Explainable artificial intelligence in forensics: realistic explanations for number of contributor predictions

by

Marthe Veldhuis

in partial fulfilment of the requirements for the degree of

**Master of Science**

in Computer Science - Data Science and Technology Track

at Delft University of Technology,

to be defended publicly on Thursday June 17th, 2021 at 13:00.

Student number: 4395778

Project duration: November, 2020 – June, 2021

Thesis committee: Dr. T. Abeel, TU Delft, Chair of the thesis committee

S. Ariëns, Netherlands Forensics Institute, Daily supervisor

Dr. C.C.S. Liem, TU Delft, External thesis committee member

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.



# Preface

The current document entails my master thesis to obtain the Master of Science degree in Computer Science, particularly, the Data Science track. This project has been a collaboration between the Technical University Delft and the Netherlands Forensics Institute (NFI).

Thesis supervisor Thomas Abeel, daily supervisor Simone Ariëns from the NFI

The main focus of the thesis is an academic paper written for FSI:Genetics, this paper and its supplementary materials includes the main work of the project.

The additional chapters describe the broader context of the project including methodology (chapter 1), surveys with users (chapter 3 and 5), and experiments that did not make it into the final product (chapter 4). This follows a mostly chronological order. Chapter 2 contains some additional information that is assumed as prior knowledge for understanding the paper.

Additional thesis chapters

[Preface 3](#_Toc73486294)

[1. Proposal and methodology 6](#_Toc73486295)

[1.1 Problem description 6](#_Toc73486296)

[1.2 Main related works 6](#_Toc73486297)

[1.3 Research questions 7](#_Toc73486298)

[1.4 Methodology and planning 7](#_Toc73486299)

[1.4.1 Risks 8](#_Toc73486300)

[2. DNA Mixture interpretation 10](#_Toc73486301)

[2.1 Short Tandem Repeat (STR) profiles 10](#_Toc73486302)

[2.2 Estimating the Number of Contributors (NOC) 10](#_Toc73486303)

[3. Survey on mixture interpretation and explanation types 13](#_Toc73486304)

[3.1 Set-up 13](#_Toc73486305)

[3.2 Question 1: workflow 13](#_Toc73486306)

[3.2.1 Analysis answers question 1 13](#_Toc73486307)

[3.3 Question 2: feature importance or counterfactual explanations 14](#_Toc73486308)

[3.3.1 Analysis answers question 2 14](#_Toc73486309)

[3.4 Question 3: features or peak information 15](#_Toc73486310)

[3.4.1 Analysis answers question 3 15](#_Toc73486311)

[3.5 Conclusion 16](#_Toc73486312)

[4. Experiments with various XAI techniques 17](#_Toc73486313)

[4.1 SHAP 17](#_Toc73486314)

[4.1.1 SHAP for multi-class classification 17](#_Toc73486315)

[4.1.2 SHAP for regression 17](#_Toc73486316)

[4.1.3 Conclusion 18](#_Toc73486317)

[4.2 Anchors 18](#_Toc73486318)

[4.2.1 Matching the counterfactual Anchor 19](#_Toc73486319)

[4.2.2 Mismatching the input Anchor 20](#_Toc73486320)

[4.2.3 Analysis of experiments 21](#_Toc73486321)

[4.2.4 User interpretation 21](#_Toc73486322)

[4.2.5 Conclusion 22](#_Toc73486323)

[4.3 Counterfactuals 22](#_Toc73486324)

[4.3.1 Existing solutions for counterintuitive counterfactuals 22](#_Toc73486325)

[4.3.2 Generalized counterfactuals 22](#_Toc73486326)

[4.3.3 Multi-objective counterfactuals 23](#_Toc73486327)

[5. Final user study 25](#_Toc73486328)

[5.1 Set-up 25](#_Toc73486329)

[5.2 Demographics 25](#_Toc73486330)

[5.2.1 Demographics results 26](#_Toc73486331)

[5.3 Introduction to explanations 27](#_Toc73486332)

[5.3.1 SHAP 27](#_Toc73486333)

[5.3.2 Counterfactual table 28](#_Toc73486334)

[5.3.3 Compound visualization 29](#_Toc73486335)

[5.4 Regulate trust 30](#_Toc73486336)

[5.4.1 Questions: increase trust 30](#_Toc73486337)

[5.4.2 Answers: increase trust 32](#_Toc73486338)

[5.4.3 Motivations: increase trust 33](#_Toc73486339)

[5.4.4 Questions: decrease trust 33](#_Toc73486340)

[5.4.5 Answers: decrease trust 36](#_Toc73486341)

[5.4.6 Motivations: decrease trust 37](#_Toc73486342)

[5.5 User friendliness 37](#_Toc73486343)

[5.6 Discussion and conclusion 38](#_Toc73486344)

[5.7 Future feature engineering 39](#_Toc73486345)

[6. Reflection on methodology 41](#_Toc73486346)

[References 42](#_Toc73486347)

# 1. Proposal and methodology

## 1.1 Problem description

One of the steps in DNA profile interpretation is determining the Number Of Contributors (NOC) to a DNA sample. The Netherlands Forensics Institute (NFI) previously developed a machine learning model to predict the NOC of a DNA sample based on features derived from Short Tandem Repeat (STR) data [1]. This random forest classifier uses 19 statistical features derived from the STR data which outperforms more conventional methods for NOC estimation. However, the only output that DNA experts can consult is the range of probabilities for each NOC. When the expert analyzes a profile and comes to different NOC than the machine learning model outputs, it is challenging to determine who is correct. No information about how the model came to this conclusion is provided, therefore not allowing experts to use this support tool optimally for their decision making. This decision can be of importance in weight of evidence calculations [2].

With the addition of eXplainable Artificial Intelligence (XAI), the NFI hopes to improve the value of their prediction tool for experts in determining the NOC. XAI has been recognized as a tool to help humans understand the *why* of outcomes in Machine Learning (ML) applications [3-7]. Many of such methods have been developed to understand the factors that influence certain decisions made by ML applications, which is what the NFI is looking for as well. For instance, if the NOC model predicts a different outcome than the expert had in mind, the expert can consult explanations of the model. In this way, the expert can make an informed decision to stick with their own conclusion if the model does not seem to have learned the correct distinctions, or choose the predicted value if the model makes a good case.

## 1.2 Main related works

Explanations have a certain scope; they can be applied to a single prediction, or to the entire ML model. This distinction is defined as local- or global explanations [3-7]. As experts are evaluating the individual predictions of a ML model, they are concerned with local explanations. Besides scope, explanation techniques can either be optimized for certain ML models, or be developed to work for any type of ML model. These are called model-specific or model-agnostic respectively [3-7]. Since the NFI has plans to keep optimizing the ML model for determining the NOC, we intend to focus on model-agnostic methods. There exist roughly two directions of generating local, model-agnostic explanations.

The first is techniques such as SHAP, which has been established as providing effective explanations in the form of the top input features that have driven the model to making a certain prediction [7]. This effectively answers the question *“Why did the model predict A?”.* Some research has implemented SHAP to real-life cases such as predicting hypoxia based on clinical data [8], and predicting the most fitting eye-surgery type [9]. They seem to have obtained valuable information for what are important factors to ML models.

The second direction of explanations is a more recent research direction, which answers the question *“Why did the model not predict class B?”*. This type of explanation is called a counterfactual, showing how the instance could have been predicted differently if certain input features were different [8, 9]. This way of reasoning is underpinned by the social sciences to be effective, as humans seek contrastive explanations [10]. Since this technique is new, numerous methods are being developed, yet none has particularly risen to the top as with SHAP [11].

The literature on generating explanations underpins the value of creating explanations that are catered towards a specific problem, as the effectiveness of explanations is highly sensitive to the audience they are presented to [10]. Therefore, it is good to explore which techniques exist, if they can be applied to this problem, and how they should be adapted to produce the best result.

## 1.3 Research questions

In this study, we aim to generate local, model-agnostic explanations for ML models that predict the number of contributors. To achieve this, we must identify the existing techniques for generating local explanations and the types of assumptions they make on the underlying data. In this way, we can decide which methods might be applicable to the specific dataset that we have available.

*How can we generate informative model-agnostic local explanations for predictions of the number of contributors (NOC)?*

1. What information do experts look at when determining the NOC?
2. What purpose does an explanation of the NOC machine learning model serve?
3. What does the NOC machine learning problem look like?
4. Which types of local explanations could work for this problem?
5. How can an explanation be presented to the users?
6. How can local explanation techniques be adapted to be applied to this problem?
7. How can we evaluate the generated explanations from a machine learning perspective?
8. How can we evaluate the generated explanations from a user perspective?

## 1.4 Methodology and planning

Originally, the thesis was planned in a linear fashion where the phases of development would be processed one by one. After considerations about maintaining quality, it was decided that an Agile approach would be more beneficial. In this way, quality can be monitored by using short cycles which includes all phases of development in a period of three weeks each. The cycles can focus more on certain stages of the process as time progresses. For instance, cycles will be more research-heavy at the earlier stages, and more evaluation-heavy at the end. At the end of each period, reflection can help adjust the course of action. Figure 1 shows the planning in the form of a Gantt chart. Table 1 contains an overview of which research questions were answered in which cycles, and what sections of the thesis contain the corresponding answers

The first cycle was designed to establish a baseline from both a user perspective as from a technological view. From the users we surveyed their usual workflow; how they use the machine learning model; and what is missing. Similarly, we applied some popular explanation techniques to the data as a baseline for the explanations. The data and machine learning model were analyzed. With this sprint, questions 1, 2, and 3 were answered, and a start with question 4 was made.

Cycle 2 was mostly concerned with exploration of several techniques such as Anchors, SHAP, counterfactuals and how these could be combined to further answer question 4. To answer question 5, the visualization was also mainly developed during this time as it was clear that the profile feature values, in combination with a counterfactual would need to be presented to the user.

From previous cycles, it was clear that a combination of SHAP values and a counterfactual could work well. As no counterfactual method was currently fully suitable for this problem, we worked on creating our own implementation, working towards the answer of question 6. This was also when extra data was sampled (see Supplementary Materials).

In cycle 4, the counterfactual method was finalized as well as the total visualization, answering questions 5 and 6. The objective evaluation functions were implemented during this time as well for question 7.

The final sprint consisted of running the objective evaluations against the state of the art, while also creating a user study to answer question 8.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cycle 1: **baseline** | Cycle 2: **exploration** | Cycle 3: **implementation** | Cycle 4: **integration** | Cycle 5: **evaluation** |
| Question 1 | P 1.1, Ch 2, 3 |  |  |  |  |
| Question 2 | P 2.3, 2.4 |  |  |  |  |
| Question 3 | P 2.1, 2.2 |  |  |  |  |
| Question 4 | P 1.2, 1.3, Ch 3 | P 1.3, Ch 4 |  |  |  |
| Question 5 |  | P 2.8 | P 2.8 | P 2.8, 3.2 |  |
| Question 6 |  |  | P 2.5 | P 2.5, 2.6 |  |
| Question 7 |  |  |  | P 2.7 | P 2.7, 3.1 |
| Question 8 |  |  |  |  | P 2.9, 3.3, Ch 5 |

Table 1: Overview of which research questions were answered in which cycles and in which sections of the thesis these can be found.

### 1.4.1 Risks

The main risks of this project relate to the techniques for explanations and user study. Some mitigation steps were defined as follows:

1. The explanation techniques are difficult to implement, slowing down the progress.
   1. Implement any techniques with available code first.
   2. Ask for help from NFI supervisor / colleagues.
2. No users are available to participate in the user study.
   1. Ask feedback from Corina Benschop to get at least one expert evaluation.
   2. Create a substitute task and ask for feedback from colleagues at the NFI.

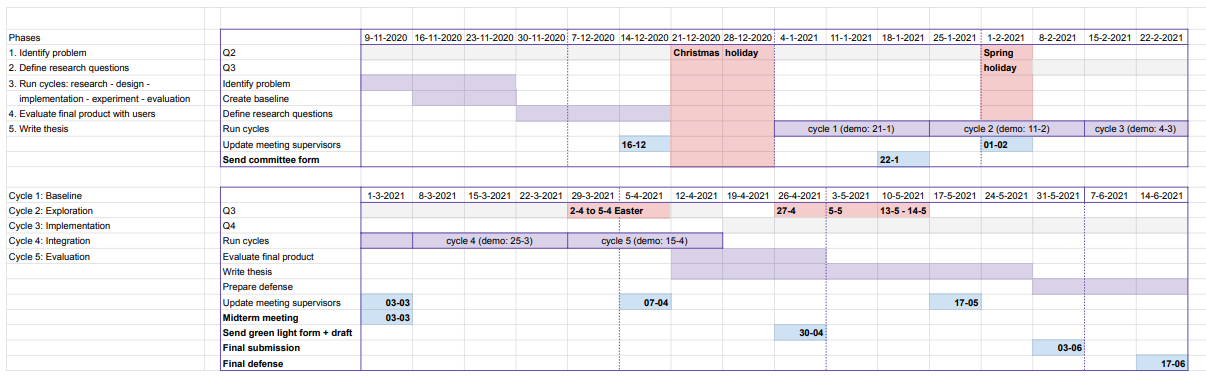


Figure 1: Gantt-chart showing the general planning of the thesis. The phases on the left show the general phases of the project; starting from identifying the problem and the research questions, then moving into cycles of development. The themes of the 5 cycles are listed below, these were adjusted during the process. Some extra time was allocated for the final evaluation of the product with users and finally writing the thesis. All phases and cycles are marked in purple and also listed inside the left-most column of the chart. Milestones and meetings are shown in blue, and holidays in red.

# 2. DNA Mixture interpretation

Experts can use DNA evidence to determine if certain people were involved in a crime by comparing the suspect DNA, victim DNA and other DNA samples to the evidence found at a crime scene. This interpretation becomes more difficult when the DNA profile consists of evidence from multiple people since information might overlap, or not every person contributed as much material. Even though software exists for analyzing this evidence, it is required that the expert inputs how many people contributed to the sample [12]. This chapter explains how to interpret a specific type of DNA profile, and gives a quick impression on how the number of contributors can be determined and which factors might influence that process.

## 2.1 Short Tandem Repeat (STR) profiles

In forensic work, DNA evidence is often analyzed using *Short Tandem Repeat (STR)* profiles. These STRs are specific tracks of repeated short DNA sequences of about two to six base pairs long that have been proven to show high variability between individuals in how many times the sequence repeats [13]. Most of these parts of the DNA or *loci* have been defined by CODIS (Combined DNA Index System), the United States national DNA database. We can capture the STR within three steps. First, a Polymerase Chain Reaction (PCR) is used to amplify the available DNA and label the target loci. Secondly, with Capillary Electrophoresis (CE), the fragments are separated based on size and label. Third and finally, this output is translated with software to an electropherogram. We will not go into detail about how these methods work specifically, as mainly the results are of interest. In Figure 2, we see a simplified result which the electropherogram can produce for locus TH01. The y-axis shows the amount of information found in Relative Fluorescent Units (RFU), which is how the machine counts the quantity of DNA found. The top of x-axis shows the number of base pairs, a measurement for the entire sequence found. Most importantly, we see two peaks, representing two alleles on this locus. These alleles are characterized by the number of repeats of the STR for locus TH01, which is [AATG]. On the right of Figure 1, we see the DNA sequence for six and eight repeats.

