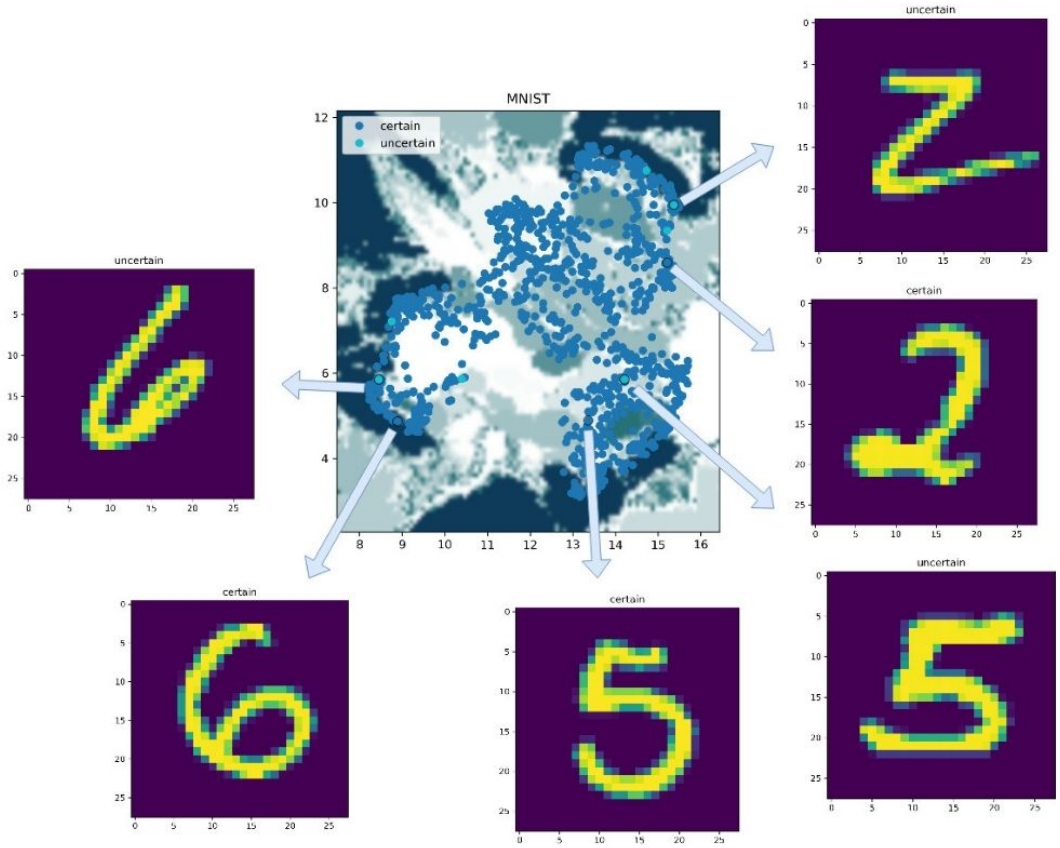


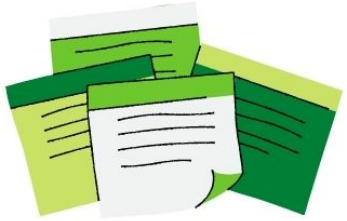
OUR WORK IN THE EXPLAINABILITY DOMAIN

- Explain individual predictions, while giving a global estimate of the uncertainty of a model [1]
- uncertainty here defined as likelihood of adversarial attack, uncertainty of the model itself or likelihood of sample not being of same model (= in supervised domain would be missclassifications)
- applicable to unsupervised learners
- applied on trained model for fashion-MNIST dataset

HOW TO QUANTIFY AND EXPLAIN UNCERTAINTY IN MACHINE LEARNING?

REFERENCES

[1] Newen C. . Müller E.(2022) Unsupervised Deepview: Global Explainability of Uncertainties for High Dimensional Data IEEE ICKG
[2] Schulz, Alexander, Fabian Hinder, and Barbara Hammer. "Deepview: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction." arXiv preprint arXiv:1909.09154 (2019).



technische universität dortmund