# 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 chapters thesis

[1. Proposal and methodology 4](#_Toc72848626)

[1.1 Problem description 4](#_Toc72848627)

[1.2 Main related works 4](#_Toc72848628)

[1.3 Research questions 5](#_Toc72848629)

[1.4 Methodology and planning 5](#_Toc72848630)

[1.4.1 Risks 6](#_Toc72848631)

[2. DNA Mixture interpretation 8](#_Toc72848632)

[2.1 Short Tandem Repeat (STR) profiles 8](#_Toc72848633)

[2.2 Estimating the Number of Contributors (NOC) 8](#_Toc72848634)

[3. Survey on mixture interpretation and explanation types 11](#_Toc72848635)

[3.1 Set-up 11](#_Toc72848636)

[3.2 Question 1: workflow 11](#_Toc72848637)

[3.2.1 Analysis answers question 1 11](#_Toc72848638)

[3.3 Question 2: feature importance or counterfactual explanations 12](#_Toc72848639)

[3.3.1 Analysis answers question 2 12](#_Toc72848640)

[3.4 Question 3: features or raw peak information 13](#_Toc72848641)

[3.4.1 Analysis answers question 3 13](#_Toc72848642)

[3.5 Concluding remarks 14](#_Toc72848643)

[4. Experiments with various XAI techniques 15](#_Toc72848644)

[4.1 SHAP 15](#_Toc72848645)

[4.1.1 SHAP for multi-class classification 15](#_Toc72848646)

[4.1.2 SHAP for regression 16](#_Toc72848647)

[4.1.3 Conclusion 16](#_Toc72848648)

[4.2 Anchors 17](#_Toc72848649)

[4.2.1 Matching the counterfactual Anchor 17](#_Toc72848650)

[4.2.2 Mismatching the input Anchor 18](#_Toc72848651)

[4.2.4 Analysis of experiments 19](#_Toc72848652)

[4.2.5 User interpretation 20](#_Toc72848653)

[4.2.6 Conclusion 20](#_Toc72848654)

[4.3 Counterfactuals 21](#_Toc72848655)

[4.3.1 Existing solutions for counterintuitive counterfactuals 21](#_Toc72848656)

[4.3.2 Generalized counterfactuals 21](#_Toc72848657)

[4.3.3 Multi-objective counterfactuals 22](#_Toc72848658)

[4.3.4 User perspective 22](#_Toc72848659)

[4.3.5 Future feature engineering 23](#_Toc72848660)

[5. Final user study 24](#_Toc72848661)

[Set-up 24](#_Toc72848662)

[Profile 1\_6B.Trace#01 25](#_Toc72848663)

[Profile 2A3.3 26](#_Toc72848664)

[Discussion user evaluation 27](#_Toc72848665)

[Reflection on methodology 28](#_Toc72848666)

[References 29](#_Toc72848667)

# 1. Proposal and methodology

## 1.1 Problem description

DNA-experts have to determine the Number Of Contributors (NOC) to a DNA sample, when it is evident that it consists of DNA from multiple people. This is required before any further analysis can be done. 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 to achieve an accuracy of about 83%. However, the only output that DNA-experts can consult is the predicted NOC. When the expert analyzes a profile and comes to different NOC than the machine learning model outputs, there is no way to determine who is correct. No information about how the model came to this conclusion is provided, therefore not allowing experts to use this tool effectively as support for their decision making. This decision is important for further calculation of evidence [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 number of contributors. 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.

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.

