Table 1: Detailed event count numbers per cohort with the maximum event number in the given cutoff written in brackets.

| | $\leq 20000$ | $\leq 30000$ | $\leq 40000$ | $\leq 50000$ |
|---|---|---|---|---|
| AML | 0 (max 0) | 0 (max 0) | 7 (max 37756) | 124 (max 50000) |
| CLL | 2 (max 16153) | 291 (max 29999) | 1249 (max 39988) | 3356 (max 50000) |
| FL | 0 (max 0) | 2 (max 29795) | 7 (max 38991) | 216 (max 50000) |
| HCL | 0 (max 0) | 0 (max 0) | 3 (max 35901) | 187 (max 50000) |
| HCLv | 0 (max 0) | 0 (max 0) | 3 (max 37997) | 54 (max 50000) |
| LPL | 1 (max 19693) | 5 (max 29814) | 22 (max 39318) | 622 (max 50000) |
| MBL | 0 (max 0) | 1 (max 29588) | 11 (max 39441) | 1458 (max 50000) |
| MCL | 2 (max 15545) | 12 (max 29887) | 62 (max 39702) | 415 (max 50000) |
| MM | 0 (max 0) | 1 (max 26217) | 2 (max 38324) | 101 (max 50000) |
| MZL | 0 (max 0) | 4 (max 28871) | 50 (max 39812) | 968 (max 50000) |
| normal | 1 (max 14598) | 1 (max 14598) | 19 (max 39860) | 8434 (max 50000) |
| PL | 1 (max 12301) | 20 (max 29810) | 132 (max 39995) | 597 (max 50000) |

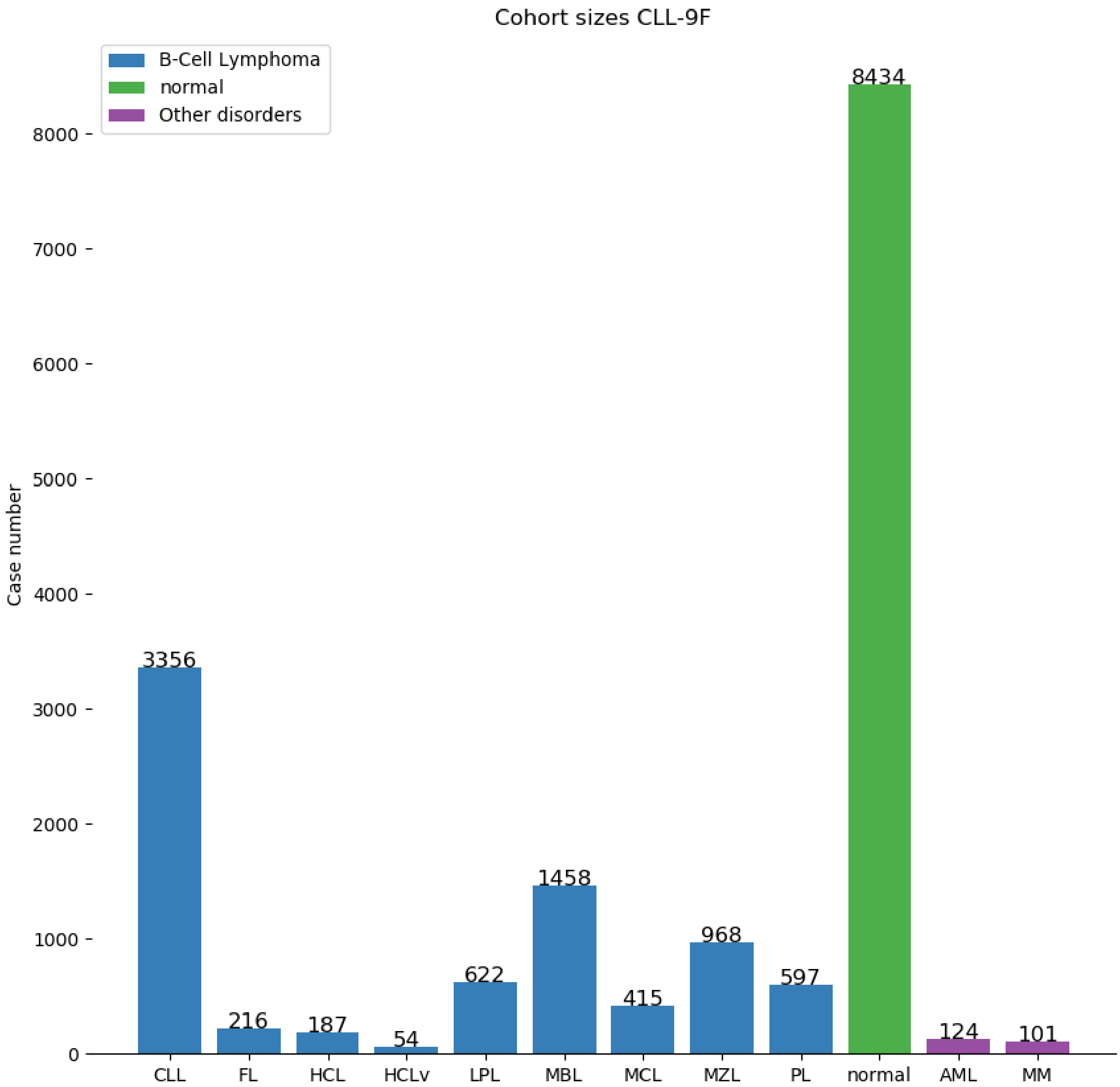



Figure 1: Overview of cohort sizes using the CLL 9F panel. These numbers include only cases with at least tube 1 and 2 of the same material and each fcs file having more than 10,000 events.

