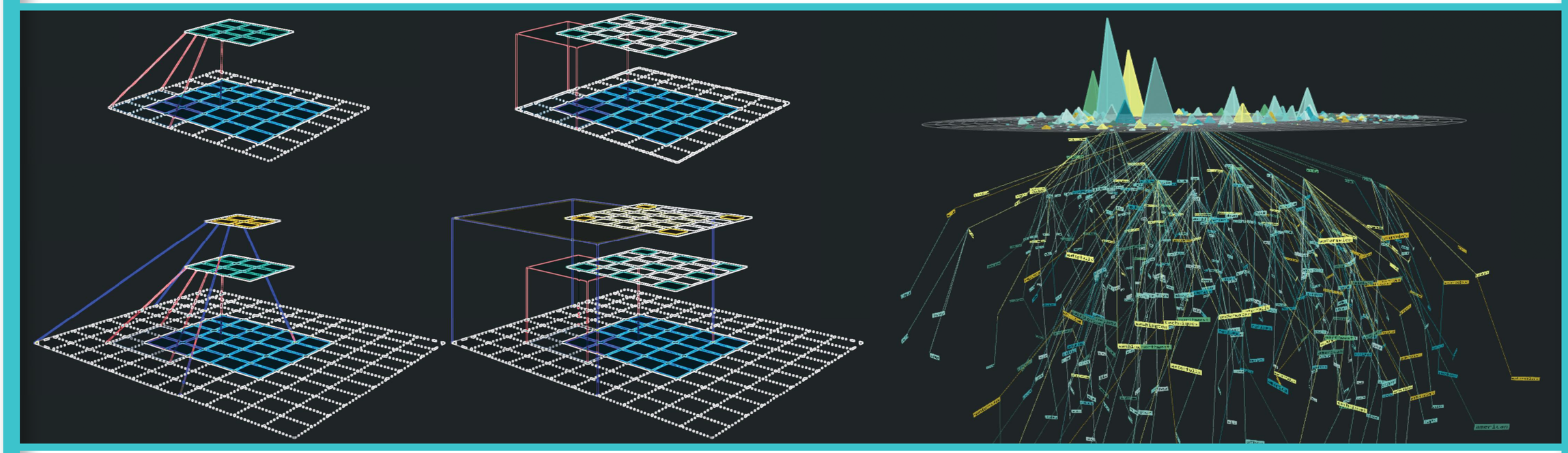
BACTERIA CLASSIFICATION BY MALDI-TOF-MS WITH CNN

Food safety accidents caused by microbial pathogens are common across the globe. The detection and identification of microorganisms in contaminated food can help provide effective information in order to implement timely control measures. Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF) is a relatively new technology which has revolutionized the way microorganisms including bacteria areidentified. The data quality of MALDI-TOF-MS is affected by noise factors such as sample preparation methods, matrix solutions, organic solvents and acquisition methods. We present an innovative deep learning method based on the application of convolution neural networks (CNN) which has shown high accuracy and robustness.

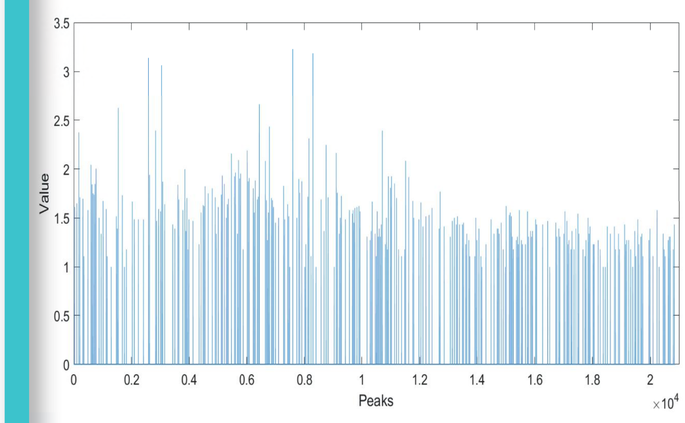
Steffen Uhlig[1], Kirsten Simon[1], Carsten Uhlig[2], Nilavra Bhattacharya[1], Kai Liu[1], Ulrike Steinacker[3], Manfred Stoyke[3] and Petra Gowik[3]
[1] QuoData GmbH, Prellerstr. 14, 01309 Dresden, Germany
[2] Akees GmbH, Ansbacher Str. 11, 10787 Berlin, Germany
[3] Federal Office of Consumer Protection and Food Safety (BVL), Mauerstr. 39-42, 10117 Berlin, Germany

THE NEW INTELLIGENCE: CONVOLUTIONAL NEURAL NETWORK



DATA AND NOISE SIMULATION

▼ MALDI-TOF Data for one sample (sparse by 35 times)



Data set (raw MALDI-TOF data without pretreatment):

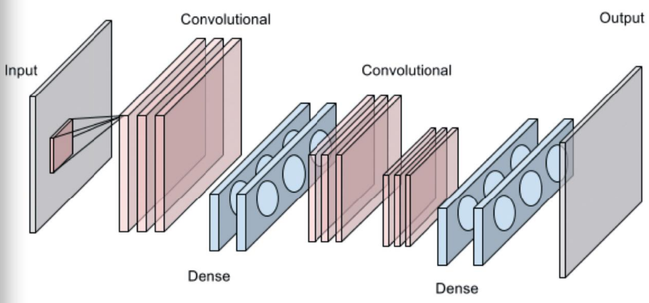
- 3 Species: Staph. aureus, Staph. intermedius, Miscellaneous
- 20,893 peaks per sample.

