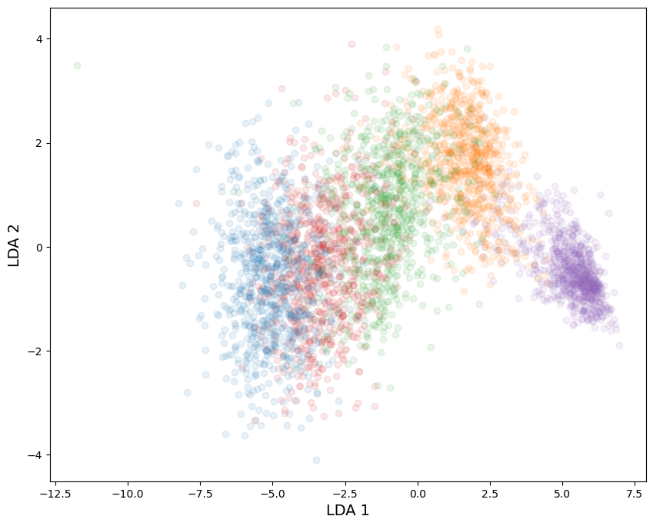
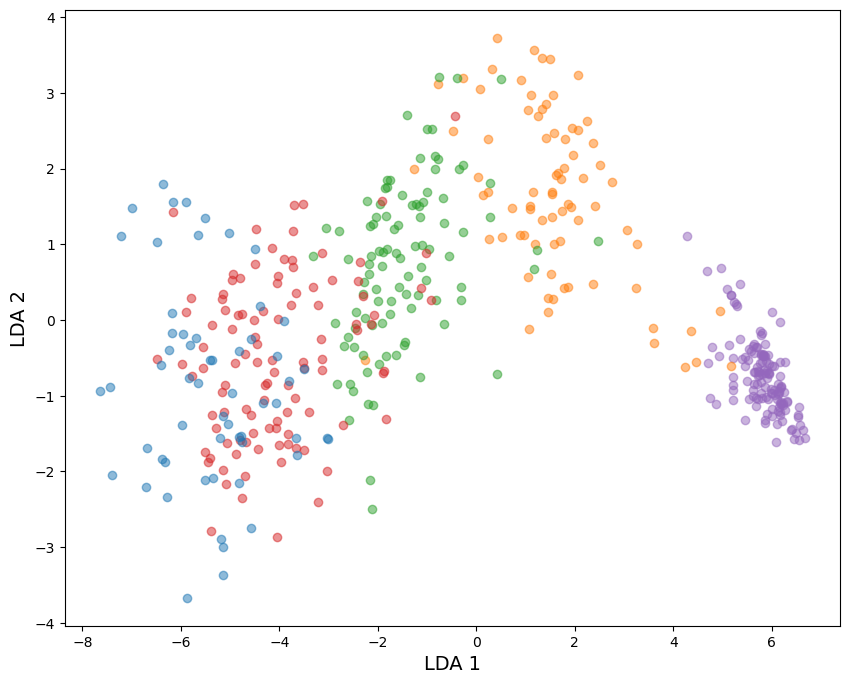
Currently, we have 4 datasets:

1. Features590\_278\_original.txt (590 samples, 278 features)
   1. The original Features590 files
2. Features590\_278\_match\_prob.txt (590 samples, 278 features)
   1. For consistency changed: *MatchKans* to *MatchProbability*
3. Features590\_278\_edited.txt (590 samples, 278 features)
   1. Corrected feature *PercAF\_<locus>* (alleles at locus / alleles in AF file for locus)
4. Features5000\_278.txt (5000 samples, 278 features)
   1. Same as 3, but the samples are created in <program>, 1000 for each NOC
5. Features590\_19.txt (590 samples, 19 features)
   1. Same as 3, but selected the 19 features
6. Features5000\_19.txt
   1. Same as 4, but selected the 19 features

