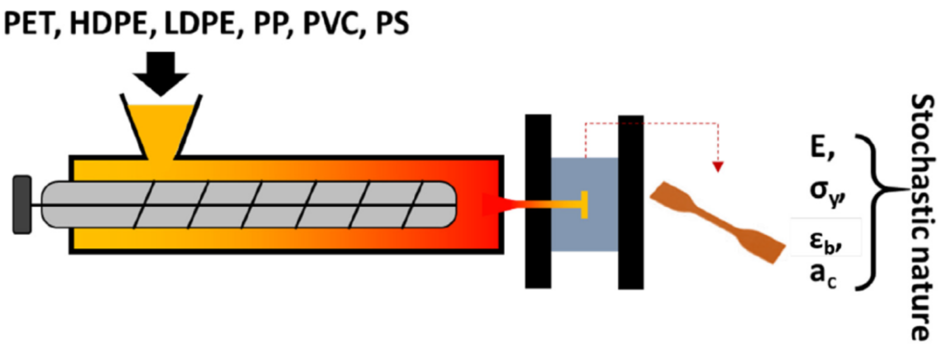


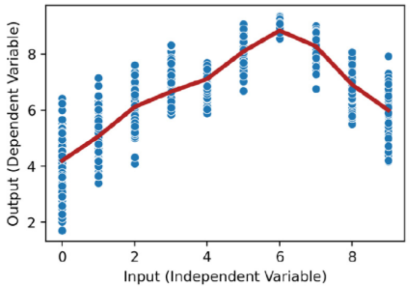
Problem Domains
Maintenance
Production
Safety
Supply Chain
Solutions
Development Aspects
Assets
People
Feedstock
Energy
Spare Parts
Site
Use Cases
Challenges
Digital Technologies
Big data and analytics
Extended reality
Cybersecurity
Autonomous robots
Simulation
5G
Cloud computing
Artificial intelligence
Additive manufacturing
Industrial internet of things
Horizontal and vertical integration
Application
Maastricht University

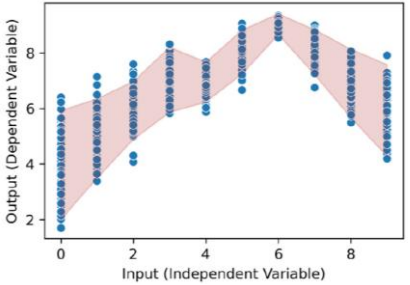
Capturing variability in material property predictions for plastics recycling via machine learning

Motivation

- Blends of waste plastics → **nonlinear and stochastic properties**
- Traditional ML models predict **single values** (point estimates) → insufficient to capture uncertainty
- **Key research gap:** Need methods that predict **ranges of properties** (intervals) to account for variability

