Classification of heavy metal contaminated samples based on micro-CT images using machine learning algorithms

Dominik Panek[a], Bartosz Leszczyński[a], Dorota Wojtysiak[b], Ewa Stępień[a]

[a]M. Smoluchowski Institute of Physics, Jagiellonian University ul. Łojasiewicza 11, 30-348 Kraków, Poland
[b]Department of Animal Genetics, Breeding and Ethology, Faculty of Animal Sciences, University of Agriculture in Kraków, Poland

INTRODUCTION

Being able to predict whether a sample is contaminated with any toxic substances or not is crucial, especially when it comes to health, and this is where machine learning algorithms are essential [1-2]. Based on the features derived from the micro-computed tomography it is possible to classify the samples with great accuracy.

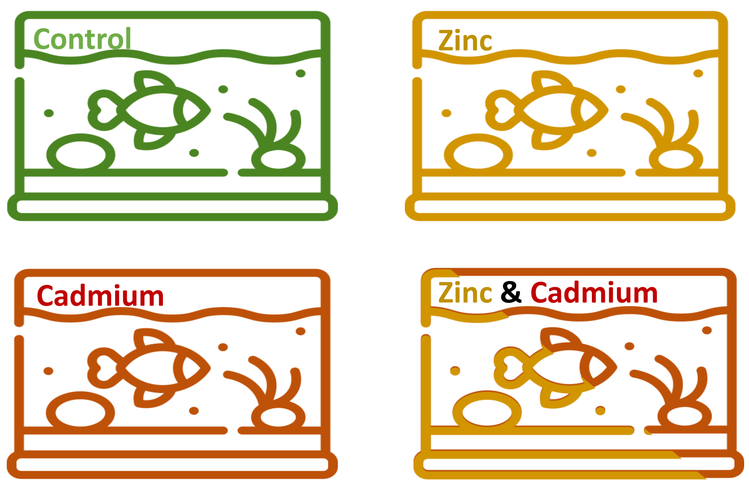

Figure 1. Representation of fish bred in different environments in exposure on different heavy metals with concentration of 4 mg/ml

METHODS

Samples (operculum) were derived from *Carassius Gibelio* fish and were contaminated with two types of heavy metals. Zinc (Zn), cadmium (Cd), and their mixture (Zn+Cd), also control group with no exposure to heavy metal was cultured (Fig. 1). Then imaging procedure using micro-computed tomography was performed. Reconstructed images delivered data, which enabled us to extract key features (like max. greyscale value, masses of the samples, mean greyscale value, area under the histograms, etc.) based on which algorithms learn. In this research following machine learning algorithms were applied: logistic regression, supporting vector machine (SVM), decision trees and k-nearest neighbors (KNN). Applied workflow can be seen in Figure 3.

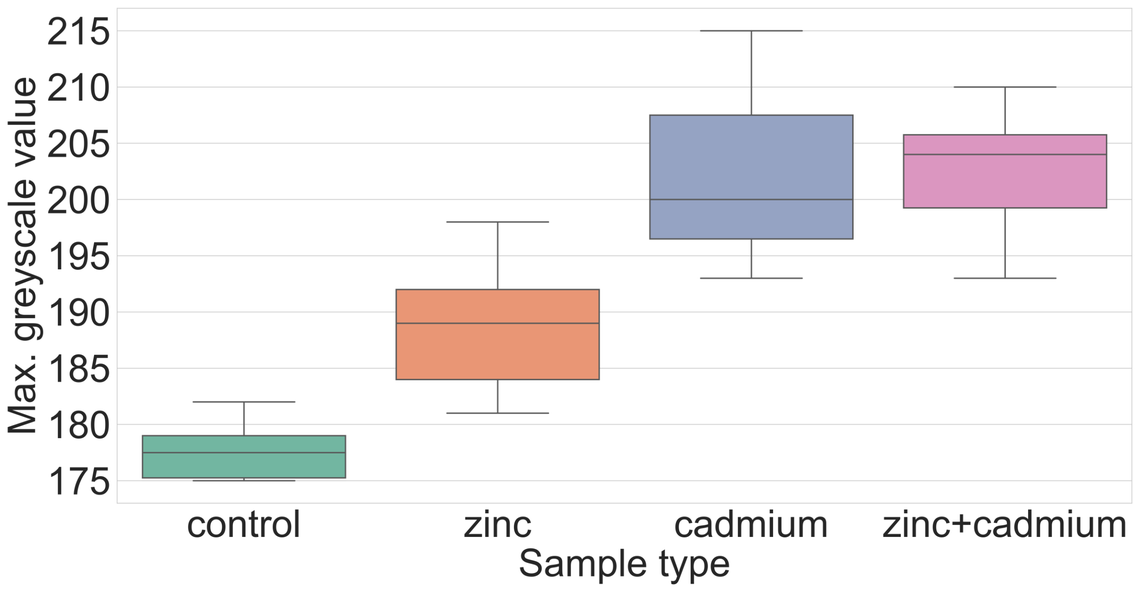

Figure 2. Distribution of the greyscale value among the different groups.

RESULTS

The best results were achieved for the simplest logistic regression, where the overall accuracy was 90%, a second-best algorithm was KNN with an accuracy of 71% for k = 1 and 86% for k = 4 (Fig. 4), next were decision trees with an accuracy of 80% and SVM with an overall accuracy of 50%. Table with exact result for all applied machine learning models can be seen below (Tab. 1).