Table 1: Overview of which research questions were answered in which cycles.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cycle 1: baseline | Cycle 2: exploration | Cycle 3: implementation | Cycle 4: integration | Cycle 5: evaluation |
| Question 1 |  |  |  |  |  |
| Question 2 |  |  |  |  |  |
| Question 3 |  |  |  |  |  |
| Question 4 |  |  |  |  |  |
| Question 5 |  |  |  |  |  |
| Question 6 |  |  |  |  |  |
| Question 7 |  |  |  |  |  |
| Question 8 |  |  |  |  |  |

### 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.

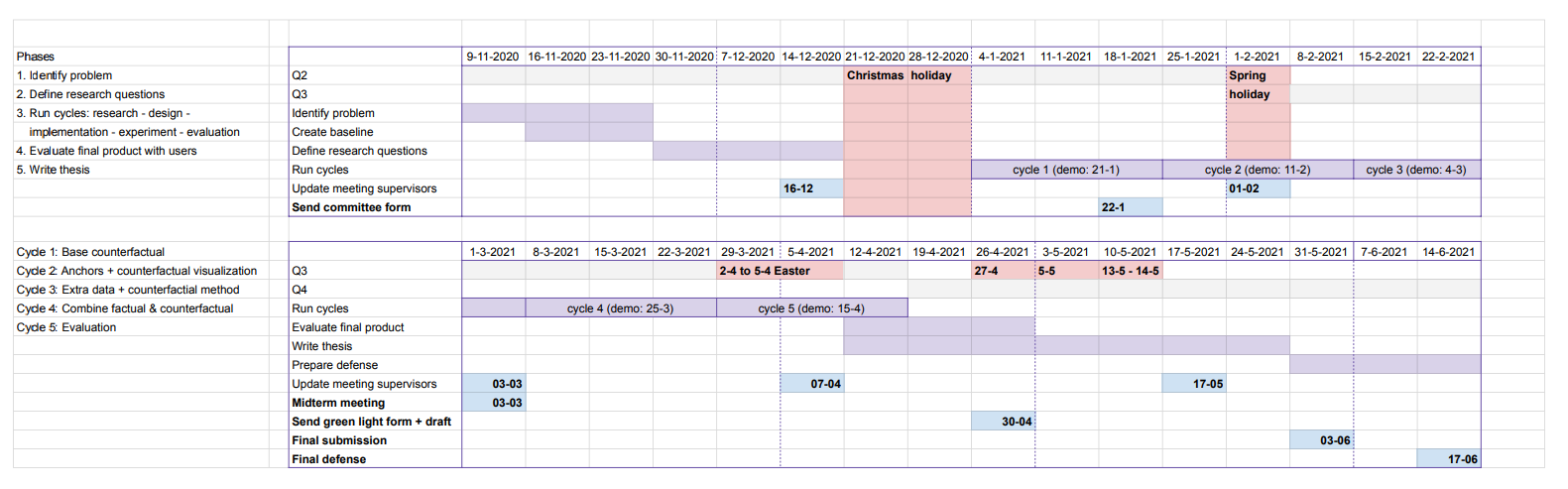


Figure 1: Gantt-chart showing the general planning of the thesis. The phases on the left show the general phases of the project; starting from identifying the problem and the research questions, then moving into cycles of development. The themes of the 5 cycles are listed below, these were adjusted during 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 in the left side of the chart. Milestones and meetings are shown in blue, and holidays in re

# 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 STR 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, the United States national DNA database. We can capture the STR with a process called electrophoresis, which produces an electropherogram. In Figure 2, we see a simplified result that 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.