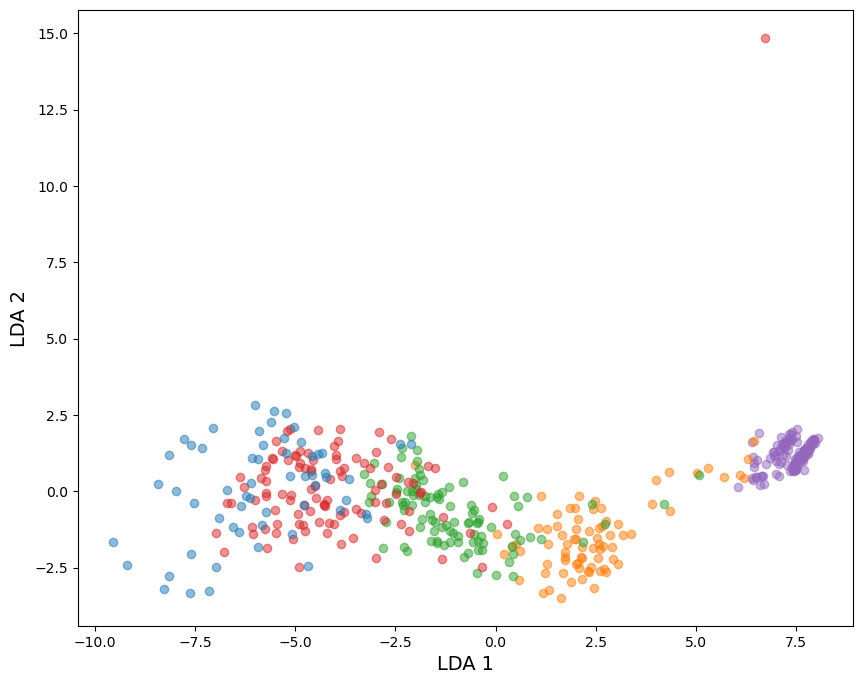
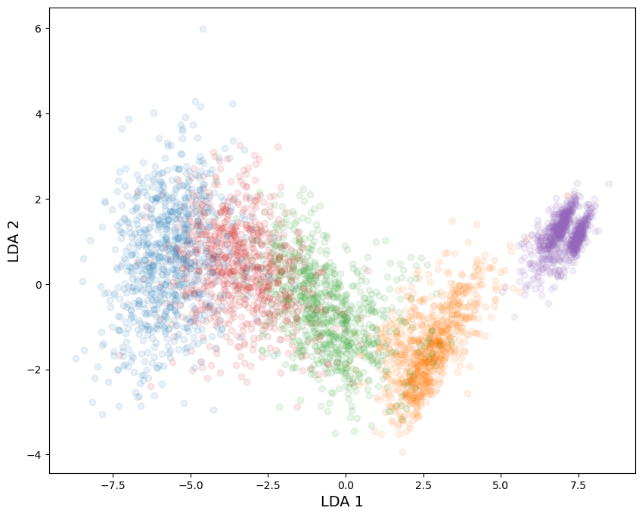
5000 samples generated using the following parameters:

* Populaties: Fusion\_6C\_Holland2
* Kit: PPF6C with high dye-specific detection thresholds
* Drop-in parameters: prC=0.05, lambda=0.01
* NOC: 1 t/m 5, 1000 profiles each
* All real donors must have a minimum LR of 1000 (without conditioning on others)
* Average peak heights: random uniform [100..20000]
* Variation coefficient peak heights: random uniform [0.1..1.0]
* Degradation: random uniform [0.4..1.1]

Running LDA on the original dataset and applying it to the 5000 samples:



Running LDA on the 5000 samples and applying it to the 590 samples:



Fitting random forest regressor on the 590 samples yields an accuracy of 99%. That same one gives ~78% on the 5000 samples test set.

Fitting random forest regressor on the 5000 samples yields accuracy of 100%. That same one gives ~77% on the 590 samples test set.

From the LDA it seems that the 590 dataset explains the variance of the 5000 samples quite well, though it seems that the other way around does worse, observing the worse spread on the 590 samples using the 5000 samples-fitted LDA. The regressors seem to perform practically the same.

Putting the two datasets together..

Ideally, Features590\_19.txt and Features5000\_19.txt should follow identical distributions for each feature. Using ks\_2samp, we determine whether that is the case. “This tests whether 2 samples are drawn from the same distribution.” https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks\_2samp.html

Note that because the 590 dataset does not have a balanced amount of each NOC, it does not line up correctly. Therefore, we run the same test for each NOC.