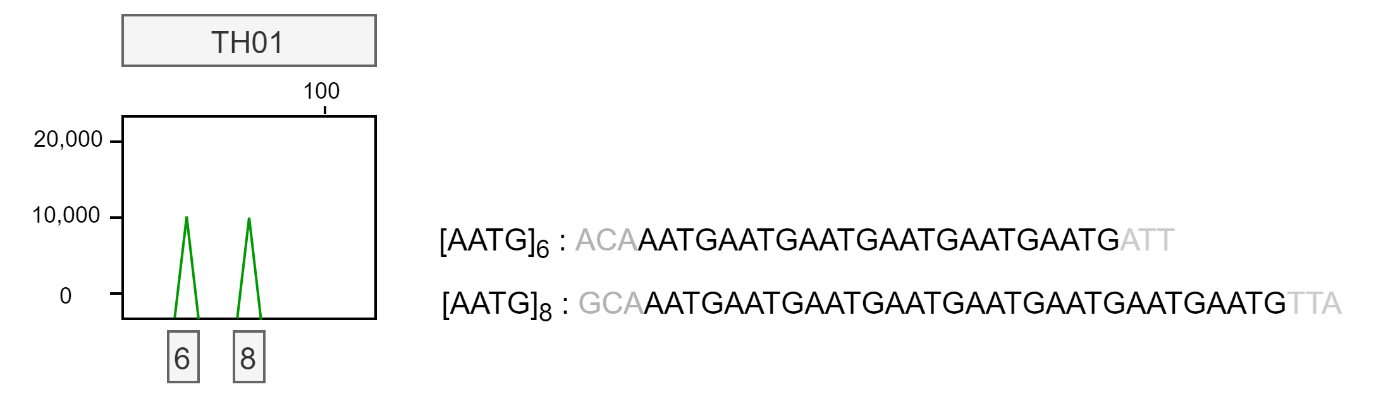


Figure 2: Simplified electropherogram result for locus TH01 showing two alleles with six and eight repeats each. The repeat sequence is shown on the right with arbitrary flanking regions that do not represent reality.

One individual can have two different alleles for a single locus; one inherited from the mother, and one from the father. It is also possible that a person inherited the same allele from both of their parents, this means that they are homozygous at that locus. The peak will then be twice as large. We will now get into more detail of how to derive the number of contributors from an STR profile.

## 2.2 Estimating the Number of Contributors (NOC)

The first step of DNA STR profile interpretation is to determine whether a sample has originated from a single source, or if the sample is a mixture [14]. This often easily discerned by checking whether or not there are loci with more than two alleles present. As we saw in Figure 2, a single person can contribute a maximum of two alleles per locus, so profiles with more alleles are usually considered a mixture. The next step is to determine the number of contributors (NOC). This step is necessary for DNA analysis software to calculate the *weight of the evidence* found [15]. An incorrect NOC can have an effect on this weight of evidence [2]. In extreme cases it could make the difference between support for the proposition that a person of interest (POI) contributed to the evidence or the support for the alternative, that an unknown person, unrelated to the POI, contributed to the evidence.

Determining the exact number of contributors can be challenging. There are several obscuring factors that could make an expert underestimate the number of donors, especially when the number of donors increases [14, 16]. While the two left-most pictures in Figure 3 are quite simple to interpret, the two on the right are somewhat ambiguous because of the factors described below.

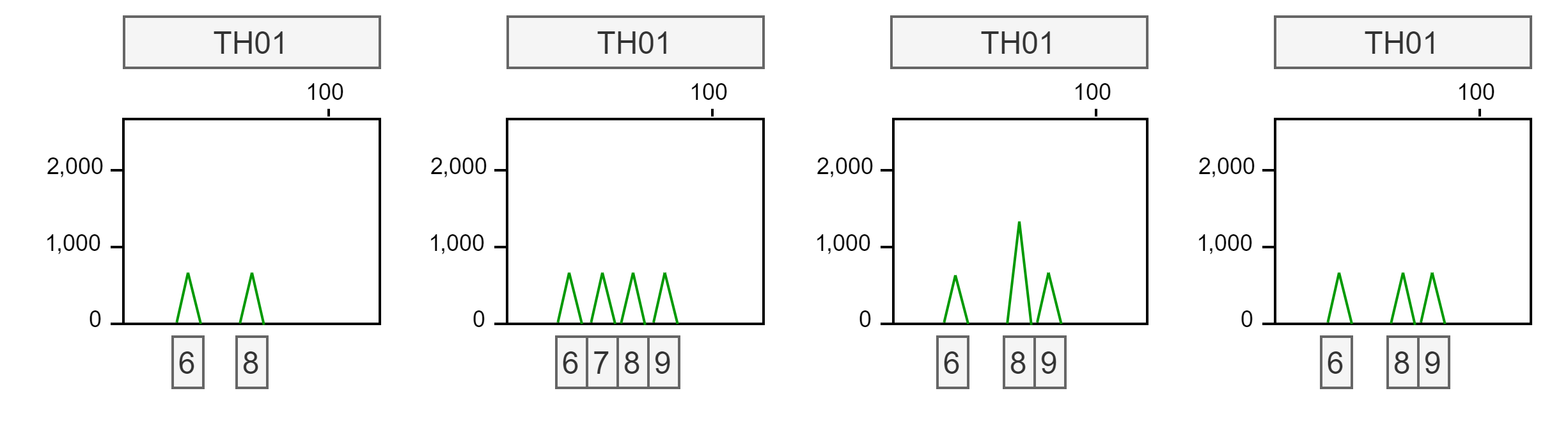


Figure 3: Four simplified electropherogram results for locus TH01. From left to right: Example of a simple single donor profile; Example of a simple 2-person mixture profile; Example of a 2-person mixture profile with allele sharing or a homozygous allele, peak 8 is twice as high compared to peak 6 and 9; Example of a 2-person mixture profile with drop-out, one peak has likely not been detected.

* Allele sharing: If two donors have the same allele at a locus, this is called allele sharing. It frequently occurs when donors are relatives, since siblings share a lot of DNA. It might be difficult to distinguish if an allele is shared between donors, or if a single donor simply is homozygous for this allele; in both cases, the peak height for that allele is higher. This can be seen in the third picture from the left of Figure 3; allele 8 is twice as high as alleles 6 and 9.
* Allele drop-out: If the DNA was degraded, for example due to sunlight, some parts of the DNA might not be present in the sample to measure. It is also possible that the amount of available DNA is so small, that the alleles fall below a certain detection threshold. Because of this low quality or quantity of DNA, some allele fragments might not show up in the profile at all, which is called drop-out. This can be seen in the right-most picture of Figure 3; only 3 alleles are found.

These factors can decrease the number of alleles found in a certain profile, which could lead to an underestimation of the number of contributors. There are also factors that could lead to an overestimation of alleles present in a sample, and thus an overestimation of the NOC [14]. The two right-most images in Figure 4 demonstrate these phenomena which are also described below.

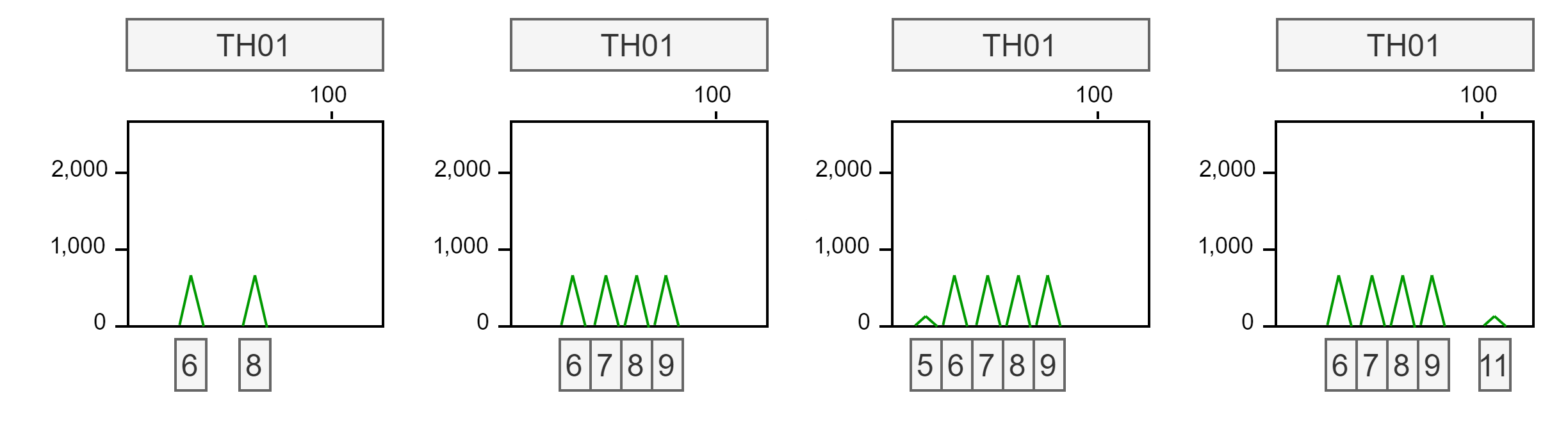


Figure 4: Four simplified electropherogram results for locus TH01. From left to right: Example of a simple single donor profile; Example of a simple 2-person mixture profile; Example of a 2-person mixture profile with a stutter peak at allele 5 caused by the folding of a STR with 6 repeats; Example of a 2-person mixture profile with a noise peak at location 11 caused by an error in reading or low-level contamination event.

* Stutter: During the process of measuring the STRs, a STR fragment can accidentally fold over itself. This could cause the electropherogram to measure this strand to have one repeat fewer, since the folded part of the fragment is not correctly measured. In this way, a small stutter peak is found in the profile just before the valid peak. In Figure 4, this is shown by the small peak at allele 5, caused by the folding of the STR with 6 repeats. Note that also +1, -2, or even complex +/-0.5 stutter peaks can occur.
* Allele drop-in or other noise: The measuring process is not perfect, so some random noise or low-level contamination might show up in the electropherogram, that does not contain any information about the DNA. In Figure 4, we can see that the rightmost image has a small peak at allele 11. Since it is not close to another allele, it is likely not a stutter peak.

Stutter peaks and noise are often filtered out using certain thresholds in the profile analysis software. As a result, some DNA information might also be lost if there is little material available. All of the previously discussed events are more troublesome and prevalent when there is little information available (low template).

The simplest method to get an estimate of the NOC is by using the Maximum Allele Count (MAC)-method [14, 17]. By taking the locus with the most alleles present, dividing that number by two, and rounding up, we can get an idea of the minimum NOC. Though this method is simple, it does not take into account the factors of allele sharing, drop-out, etc. Therefore, the performance is quite poor with 3 or more contributors, when there is a lot of allele sharing, or when the quality of the profile is low [18, 19]. When assessing mixtures between 2-5 contributors, depending on the dataset, the MAC can obtain correct predictions for about 60-70% of samples [1, 19, 20].

Besides the MAC, the Total Allele Count (TAC) can also give an indication for the NOC. This is a count of all the found alleles of all loci, which gives a more general overview of the profile as a whole [14]. A combination of these two measures can give a better impression of the entire profile [14, 20].

# 3. Survey on mixture interpretation and explanation types

From the background information, we obtained a good grasp of how the NOC can be determined. This survey was then run to verify that the experts at the NFI had a similar workflow, thought process and looked at similar data. There were 12 responses in total.

## 3.1 Set-up

The survey was structured according to three main questions:

1. What is the normal workflow of experts when estimating the NOC?
2. What type of explanation do experts prefer to help them make a decision (feature importance or counterfactual)?
3. What type of data do experts prefer to help them make a decision (features or peak information)?

Question 1 verifies the workflow of the experts to see if any information was overlooked. Questions 2 and 3 relate to the possible types of explanations that could be implemented. For explaining single predictions in a model-agnostic fashion, there are two main approaches that work well for tabular data; feature importance methods and counterfactual explanations (see Literature Survey x). To confirm that these are valuable for this specific problem, this survey contrasted these types of explanations. The type of data that is presented to the user is also important. Currently, the machine learning models crafted by the NFI are based on features concerning summary statistics of the profile (such as the TAC and the MAC). However, it would also be possible to train (deep) models on the peak information to create predictions. This information concerns the peak location and size for all loci per profile. With this question, we intended to find any preference regarding the data.

## 3.2 Question 1: workflow

This question describes an average workflow as interpreted by the literature. The users were asked to write any missing steps. In summary:

* Inspect general information about the profile (peak heights, TAC, MAC, NOC tool prediction);
* Check the locus with the MAC to see if all peaks can be explained with the expected number of donors;
* Check for stutter peaks / extra peaks from another donor.

### 3.2.1 Analysis answers question 1

The following remarks were reported to be missing from the workflow:

* Check the number of peaks below the detection threshold (6x). This gives an indication of the DNA quality (1x) / the amount of dropout (3x).
* Experts can often not make a reliable choice between 4 or 5 donors based on the information (1x).
* Locus SE33 (1x).
* None (3x).

In summary, missing information concerns the number of peaks below the detection threshold which is not available to the machine learning model, so this information could not be incorporated. The remark about 4 or 5 donors demonstrates the difficulty of making decisions with more donors. Locus SE33 has the largest variety of alleles, which is why one user finds it more informative. The remarks suggest that **more and/or different features could be useful for the machine learning model in the future.**

## 3.3 Question 2: feature importance or counterfactual explanations

This question describes two types of explanations for the same prediction of a profile. Option A are SHAP values, while option B is a counterfactual explanation. The users were asked to choose which explanation would be most helpful to make a decision between two NOC values (4 or 5). They could also pick both options. The question is shown in Figure 5 and the results are presented in Figure 6.



Figure 5: Question 2 with option A showing feature importance values and option B a counterfactual explanation.

### 3.3.1 Analysis answers question 2

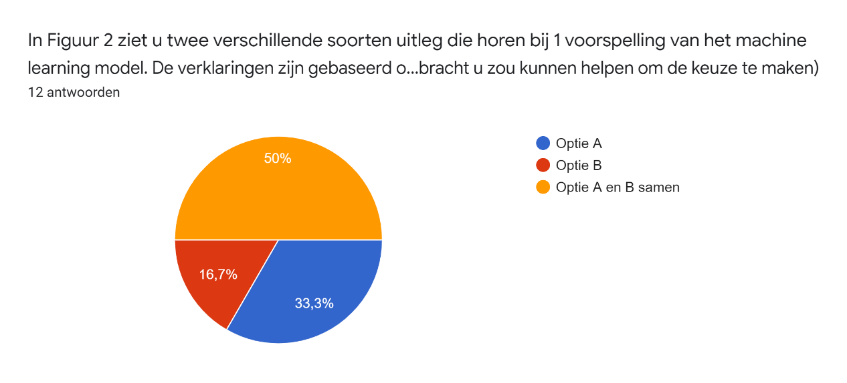


Figure 6: Pie chart of answers to question 2.

Motivation for choosing A:

* You want to know why the model predicted its result (2x)
* Easier to understand (1x)
* Option B is also good, but a visual explanation is better (1x)
* Option B is also good, but can the model know if the expert is interested in 4 or 6? (2x)

Motivation for choosing B:

* Is more specific for comparing one to another (2x), which is relevant for criminal case investigations (1x), option A is more background information.
* Option A and B together is also a good option.