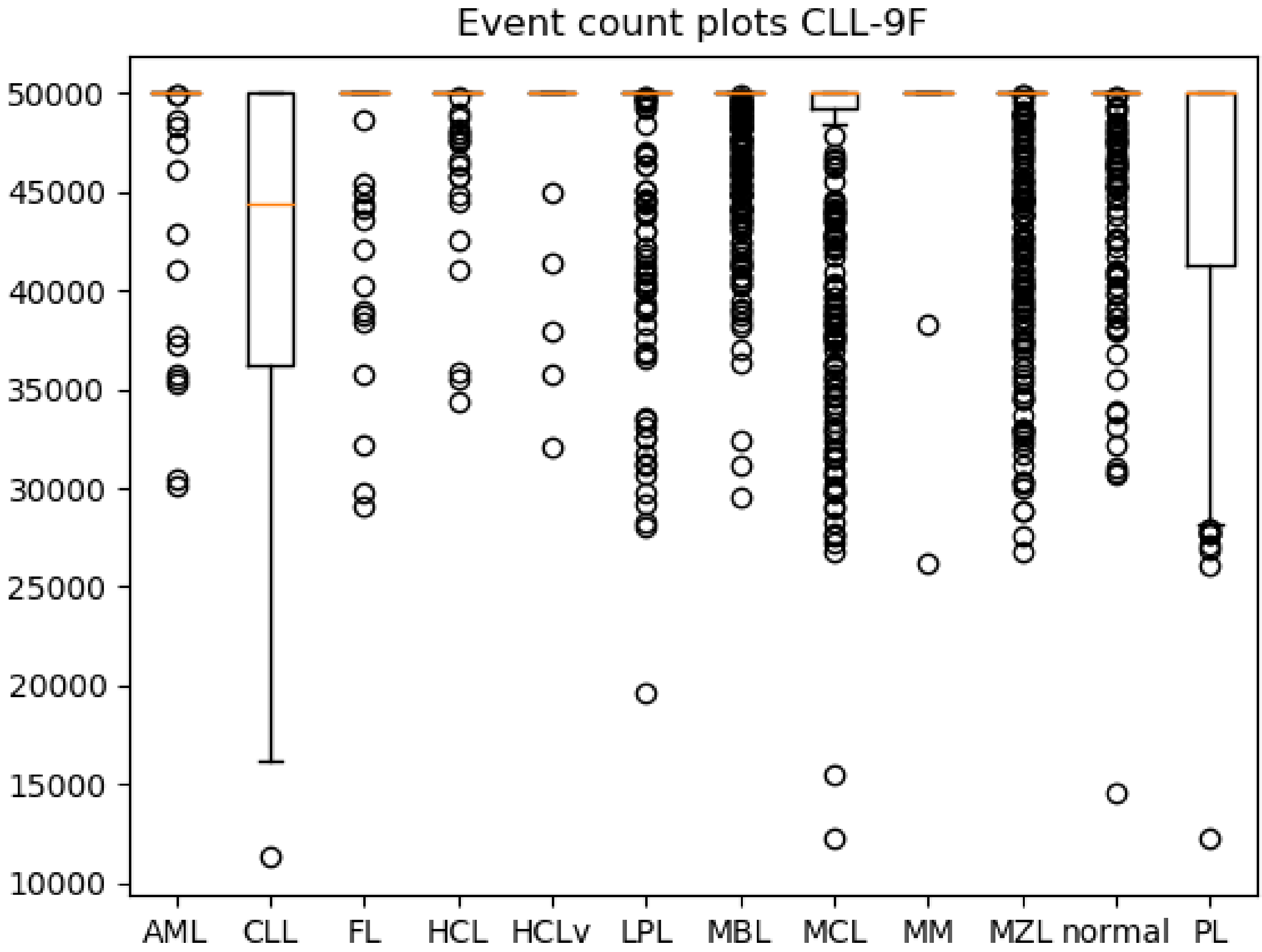



Figure 2: Number of events in each fcs file in tube 1 for each cohort. The whiskers represent 25th and 75th percentile. Numbers outside these ranges are represented as individual dots.