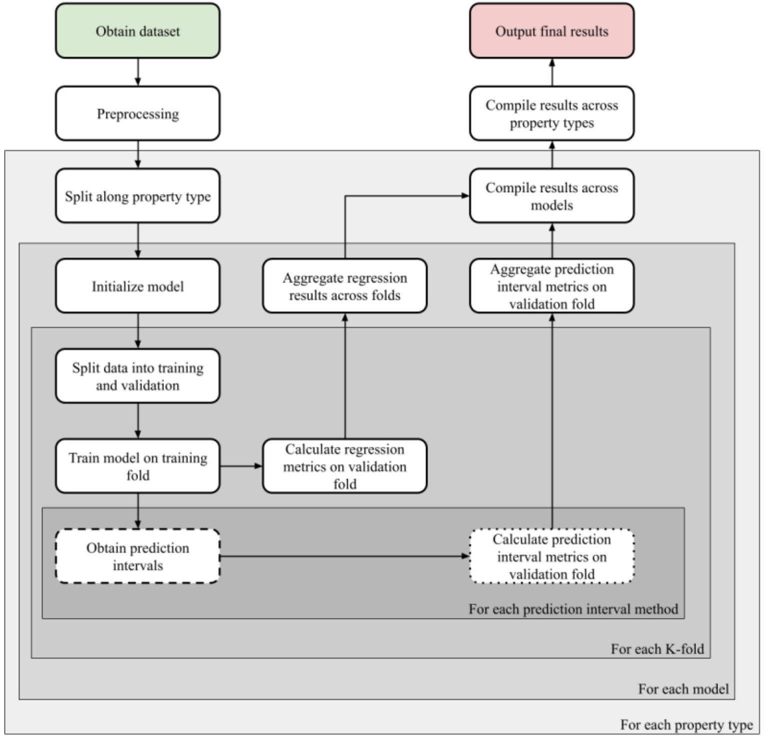






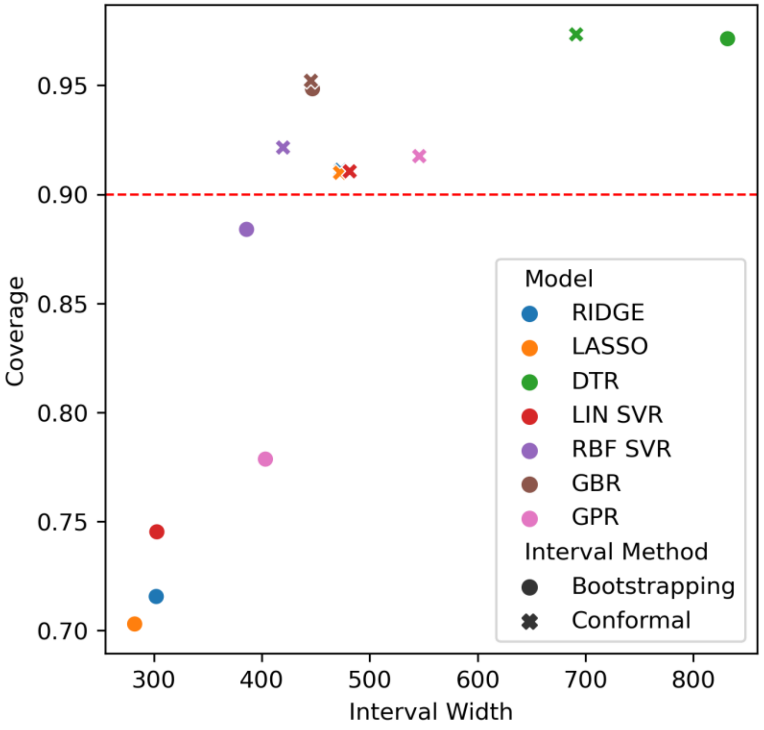
Methodology

- Dataset: **3352 samples**, 10 virgin polymers, tested in blends
- Target properties: Elastic modulus (E), Yield strength (σy), Strain at break (εb), Impact strength (ac)
- Models tested: **Ridge, Lasso, SVR, Decision Trees, Gradient Boosting, Gaussian Process**
- Interval methods: **Residuals, Bootstrapping, Conformal Predictions**
- Evaluation: RMSE, Coverage, Interval Width



Results

- **Gradient Boosting and SVR** best for **point predictions**
- **Conformal predictions** best for **interval accuracy**
- Demonstrates the use of interval-based machine learning to capture the variability in the mechanical properties of heterogeneous plastic blends



Data-driven model for fingerprinting plastic waste material using low-cost spectroscopy

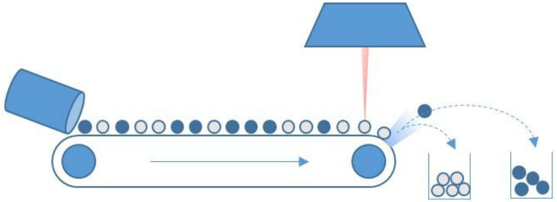
Maastricht University

The Plastic Waste Problem

- Plastics are ubiquitous: durable, lightweight, cheap
- But…
  - 12 billion tonnes projected in landfills and ecosystems by 2050
  - Only 9 percent of plastics globally are recycled with over 70 percent ending up in sanitary landfills and uncontrolled dumpsites (OECD)
  - Mismanaged plastic is a major contributor to the environmental crisis
- Mechanical recycling = scalable circular economy approach
- **Challenge:** plastics are chemically diverse and thus need **accurate separation**
  - Poor sorting → contamination → degraded material quality → reduced economic viability

Spectroscopy in recycling

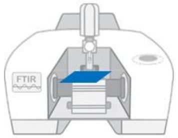
- Vibrational spectroscopy → unique **spectral fingerprints** of polymers
- Advantages: fast, non-destructive, automated
- Near-Infrared (NIR, 1000–2500 nm) widely used with ML
- State of the art → high-cost, wide-band spectrometers
- **Gap:** limited research on low-cost, narrow-band devices (<1100 nm)



Conveyor Spectral Sorting

Pros
- Very fast scanning speed
- Direct sorting of materials

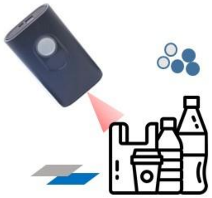
Cons
- Large monetary investments
- Practically immovable