Motivation for choosing A and B:

* Option B can provide very specific information (1x) (e.g. if the allele count on one locus were lower to get a different NOC, and it could be explained by stutter).
* Option B is relevant when you came to a different NOC than the tool outputs (2x)
* Option A tells you why it came to its result in the first place (4x).
* Option B tells you where the threshold values lie (1x).
* Option B tells you if the predictions were close together (1x).
* More convincing (1x)
* Combination of information makes the decision complete (1x)

In summary, **most users liked the** **combination of explanations to form a complete picture**. People that picked one option, often also mention they liked the other as well. **They enjoyed the general information of the feature attributions, and the specific values of the counterfactuals.** The counterfactual seemed to provide extra information such as giving an impression of the threshold values and how close the decision is.

Since option A had a visualization, as opposed to option B, it could have induced some presentation bias as seen in one of the responses.

## 3.4 Question 3: features or peak information

This question describes two counterfactual explanations based on different types of data. Option A consists of the features that are currently used by the machine learning models. These are mainly summary statistics that describe aspects of the profile. Option B shows information about peak heights. The users were asked to choose which explanation would be most helpful to make a decision between two NOC values (4 or 5). The question can be found in Figure 7 and its answers in Figure 8.

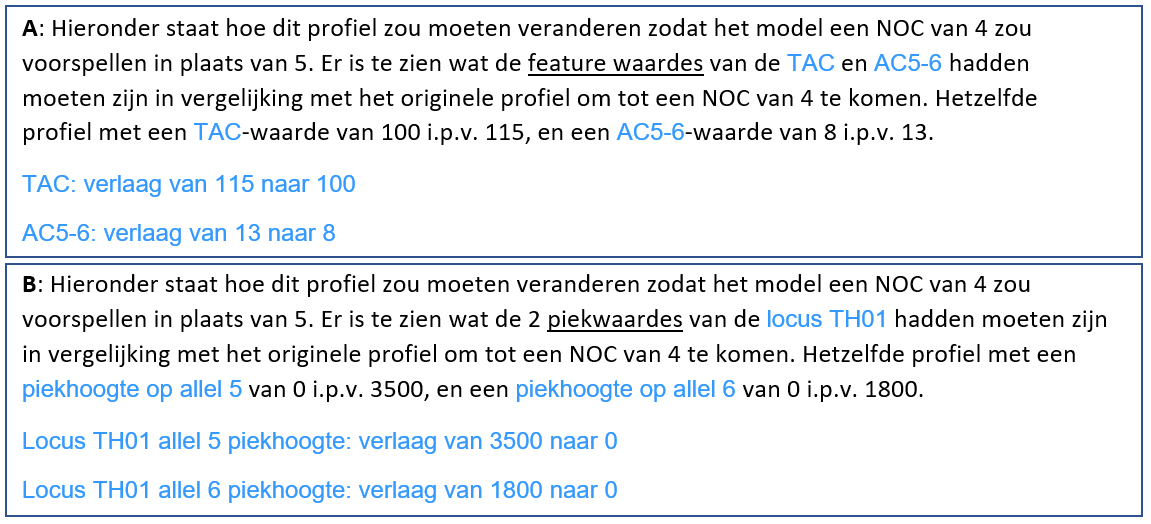


Figure 7: Question 3 with option A showing an explanation based on features and option B showing an explanation using peak height information.

### 3.4.1 Analysis answers question 3

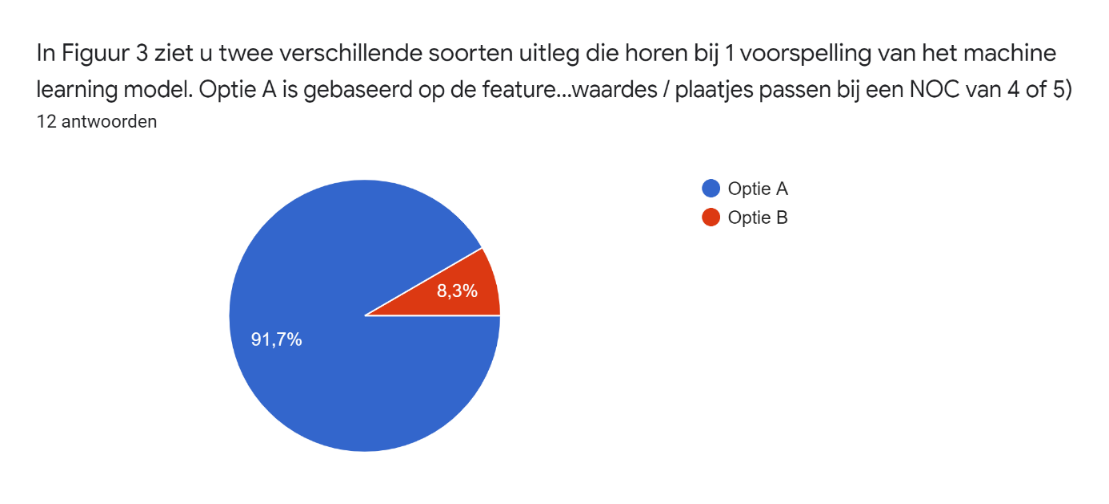


Figure 8: Pie chart of answers for question 3.

Motivations for option A:

* When you change the number of peaks or peak heights, you essentially change the TAC / MAC / other features (1x).
* Changing peak heights / removing peaks at a certain locus does not make sense (2x).
* Option B seems too trivial, it sounds like the prediction is based on only one locus (2x).
* Prefer to look at TAC / MAC (profile-level), not the peak heights at one locus (3x).
* Option B would require a lot of research into individual loci, option A is enough (1x).
* Option A gives more information than the expert can see, whereas option B the expert probably already noticed (2x).
* Peak heights are not stable for the PPF6C kit (1x).

Motivations for option B:

* Option B could be useful to characterize the imbalance between peak heights. They also mention that they would like to see information such as stutter levels, drop-ins, TAC, MAC, and mention that a combination would be ideal.

In summary, **every participant sees the** **value in profile-wide features**. Most people mentioned that they mainly **consider the profile as a whole, and would not consider peak heights at a single locus informative**. On top of that, one expert found that the peak heights in the used kit are not stable so making a decision on that information might not be a good idea. Features also give new information, whereas the peak heights are already available to the expert. The one person who picked option B mentions that there could be value in the ratio between peak heights, while also expressing their interest in the features. It could be interesting to **encode this ratio into a new feature.**

Because option B only adjusted peaks at one locus, a lot of experts expressed that they would never base their decision on a single locus and were therefore a bit confused. It would have been better to have presented multiple changes at different loci to mitigate that.

## 3.5 Conclusion

The workflow matched well with our expectations, there were no unexpected answers. Regarding the explanations, it was interesting to see that the experts found most value in the combination of the two types. Where feature attributions give a general impression of the prediction, the counterfactual provides more specific information showing where the threshold of the prediction lies. It was surprising to see that almost nobody found the peak data informative to demonstrate with the explanation. This could be attributed to the fact that we only presented information about a single locus, or because this peak information is already available to the experts when they analyze a profile, therefore not resulting in new insights.

# 4. Experiments with various XAI techniques

From the user survey and an additional brainstorming session with the DNA experts, we determined that there were two questions that required an answer:

1. *What were the main reasons for the model to reach the current prediction?*
2. *With which feature changes could the model have arrived at a different prediction?*

It seems that for question 1, a general overview of feature importance is adequate. Then to answer question 2, a specific example is fitting. In this way, there is both general information about the current prediction, as well as a specific example of a close different prediction. For both questions 1 and 2, we wanted a technique that is both model-agnostic, and presents an explanation per prediction, so is local.

The following sections show several experiments that ultimately led to the requirements and implementation found in the paper. These experiments show how several techniques were considered.

## 4.1 SHAP

SHAP is one of the most established feature importance methods with a solid theoretical foundation [21]. From question 2 of the user survey, we received positive feedback on the general information that the SHAP values can provide about a prediction, which is why we wanted to incorporate it into our explanation. As we were on the verge of deciding between classification and regression, we wanted to explore what information SHAP values can provide in both settings, and which version the users prefer. SHAP values can be presented in various ways, but we mainly wanted to focus on the information that it conveys. Therefore, we used one of the simplest visualizations that the package offers, *force plots*.

### 4.1.1 SHAP for multi-class classification

Figure 9 shows an example of the explanation for a 4-person mixture profile, which is predicted as such by the original classification model with a probability of 0.67. Note that this figure only shows the explanation for class ‘4’, with red bars representing feature values that make this prediction more certain, and blue bars representing feature values that make this prediction less certain. The force plot is two-dimensional, so the red and blue bars can only represent two directions such as “certain” and “less certain”. If the expert were to explore the full range of the prediction, that means they would need to look at five separate force plots.

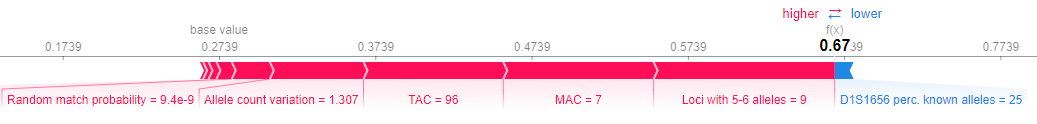


Figure 9: SHAP force plot for classification.

The experts liked that this explanation gives them access to the probability of the prediction, and they enjoyed the overall visualization. However, they would not want to look at multiple figures because that takes too much time and effort to understand and compare. They also thought it was confusing that the same feature values can contribute positively to multiple classes. In essence, they do not want to extrapolate the relevant information themselves.

### 4.1.2 SHAP for regression

By using SHAP force plots in the context of a regression model, we obtain a more concise report. For example, in Figure 10 the same profile as in Figure 9 is predicted with a regression model as 3.96. The red bars now represent the feature values that push the prediction “up”. If there were blue bars, these would push the prediction “down”. The two-dimensional scale can now represent the entire range of the outputs from 1-5, meaning that only one figure is needed to show all possible values of the output.

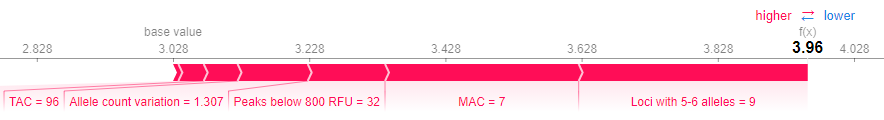


Figure 10: SHAP force plot for regression.

Users generally liked this version better, because they would only need to consult one image and it provides a full summary of what is going on. They also mentioned that handling the NOC problem with regression felt more natural that multi-class classification, as the output is on an ordinal scale. With a bit of explaining, the users could understand that a prediction of 3.96 is more certain than for example 3.55. We also demonstrated that if many red and blue bars are working against each other, that might also indicate the uncertainty of the model.

### 4.1.3 Conclusion

Users seemed to prefer SHAP values within the context of regression since they only have to analyze one consistent image which represents the entire range of output values. With some coaching, we believe the experts will be able to interpret the SHAP values well enough to provide them some insight into the prediction.

## 4.2 Anchors

From the initial user study, we found that experts valued the specific values of counterfactuals. For that reason, Anchors seemed like a good method to generate the *factual* side of the explanation. This is because Anchors consists of feature value ranges, that “anchor” a prediction in a certain region of the feature space [22]. In other words, if the Anchor holds, the output can be predicted with high probability. This means that the features not included in the Anchor can vary and the prediction will stay the same. Another advantage from this approach was that these Anchors can be presented in a visually attractive way on the base visualization that we already had designed (see Figure 11). The idea was to combine both Anchors with counterfactuals to incorporate factual- and counterfactual information into one picture.

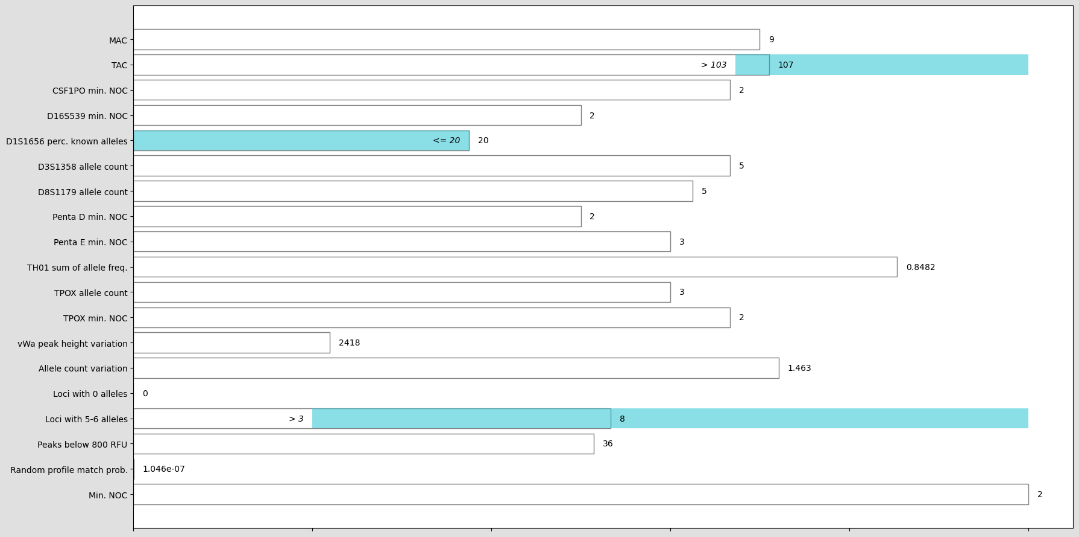


Figure 11: Anchors visualization. The light blue shaded values represent feature ranges for which the same prediction holds.

The definition of Anchors also seemed to lend itself to create sparse counterfactuals from. Instead of showing the differences in feature values between the input and counterfactual, we could present the differences in their Anchors. As Anchors only consist of a handful of rules, this would create a sparser result.

4.2.1 Matching the counterfactual Anchor

The first experiment explored the idea whether valid counterfactuals could be produced when derived from these differences in Anchors. We used the following steps for some initial experimentation:

1. Generate an Anchor for the input (left side of Table 2).
2. Find the closest training point with a target prediction to the be counterfactual.
3. Generate an Anchor for that counterfactual (right side of Table 2).
4. Change the input to match the counterfactual Anchor (Table 3).
5. Check that the prediction changes to the target.

|  |  |
| --- | --- |
| Anchor for the input instance | Anchor for the counterfactual instance |
| *The model will predict* ***3 contributors 96%*** *of the time when ALL the following rules are true:*   * *TAC <= 79.00* * *Loci with 5-6 alleles > 1.00* * *D8S1179 allele count <= 3.00*   *These rules hold for the original data with a probability of 0.02* | *The model will predict* ***2 contributors 97%*** *of the time when ALL the following rules are true:*   * *Allele count std. <= 0.93* * *Random match probability <= 0.00* * *D8S1179 allele count <= 2.00*   *These rules hold for the original data with a probability of 0.02* |

Table 2: Two Anchors for Experiment 1 that led to a successful result.

|  |  |  |
| --- | --- | --- |
|  | Input value | New value |
| Allele count std. | 1.03 | 0.70 (<= 0.93) |
| D8S1179 allele count | 3.00 | 2.00 (<= 2.00) |
| Prediction | **3.0** | **2.0** |

Table 3: Changing the input features that do not match the counterfactual Anchor caused the prediction to be **2.0**.