MAC (statistic=0.12147457627118644, pvalue=3.028903186041987e-07)

TAC (statistic=0.09536949152542373, pvalue=0.00012468202041959753)

CSF1PO min. NOC (statistic=0.040901694915254236, pvalue=0.33060270450852813)

D16S539 min. NOC (statistic=0.09043050847457627, pvalue=0.00033059857811390536)

D1S1656 perc. known alleles (statistic=0.6253762711864407, pvalue=4.707626222937479e-200)

D3S1358 allele count (statistic=0.059108474576271185, pvalue=0.047902519353861805)

D8S1179 allele count (statistic=0.0640135593220339, pvalue=0.025216095365568902)

Penta D min. NOC (statistic=0.04047457627118644, pvalue=0.3428814446282291)

Penta E min. NOC (statistic=0.06196610169491525, pvalue=0.03316633746827513)

TH01 sum of allele freq. (statistic=0.06589491525423728, pvalue=0.019449548628873492)

TPOX allele count (statistic=0.05329491525423729, pvalue=0.09592830798488095)

TPOX min. NOC (statistic=0.015416949152542372, pvalue=0.999441638443494)

vWa peak height variation (statistic=0.41530169491525426, pvalue=5.774705461364854e-83)

Allele count variation (statistic=0.12691186440677965, pvalue=7.19227521983612e-08)

Loci with 0 alleles (statistic=0.04828135593220339, pvalue=0.1647851411722494)

Loci with 5-6 alleles (statistic=0.053813559322033896, pvalue=0.09042907419699797)

Peaks below 800 RFU (statistic=0.1944813559322034, pvalue=5.70838945492804e-18)

Random profile match prob. (statistic=0.09674576271186441, pvalue=9.413882284037278e-05)

Min. NOC (statistic=0.04828135593220339, pvalue=0.1647851411722494)