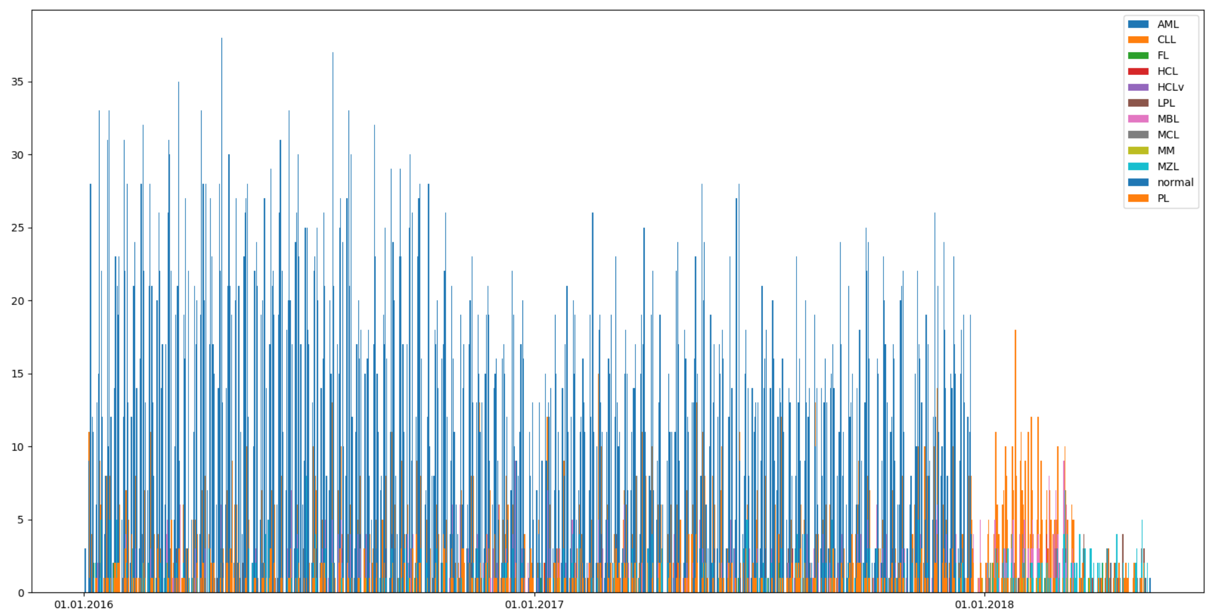



Figure 3: Time-histogram of case date over time. This visualization can be used to spot skewed distributions in individual cohorts.

Table 2: Overview of classification runs so far and the different applied processing steps in clustering. The mean accuracy has been calculate counting single cases (micro-average), taking class-imbalances into account.

(a) CLL, PL, FL, HCL, HCLv, LPL, MBL, MCL, MZL, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| initial_comp_all_groups | indiv_pregating | random | 1 | 0.71 | 0.01 |
| | indiv_pregating_exc | random | 1 | 0.70 | 0.01 |
| | normal | random | 1 | 0.68 | 0.02 |
| | normal_exc | random | 1 | 0.68 | 0.01 |

(b) CLL, PL, FL, HCL, LPL, MBL, MCL, MZL, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| abstract_single_groups | normal | random | 1 | 0.72 | 0.02 |
| | pregated | random | 1 | 0.74 | 0.01 |
| | somgated | random | 1 | 0.77 | 0.00 |
| abstract_single_groups_cbrt | normal | random | 1 | 0.77 | 0.03 |
| | somgated | random | 1 | 0.79 | 0.00 |
| abstract_single_groups_log1p | normal | random | 1 | 0.72 | 0.02 |
| | somgated | random | 1 | 0.77 | 0.00 |
| abstract_single_groups_sqrt | normal | random | 3 | 0.77 | 0.03 |
| | somgated | random | 3 | 0.79 | 0.00 |

(c) CLL, PL, FL, LPL, MBL, MCL, MZL, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| abstract_single_no_hcl | normal | random | 1 | 0.72 | 0.02 |
| | somgated | random | 1 | 0.77 | 0.00 |

(d) CM, FL, HCL, HCLv, LMg, MtCp, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| initial_comp_more_merged | indiv_pregating | random | 1 | 0.80 | 0.01 |
| | indiv_pregating_exc | random | 1 | 0.80 | 0.01 |
| | normal | random | 1 | 0.76 | 0.02 |
| | normal_exc | random | 1 | 0.77 | 0.02 |

(e) CM, FL, HCL, LMg, MtCp, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| abstract_merged_hzl | somgated | random | 3 | 0.83 | 0.01 |
| | somgated_equal | random | 1 | 0.74 | 0.01 |
| abstract_merged_hzl_cbrt | normal | random | 1 | 0.84 | 0.03 |
| | pregated | random | 1 | 0.85 | 0.01 |
| | somgated | random | 1 | 0.87 | 0.00 |
| abstract_merged_hzl_log1p | normal | random | 1 | 0.80 | 0.02 |
| | pregated | random | 1 | 0.83 | 0.01 |
| | somgated | random | 1 | 0.86 | 0.01 |
| abstract_merged_hzl_sqrt | normal | random | 1 | 0.84 | 0.03 |
| | pregated | random | 1 | 0.86 | 0.01 |
| | somgated | random | 1 | 0.87 | 0.00 |
| hcl_included | merged | random | 1 | 0.85 | 0.00 |
| somgated_fast | somgated_fast | random | 2 | 0.88 | NaN |
| somgated_fast_cnn | somgated_fast_cnn | random | 2 | 0.76 | NaN |
| somgated_fast_mi | somgated_fast_mi | random | 3 | 0.88 | NaN |
| somgated_fast_micl | somgated_fast_micl | random | 2 | 0.85 | NaN |
| somgated_fast_nn | somgated_fast_nn | random | 2 | 0.76 | NaN |
| somgated_fastcbrt | somgated_fastcbrt | random | 1 | 0.87 | NaN |
| somgated_fastconv | somgated_fastconv | random | 3 | 0.81 | NaN |
| somgated_fastlog1p | somgated_fastlog1p | random | 1 | 0.82 | NaN |