As can be seen in Table 3, the prediction did change after matching the counterfactual Anchor. However, problems occurred when:

* The input instance already fit the Anchors of the counterfactual.
* Changing the input to fit the Anchors of the counterfactual did not lead to a change in prediction. An example of this is shown in Tables 4 and 5.

|  |  |
| --- | --- |
| Anchor input | Anchor counterfactual |
| *The model will predict* ***1 contributors 93%*** *of the time when ALL the following rules are true:*   * *TPOX min. NOC <= 1.00* * *Allele count std. <= 0.65*   *These rules hold for the original data with a probability of 0.29.* | *The model will predict* ***2 contributors 100%*** *of the time when ALL the following rules are true:*   * *Allele count std. <= 0.83* * *TAC <= 79.00* * *vWa peak height std. > 3024.43*   *These rules hold for the original data with a probability of 0.01.* |

Table 4: Two Anchors for Experiment 1 that did not lead to a successful result.

|  |  |  |
| --- | --- | --- |
|  | Input value | New value |
| vWa peak height std. | 108 | 5000 (> 3024) |
| Prediction | **1.0** | **1.0** |

Table 5: Changing the input features that do not match the counterfactual Anchor caused the prediction to stay 1.0.

So even though the Anchors for the counterfactual prediction were fulfilled, the prediction stayed the same. Perhaps it was caused by the fact that the Anchors for the input also still held. Therefore, we conducted a different experiment.

### 4.2.2 Mismatching the input Anchor

The procedure is similar to the first experiment with the exception of step 4; here we also want to ensure that the instance no longer matches the input Anchor. We used the same profile as in Table 5.

1. Generate an Anchor for the input (left side of Table 4).
2. Find the closest training point with a target prediction to the be counterfactual.
3. Generate an Anchor for that counterfactual (right side of Table 4).
4. Change the input to match the counterfactual Anchor **and to no longer match the input Anchor** (Table 6).
5. Check that the prediction changes to the target.

|  |  |  |
| --- | --- | --- |
|  | Input value | New value |
| vWa peak height std. | 108 | 5000 (> 3024) |
| Allele count std. | 0.46 | **0.79 (<= 0.83** **but not <= 0.65**) |
| Prediction | **1.0** | **2.0** |

Table 6: Changing the input features that do not match the counterfactual Anchor, and did match the original Anchor caused the prediction to change to 2.0.

However, it apparently matters how much this value is changed. In Table 7, we see that it does not always work.

|  |  |  |
| --- | --- | --- |
|  | Input value | New value |
| vWa peak height std. | 108 | 5000 (> 3024) |
| Allele count std. | 0.46 | **0.70** (<= 0.83 but not <= 0.65) |
| Prediction | **1.0** | **1.0** |

Table 7: By changing the value of Allele count std. slightly less that in the previous attempt, the prediction stayed at 1.0.

Perhaps this has to do with how Anchors are binned; all data is discretized before Anchors are generated. However, looking at the bins for feature Allele count std., 0.79 and 0.70 do not belong to a different bin.

* 0:'Allele count std. <= 0.38'
* 1:'0.38 < Allele count std. <= 0.45'
* 2:'0.45 < Allele count std. <= 0.65'
* **3:'0.65 < Allele count std. <= 0.83'**
* 4:'0.83 < Allele count std. <= 0.93'
* 5:'0.93 < Allele count std. <= 1.05'
* 6:'1.05 < Allele count std. <= 1.17'
* 7:'1.17 < Allele count std. <= 1.28'
* 8:'1.28 < Allele count std. <= 1.44'
* 9:'Allele count std. > 1.44'

This seemed strange; the Anchor on the right side of Table 4 states that it will predict 2 contributors 100% of the time when the stated rules are true. This started a suspicion that this does not mean that all other feature values can have any random value. Considering this same counterfactual Anchor from Table 4, setting all other feature values to zero results in a prediction of 1.0. The same prediction apparently does not apply when arbitrary values for the features not included in the Anchor’s rules are chosen.

### 4.2.3 Analysis of experiments

From the experiments, some information had to be uncovered in some more detail.

First of all, though it is described that when an Anchor holds “changes to the rest of the feature values of the instance do not matter” [22], we have seen that is not true for *all* of the feature space. Within the more formal description of the method, one can derive that an Anchor only holds for a sampled subspace around the input instance.

Secondly, this subspace is generated by randomly sampling all feature values not included in the Anchor, which we only uncovered by analyzing the source code. This process likely cannot generate realistic data points for our dataset of highly correlated features. The method description does not clearly specify how sampling is done or how they determine what is still considered local. It is therefore quite difficult to determine for which part of the data the explanation holds. This provides the explanation as to why setting all feature values excluded from the Anchor to zero did not work; an instance where most of the feature values are zero are not part of the local neighborhood of a normal input instance. The precision and coverage that they denote with each Anchor are also based on this perturbation space. Looking at the counterfactual Anchor in Table 4, it holds with a probability of 0.01 for the sampled data. If that Anchor holds, then the prediction of 2 contributors is 100% certain. This does not mean that it applies in the same way for the original data. This could explain why the counterfactual Anchors did not work out:

* The Anchors for seemingly similar instances did not fall into the same *local neighborhood*.
* The sampled data does not represent the dataset accurately, especially since they use *random sampling*, and therefore the specific values of the Anchors do not hold for the dataset.
* Our approach of changing the input instance to match the counterfactual Anchor creates an instance that is unrealistic and/or does not match the perturbed data.

Lastly, Anchors do not inform the user about the most *influential features* for the current prediction which was the initial goal of using Anchors; feature importance with the addition of ranges. 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. In this way, major contributing features could be excluded from the explanation.

### 4.2.4 User interpretation

We presented the picture in Figure 11 to the DNA experts in a brainstorming session about the visualization and asked how they interpreted it. Most users saw the Anchors as the most important features for the current prediction. As we determined, this is not the case due to the stochasticity with which the features are added to be part of an Anchor. This stochasticity has another undesirable effect; if the same profile is explained multiple times in a row, different Anchors will be shown. This not help with user interpretation. Other users have even interpreted that the Anchors represent the only features that *could* be varied, while others should remain the same. Lastly, the precision of the Anchor was mostly misinterpreted as the certainty of the model’s current prediction.

After attempting to explain how Anchors are generated, we noticed that idea of a local neighborhood cannot be translated to layperson-terms. Since the perturbation space cannot be communicated to the user, the context of this explanation is unclear.

The one positive remark about Anchors was the fact that it shows a range. One expert mentioned that in the example of Figure 11, if more alleles were present in the profile by for instance lowering the detection threshold, the TAC would increase. However, as the Anchor specifies that any TAC > 103 holds for this prediction, the same output would occur. This could also be solved by simply inputting a different value into the model, and seeing if the same prediction holds.

### 4.2.5 Conclusion

The current implementation of Anchors is not very useful for this problem because:

* Anchors do not represent the most important feature values that have led the model to the current prediction, even though users will interpret it as such.
* Anchors can differ between multiple runs, which is confusing for users.
* Anchors do not hold for a real dataset with correlated features because the random sampling process does not generate realistic data.
* Anchors only hold for a local neighborhood, which cannot be communicated clearly to users.

## 4.3 Counterfactuals

At first, a base counterfactual was implemented. This was simply the closest counterfactual instance from the training data. This instantly uncovered one of the fundamental problems with presenting an instance from the training data; not all of the presented differences are relevant, some might even counteract the prediction. For example, in one of the instances it showed to *increase* the TAC to *decrease* the NOC. This is counterintuitive since more alleles correspond to more contributors. To solve this issue, we considered several options before landing at the final ReCo implementation.

### 4.3.1 Existing solutions for counterintuitive counterfactuals

One way to mitigate counterintuitive examples, is to show multiple (diverse) profiles so that a more generalized picture is painted to the user [23-25]. However, in this way the users are burdened with having to extrapolate which information is relevant by themselves.

Similarly, by showing the distribution of multiple profiles you can get the same effect [26]. However, distributions are not informative to users that are not familiar with them [10] . They can also occlude local effects that are not visible from a global perspective.

Finally, a suggestion was made to enhance sparsity in training data counterfactuals by introducing a matching tolerance [27]. This means that if the difference between a feature value from the input and the counterfactual is small, for example less than 5%, they would be considered equal. However, we argue that this might overlook the exact threshold values on which a model makes a decision. This could thus lead to missing the target prediction.

### 4.3.2 Generalized counterfactuals

To counteract the previous issue related to presenting an example data point from the training data, we brainstormed some ideas about how to generalize the counterfactuals.

The first idea was to **cluster** the training points, and then present the median profile of the closest cluster with respect to the input as a counterfactual. This would have the benefit that since it the median of the cluster, it probably has feature values that are more in consensus with this group of instances, therefore presenting a more generalized counterfactual. However, since there are 19 features per profile, counter-intuitive feature values can still occur. It is also not clear how choose the extra parameters that come with clustering (such as when to stop) obtain an optimal result. Lastly, since the counterfactual is in the middle of a cluster, it is likely quite different from the input.

Inspired by the clustering idea, we implemented a **distance kernel** approach. What this entails is that a new instance is generated by taking the average of all training data points, weighted by (1 - their distance to the input profile). In this way, feature values of multiple instances are summarized, but the values of closer instances are incorporated more. This mitigates a lot of the issues with the clustering approach, such as complicated optimization and the distance to the original instance. As such, we implemented the distance kernel and compared the results on our training data to those we obtained from the base counterfactuals. These results are summarized in Table 8.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Score distance kernel** | **Score closest training data** |
| Mean number of feature differences | 13 (/19) | 8 (/19) |
| Mean distance to the input | 0.123 (/1) | 0.039 (/1) |

Table 8: Two metrics measured on the training data from the distance kernel implementation of counterfactuals, in comparison to simply selecting the closest counterfactual training data point (baseline counterfactual).

We can see that the distance kernel already performs worse than the baseline, since both distance to the input and the number of feature differences between the input and counterfactual are high. What we have obtained with the distance kernel is a new data point, which is a summary of multiple other instances. Because it incorporates information from many instances, it is going to have lots of differences in comparison to our input. These many differences also translate into a larger distance.

In conclusion, generalized counterfactuals stray too far from the concept of a counterfactual explanation which is to find an *example* which is the most similar to the input. By presenting a data point that is an amalgamation of other instances, we lose the quirks of what makes an example profile unique. Generalized counterfactuals also stray too far from the input in a literal sense; the distance and number of feature changes is larger than the baseline counterfactual. Presenting a counterfactual profile that has similar quirks to the input, even though those do not fit the average, is preferred. The counterintuitive differences had to be solved from a different perspective.

### 4.3.3 Multi-objective counterfactuals

As a way to find sparser training data counterfactuals, we determined that it could be beneficial to select these instances not only based on distance, but on the number of feature differences as well. Minimizing two scores at the same time can be solved with multi-objective optimization. For solving multi-objective optimization problems, there exist several approaches with their own advantages and disadvantages. We considered two simple strategies; weighted-sum and non-dominated.

The simplest solution is considered to be the **weighted-sum** in which scores are collapsed into a single objective by adding them together with predefined weights. This method’s main disadvantage is that it is difficult to balance objectives properly. In our case, we only have two objectives; the distance and the number of feature differences. Though both scores lie between zero and one, their medians and variance still differed which can be derived from Table 9. We thought that it would suffice to assign the weights based on the median of both scores that we obtained from the training data. However, we had no clear perception of how these weights influenced the obtained points. The weighted-sum method has no optimality guarantees. If a new objective was added, or the training data was expanded, the weights would require re-adjusting as well. Even though the method is fast and easy to implement, we did not find it adequate to this problem.

|  |  |  |
| --- | --- | --- |
|  | **Distance score** | **Number of feature differences score** |
| Median | 0.125 | 0.684 |
| Mean | 0.128 | 0.704 |
| Maximum | 0.426 | 1.000 |
| Minimum | 0.004 | 0.158 |

Table 9: Median, mean, maximum and minimum of the two scores that we want to minimize. These are calculated from the training data.

For this reason, we switched to the non-dominated strategy. Compared to the weighted-sum, this technique is guaranteed to find the Pareto-optimum set of solutions [28, 29]. For the sake of the NFI’s future plans, more objectives can also be added. One disadvantage is the computational effort, since many comparisons need to be made in order to identify the non-dominated set. For this dataset and problem, this was hardly noticeable in practice.

# 5. Final user study

It was important to do a soft evaluation get the DNA experts’ opinion of the explanation. We also wanted to test if our explanation could help users gain some insight into how the model makes predictions. Therefore, a final survey was created (implemented with Google Forms). There were 8 responses in total, of which 7 were useable.

## 5.1 Set-up

The main goal of the survey was to establish if our visualization can help users gain insight into how the model makes predictions. By extension, we tested if this information helps regulate the users’ trust in the model. This means that when the model is quite sure about a prediction, that this influences the users to trust the prediction more. Conversely, if the model is unsure or incorrect, the explanation influences the users to doubt the model. For this exercise, we compared against two other state-of-the-art explanations.

Besides this regulation of trust, we also wanted to determine if the visualization is user-friendly. The NFI placed emphasized this latter part, as the user need to want to use it, before they would invest more time into it.

The survey was structured into 5 sections:

1. Demographics.
2. Introduction to the different explanations.
3. Can our visualization increase trust in the prediction when the model seems certain?
4. Can our visualization decrease trust in the prediction when the model seems uncertain or wrong?
5. Which explanation is most user-friendly?

Demographics are important to help understand the background of each participant in case this has an influence on the outcome. Section 2 was essential for users to learn what the different visualizations encode and how to interpret them. In sections 3 and 4 we tested if the visualization can regulate trust, and in section 5 user preferences were evaluated.

## 5.2 Demographics

There were three relevant demographics questions for the users; age, English reading / listening level, and level of openness towards using machine learning in their decision-making process. Age can influence how accustomed someone is towards technology, or how well they can learn new things [30]. Therefore, we found it useful to gather this information. The survey was created in English to help with the analysis of the results. We were informed that the users understand English well, as they read scientific literature. To ensure that this is indeed the case, we added a question about their English reading and listening level. Lastly, colleagues remarked that some people might be morally opposed to any sort of AI, and therefore might fill in the survey with a negative attitude. Therefore, we asked about their attitude towards using AI beforehand to map out any bias they might have.

### 5.2.1 Demographics results

A variety of different people participated, with ages spread over the youngest four groups as can be seen in Figure 12.