Benchtop Spectroscopy

Pros
- High spectral quality
- Transportable device

Cons
- Inflexible operation
- Sample preparation often needed



Handheld Spectroscopy

Pros
- Low-cost implementation
- Highly portable for many applications

Cons
- Slight lower data quality

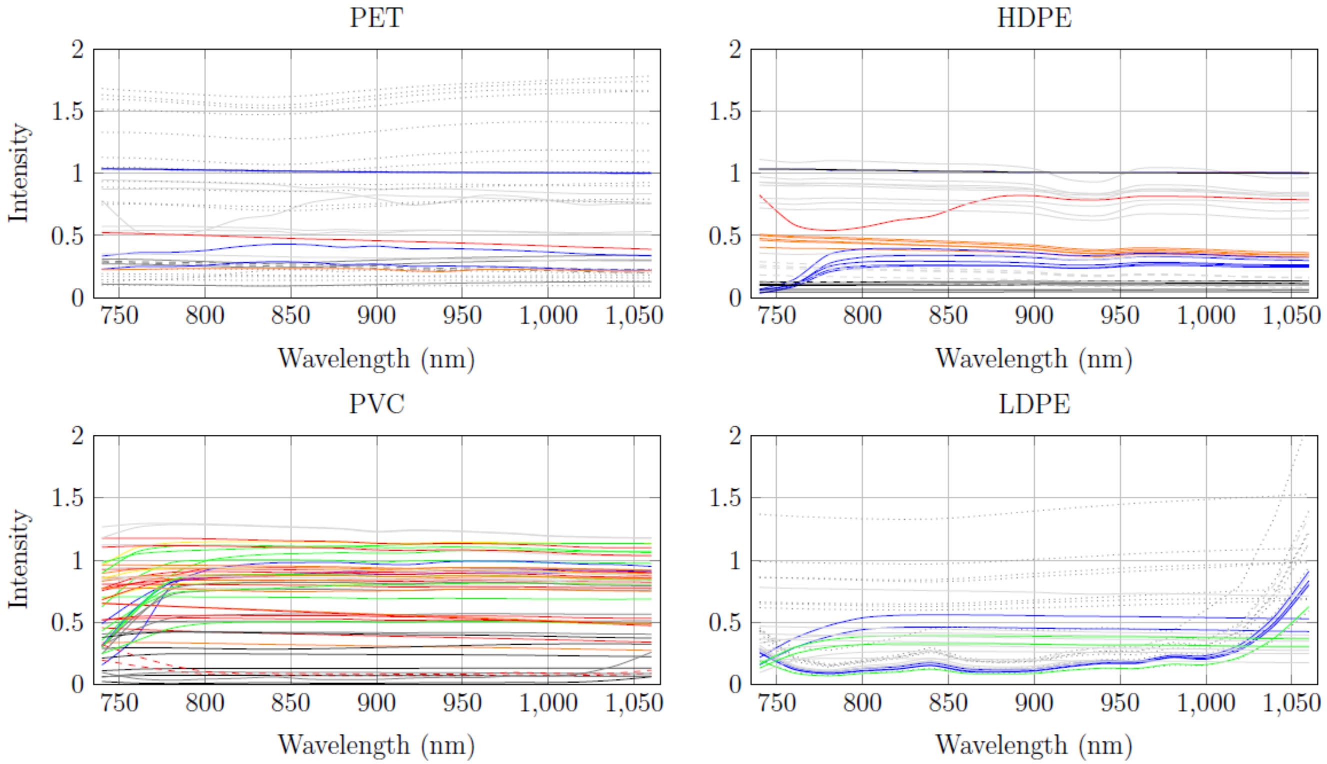
Objectives

- Assess feasibility of low-cost, narrow-band NIR spectroscopy (740–1070 nm) for classifying plastics.
  - Make classification accessible in **resource-limited environments**
  - Potential to enable recycling **at point of collection** → less contamination downstream
- **Research question:** *Can consumer-grade, narrow-band NIR spectroscopy + ML accurately classify plastic waste?*
- **Tasks:**
  - Collect **novel dataset** of household plastics (7 types).
  - Develop **machine learning** pipeline for classification.
  - Benchmark against existing approaches.

Data collection

- Data collection performed by students on-site at Chemelot
- Househouse plastic waste
- Device: **SCiO Mini spectrometer** (740–1070 nm)
- Samples: **363 data points** (7 classes: PET, HDPE, PVC, LDPE, PP, PS, Other)
- Features:
  - **Spectral** (wavelength intensities)
  - **Categorical** (color, transparency)

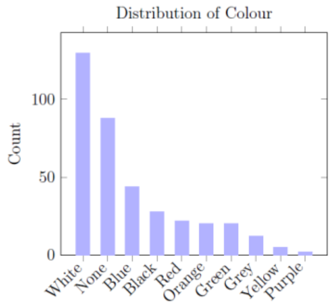
Data collection

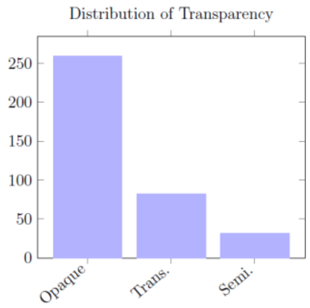


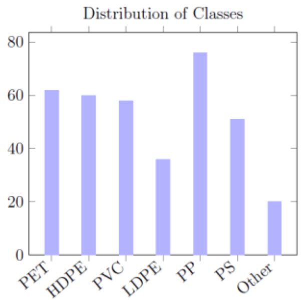
Data insights

- Imbalance observed:
  - Color → white/transparent/opaque overrepresented
  - Class sizes uneven

	White	None	Blue	Black	Red	Orange	Green	Grey	Yellow	Purple	Opaque	Trans.	Semi.
PET	15	36	4	0	1	1	0	5	0	0	26	33	3
HDPE	25	0	17	8	1	9	0	0	0	0	55	0	5
PVC	9	0	3	6	14	7	9	7	3	0	56	0	2
LDPE	14	14	5	0	0	0	3	0	0	0	21	14	1
PP	28	12	13	5	3	2	7	0	4	2	61	10	5
PS	17	23	1	9	0	1	0	0	0	0	28	23	0
Other	19	1	0	0	0	0	0	0	0	0	7	0	13