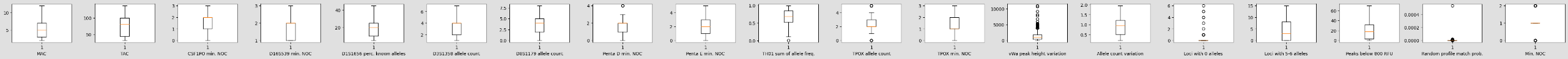


Figure 3: Boxplots 590

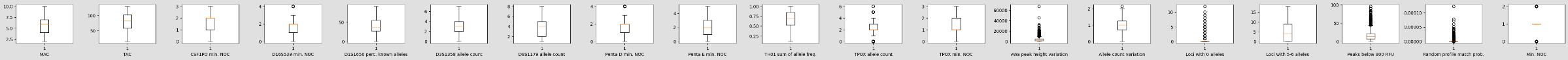


Figure 4: Boxplots 5000

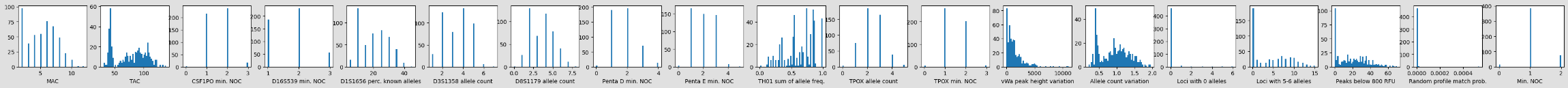


Figure 5: Histograms 590

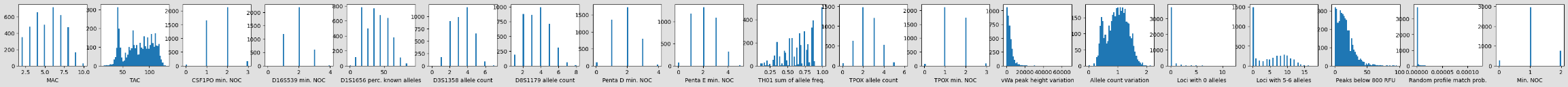


Figure 6: Histograms 5000

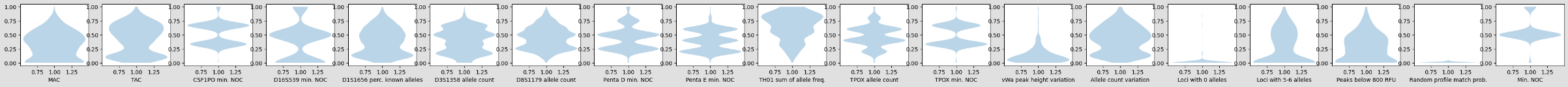


Figure 7: Violin plots 590

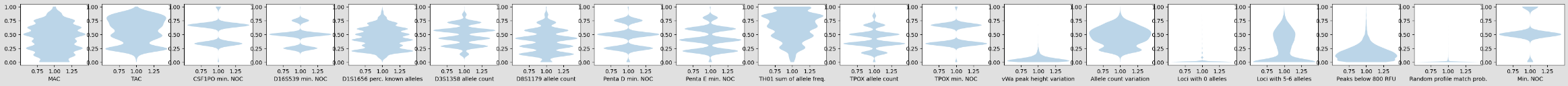


Figure 8: Violin plots 5000

Per true label

NOC 1

MAC (statistic=0.3104545454545454, pvalue=1.3408163468398016e-12)

TAC (statistic=0.11766666666666667, pvalue=0.036105559192703196)

CSF1PO min. NOC (statistic=0.025, pvalue=0.9999748040609778)

D16S539 min. NOC (statistic=0.041939393939393936, pvalue=0.9542546750128711)

D1S1656 perc. known alleles (statistic=0.878, pvalue=0.0)

D3S1358 allele count (statistic=0.039, pvalue=0.9758071091355918)

D8S1179 allele count (statistic=0.05, pvalue=0.8506896274717889)

Penta D min. NOC (statistic=0.063, pvalue=0.6020836981116429)

Penta E min. NOC (statistic=0.049, pvalue=0.8667275643162612)

TH01 sum of allele freq. (statistic=0.10881818181818181, pvalue=0.06424370915331801)

TPOX allele count (statistic=0.03193939393939394, pvalue=0.9976778956413546)

TPOX min. NOC (statistic=0.03193939393939394, pvalue=0.9976778956413546)

vWa peak height variation (statistic=0.36333333333333334, pvalue=0.0)

Allele count variation (statistic=0.29384848484848486, pvalue=2.5879631770919787e-11)

Loci with 0 alleles (statistic=0.12169696969696969, pvalue=0.027359581868654392)

Loci with 5-6 alleles (statistic=0.0, pvalue=1.0)

Peaks below 800 RFU (statistic=0.4645151515151515, pvalue=0.0)

Random profile match prob. (statistic=0.17863636363636365, pvalue=0.0001988929122712646)

Min. NOC (statistic=0.12169696969696969, pvalue=0.027359581868654392)

NOC 2

MAC (statistic=0.14591666666666667, pvalue=0.0429701037232485)

TAC (statistic=0.13975, pvalue=0.05886882824077189)

CSF1PO min. NOC (statistic=0.014083333333333333, pvalue=1.0)

D16S539 min. NOC (statistic=0.11716666666666667, pvalue=0.16594971517030632)

D1S1656 perc. known alleles (statistic=0.8455833333333334, pvalue=1.425390434911319e-71)

D3S1358 allele count (statistic=0.08083333333333333, pvalue=0.58591303811435)

D8S1179 allele count (statistic=0.10858333333333334, pvalue=0.23442723376614938)

Penta D min. NOC (statistic=0.034583333333333334, pvalue=0.9998238334855353)

Penta E min. NOC (statistic=0.0605, pvalue=0.8848144722515316)

TH01 sum of allele freq. (statistic=0.06541666666666666, pvalue=0.8221578164761633)

TPOX allele count (statistic=0.11566666666666667, pvalue=0.17662161407448307)

TPOX min. NOC (statistic=0.11566666666666667, pvalue=0.17662161407448307)

vWa peak height variation (statistic=0.4305, pvalue=4.551914400963142e-15)

Allele count variation (statistic=0.1285, pvalue=0.10093634226851511)

Loci with 0 alleles (statistic=0.10675, pvalue=0.25148749448488084)

Loci with 5-6 alleles (statistic=0.14591666666666667, pvalue=0.0429701037232485)

Peaks below 800 RFU (statistic=0.338, pvalue=2.0461448091424472e-09)

Random profile match prob. (statistic=0.17641666666666667, pvalue=0.0074043150586476925)