Noise Type:

- Spectrum shift (by 44 steps which equals to CO_2 mol. weight) from a randomly selected position in the spectrum for selected samples
- Magnify all peaks of randomly selected samples (1.3 times)
- Add normal noise to all peaks of randomly selected samples (30%)

The following diagrams show the effect of the three noise types when the model developed on the basis of the training set is applied:

- To the original test set
- To the "noisy" test set, i.e. where all samples of the test set are subjected to the same type of noise as the randomly selected samples.



CNN Hidden layers:

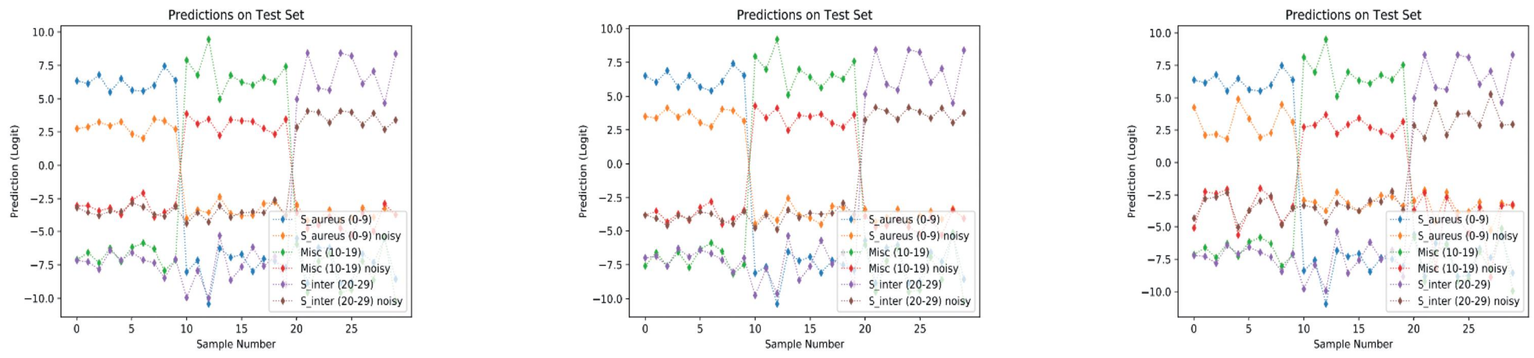
1. Convolutional layers
2. Pooling layers
3. Fully connected layers
4. Normalization layers

Basic principle of neural networks:

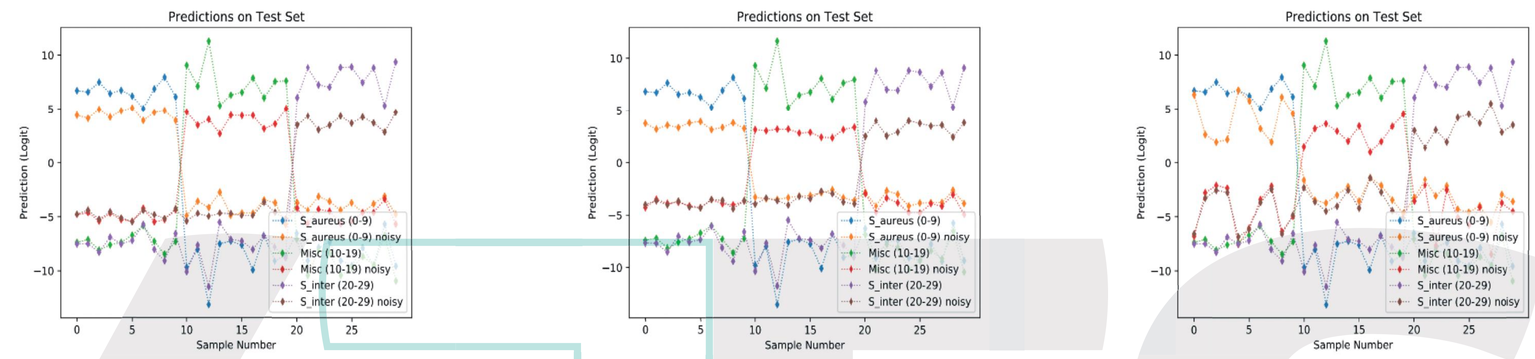
- Minimize the difference between CNN prediction and true classification
- Adjust parameters (weights, etc.)
- Iterate through samples
- Estimate "Error" through loss function.

Reference:
Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., & Martens, J. (2015). Adding gradient noise improves learning for very deep networks. arXiv preprint arXiv:1511.06807.
De Bruyne, K., Slabbinck, B., Waegeman, W., Vauterin, P., De Baets, B., & Vandamme, P. (2011). Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. Systematic and applied microbiology, 34(1), 20-29.

▼ Percentage of training samples affected by noise: 1%



▼ Percentage of training samples affected by noise: 10 %



Noise type:	Shift by 44 peaks	Magnify peaks 1.3 times	Add 30% normal noise

Conclusions:

- CNN displays high accuracy at a lower computational cost even in the presence of noise.
- Adding noise during the training process makes the model more robust: the difference between the results for the original and "noisy" test data is reduced.
- Normal noise during the training process has a greater effect on the "noisy" test set than that of peak shifts and peak magnification.

GET IN TOUCH WITH US! FIND WHAT MATTERS IN YOUR DATA!

AKEES Data Intelligence GmbH, Ansbacherstr. 11, 10787 Berlin, T +49 (0) 8161 – 57 70 17 75 – 2, info@akees.com, www.akees.com