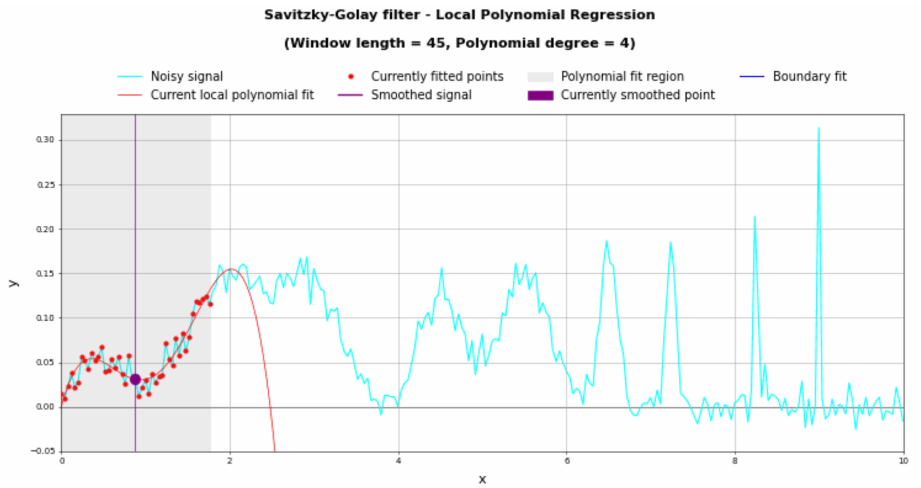


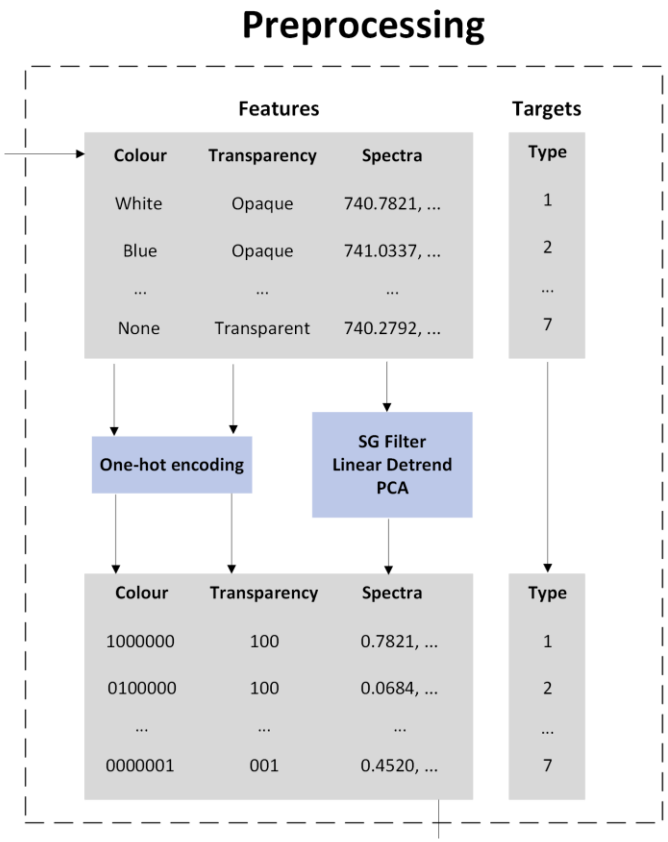




Data preprocessing

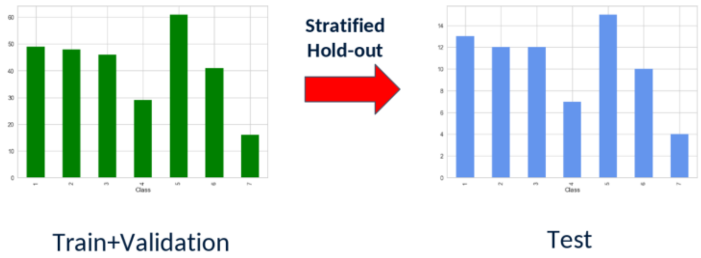
- Categorical features:
  - **One-hot encoding**
- Spectra:
  - **Savitzky-Golay filter** → smooth and enhance spectral features
  - **Linear detrending** → remove baseline shifts
  - **PCA** → dimensionality reduction (model-dependent benefits)