Min. NOC (statistic=0.10675, pvalue=0.25148749448488084)

NOC 3

MAC (statistic=0.13441935483870968, pvalue=0.0334266417391641)

TAC (statistic=0.18803225806451612, pvalue=0.0006849656558992612)

CSF1PO min. NOC (statistic=0.05554838709677419, pvalue=0.8644406425011947)

D16S539 min. NOC (statistic=0.0332258064516129, pvalue=0.9993523847909409)

D1S1656 perc. known alleles (statistic=0.863, pvalue=7.771561172376096e-16)

D3S1358 allele count (statistic=0.1988709677419355, pvalue=0.0002660726336689567)

D8S1179 allele count (statistic=0.10145161290322581, pvalue=0.1914862491142464)

Penta D min. NOC (statistic=0.10664516129032257, pvalue=0.15027263442865202)

Penta E min. NOC (statistic=0.1383548387096774, pvalue=0.026267042103007898)

TH01 sum of allele freq. (statistic=0.12529032258064515, pvalue=0.05692082340686477)

TPOX allele count (statistic=0.04306451612903226, pvalue=0.9805675205348048)

TPOX min. NOC (statistic=0.04306451612903226, pvalue=0.9805675205348048)

vWa peak height variation (statistic=0.5628387096774193, pvalue=7.771561172376096e-16)

Allele count variation (statistic=0.06690322580645161, pvalue=0.6787990561755759)

Loci with 0 alleles (statistic=0.030258064516129033, pvalue=0.9998853977344918)

Loci with 5-6 alleles (statistic=0.14474193548387096, pvalue=0.01750073737298674)

Peaks below 800 RFU (statistic=0.4228387096774194, pvalue=7.771561172376096e-16)

Random profile match prob. (statistic=0.14787096774193548, pvalue=0.014247473766676344)

Min. NOC (statistic=0.03583870967741935, pvalue=0.9978771977064038)

NOC 4

MAC (statistic=0.23893939393939395, pvalue=2.4299467826782717e-06)

TAC (statistic=0.16145454545454546, pvalue=0.003985405929092689)

CSF1PO min. NOC (statistic=0.03636363636363636, pvalue=0.996244478427466)

D16S539 min. NOC (statistic=0.04845454545454545, pvalue=0.9332103655696969)

D1S1656 perc. known alleles (statistic=0.9351212121212121, pvalue=1.2212453270876722e-15)

D3S1358 allele count (statistic=0.2043030303030303, pvalue=9.559372766487773e-05)

D8S1179 allele count (statistic=0.07454545454545454, pvalue=0.5097558675549618)

Penta D min. NOC (statistic=0.0383939393939394, pvalue=0.992549286530408)

Penta E min. NOC (statistic=0.054696969696969695, pvalue=0.8551984143751674)

TH01 sum of allele freq. (statistic=0.061242424242424244, pvalue=0.7482507039013719)

TPOX allele count (statistic=0.022454545454545456, pvalue=0.9999999192779455)

TPOX min. NOC (statistic=0.022454545454545456, pvalue=0.9999999192779455)

vWa peak height variation (statistic=0.41903030303030303, pvalue=1.2212453270876722e-15)

Allele count variation (statistic=0.21196969696969697, pvalue=4.462413611650007e-05)

Loci with 0 alleles (statistic=0.014, pvalue=1.0)

Loci with 5-6 alleles (statistic=0.029787878787878787, pvalue=0.9998570372418578)

Peaks below 800 RFU (statistic=0.27493939393939393, pvalue=2.8312494348980977e-08)

Random profile match prob. (statistic=0.13384848484848486, pvalue=0.027591524259529376)

Min. NOC (statistic=0.030333333333333334, pvalue=0.9997963874718985)

NOC 5

MAC (statistic=0.15556164383561644, pvalue=0.0660756442096675)

TAC (statistic=0.22365753424657533, pvalue=0.001791210188627601)

CSF1PO min. NOC (statistic=0.010904109589041096, pvalue=1.0)

D16S539 min. NOC (statistic=0.13532876712328767, pvalue=0.15013171469737518)

D1S1656 perc. known alleles (statistic=0.966, pvalue=2.220446049250313e-16)

D3S1358 allele count (statistic=0.11004109589041096, pvalue=0.3546070536859789)