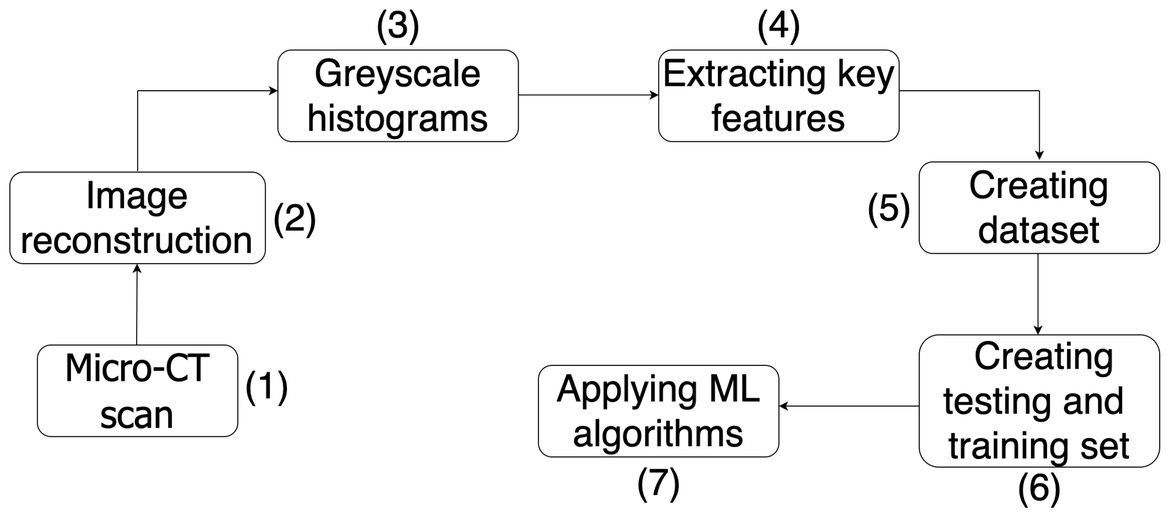

Figure 3. Workflow performed during the experiment. The most important part was the micro-CT scan, where the parameters of the scan had to be exactly identical for all the samples. Numbers in the parentheses indicate the order of action.

Table 1. Results of classification (precision and recall) using different methods (0 – not contaminated samples, 1 - contaminated samples).

Model	Class	Precision	Recall
Logistic regression	0	100%	80%
	1	83%	100%
Decision Trees	0	100%	25%
	1	67%	100%
KNN (k=1)	0	100%	33%
	1	67%	100%
SVM	0	0%	0%
	1	50%	100%

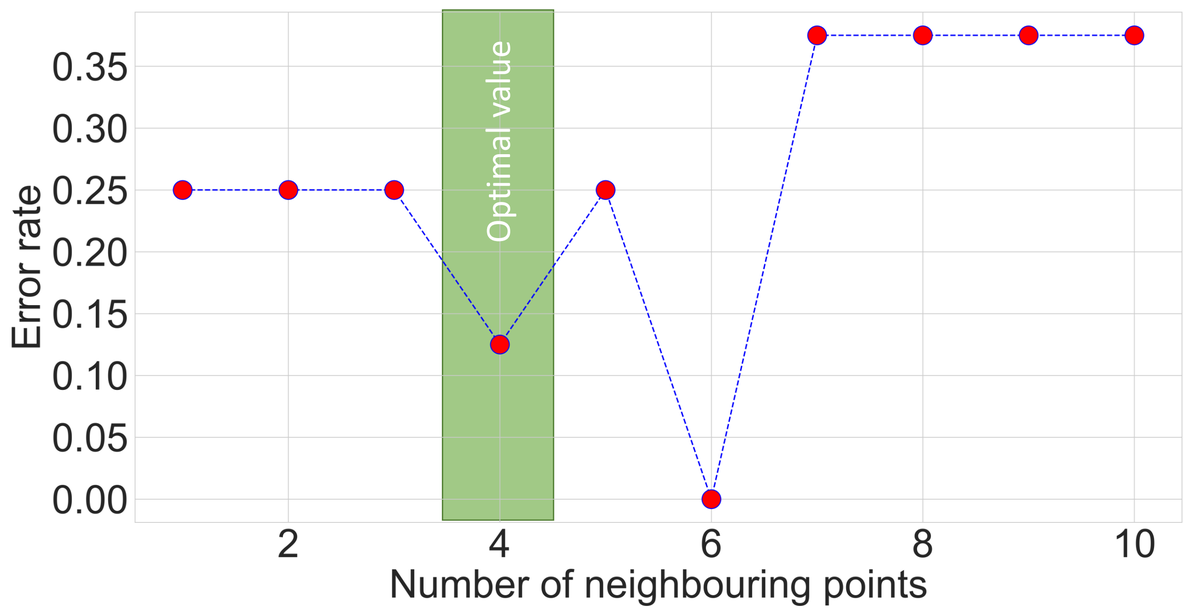

Figure 4. Changing value of error rate depending on the numbers of neighboring points

CONCLUSIONS

Most of the applied machine learning models are very good when it comes to classification, exceeding 80% of accuracy (for KNN, DT and logistic regression). With some adjustments, in the future, it is possible to achieve even more prominent results. Furthermore, using more advanced machine learning algorithms like neural networks the results could be even more accurate.

REFERENCES

[1] Heavy metals in suspended matters during a tidal cycle in the turbidity maximum around the Yangtze Estuary, Huaijing Zhang et al., Acta Oceanologica Sinica 34, 36–45(2015).

[2] Toxicity, mechanism and health effects of some heavy metals, Monisha Jaishankar, Interdiscip Toxicol. 2014 Jun; 7(2): 60–72.