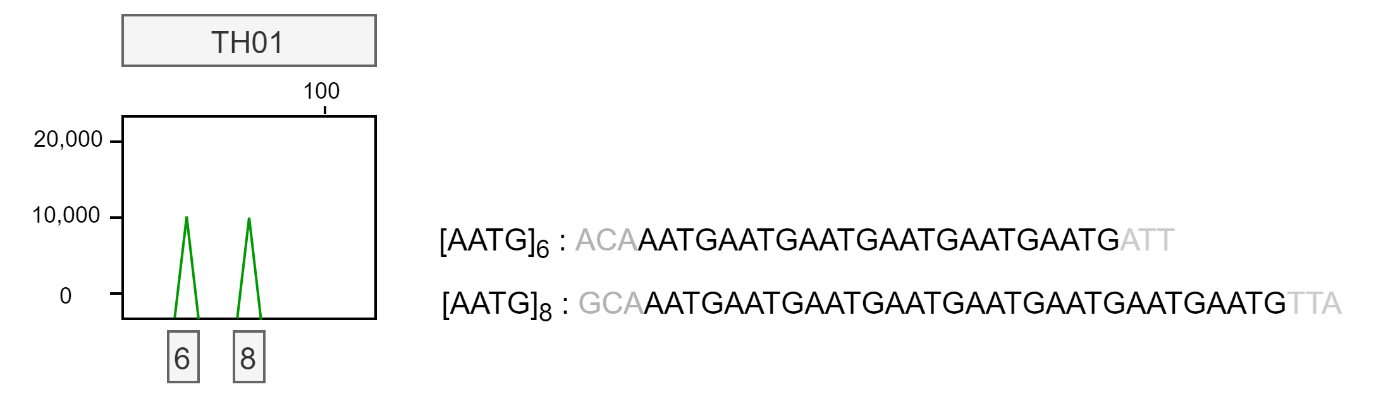


Figure 2: Simplified electropherogram result for locus TH01 showing two alleles with six and eight repeats each. The repeat sequence is shown on the right with arbitrary flanking regions that do not represent reality.

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 1, 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]. When an incorrect NOC is used for further analysis involving the investigation of the DNA profiles, the results are unreliable [2]. It could make the difference between whether or not a person of interest is included in the evidence or not.

Determining the exact number of contributors is difficult. 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.

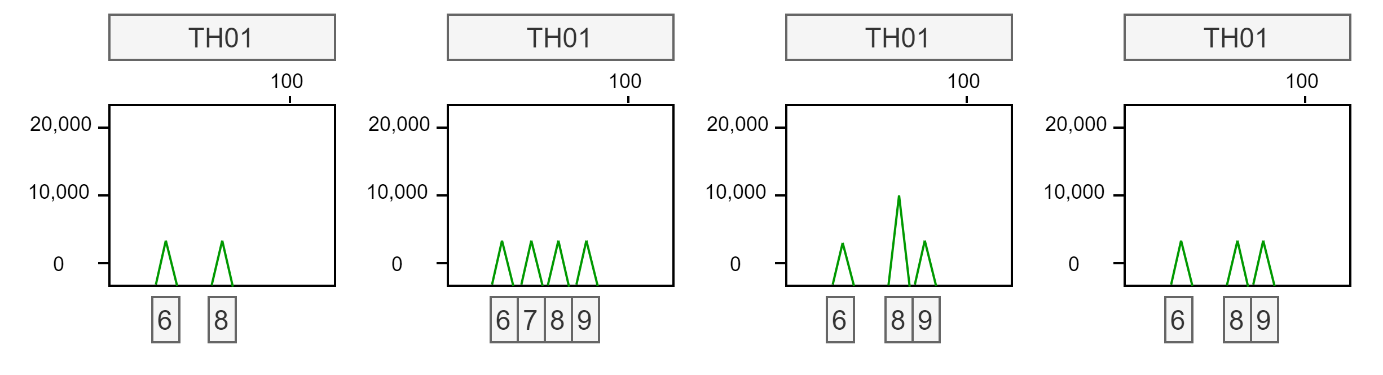


Figure 3: Four simplified electropherogram results for locus TH01. From left to right: Example of a simple single donor profile; Example of a simple 2-person mixture profile; Example of a 2-person mixture profile with allele sharing or a homozygous allele, peak 8 is twice as high compared to peak 6 and 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 has twice as much information 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 noise filter. 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 right-most two images in Figure 4 demonstrate these phenomena which are also described below.