(f) CLL, FL, LPL, MBL, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| num_cases_single | c1 | random | 1 | 0.89 | 0.0 |

(g) CM, FL, LMg, MtCp, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| abstract_merged | normal | random | 1 | 0.80 | 0.02 |
| | pregated | random | 1 | 0.84 | 0.01 |
| | somgated | random | 1 | 0.86 | 0.00 |
| infiltration | normal | random | 1 | 0.52 | 0.02 |
| | pregated | random | 1 | 0.59 | 0.02 |
| | somgated | random | 1 | 0.63 | 0.01 |

(h) CM, LMg, MtCp, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| comp_pregating | always_som | random | 1 | 0.83 | 0.01 |
| | pregated_combined | random | 1 | 0.85 | 0.01 |
| | som | random | 1 | 0.83 | 0.01 |
| | som_combined | random | 1 | 0.87 | 0.00 |
| | somgated | random | 1 | 0.87 | 0.00 |
| initial_comp | indiv_pregating | random | 3 | 0.85 | 0.01 |
| | indiv_pregating_exc | random | 3 | 0.84 | 0.01 |
| | normal | random | 3 | 0.80 | 0.03 |
| | normal_exc | random | 3 | 0.81 | 0.02 |
| initial_comp_selected | indiv_pregating | random | 1 | 0.84 | 0.01 |
| | indiv_pregating_exc | random | 1 | 0.85 | 0.02 |
| | normal | random | 1 | 0.83 | 0.02 |
| | normal_exc | random | 1 | 0.82 | 0.02 |

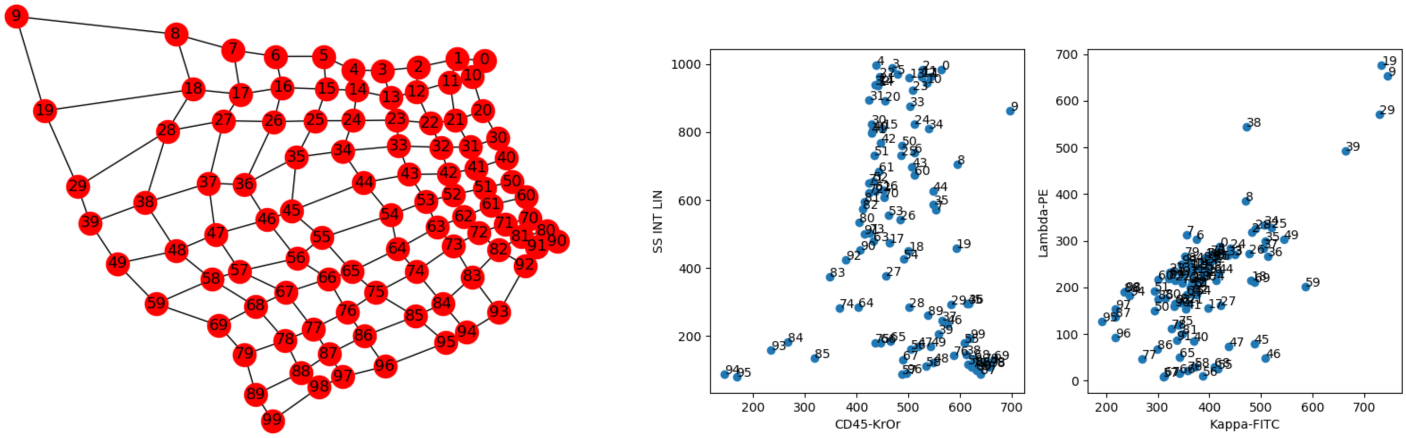
(i) AML, MM, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| exotic | exotic | random | 1 | 0.79 | NaN |
| exotic_sqrt | exotic_sqrt | random | 1 | 0.88 | NaN |