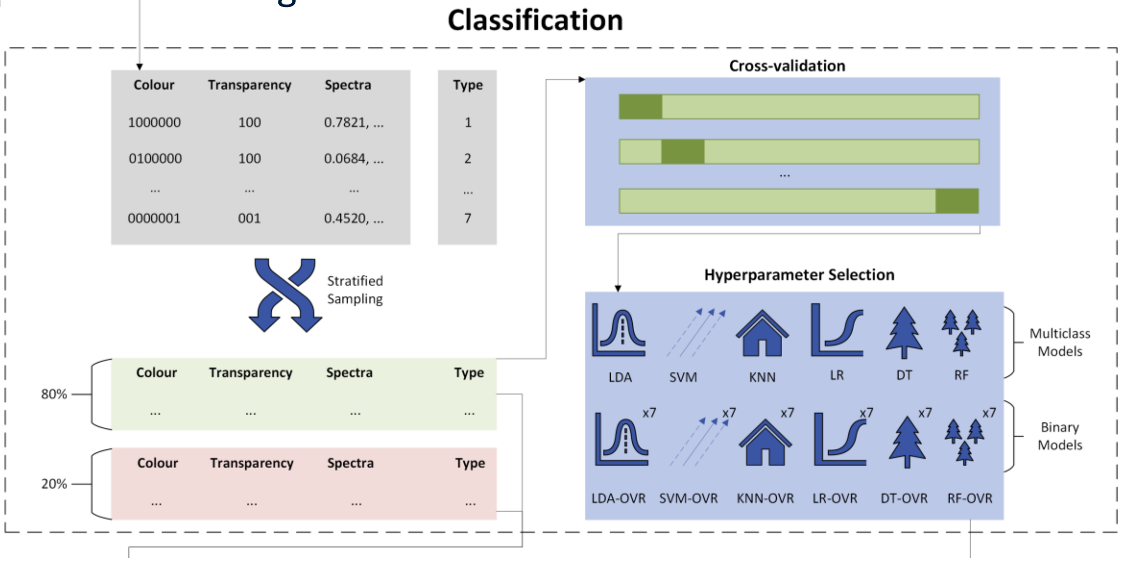




Classification paradigms

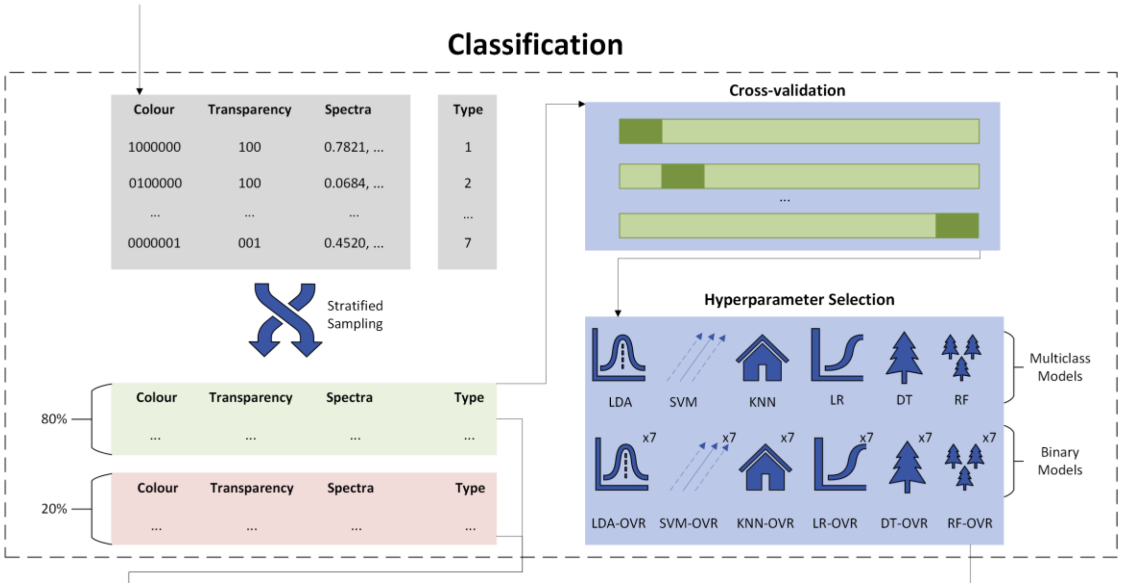
- **Multiclass Classification**: predict 1 of 7 classes directly
- **One-vs-Rest (OVR)**: 7 binary classifiers, combined
- Tradeoff:
  - Multiclass = simpler, faster
  - OVR = sometimes better accuracy and robustness
- **10-fold Cross-Validation** for hyperparameter tuning





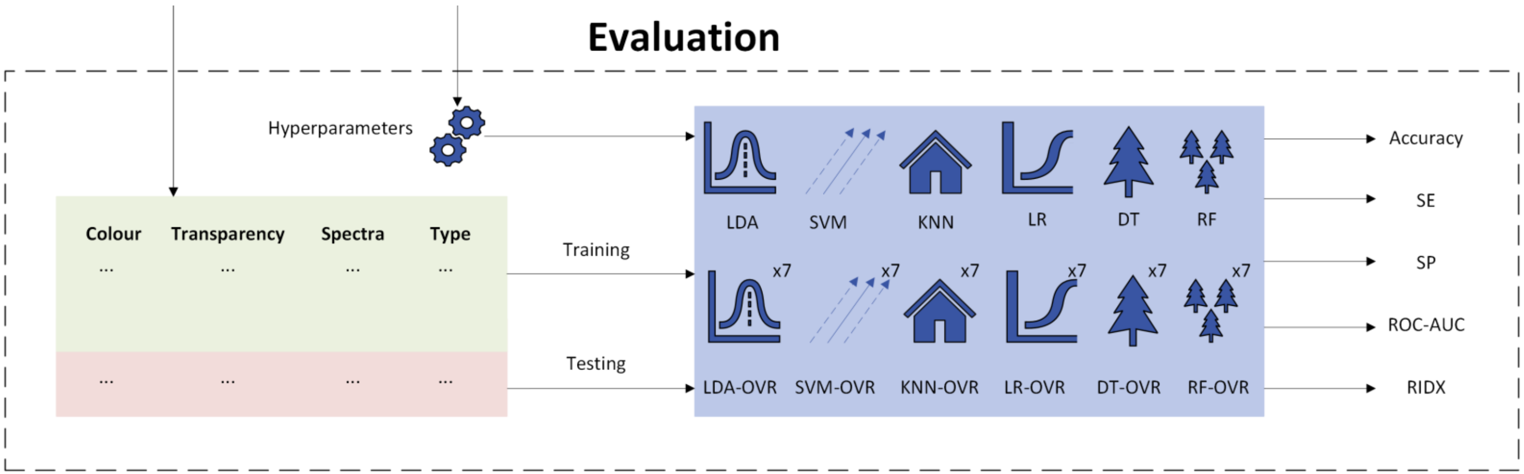
Classification models

- Linear Discriminant Analysis (LDA)
- Support Vector Machines
- K-Nearest Neighbours (KNN)
- Logistic Regression (LR)
- Decision Trees (DT)
- Random Forest (RF)