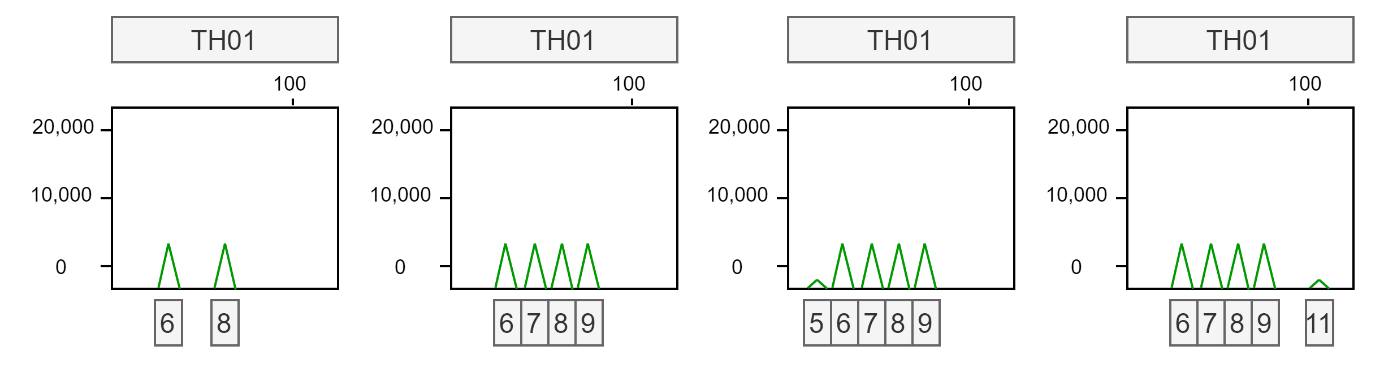


Figure 4: Four simplified electropherogram results for locus TH01. From left to right: Example of a simple single donor profile; Example of a simple 2-person mixture profile; Example of a 2-person mixture profile with a stutter peak at allele 5 caused by the folding of a STR with 6 repeats; Example of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

* 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.
* Allele drop-in or other noise: The measuring process is not perfect, so some random noise 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. As a result, some DNA information might also be lost if there is little material available.

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 is unreliable due to the factors of allele sharing, drop-out, etc. Performance is quite poor, especially with 3 or more contributors, when there is a lot of allele sharing, or when the quality of the profile is low [18, 19]. On average, when assessing mixtures between 2-5 contributors, the MAC obtains correct predictions for about 60-70% of samples [1, 19, 20].

Besides the MAC, the Total Allele Count (TAC) is also frequently used. 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 raw 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 (Appendix 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 raw peak information to create predictions. This raw 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 should be used 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 were SHAP values, while option B was 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.



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

### 3.3.1 Analysis answers question 2

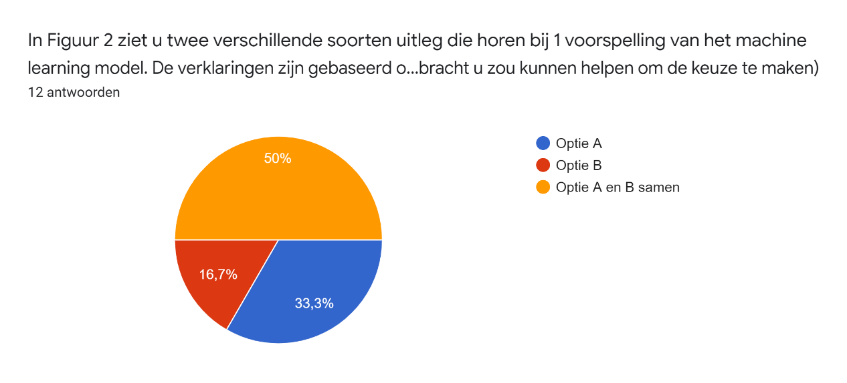


Figure 6: Pie chart of answers to question 2.

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 raw 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).