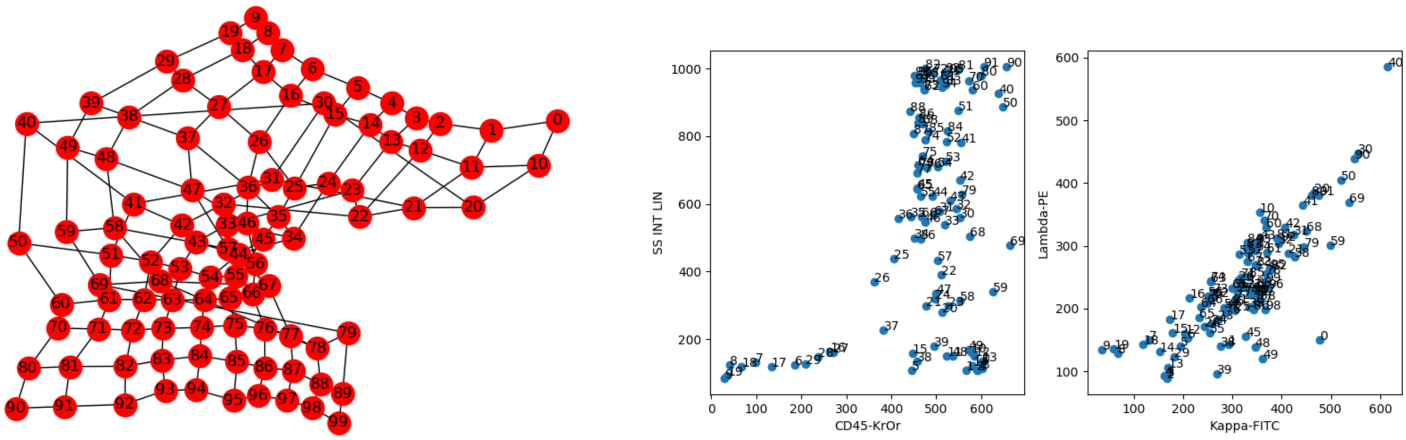
(j) CD5neg, CD5pos, normal

| set | name | type | count | f1 | std |
|---|---|---|---|---|---|
| cd5_threeclass | normal | random | 1 | 0.84 | 0.02 |
| | pregated | random | 1 | 0.87 | 0.01 |
| | somcombined | random | 1 | 0.89 | 0.00 |
| cd5_threeclass_sqrt | normal | random | 1 | 0.86 | 0.02 |
| | pregated | random | 1 | 0.88 | 0.01 |
| | somcombined | random | 1 | 0.89 | 0.00 |
| hcl_included | cd5 | random | 1 | 0.89 | 0.00 |

(k) CLL, normal




(a) One case per cohort.




(b) Five cases per cohort.

Figure 4: Grid visualization of consensus SOM after two training iterations. Using cases in CLL, MBL and normal.

Table 3: Significance tests for some implemented adaptations. Dataset a (normal) is always compared against b (with modifications) with the number of results used for analysis given. p-values are calculated using Welch's t-test. Primarily because of possible differences in variance depending on modifications to the consensus SOM generation. An unjoined 9-class analysis (without HCLv) is always used for analysis unless stated otherwise. Global top 1 accuracies counting single cases are used as the metric.

| | mean_a | n_a | mean_b | n_b | p_value |
|---|---|---|---|---|---|
| normal vs somgated | 0.675281 | 10 | 0.738160 | 10 | 0.000011 |
| normal vs sqrt_transformed | 0.792941 | 1 | 0.872941 | 1 | NaN |

Table 4: Comparison of histogram transformations. This table contains informations on sqrt transformations in our different group configurations. When using sqrt Transformation, the root of each number in the histogram is taken.

| | Acc normal | №normal | Acc sqrt | №Acc | p-Value |
|---|---|---|---|---|---|
| 6-class merged somgated | 0.803632 | 10 | 0.871953 | 10 | 3.256652e-11 |
| 9-class normal | 0.741198 | 10 | 0.777630 | 10 | 8.102392e-04 |
| 9-class somgated | 0.777684 | 10 | 0.797946 | 10 | 3.988427e-10 |
| CD5 - normal | 0.841617 | 10 | 0.864223 | 10 | 2.766266e-02 |
| CD5 - somgated | 0.886777 | 5 | 0.890286 | 10 | 6.361855e-03 |