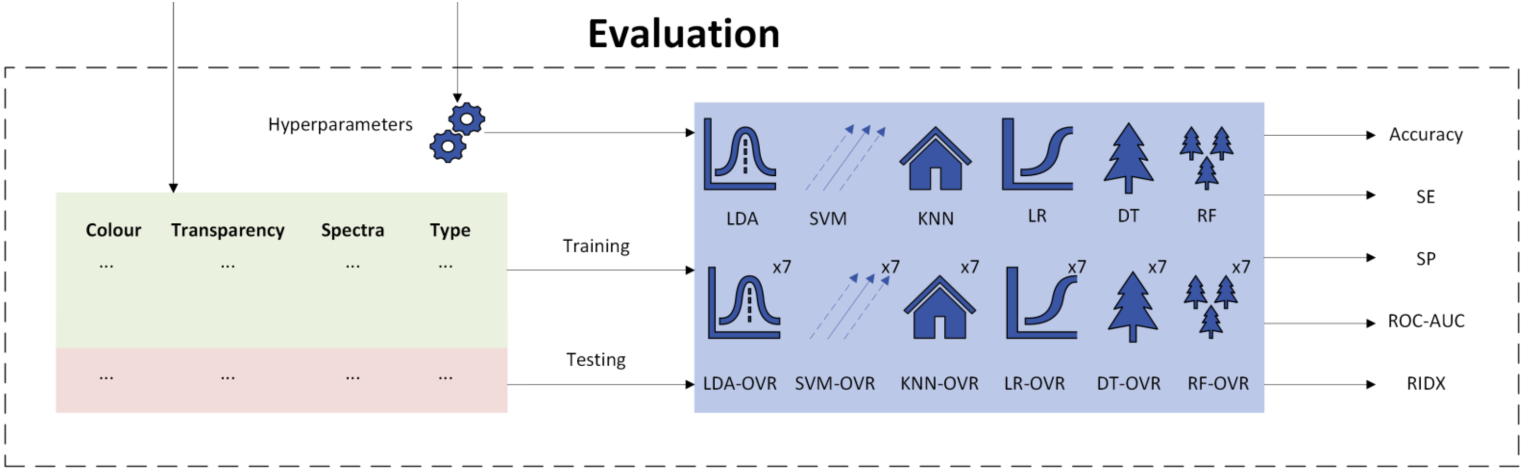
Evaluation setup

- **Holdout test set** (20%) for final evaluation.
- Stratified sampling to preserve class proportions.
- Metrics: Accuracy, Sensitivity, Specificity, ROC-AUC, **RIDX**.



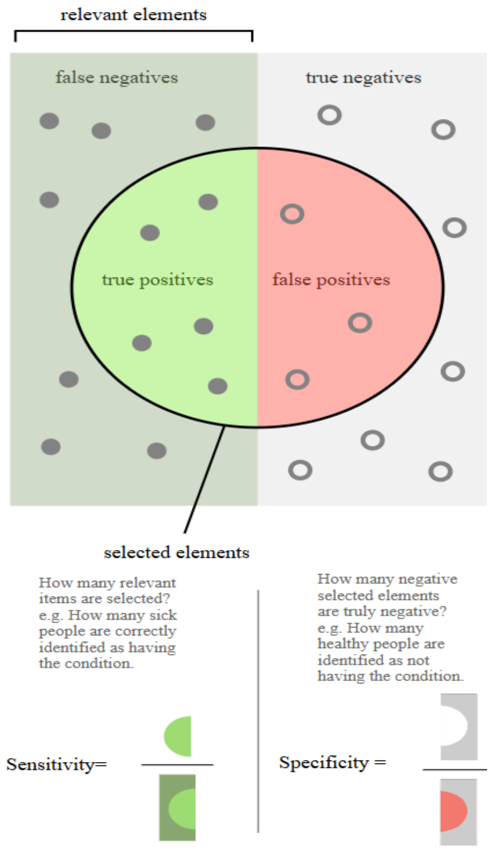
Why evaluation metrics matter?