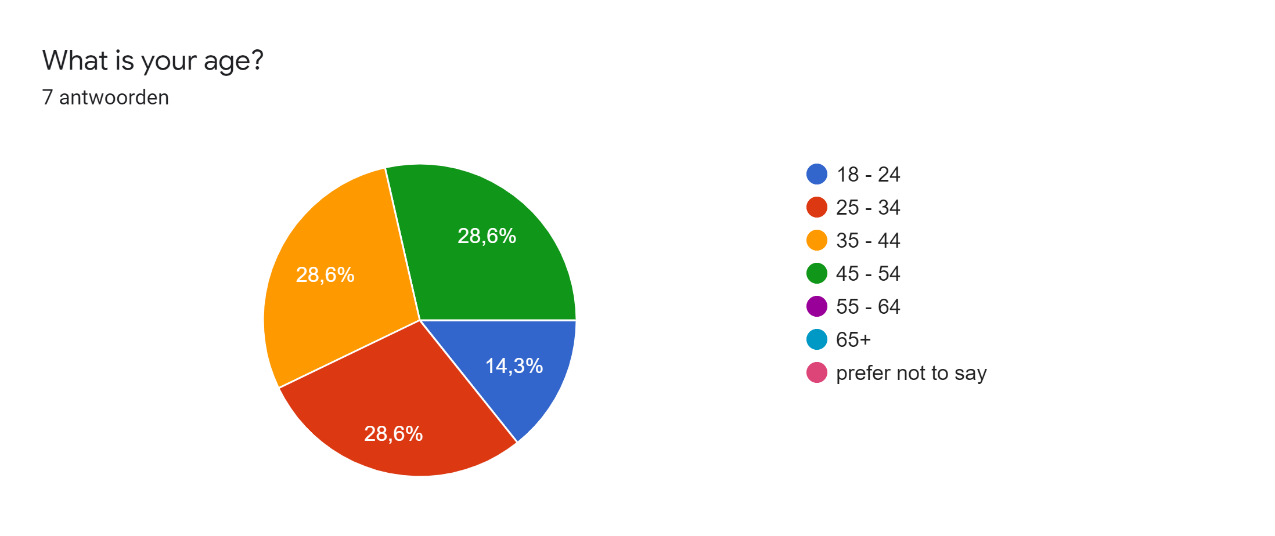


Figure 12: Age demographic answers.

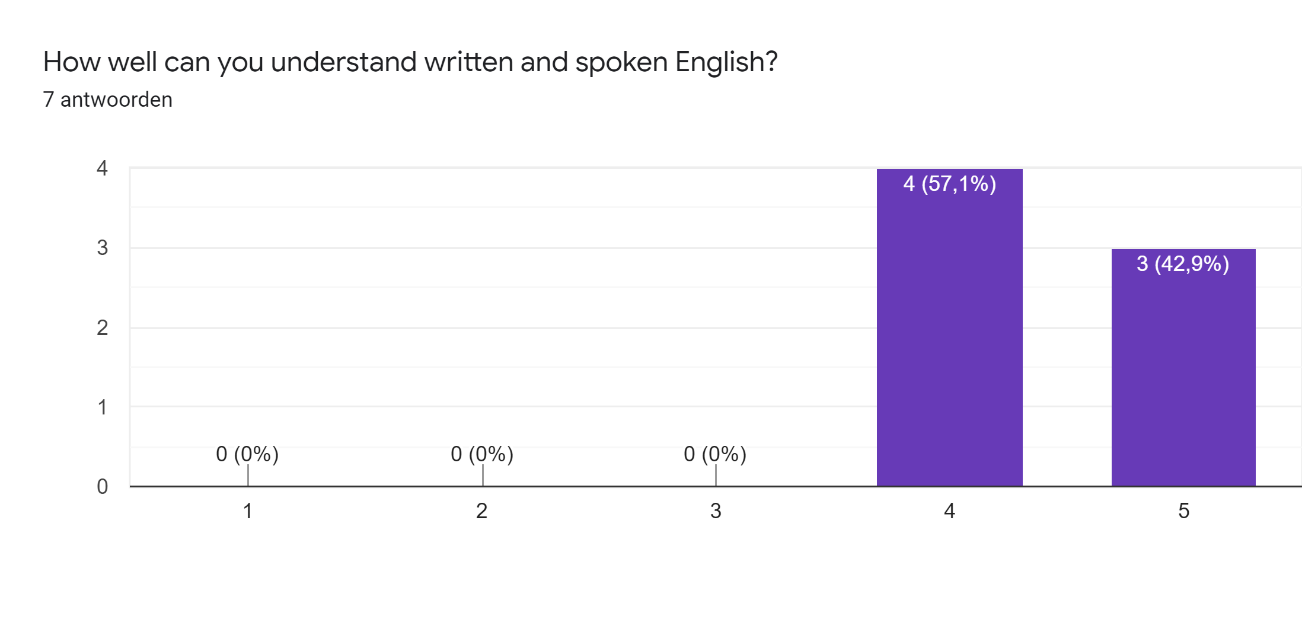
When it comes to English reading and listening ability, every person deemed themselves at least above average which we think is suitable for the survey (see Figure 13).

Figure 13: English level answers.

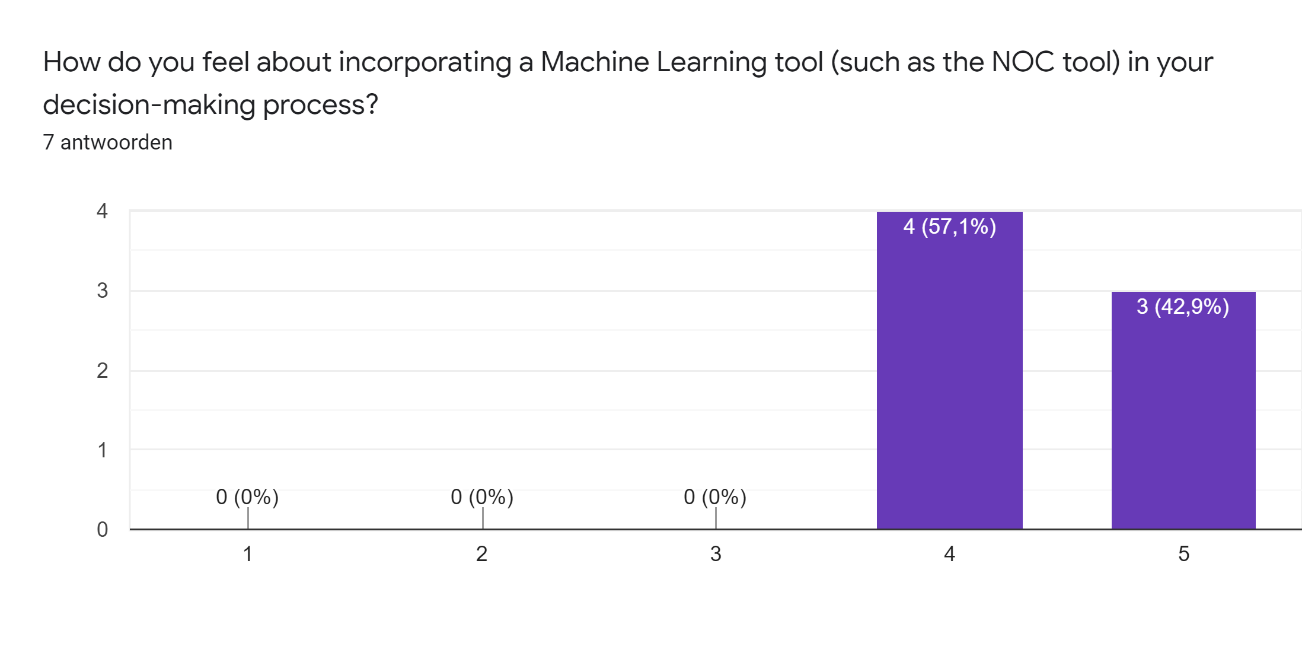
All participants have a positive perception of machine learning as a support tool for their decision-making process, so no upfront negativity was measured (see Figure 14).

Figure 14: Attitude towards machine learning answers.

No surprising results or particular correlations between the demographics and given answers were found.

## 5.3 Introduction to explanations

This section was used to introduce the SHAP force plot, counterfactual table, and compound explanation to the users. It included two disclaimers.

The first disclaimer concerns the fact that these explanations were meant to be consulted in a standalone fashion. They do not test how well the user can determine the NOC; they test if users can understand the way the machine learning model makes decisions. However, we understood that some users have some aversion to solely rely on the machine learning model and explanations, and might want to look at the profile anyways. For these users, we added the profiles in their analysis software DNAxs and told them they were free to consult the profile if they do so desire.

Secondly, the DNA experts have repeatedly expressed that they do not understand many of the current 19 features, or how they contribute to making a prediction of the NOC. As such, we had to instruct the users to focus only on a few features. Some of the features are familiar to the users; the *TAC* and the *MAC*. Others are quite simple to understand, such as *Loci with 5-6 alleles*, which represents the number of alleles with 5 or 6 alleles. Some features give an impression of quality of the profile; more *Peaks below 800 RFU* and *Allele count variation* could indicate lower quality, and/or more drop-out. For all locus-specific features, we told the users to view all of those as indications of the amount of information at each of those loci.

After the two disclaimers, users were presented with an introduction of the three explanations. For each type, we presented an image, an explanatory video, and a short summary of bullet points.

### 5.3.1 SHAP

Figure 15 shows the SHAP force plot used for introduction.

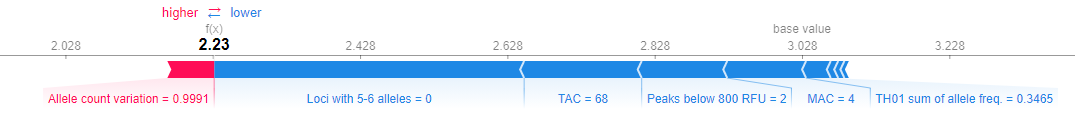


Figure 15: SHAP force plot used for introduction.

The introductory video can be watched from the following YouTube link: <https://www.youtube.com/watch?v=lysnLemJTfg> .

The summary of the SHAP explanation consists of the following points:

* Shows which feature values of this profile the NOC tool used to make this prediction.
* Some values push the prediction down towards 1 or 2 (blue bars), while other values push the prediction up towards 4 or 5 (or more) (red bars). This is relation to the base value of 3.
* Shows only information about the current profile and prediction.

The control question verified that users had read the summary and/or watched the video and understood that the values with red bars push the prediction up.

What is true about the SHAP explanation example we saw?

1. **The value of "Allele count variation = 0.9991" of this profile has caused the NOC tool to make a slightly higher prediction.**
2. The value of "Allele count variation = 0.9991" of this profile has caused the NOC tool to make a slightly lower prediction.

All participants answered this question correctly.

### 5.3.2 Counterfactual table

Figure 16 shows the counterfactual (CF) table used for introduction.

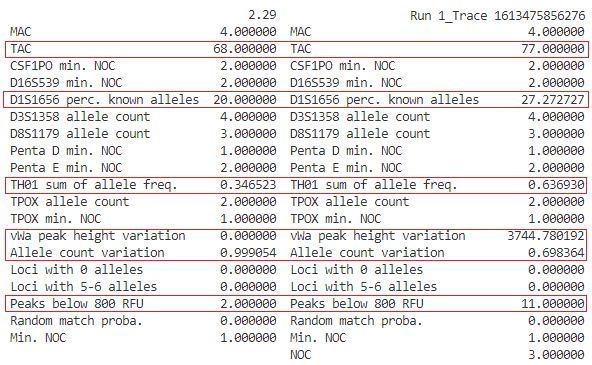


Figure 16: Counterfactual table used for introduction.

The introductory video can be watched from the following YouTube link: <https://www.youtube.com/watch?v=-VRIsHA8Sq4> .

The summary of the CF table explanation consists of the following points:

* Shows a comparison of the current profile with a profile that had a different rounded-off prediction.
* Differences are highlighted in red boxes.
* We do not know which differences are relevant to arrive at another prediction.

The control question verified that users had read the summary and/or watched the video and understood that the values with highlighted in red boxes are simply the differences between two example profiles and they are not all necessarily relevant to change the prediction.

What is true about the CF table explanation example we saw?

1. The feature values highlighted in red boxes have the most influence to change the prediction from 2 to 3 contributors.
2. **The feature values highlighted in red boxes are differences between two example profiles with a prediction of 2 and 3 contributors.**

All participants answered this question correctly.

### 5.3.3 Compound visualization

Figure 17 shows our compound visualization used for introduction.

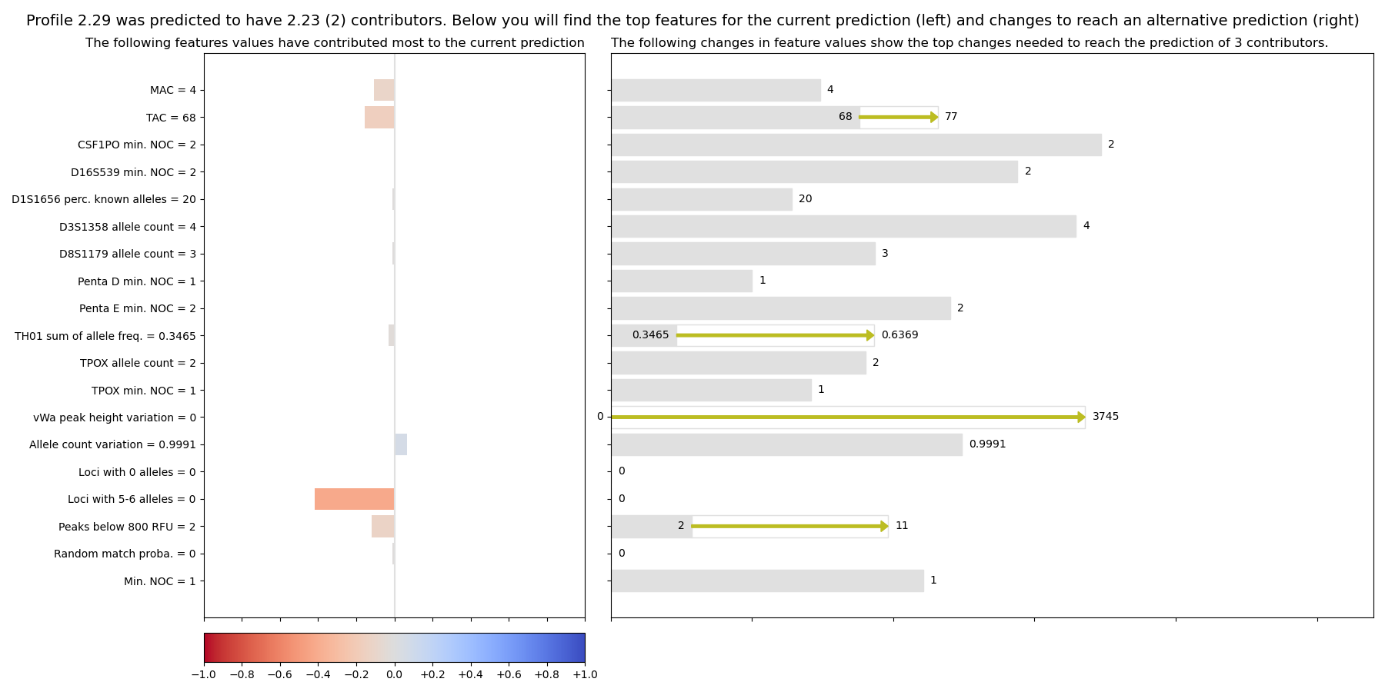


Figure 17: Compound visualization used for introduction.

The introductory video can be watched from the following YouTube link: <https://www.youtube.com/watch?v=rz3zm5AQ94c> .

The summary of the compound visualization explanation consists of the following points:

* Consists of two parts; one showing how the feature values have influenced the prediction, one showing the relevant feature value changes needed to reach a different prediction. The only connection that the two parts share, are the feature values.
* Shows how high each feature value (change) is compared to the entire range of that feature.
* By looking at the right side; seeing how large and relevant the necessary feature value changes are, you could determine how clear the NOC tool is about the prediction.

The control questions verified that users had read the summary and/or watched the video and understood that the left side consists of the same values as the SHAP explanation, but the right side is different from the counterfactual table, as only relevant feature value changes are shown. The two sections are not directly connected; we do not know what happens when we change one value.