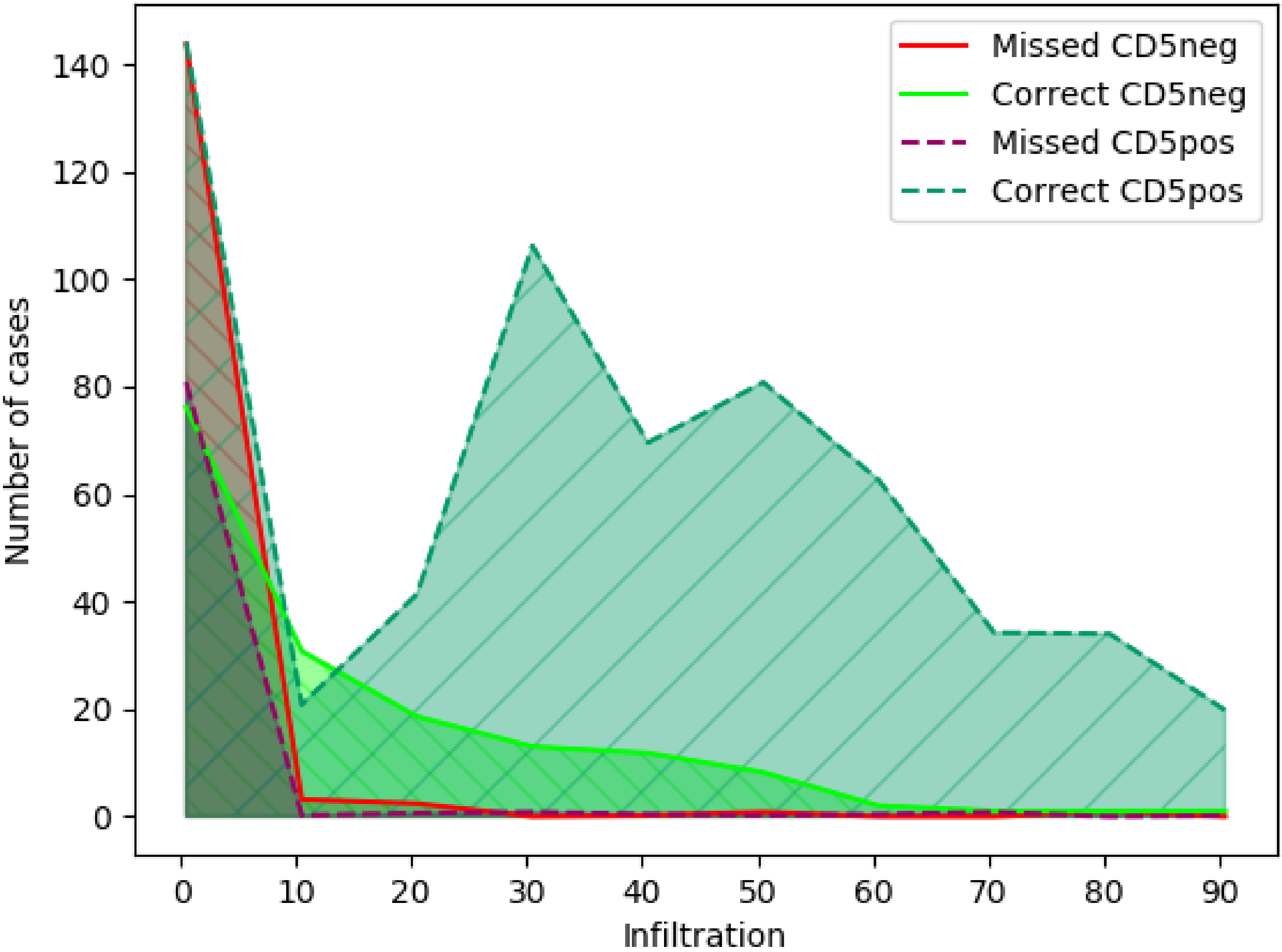



Figure 5: Binned histogram visualization of infiltration percentages for misclassified and non-misclassified cases for each cohort.

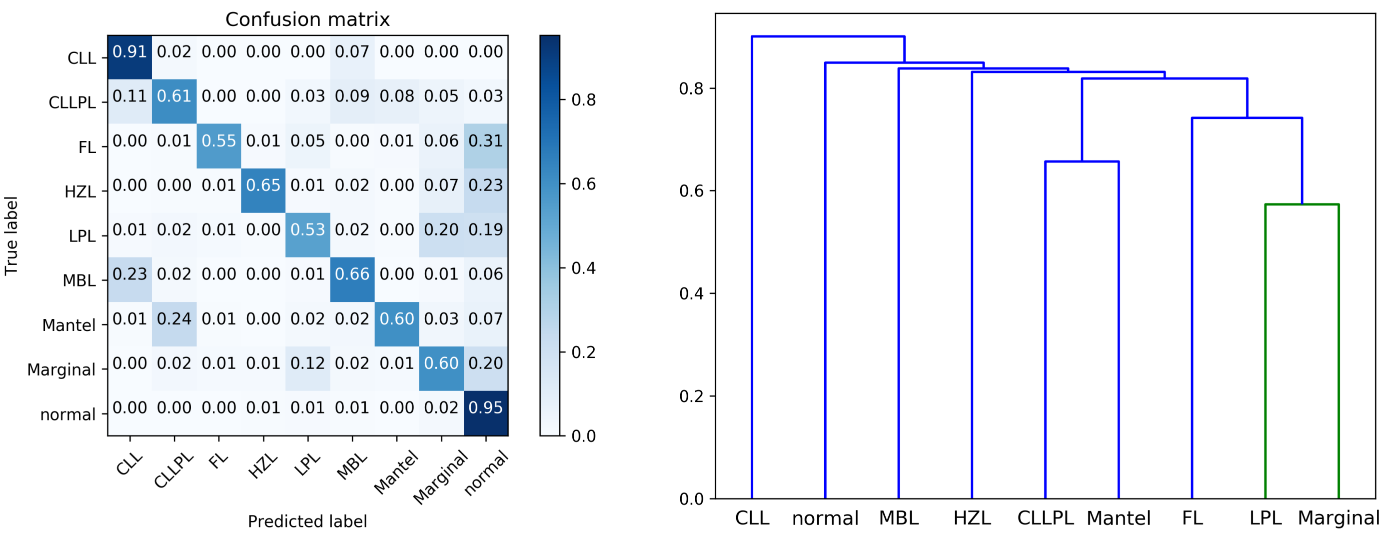



Figure 6: Hierarchical clustering using pairwise distances in the normalized misclassification table for the 9-class classification using the somgated approach in consensus SOM generation and upsampling.