D8S1179 allele count (statistic=0.078, pvalue=0.7707264263216154)

Penta D min. NOC (statistic=0.09528767123287671, pvalue=0.5334472714536023)

Penta E min. NOC (statistic=0.20771232876712328, pvalue=0.004705552151882841)

TH01 sum of allele freq. (statistic=0.11294520547945205, pvalue=0.3243601843342745)

TPOX allele count (statistic=0.17782191780821918, pvalue=0.023528147475631034)

TPOX min. NOC (statistic=0.0379041095890411, pvalue=0.9999000146205534)

vWa peak height variation (statistic=0.346013698630137, pvalue=9.010346380122769e-08)

Allele count variation (statistic=0.12746575342465755, pvalue=0.1999205641143328)

Loci with 0 alleles (statistic=0.003, pvalue=1)

Loci with 5-6 alleles (statistic=0.11208219178082192, pvalue=0.3330188018385455)

Peaks below 800 RFU (statistic=0.25215068493150683, pvalue=0.0002665865086396435)

Random profile match prob. (statistic=0.19595890410958905, pvalue=0.009112864475040672)

Min. NOC (statistic=0.003, pvalue=1)

When running a randomforestregressor 50 times on the 590 dataset (edited one), and 10 times on the 5000 dataset, as well as the partial correlation for both. The top 25 features are the following:

|  |  |  |  |
| --- | --- | --- | --- |
| RandomForestRegressor top features (590) | RandomForestRegressor top features (5000) | Partial Correlation top features (590) | Partial Correlation top features (5000) |
| **MAC5-6** | **MAC5-6** | **MAC** | **MAC** |
| **TAC** | **meanAllele** | **TAC** | **TAC** |
| **meanAllele** | **TAC** | **stdAllele** | **stdAllele** |
| **MAC1-2** | **MAC3-4** | **MAC5-6** | **MAC5-6** |
| **stdAllele** | **MAC** | LowAF\_Penta E | **peaksAboveRFU** |
| **MAC** | **stdAllele** | ***MatchProbability*** | **MAC7-8** |
| **MAC3-4** | **peaksBelowRFU** | **PercAF\_TH01** | stdHeight |
| ***MatchProbability*** | **MAC7-8** | stdHeight\_vWA | PercAF\_Penta E |
| maxHeight\_D2S441 | Below/AboveRFU | **peaksAboveRFU** | medianHeight |
| **minHeight\_D12S391** | Above/BelowRFU | **peaksBelowRFU** | **MAC0** |
| **stdHeight\_D2S441** | **MAC1-2** | **MAC0** | **minHeight** |
| SumAF\_D10S1248 | minHeight\_D1S1656 | **MAC1-2** | MAC9 |
| **SumAF\_D18S51** | minHeight\_D21S11 | MinNOC\_Penta D | **MinNOC\_D13S317** |
| SumAF\_SE33 | minHeight\_D2S441 | **stdHeight\_D2S441** | LowAF\_D2S1338 |
| minHeight\_Penta E | **minHeight** | maxHeight\_CSF1PO | AlleleCount\_SE33 |
| medianHeight\_D7S820 | minHeight\_FGA | maxHeight\_TH01 | **MAC1-2** |
| SumAF\_D12S391 | minHeight\_vWA | **SumAF\_vWA** | MinNOC |
| **minHeight\_D18S51** | minHeight\_D8S1179 | medianHeight\_D21S11 | medianHeight\_D8S1179 |
| SumAF\_D1S1656 | medianHeight\_vWA | stdHeight\_D16S539 | ***MatchProbability*** |
| SumAF\_D21S11 | minHeight\_D2S1338 | **medianAllele** | **PercAF\_TH01** |
| medianHeight\_Penta E | **minHeight\_D12S391** | AlleleCount\_D8S1179 | PercAF\_D3S1358 |
| SumAF\_D16S539 | **minHeight\_D18S51** | PercAF\_D8S1179 | AlleleCount\_D16S539 |
| maxHeight\_vWA | minHeight\_SE33 | medianHeight\_D5S818 | **SumAF\_vWA** |
| minHeight\_TH01 | **SumAF\_D18S51** | MinNOC\_D2S441 | **medianAllele** |
| SumAF\_TH01 | minHeight\_D7S820 | **MinNOC\_D13S317** | LowAF\_D18S51 |

5 features are always important: MAC, TAC, MAC5-6, MAC1-2, stdAllele

1 feature that was in methods ¾: MatchProbability

# Notes brainstorm & feedback session

## Need for explanations

The need for explanations stems mainly from the fact that the NOC tool is a black-box. Transparency into individual predictions is interesting to inform the user why a certain output is generated. In this way, the system can be understood more clearly. While some experts use the NOC tool every time to help make a decision, others only use it in selected cases.