- Accuracy **NOT** enough: doesn't capture *what* was misclassified
- In recycling, **cost of errors** differs:
  - Misclassifying PVC as PET is much worse than misclassifying PET as HDPE
- Need metrics that reflect **real-world recycling impact**



Standard metrics used

- **Accuracy** – proportion of correct predictions
- **Sensitivity (Recall)** – correctly identified positives
- **Specificity** – correctly identified negatives
- **ROC-AUC** – ability to separate classes across thresholds



Recyclability Index (RIDX)

- Novel metric proposed in this work
- Captures **cost of misclassification** based on recycling compatibility

$$\mathrm{RIDX} = 1 - \frac{\sum F_{ij} \times w_{ij}}{\sum F_{ij} \times \max(w_{ij})}$$

- Values between **0 (worst)** and **1 (perfect classification)**
- Rule: **Misclassifying harder-to-recycle plastics as easier ones = higher penalty**

		Predicted						
		PET	HDPE	PVC	LDPE	PP	PS	Other
Actual	PET	0	1	2	2	2	2	2
	HDPE	1	0	2	2	2	2	2
	PVC	6	6	0	6	6	5	5
	LDPE	4	4	4	0	3	4	4
	PP	4	4	4	3	0	4	4
	PS	6	6	5	6	6	0	5
	Other	6	6	5	6	6	5	0

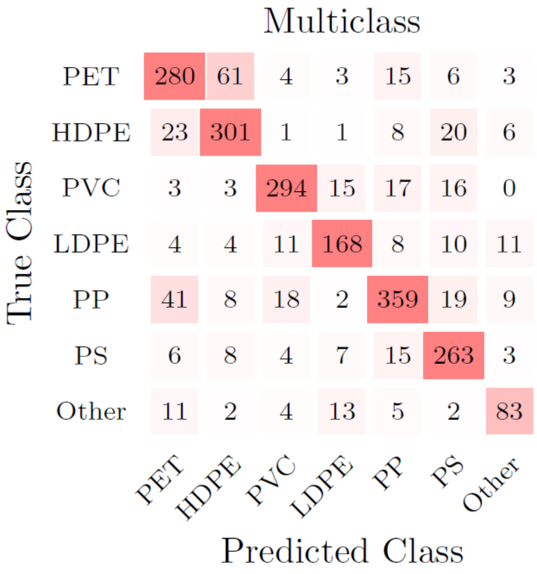
Results

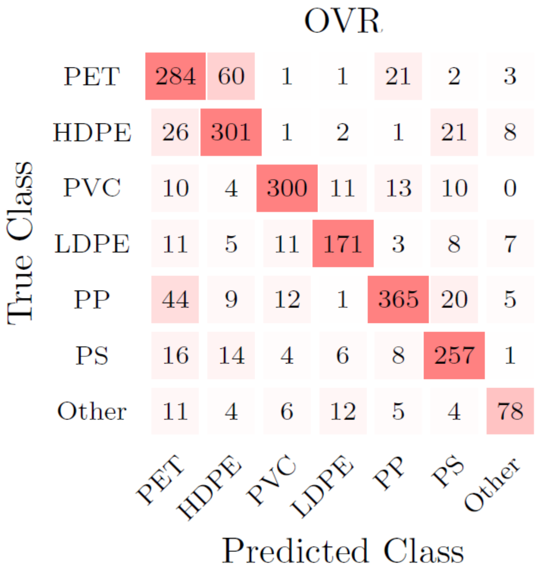
- Best performer: **SVM (One-vs-Rest)**
- Accuracy: **87%**
- Both multiclass and OVR similar structure, OVR slightly better

Model	Accuracy	SE	SP	ROC-AUC	RIDX
LDA	0.7904	0.7782	0.9650	0.9379	0.8726
LDA - OVR	0.7763	0.7574	0.9624	0.9365	0.8568
SVM	0.8404	0.8459	0.9731	0.9822	0.9124
SVM - OVR	**0.8676**	**0.8656**	**0.9778**	0.9730	**0.9297**
KNN	0.8016	0.7916	0.9662	0.9578	0.8713
KNN - OVR	0.8096	0.7968	0.9677	0.9600	0.8776
LR	0.8347	0.8401	0.9721	0.9687	0.9095
LR - OVR	0.8512	0.8412	0.9747	0.9702	0.9150
DT	0.6884	0.6579	0.9470	0.8397	0.7960
DT - OVR	0.6910	0.6675	0.9470	0.9267	0.7972
RF	0.8593	0.8464	0.9761	0.9751	0.9200
RF - OVR	0.8399	0.8247	0.9732	**0.9770**	0.9075

Results

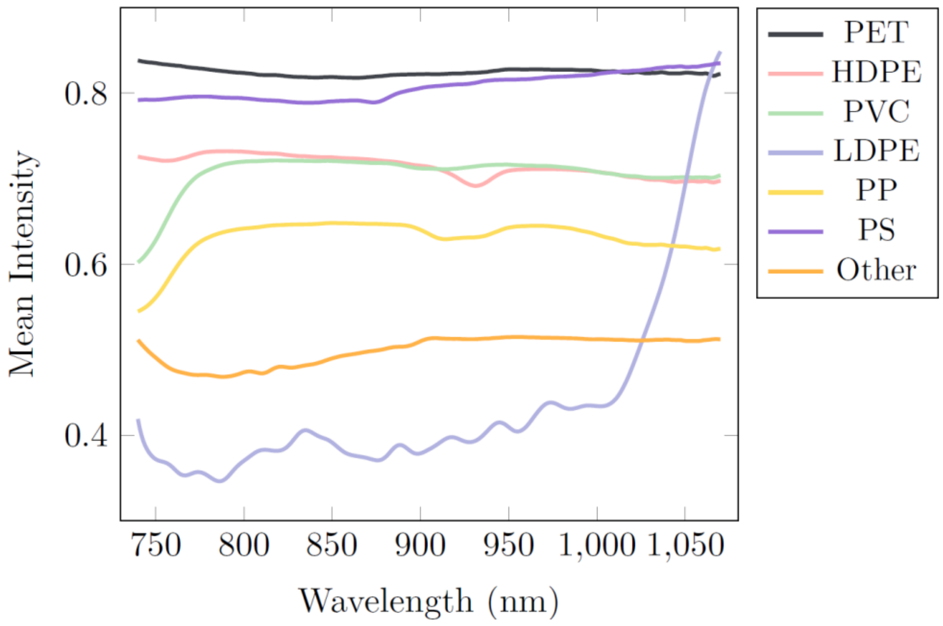
- Misclassifications observed:
  - PET ↔ HDPE (most common)
  - PP misclassified as PET