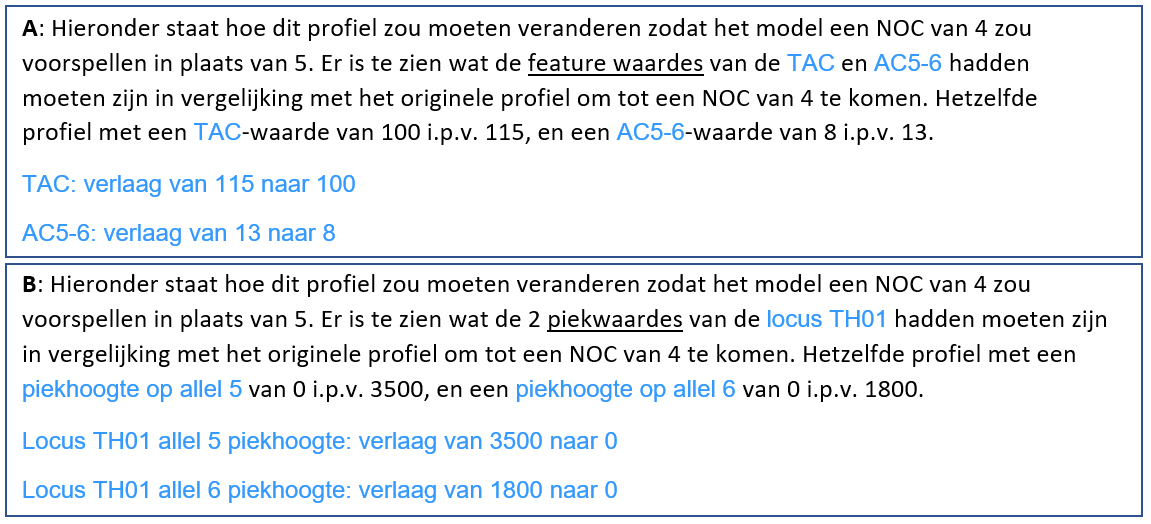


Figure 7: Question 3 with option A showing an explanation based on features and option B showing an explanation using peak height information.

### 3.4.1 Analysis answers question 3

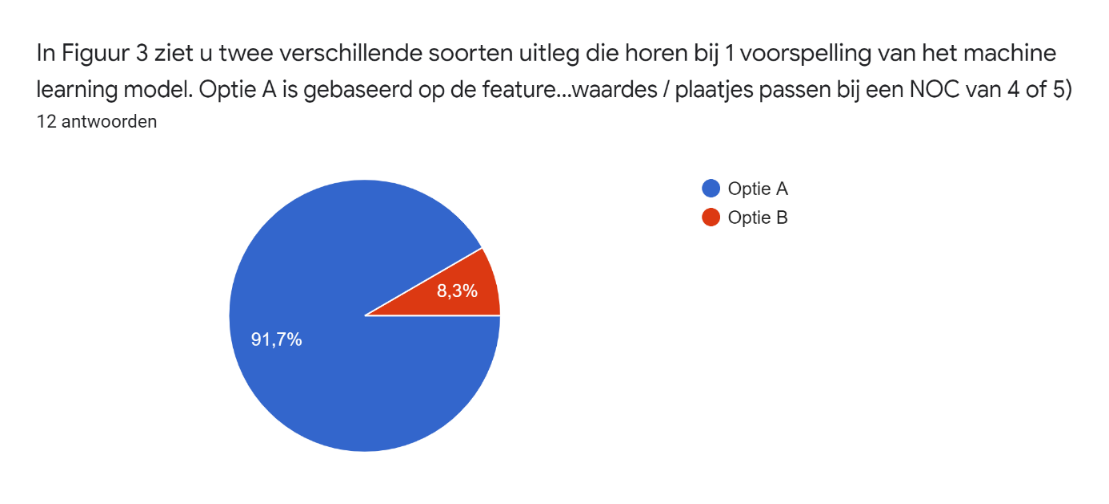


Figure 8: Pie chart of answers for question 3.