The tool is mainly used as a support system. When the expert comes to a certain decision, and the tool outputs a different NOC, this discrepancy must be understood. Perhaps the expert missed some information, or conversely, the model made a mistake. For case work, it is important to know which factors have contributed most when a decision is made.

The experts generally look at the electropherogram on its own first, and later confirm their ideas with the tool. This leaves some room for computational time.

## 19 Features

The current features are quite difficult to understand, and its values are not immediately clear as to how they contribute to the end result. Experts would not look at all the features, maybe the top ones. The features used do not seem to provide the bigger picture of a profile, since a selection is made as well (e.g. loci with 5-6 alleles, but not loci with 7-8 alleles). With the help of the visualization, you do grasp the image of a profile better. In this way, a large amount of data is available, in the simplest representation. It will take some getting used to.

For the future, the model should incorporate more information such as replicate runs, major and minor contributors, conditioning on a known contributor, quality of certain channels (the green channel apparently performs worse). Another interesting feature would be the number of peaks at stutter positions. The MAC and TAC values should not be omitted because they are familiar features and the experts like seeing how they fit into the entire explanation. Also, with different kits and thus different datasets, these values can be wildly different to what they are used to. It is nice to “get a feel” for these values with the help of the explanation.

## Visualization profile

It would improve readability if features were grouped logically. For example, put all profile-specific values first, then locus-specific ones.

## Anchors

Anchors do not seem to have the intended effect. Firstly, the idea of a local neighborhood cannot be translated to layperson-terms. This could make users assume that any time a profile comes in for which this Anchor holds, it will be classified accordingly. However, this is not the case. It only holds for the perturbation space. This perturbation space cannot simply be communicated to the user, so the premise of this explanation does not hold. Secondly, Anchors do not inform the user about the factors contributing to the current prediction. The rules that are included in the Anchor are generated stochastically until a certain precision value is reached. This means that the features included in the Anchor are not necessarily the most important ones, but important enough to reach a certain precision. This means that major contributing features could be excluded from the explanation. Users have expressed that the features in the Anchor must be the top ones. Others have even claimed the opposite, since they interpreted these rules to represent the only features that *could* be varied, while others should remain the same. The precision of the Anchor was also often interpreted to be the certainty of the model’s current prediction.

The one positive remark about Anchors was the fact that it shows leeway. For example, if the expert noticed a few unnamed peaks in the electropherogram, they keep in mind that this might raise the true TAC. This Anchor then shows that that would not change the prediction since the rule states any TAC > 103. This shows how Anchors answer the question “*What if value x1 was value x2?*”

## SHAP for multi-class classification

The probabilities that a multi-class classification model outputs are helpful in making an informed choice between multiple options. However, having to analyze multiple images is takes too much effort. It is especially confusing when the same feature values contribute to multiple classes (e.g. TAC = 107 contributes positively to both NOC = 4 and NOC = 5). The responses were mostly positive or otherwise directed at the features, not the type of explanation.

## SHAP for regression

Using regression, we get a more concise report with only one figure. This seemed to lead to generally more positive responses, since it provided a summary of what is going on. They remarked that a decision is usually made between two options anyways. The participants also seemed to agree that the problem is more suited to be a regression problem than a multi-class classification problem.

They asked how to see probabilities or levels on uncertainty of the model with this approach. We demonstrated that if there are many feature values involved, that are dragging the prediction both higher and lower, that indicates that the model sees multiple feature values that are contradicting each other. Also, there is a difference in certainty between a profile that is predicted as 4.1, and another as 4.4. The latter being more unsure than the prior.

## Counterfactual

The users showed interest in knowing why a different prediction was not made, which is the answer to the question “*Why not z?*”. They were interested in such a tool if they were able to input another NOC themselves, since they would use it primarily when they do not agree with a prediction.