What is true about the compound explanation example we saw?

1. **The left side of the explanation shows the same information as the SHAP explanation, but the right side shows only relevant feature value changes, which is different from the Counterfactual table.**
2. The left side of the explanation shows the same information as the SHAP explanation, and the right side shows the same information as the Counterfactual table explanation.

What is true about the compound explanation example we saw?

1. If we change the TAC value from 68 to 77, we know for certain that the TAC value's red bar on left would become blue.
2. **If we change the TAC value from 68 to 77, we do not know for certain how that will affect the red and blue bars on the left, because they only give information about the current profile's feature values.**

One participant answered both of these questions wrong, which could mean that they did not watch the video and/or read the summary well. From the open-ended questions in the next section, we could also derive that this participant did not read all the text in the survey; they filled in that they would need to see the profile before they could make any decisions, yet on every page there were instructions on how to look up the profile if they felt inclined to do so. We therefore had to remove their answer from the responses.

## 5.4 Regulate trust

At the end of the introduction, we presented the users with the overview in Figure 18 to help them understand this section of the survey. The schematic shows that two profiles will be presented. These two profiles correspond to the two use-cases we described before; one profile was fairly simple for the model to predict and thus the explanation could show this to the user. On the other hand, the second profile was difficult to the model and even led to an incorrect prediction. Per profile, the prediction from the model is presented. This prediction serves as the baseline; there is no explanation. Then, two explanations are presented. For both of these profiles, we will show our compound visualization and another state-of-the-art explanation that is fit for each use case. For profile 1, we compare against a SHAP force plot, while for profile 2 a counterfactual table was used for comparison. Initially, we had planned to compare our visualization against both the SHAP plot and the counterfactual table for each profile, but the survey became too long. As SHAP is originally designed to understand “why a model makes a certain prediction” [21], we deemed it fit for the goal of increasing trust. In contrast, counterfactuals show how a different prediction can be reached. If little change is required for that change to occur, this can decrease trust in the original prediction.

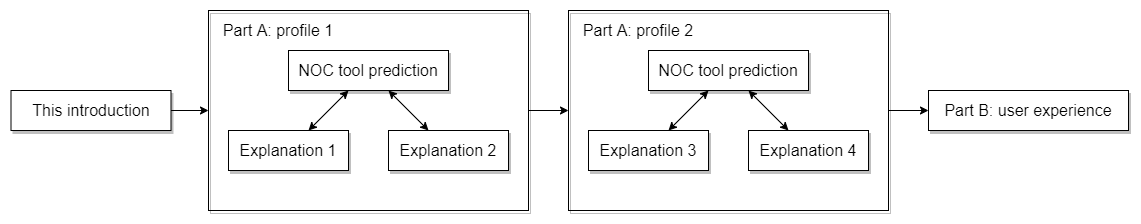


Figure 18: Overview of the survey.

### 5.4.1 Questions: increase trust

The first use-case that we put to the test was to see if our visualization can increase trust in the prediction when the model seems fairly certain. For this aim, we chose to show profile 1\_6B.Trace#01. This profile was chosen because it was difficult in an old NOC interpretation training by the NFI; DNA experts would define this profile as 2/3, 3, 3/4, 4, 4/5. It has a lot of missing alleles, which is why it proved difficult. However, the model correctly identifies it as a 3-person mixture.

The following questions were asked:

1. *Please select all number(s) of contributors you would consider after seeing a prediction of 3 (3.22).*
2. *Do you think the prediction of 3 (3.22) is correct?*
3. *Please select all number(s) of contributors you would consider when you can consult this explanation (SHAP). Do you consider the same, or less options than in question 1?*

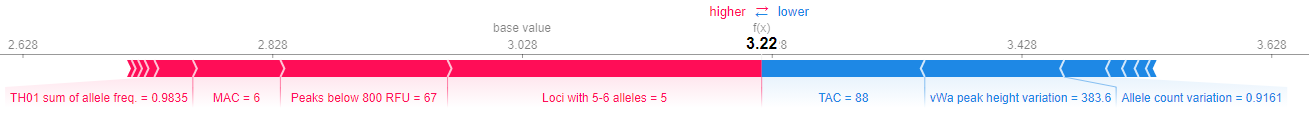


Figure 19: SHAP explanation for profile 1\_6B.Trace#01 used in the final survey for question 3.

1. *Do you think the prediction of 3 (3.22) is correct?*
2. *Can you explain why you have answered questions 3 and 4 differently (or not) than questions 1 and 2 after looking at the SHAP explanation, in comparison to only seeing the prediction of 3 (3.22)?*
3. *Please select all number(s) of contributors you would consider when you can consult the explanation (Compound explanation). Do you consider the same, or less options than in question 1? Note: we are comparing with predictions of 2 and 4 donors. Normally, you would be able to choose which comparison you would like to make.*

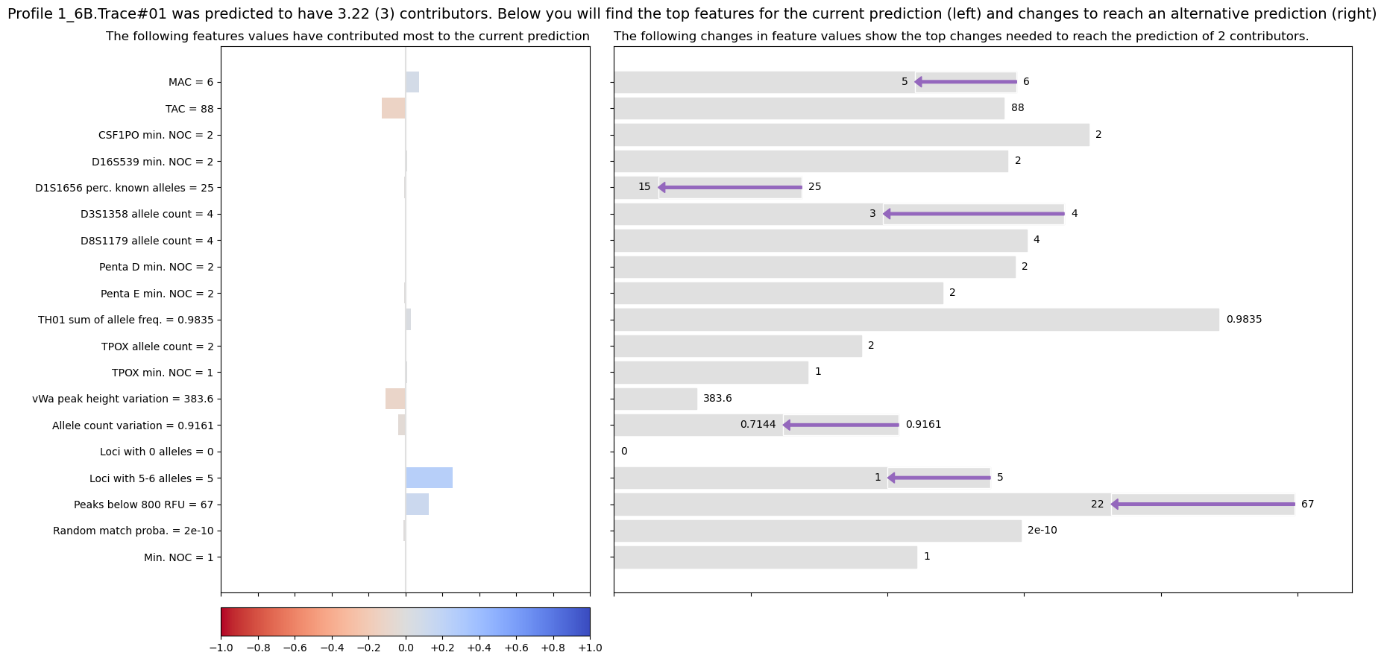
**

Figure 20: Our visualization for profile 1\_6B.Trace#01, comparing to a NOC of 2, used in the final survey for question 6.

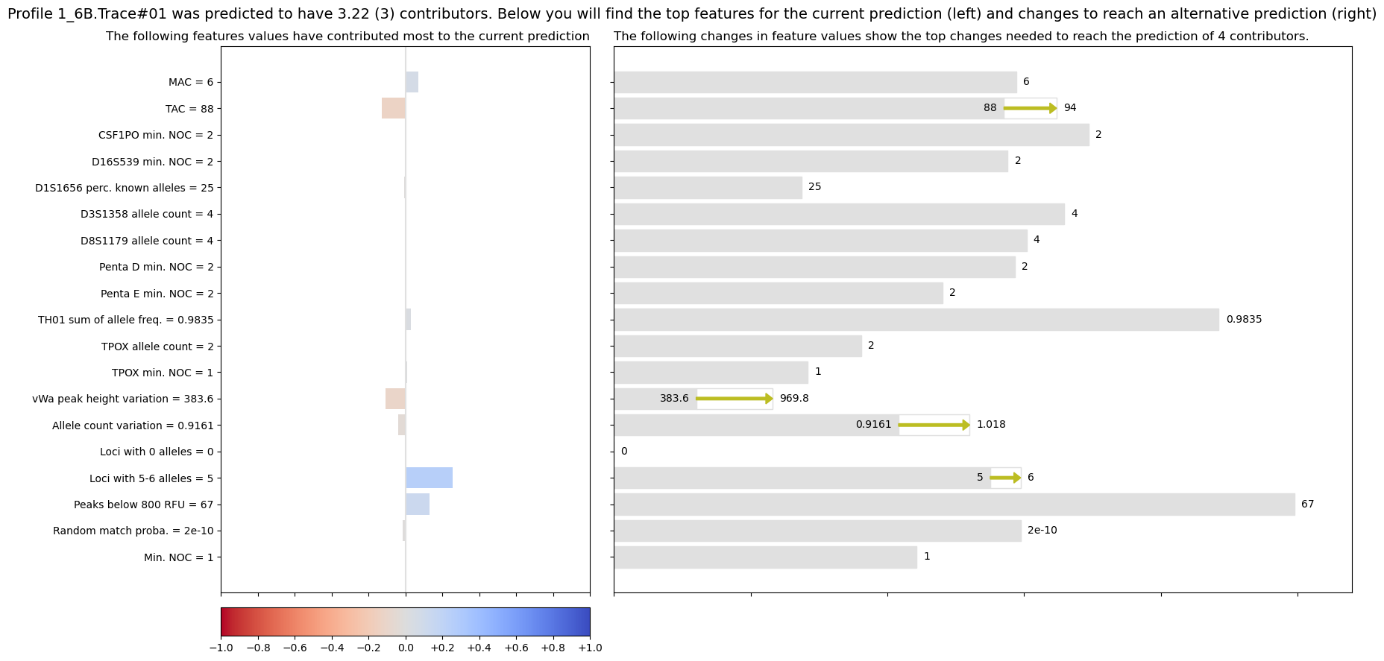


Figure 21: Our visualization for profile 1\_6B.Trace#01, comparing to a NOC of 4, used in the final survey for question 6.

1. *Do you think the prediction of 3 (3.22) is correct?*
2. *Can you explain why you have answered questions 6 and 7 differently (or not) than questions 1 and 2 after looking at the Compound explanation, in comparison to only seeing the prediction of 3 (3.22)?*

### 5.4.2 Answers: increase trust

We wanted to see if our visualization could take away some doubt about the prediction. We compared against the SHAP explanation. Increased trust means that **less options for the NOC** are considered, and that **users think the prediction is correct**.

The answers to questions 1, 3, and 6 are summarized in Figure 22. The influence that the SHAP explanation has on which NOC is considered is shown in the top graph. We can see that most people switched from considering 2, 3 and 4 contributors, to only 3 and 4. The others stayed with the same as what they chose after only seeing the prediction, or even started to consider 2 after previously only considering 3 and 4. There is quite a range of different answers.

Similarly, the influence of our visualization on which NOC is considered is shown in the bottom graph. Most people switched from considering 2, 3 and 4 contributors, to only 3 and 4. The others stayed with their original consideration which lies close to the prediction.

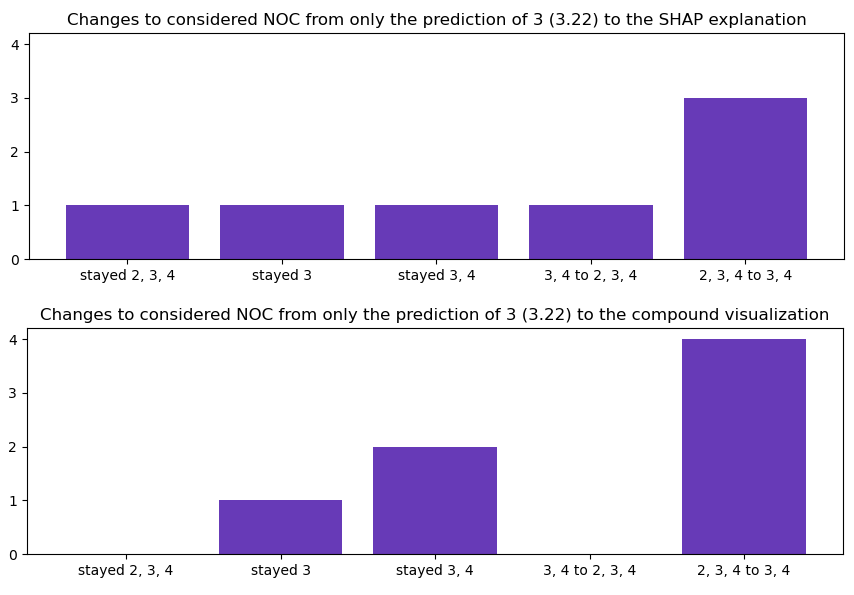


Figure 22: Changes in answers from questions 1 to 3 in the top graph (prediction to SHAP) and from questions 1 to 6 in the bottom graph (prediction to visualization).

A similar trend can be seen in Figure 23 for questions 2, 4 and 7. In general, more people seem to think the prediction is correct after seeing the SHAP explanation (going from the first to the second pie chart). Our visualization has a comparable effect (going from the first to the third pie chart).

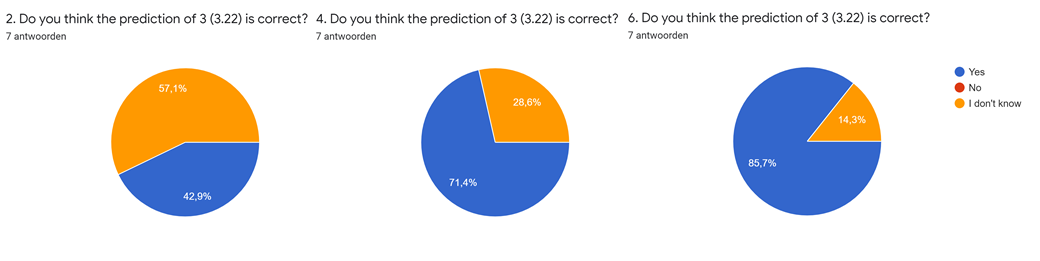


Figure 23: Do participants think that the prediction of 3 (3.22) is correct after question 2 (only the prediction), question 4 (SHAP) and question 7 (compound visualization).

### 5.4.3 Motivations: increase trust

The motivations that the participants gave in question 5 concern why they considered different answers after seeing the **SHAP explanation in comparison to no explanation**.