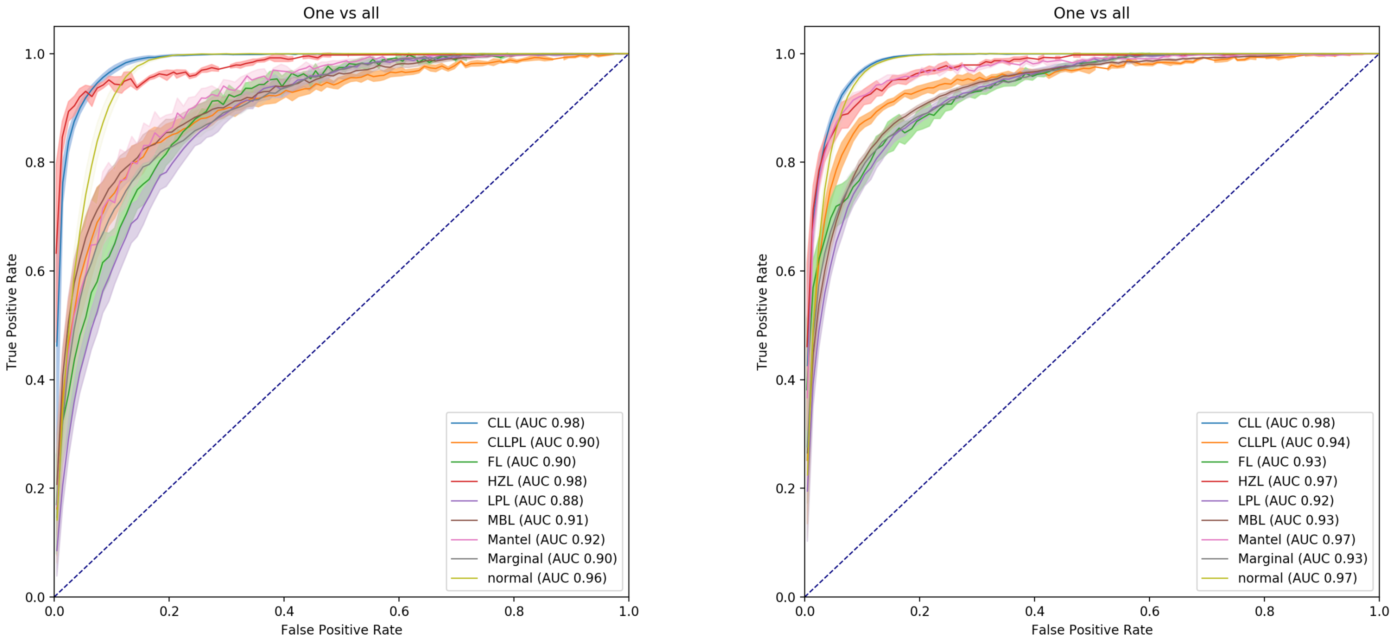



Figure 7: ROC curve for one-vs-all comparisons between a 9-class classification with a normal pipeline (left) and a somgated pipeline (pregating and subsequent generation of consensus SOM using SOM node weights instead of raw fcs data — right). Curves are averaged between all runs through binning. The colored area around a graph represents the standard deviation.