Results

- Mean aggregated spectra → good separability, but **sample-level variance matters**



Results

- Tested dropping key pipeline components:
  - **Categorical features (color, opacity):** performance drops sharply
  - **Preprocessing (SG filter + detrending):** essential for SVM, KNN
  - **PCA:** mixed results → helps some models, hurts others
- Conclusion: **categorical features and preprocessing are critical**
- Future direction: **multi-sensor fusion** (cheap colorimeters + spectrometers)

	Ablation					
Model	None	CAT	PR	PCA	CAT&PCA	PR&PCA
LDA	0.7904	0.7410	0.7874	**0.7959**	0.7495	0.7548
LDA - OVR	**0.7763**	0.7381	0.7819	0.7683	0.7244	0.7352
SVM	0.8404	0.7964	0.5675	**0.8586**	0.7906	0.5620
SVM - OVR	**0.8676**	0.7769	0.5809	0.8511	0.7709	0.5698
KNN	0.8016	0.7602	0.6026	**0.8155**	0.7736	0.6026
KNN - OVR	**0.8096**	0.7518	0.6054	0.7931	0.7576	0.6026
LR	0.8347	0.7576	0.7656	**0.8514**	0.7658	0.7655
LR - OVR	0.8512	0.7654	0.7488	**0.8594**	0.7629	0.7543
DT	0.6884	0.6060	0.7076	**0.7577**	0.7353	0.4732
DT - OVR	**0.6910**	0.6247	0.6498	0.6806	0.6783	0.5261
RF	0.8593	0.8398	0.8538	**0.8597**	0.8322	0.5262
RF - OVR	0.8399	0.8263	0.8593	**0.8677**	0.8372	0.5017

Comparison with prior work

- Wide-band devices (>1200 nm) → accuracies **>95%** (CNN, SVM, etc.)
- Despite limits → achieved **best reported performance below 1100 nm**

Citation	Device	Type	Band (nm)	Model	# Samples	# Classes	Accuracy
2020 [75]	Pynect NIR-S-G1	Handheld	900 - 1700	SVM	1800	7	1.000
2021 [35]	NIRQuest 512-1.7	Benchtop	900 - 1700	CNN	512	5	0.984
2021 [34]	NIR512	Benchtop	900 - 1700	Combined	200	7	0.957
2021 [76]	ASD FieldSpec 4	Handheld	350 – 2500	PLS-DA	36	2	0.806
2021 [36]	MicroNIR Pro ES 1700	Handheld	950 - 1650	CNN	759	7	0.980
2022 [37]	NIR17S	Benchtop	900 - 1700	CNN	650	3	0.978
2023 [38]	Helios-G2–320	Conveyor	930 - 1700	CNN	31500	7	0.999
2023 [44]	IAS-5000	Handheld	900 - 1700	SVM	60	4	0.980
2023 [41]	Resonon Pika NIR-320	Conveyor	900–1700	ANN	1080	4	0.895
2023 [58]	MicroNIR Pro OnSite	Handheld	950 - 1650	Combined	1625	4	0.990
2023 [49]	Custom	Handheld	750 - 1050	CNN	64	5	0.621
2024 [39]	AS7265X	Handheld	410 - 940	CNN	1200	6	0.725
2024 [77]	IS10	Benchtop	900 - 1700	PLS-DA	298	5	0.975
2025 [42]	IS10	Benchtop	900 - 1700	GAN	250	5	0.980
2025 [48]	Bruker Vector 22/N NIR	Benchtop	800 - 1700	KNN	142	5	1.000
2025 [28]	Avaspec-NIR256-2.0TEC	Benchtop	1000 - 2000	RF	813	7	0.974
Our work	SCiO Mini	Handheld	740 - 1070	SVM - OVR	363	7	0.867

Summary and future work

- **ML + NIR is feasible:** even low-cost consumer devices can classify plastics
  - Data fusion of categorical features improves results

- Only **household plastics** included → may not generalize to industrial waste
- **Small dataset (363 samples):** limits use of deep learning
  - **Expand dataset** → more samples, more diverse plastics
  - **Deep learning** (CNN, GANs) when data volume allows
- **Multi-sensor fusion:** combine cheap sensors (colorimeter, opacity)

- **Next step:** finish paper for Waste Management Journal

Thank you

Contact: marcin.pietrasik@maastrichtuniversity.nl