There were a few users that expressed to no longer consider 2 contributors because of feature values. One noted the MAC and TAC in combination with the high number of peaks below 800 RFU would not suit 2 donors, another mentioned the high number loci with 5-6 alleles. Another participant came to the same conclusion because there are more values pushing the prediction up than down.

Interestingly, one user actually switched from only considering 3 and 4, to also including 2 because the red blocks are located all the way into the 2.6 area. Another user became similarly confused as the red and blue bars seem to pull the prediction to both sides. They both answered “I don’t know” instead of “yes” to whether they think the prediction is correct after seeing the SHAP explanation.

The two participants that stayed with their original answers of 3 and 2, 3, 4 though they had similar reasoning where they expressed to trust the prediction as many changes would need to be made to reach a different prediction. One person seemed more inclined to just take the multiple options anyways.

The motivations noted from question 8 show why participants gave different answers after seeing our **compound visualization in comparison to no explanation**.

Most users noticed that to reach a prediction of 2 contributors, a lot of feature values needed to change, and also by a large extent. Especially in comparison to what needs to be altered to reach a prediction of 4 contributors. That is why they no longer considered 2 contributors. One user came to this conclusion because they thought the compound visualization indicates some of the change that they considered themselves (lower MAC and peaks below 800 RFU).

The person who stayed with 3 and 4 mentioned that they would like to look at the EPG before making a decision. The person to remained at 3 noted that they report a minimum NOC, so 3 would be certain enough.

### 5.4.4 Questions: decrease trust

The second use-case that we put to the test was to see if our visualization can make users trust the prediction less when the model seems uncertain or is simply wrong. For this aim, we chose to show profile 2A3.3. This 3-person mixture profile was predicted by the model to have 4 contributors. As the output is 3.53, we can tell that the model is unsure about this profile.

The following questions were asked:

1. *Please select all number(s) of contributors you would consider after seeing a prediction of 4 (3.53)*
2. *Do you think the prediction of 4 (3.53) is correct?*
3. *Please select all number(s) of contributors you would consider when you can consult this explanation (Counterfactual table). Do you consider the same, or less options than in question 1? Note: we are comparing with example profiles with a prediction of 3 and 5 donors. Normally, you would be able to choose which comparison you would like to make.*

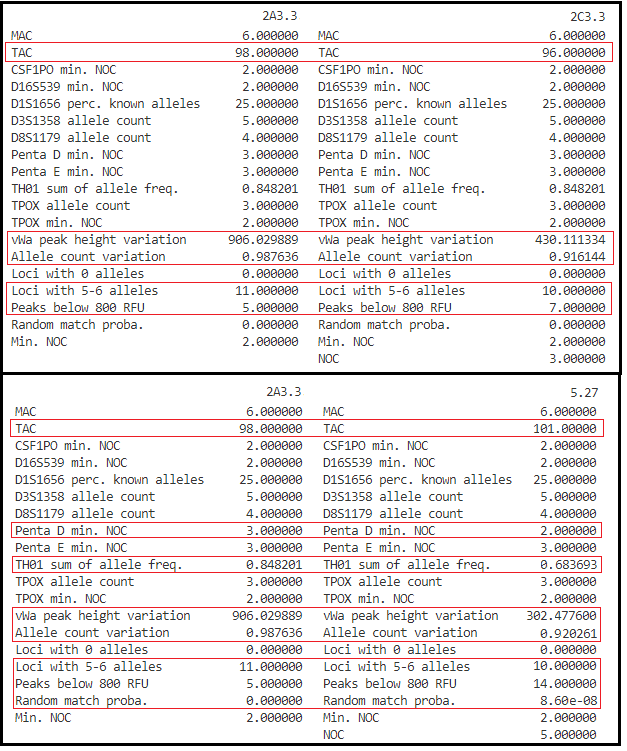


Figure 24: CF table explanation for profile 2A3.3, comparing to a NOC of 3 and 5, used in the final survey for question 3.

1. *Do you think the prediction of 4 (3.53) is correct?*
2. *Can you explain why you have answered questions 3 and 4 differently (or not) than questions 1 and 2 after looking at the Counterfactual table explanation, in comparison to only seeing the prediction of 4 (3.53)?*
3. *Please select all number(s) of contributors you would consider when you can consult the explanation below (Compound explanation). Do you consider the same, or less options than in question 1? Note: we are comparing with predictions of 3 and 5 donors. Normally, you would be able to choose which comparison you would like to make.*

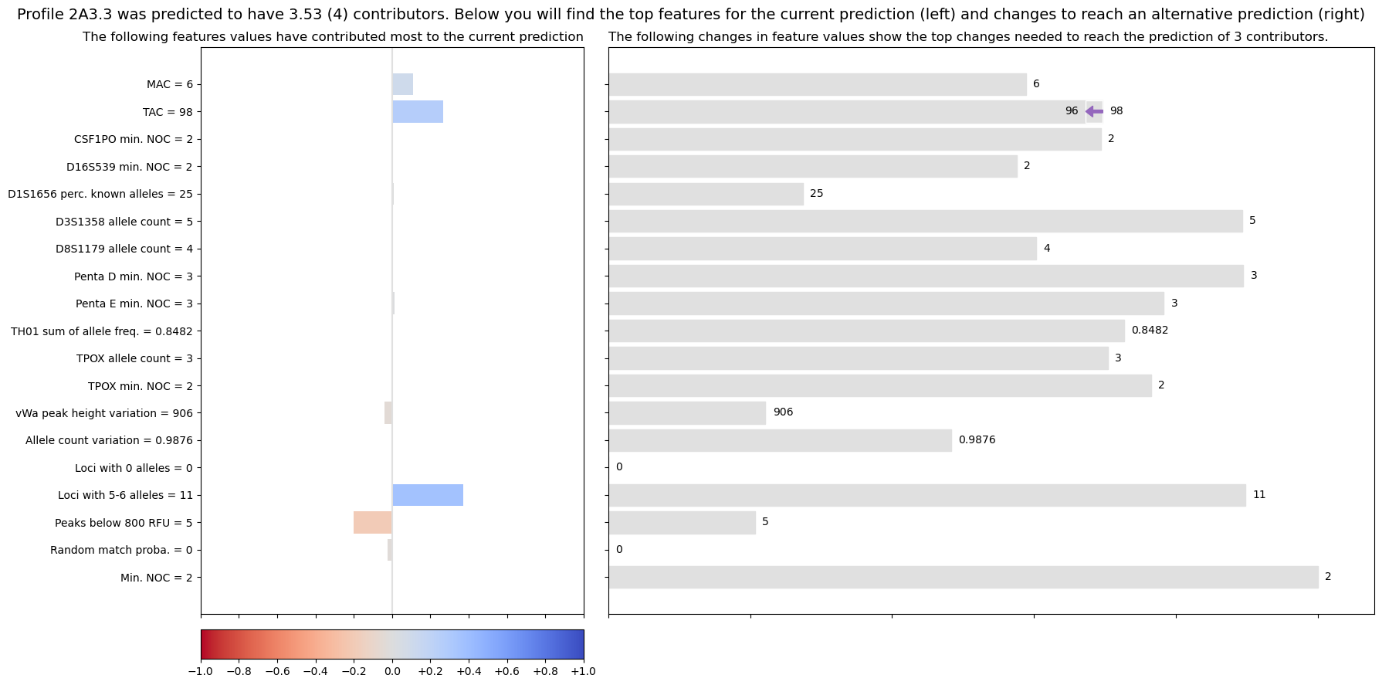


Figure 25: Our visualization for profile 2A3.3, comparing to a NOC of 3, used in the final survey for question 6.

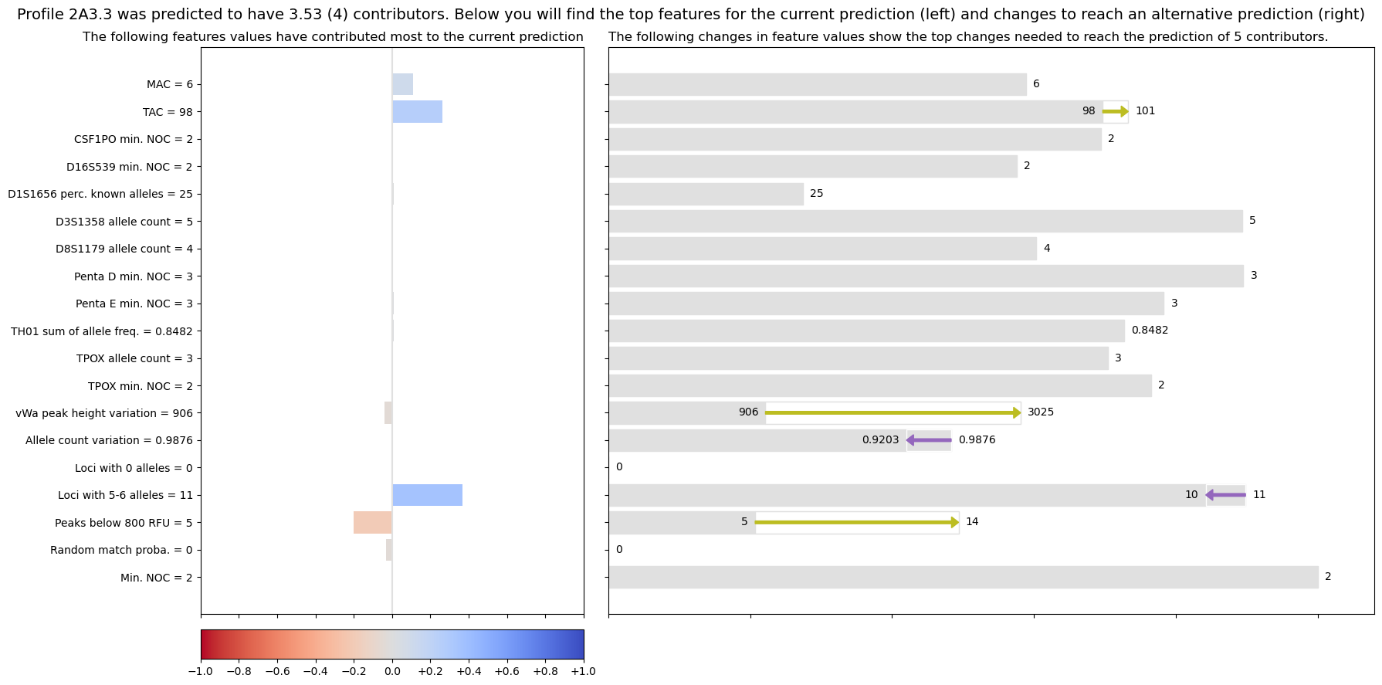


Figure 26: Our visualization for profile 2A3.3, comparing to a NOC of 5, used in the final survey for question 6.

1. *Do you think the prediction of 4 (3.53) is correct?*
2. *Can you explain why you have answered questions 6 and 7 differently (or not) than questions 1 and 2 after looking at the Compound explanation, in comparison to only seeing the prediction of 4 (3.53)?*

### 5.4.5 Answers: decrease trust

We wanted to see if our visualization can decrease trust in the prediction. We compared against a counterfactual (CF) table explanation. Decreased trust means **no less options for the NOC** are considered, and that **users doubt that the prediction is right**.

The answers to questions 1, 3, and 6 are summarized in Figure 27. The influence that the CF table explanation has on which NOC is considered is shown in the top graph. We can see that most people stayed with their initial estimation of 3 or 4 (or 5) contributors. One person dropped their consideration of 2 and 5 donors to match the other answers of 3 and 4 contributors. Only one person was drastically different as they changed their answer to match the prediction of only 4.

The answers are largely the same for the compound visualization. The person that thought it was 4 contributors after the CF table, changed it back to 3 and 4. One of the people who first considered 3 and 4, though it was 3 instead; they went against the prediction.

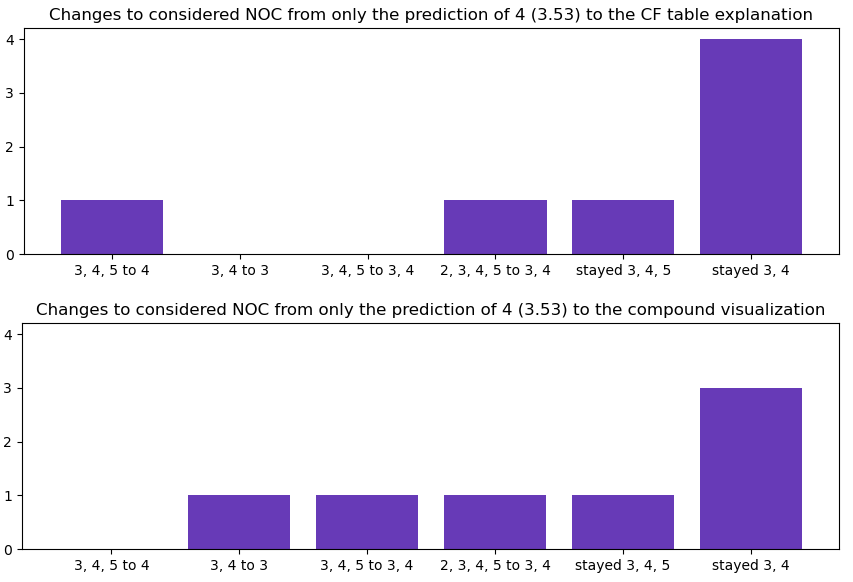


Figure 27: Changes in answers from questions 1 to 3 in the top graph (prediction to CF table) and from questions 1 to 6 in the bottom graph (prediction to visualization).

A similar trend can be seen in *Figure 28* for questions 2, 4 and 7. Most people do not know if the prediction is correct or not. The CF table has no influence on this trust (going from the first to the second pie chart). With our visualization, one person changed their answer from “I don’t know” to “No”, and another changed from “Yes” to “I don’t know” (going from the first to the third pie chart).

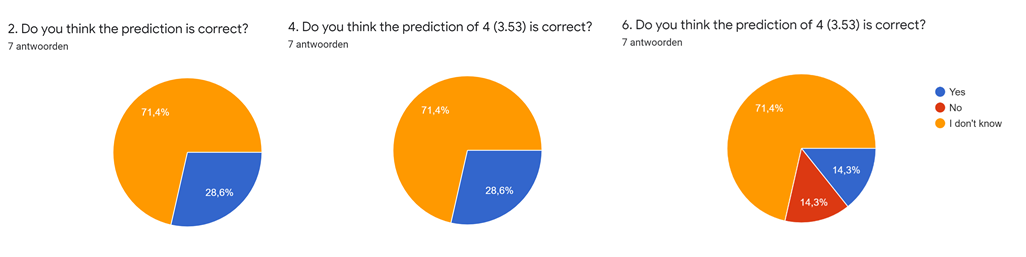


Figure 28: Do participants think that the prediction of 4 (3.53) is correct after question 2 (only the prediction), question 4 (CF table) and question 7 (compound visualization).

### 5.4.6 Motivations: decrease trust

The motivations that the participants gave in question 5 concern why they considered different or similar answers after seeing the **CF table in comparison to no explanation**.

Two users expressed that the model seemed to be uncertain when looking at the prediction of 3.53 and therefore thought it should be 3 or 4. This means that the CF table did not influence their decision in any other direction as opposed to seeing only the prediction. The other two users that picked 3 or 4 contributors, noted that in the CF table there are less changes to be made to reach 3 contributors than to reach 5.