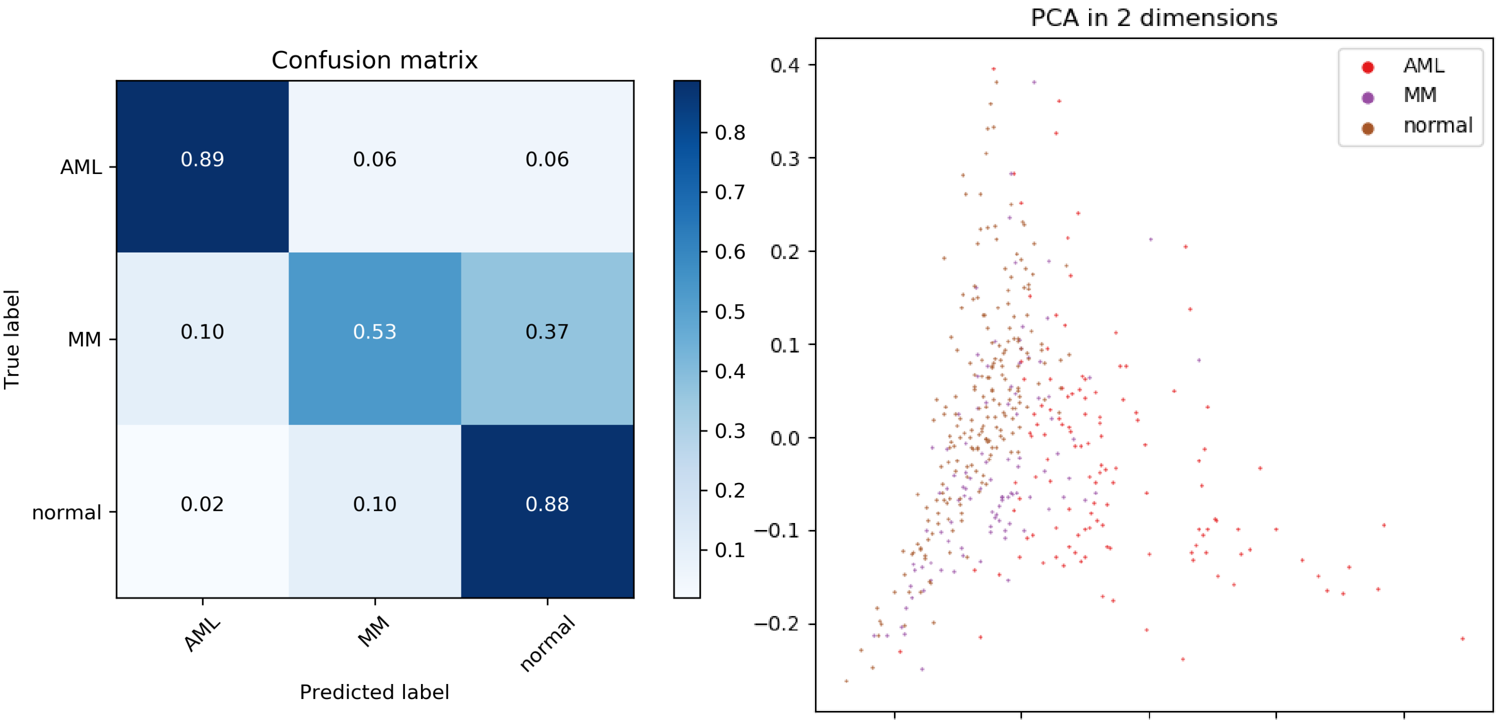



Figure 8: Processing of diagnoses outside the scope of the CLL 9F panel, such as acute myeloic lymphoma (AML) and multiple myeloma (MM). Their pathogenic cell populations are not well captured by the panel itself, making them good targets to measure the effect of foreign cohorts on classification outcome. Clustering did not use any additional preprocessing. The consensus SOM was generated using normal and B-Cell lymphoma cohorts. AML and MM were not used in the consensus SOM generation, but only utilized it for upsampling. Classification was done with the entire AML and MM cohorts vs 200 randomly sampled cases from the normal cohorts as a comparison.

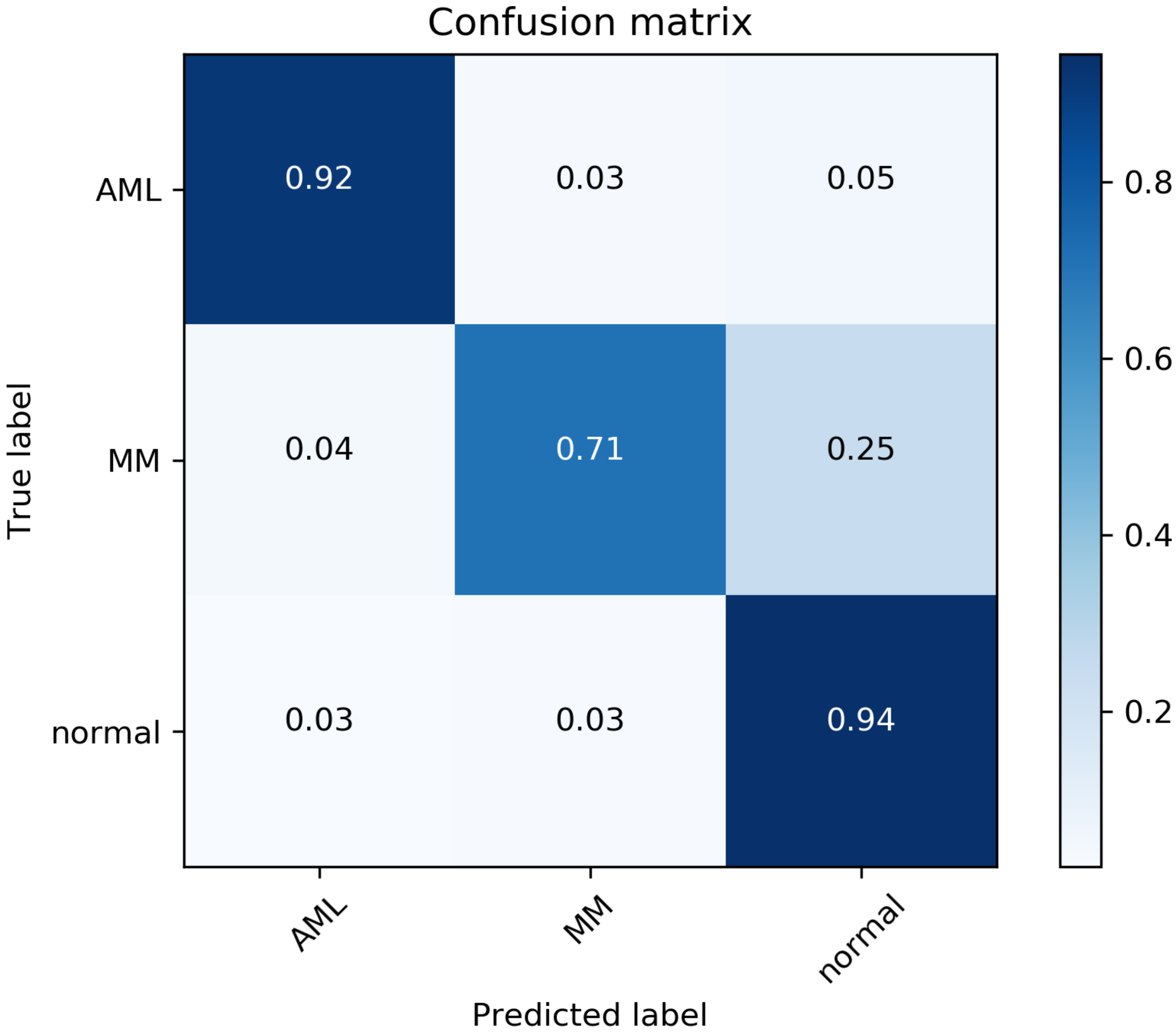



Figure 9: Classification accuracy decreases for smaller populations and low infilration rates. Non-linear transformations could improve the classification accuracy, such as taking the square root of all infiltration numbers prior to training and prediction using the neural network.