[1] Has identified which types of explanations work best for which types of questions. According to them, our experts would like to have “WH-X” and “WH-NOT-Y” questions answered. These correspond to “Why does the model think the NOC is this value?” and “Why does the model think the NOC is not this other value?”. These are respectively best answered with a factual explanation that denotes what in this profile is causing the prediction, and a counterfactual explanation demonstrating what in this profile would need to change the prediction. Both of the questions are classified to be *contrastive*, meaning that they compare something. For the WH-NOT-Y case that corresponds to “Why not Y”, and WH-X to “Why X rather than not X”.

Notes midterm

Why is the Intuitive/realistic desiderata not for factual explanations?

It is fine to use SHAP as a component of your product. You are not going to reinvent the wheel if you do not need to.

For the additional chapters it is important to create a logical storyline; “first this, then that”. Though that does not have to be chronological per se. For example; do not put all the intermediate visualizations of counterfactuals in the chapter; do put the feedback I got as prerequisites / requirements to the visualizations when presenting the final visualization. For Anchors; show why it was considered, and then how I found out that it fails to meet the requirements and thus why it was excluded.

It is important to make clear what the distance function represents (and of course to include it in the report). This distance function must have some sort of meaning (closest world?).

For the counterfactuals; instead of using a median, use a weighted average with a kernel. This is sort of a merged idea of my clustering idea and the naïve version.

Problems

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Naïve CF | Naïve CF+correction | Clustering CF | Kernel CF |
| Distance | ++ (closest) | ++ (closest) | +/- (middle of a cluster) | + (somewhere in between) |
| Realism | ++ (real point) | + (some features are altered) | ++ (real point) | + (new point, but derived from) |
| Intuition | - (can be far from bulk) | +/- | +/- | + |
| Parameters | 0 | Top n closest | Unsolved problem of when to stop | 0 |

Another idea is to draw a line from factual to counterfactual: to find a minimum CF. When do I cross the decision boundary, by linearly stepping 19 features at a time.

For the visualization; organize it so that the same information can be found in the same place. So make the SHAP values be put on the left of the labels that are there. Waterfall plot maybe? Make sure that this new information is encoded differently than the counterfactual. For instance, different colors.

For the evaluation; do not compare my 2-part explanation as two separate explanations. It is meant to be used as a single unit. Instead compare it to the SOTA; SHAP / other CF method. It is not realistic to evaluate an integrated method as separate pieces; they will not be used on their own, they will be used as a unit. It is not relevant to reality to evaluate separate pieces.

Start mixing evaluation and implementation, start writing thesis. In this way, you can add numbers to “how well something works”.

Focus paper on the journal audience.

Notes update

pieken onder detectie-threshold zijn te simuleren in EuroForMix; iets voor de toekomst.

Voor de intuïtie:

* Het is niet gegarandeerd dat je een outlier laat zien met de huidige CF+corr methode.
* Outlier moet je niet laten zien; zorg voor een profiel dat in de buurt ligt van andere profielen
* Waardevol om een echt profiel te laten zien vanwege de sterk-gecorreleerde features
* Het model is niet altijd juist;

Het is waardevol als gebruikers leren over hoe het model werkt. Dus dat die niet-intuïtieve features soms niet erg zijn. Bijvoorbeeld ook omdat ze weten dat er nog andere features zijn.

Vis:

* De pijltjes zijn verwarrend; voor SHAP duwt het de voorspelling omhoog en omlaag, voor CF wat er moet veranderen.
* Het zijn voor SHAP de huidige feature waardes die de voorspelling omhoog/omlaag duwen, niet “een hoge waarde van feature f leidt tot een hogere voorspelling”
* Kleurtjes voor hoe sterk!! (grijswaarden)
* Barchart!

Mening is belangrijk!! Beter inzicht in voorspelling en model, ookal doen ze de taak minder goed, is hetgene wat belangrijk is!

Zet erin dat XAI literatuur alleen op de cijfertjes let, en de mening van de gebruikers waar het nou eenmaal voor is bedoeld achterwege laat.

1. Akula, A.R., et al. *Natural Language Interaction with Explainable AI Models*. in *CVPR Workshops*. 2019.