The person that ruled out 2 and 5 contributors, based this decision on the feature values, as 2 donors with this MAC, TAC and peaks below 800 RFU seems improbable, and 5 donors are unlikely based on vWa and peaks below 800 RFU values.

The one person that chose 4 contributors noted that to reach 3 or 5, lots of change was required.

One user still considered their original three options as they noted that the counterfactual tabel does not show which feature differences are actually relevant.

The motivations noted from question 8 show why participants gave different answers after seeing our **compound visualization in comparison to no explanation**.

Five users noted that they were starting to doubt the model’s prediction of 4, as only minor changes were required to reach the prediction of 3. One participant noted that for a TAC of 98, there can be 2 artefact peaks. They therefore thought the prediction was incorrect.

One person noticed how it is easier to determine to what extend features need to change to reach the alternative predictions as compared to the CF table.

## 5.5 User friendliness

As subjective evaluation of user friendliless, the users were asked to pick which explanations they preferred in three categories:

* Ease of use (how easily they could find the relevant information)
* Appeal (how nice they thought it was to use)
* Completeness (how well they could form a total picture of the prediction)

For each category, only one explanation could be selected. The results can be found in Figure 29.

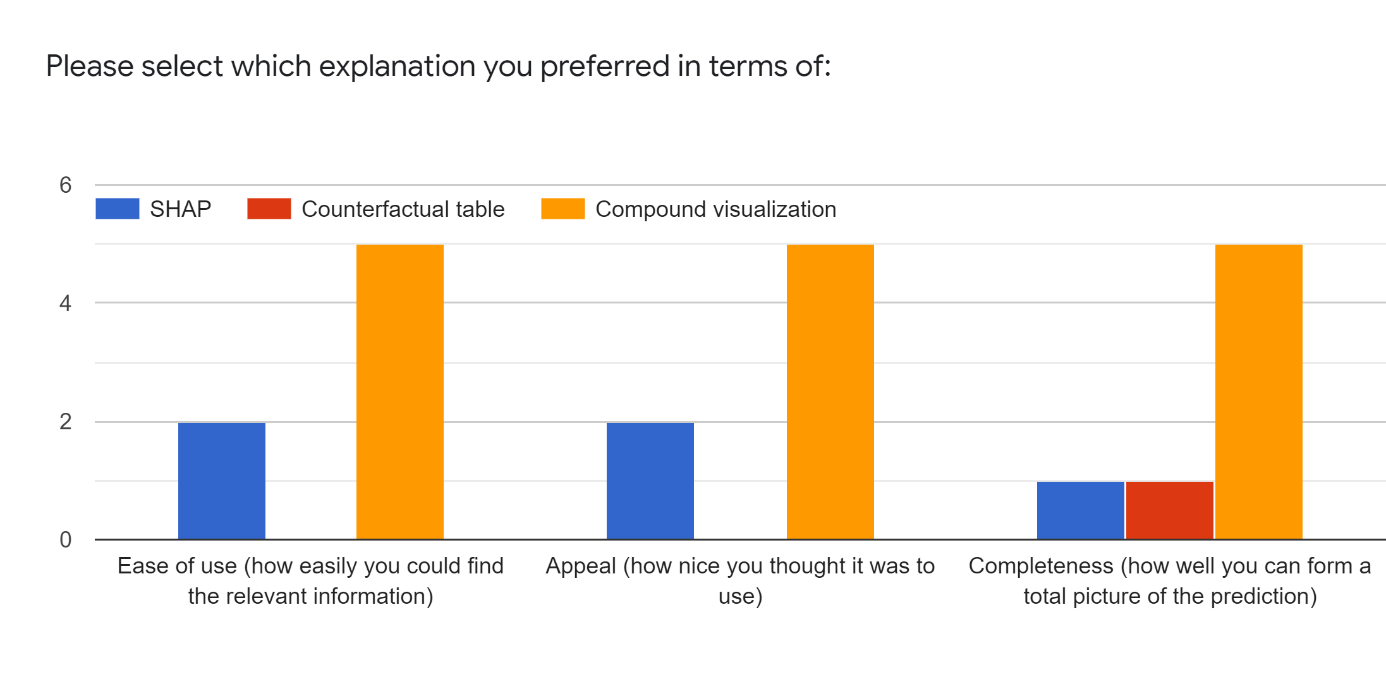


Figure 29: Results of subjective evaluation of the 3 presented explanations.

In general, the compound visualization scored well on all three categories. The main reasons for picking the visualization were that:

* There was more information (2x)
* There was less information, but the information was more relevant (1x)
* The information was easy to find (3x)
* It was presented in a visual way (5x)
* It was easier to understand (1x)
* It was more like how I would think about this problem (1x).

One user admitted to being a fan of the SHAP method and therefore chose it for all three categories. They expressed it was easy to use, the information was easy to find and the results were presented in a visual way.

## 5.6 Discussion and conclusion

Before we conclude, we note that there were some threats to validity that could have influenced the results.

First of all, the *ordering* of the explanations was SHAP first, CF table second, and our visualization last for all participants. This order was chosen as such because the compound visualization consists of SHAP values, and a modified counterfactual. It therefore makes sense to first explain the components, before explaining the whole. This could have introduced some bias in favor of SHAP and against our visualization, though we did not notice this in the results.

Secondly, we missed the fact that when the first explanation was shown for both profiles, this was also the first time that the participants came into contact with the feature values of the profiles. Some users focused on the *feature values* instead of on what the explanation was trying to communicate. With the features being quite difficult to understand, the translation into simpler terms was required for the users to have some idea of what they meant. However, this is an extra step that users need to take before being able to understand the explanation. This could cause fatigue or make participants feel unmotivated. The same can be said for the introductory section. Even though they enjoyed the way that it was explained, it still took a toll on their capacity for the rest of the survey.

Thirdly, a threat to construct validity was the fact that DNA experts at NFI report a *minimum* NOC. Because of this fact, a participant might have been inclined to choose a NOC of 3 instead of 4, not because the explanation made the prediction of 3 seem more likely. We noticed one participant that seemed to hint at this fact.

Lastly, we only received 7 replies from a group of about 35 experts. This selection of subjects means that the survey suffered from *representation bias*.

Because of these limitations, we approached the final evaluation as a subjective collection of participants’ opinions, rather than try to perform statistical tests. The addition of textual responses was especially valuable to see how people thought about the explanations.

Within the section of increasing trust, SHAP seemed to induce some confusion with the bars pushing into each other. Most of the users that did express to drop one option, based this on the feature values, not the actual information that the SHAP plot was trying to communicate. In our visualization, users had the option to explore all feature values, as well as see how these values should be changed to reach a different outcome. The participants dropped one option because many feature values had to be changed and also by a large extent. One person noted that they dropped one option because the counterfactual section of the visualization suggested the same changes in feature values that they considered themself.

When the model made a wrong prediction, the counterfactual table did not seem influence many people’s decision. Some participants noted the feature values as reasons to drop one NOC from consideration, and noticed that the prediction value was on a threshold. These facts are not specific to the CF table. Only a few users noticed that fewer, or less important features needed to change to reach 3 donors. With the help of the compound visualization, more users started to doubt the model’s prediction, or even deemed it incorrect as no decision should be made on a difference of 2 alleles. One person noticed how it is easier to determine to what extend features need to change to reach the alternative predictions as compared to the CF table.

Despite the small group of people that participated, the uncontrolled environment, and various sources of bias, the results were insightful about the experts’ various perceptions. They could be used to guide further direction, but generally show promise that experts could use our compound visualization as a tool to gain more insight into predictions of the model. We are mainly proud that it can help pinpoint the number of contributors better, and in some cases adjust trust in the prediction appropriately. Before such explanations can be used in practice however, training is required to help users become more familiar with them in practice.

# 6. Future feature engineering

Throughout the project, many of the DNA experts expressed that they did not like the current features, as they are too complicated, are incomplete, and do not seem to represent information that the experts think relates to the number of contributors. They have given us a lot of information on how to possibly improve the features in the future. From working with the features, we also uncovered some insights related to their use from a machine learning perspective. As such, we have accumulated the following recommended changes to the features:

1. Remove redundant features. For example, MinNOC*locus* encodes the same information as AlleleCount*locus*. MinNOC*locus* is equal to AlleleCount*locus* divided by two, rounded down. The one feature is not going to give more information than the other.
2. Remove locus-specific features. The loci that are included in the 19 features appear to be the loci that most often have the MAC of the profile. It would be interesting to replace the 23 locus-specific features with the following:
   * Locus (/loci) with the maximum allele count of the profile.
   * Locus (/loci) with the minimum allele count of the profile.

The opinion of the users is that mostly profile-specific information is important. It is difficult to understand why certain loci are included in the current features and others not. This proposal might remove that confusion. The way that the original 19 features were selected, was based on little data. This could be the reason why these locus-specific features were added.

1. Make profile features complete. For instance, the features now include the number of loci with 5 or 6 alleles, but not the number of loci with 7 or 8 alleles. Including the full range helps with consistency and understanding in explanations.
2. Add more information outside of the STR profile such as:
   * Number of peaks at stutter positions.
   * Number of unnamed peaks (visible in epg, can be sampled in EuroForMix).
   * Replicate runs.
   * Major / minor contributors.
   * The quality of certain channels.

Some of these might be more viable and useful than others. They are ordered according to the expected feasibility. Most of these were brought up by the DNA experts.

1. Keep MAC and TAC features since they are very familiar to the users and give context to the complete explanation. Additionally, the TAC values can vary among different kits, with which the users might not have experience with yet. Seeing the familiar variables with new values can help users get used to a new kit.

# 7. Reflection on methodology

Despite confirming with supervisors from the NFI and TU Delft that a linear strategy would suffice, we quickly realized that this would make it difficult to guard the quality of the project. It was a good decision to switch strategy to Agile such that new goals could be set and evaluated every three weeks. In this way, small increments are made to the product and less successful endeavors do not hinder progress as much. Perhaps this shows how little the TU Delft focusses on project methodology, as we relied on past experiences to organize this work.

With similar flexibility, we adapted the main research question to better fit the needs of users. Where we originally wanted to only focus on the counterfactuals because it is an interesting field of research, the users made us realize that a combination with feature importance is more valuable.

One aspect to improve upon is to have clearer agreements about the available time from the DNA experts and how many would be willing to participate. It seems quite difficult to get a company to give some of their employees’ time to an intern. Though we identified this risk from the start (see page 8), and we agreed with the NFI upon two surveys with users (one short and one long), no specific constraints were set about how many people would participate or for what period of time. Without Covid-19 restrictions, a fixed date and time could be planned for a sit-down evaluation at a specified location. This provides a lot of control as compared to sending a survey and hoping people will take the time to respond. Such a meeting could be communicated with the DNA experts’ management as well so that this could be planned into their schedule. In this way, we could have presented each of the explanations with a suitable introduction where users could ask questions. This ensures that everyone understands the concepts before proceeding. In the final survey for example, we had to eliminate the response of one user because they answered the control questions incorrectly. If we were there to support that person, the outcome might have been different. Within Covid restrictions, a specific moment could still have been planned, but it would lose some the controlled setting.

There was quite some pressure on the work; The thesis was completed in seven months instead of the expected nine, as the defense data was quickly set for mid-June. Perhaps it would be better to establish beforehand the expectations about this set date, before committing to it fully. This might have mitigated some stress. However, we are grateful to have finished the thesis before the summer.

# References

1. Benschop, C.C.G., et al., *Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach.* Forensic Science International: Genetics, 2019. **43**: p. 102150.

2. Benschop, C.C.G., et al., *The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures.* Forensic Science International: Genetics, 2015. **19**: p. 92-99.

3. Adadi, A. and M. Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).* IEEE Access, 2018. **6**: p. 52138-52160.

4. Barredo Arrieta, A., et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.* Information Fusion, 2020. **58**: p. 82-115.

5. Carvalho, D.V., E.M. Pereira, and J.S. Cardoso, *Machine learning interpretability: A survey on methods and metrics.* Electronics (Switzerland), 2019. **8**(8).

6. Gilpin, L.H., et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018.

7. Lipton, Z.C., *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.* Queue, 2018. **16**(3).

8. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. **2**(10): p. 749-760.

9. Yoo, T.K., et al., *Explainable Machine Learning Approach as a Tool to Understand Factors Used to Select the Refractive Surgery Technique on the Expert Level.* Transl Vis Sci Technol, 2020. **9**(2): p. 8.

10. Miller, T., *Explanation in artificial intelligence: Insights from the social sciences.* Artificial Intelligence, 2019. **267**: p. 1-38.

11. Verma, S., J.P. Dickerson, and K. Hines, *Counterfactual Explanations for Machine Learning: A Review.* ArXiv, 2020. **abs/2010.10596**.

12. Taylor, D., J.-A. Bright, and J. Buckleton, *Interpreting forensic DNA profiling evidence without specifying the number of contributors.* Forensic Science International: Genetics, 2014. **13**: p. 269-280.

13. Boavida, A., et al., *PowerPlex® fusion 6C system: internal validation study.* Forensic sciences research, 2018. **3**(2): p. 130-137.

14. Intituut, N.F., *HBS: Interpretatie van autosomale STR DNA-profielen*. 2020. p. 16.

15. Bleka, Ø., G. Storvik, and P. Gill, *EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts.* Forensic Science International: Genetics, 2016. **21**: p. 35-44.

16. Coble, M.D., et al., *Uncertainty in the number of contributors in the proposed new CODIS set.* Forensic Science International: Genetics, 2015. **19**: p. 207-211.

17. Clayton, T.M., et al., *Analysis and interpretation of mixed forensic stains using DNA STR profiling.* Forensic Science International, 1998. **91**(1): p. 55-70.

18. Benschop, C., *PowerPlex Fusion 6C Profile analysis & interpretation.* 2020.

19. Haned, H., et al., *Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood Perform Better Than Maximum Allele Count?* Journal of Forensic Sciences, 2011. **56**(1): p. 23-28.

20. Benschop, C., A. Backx, and T. Sijen, *Automated estimation of the number of contributors in autosomal STR profiles.* Forensic Science International: Genetics Supplement Series, 2019. **7**.

21. Lundberg, S. and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*. 2017.

22. Ribeiro, M.T., S. Singh, and C. Guestrin. *Anchors: High-Precision Model-Agnostic Explanations*. in *AAAI*. 2018.

23. Wachter, S., B. Mittelstadt, and C. Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.* Harvard journal of law & technology, 2018. **31**: p. 841-887.

24. Schleich, M., et al., *GeCo: Quality Counterfactual Explanations in Real Time*. 2021.

25. Russell, C. *Efficient search for diverse coherent explanations*. 2019.

26. Gomez, O., et al. *ViCE*. 2020.

27. Keane, M. and B. Smyth, *Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)*. 2020.

28. Gunantara, N., *A review of multi-objective optimization: Methods and its applications.* Cogent Engineering, 2018. **5**(1): p. 1502242.

29. Chiandussi, G., et al., *Comparison of multi-objective optimization methodologies for engineering applications.* Computers & Mathematics with Applications, 2012. **63**(5): p. 912-942.

30. Rojas-Mendez, J., A. Parasuraman, and N. Papadopoulos, *Demographics, attitudes, and technology readiness: A cross-cultural analysis and model validation.* Marketing Intelligence & Planning, 2017. **35**: p. 18-39.