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, 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 Concluding remarks

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 especially surprising to see that almost nobody found the raw peak data informative, though some of this could be attributed to the fact that we only presented information about a single locus.

# 4. Experiments with various XAI techniques

From the user survey, we determined there we two questions that required an answer.

1. *What were the main feature values that have caused the model to reach the current prediction?*
2. *With which minimum set of feature value 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 the question 2 of the user survey, we received positive feedback on the general information 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.

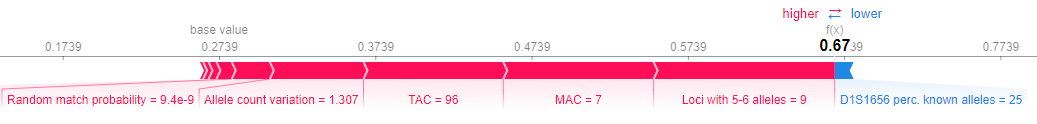


Figure 9: SHAP force plot for classification

The experts like that this explanation gives them access to the probability of the prediction, and they enjoy 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 think it is 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.

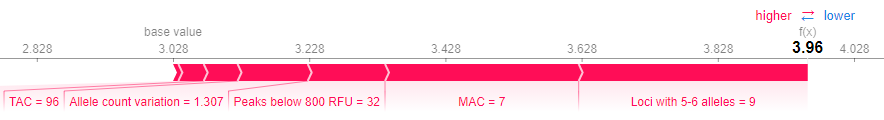


Figure 10: SHAP force plot for regression

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 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 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, they can interpret the SHAP values in relation to the model quite well.

## 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 defining 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. The idea was to combine both Anchors with counterfactuals to incorporate factual- and counterfactual information into one picture.

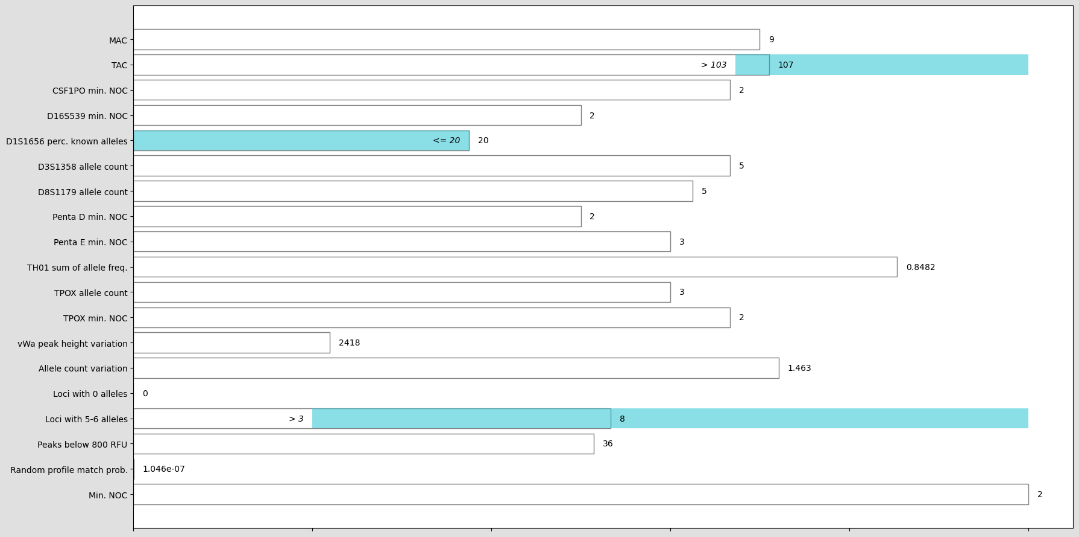


Figure 11: Anchors visualization.

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 is valid counterfactuals could be produced when derived from these differences in Anchors:

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

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

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

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

As can be seen in Table 2, the prediction did change after matching the counterfactual Anchor from Table 1. 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 Table 3 and Table 4.

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

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

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

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.

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

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

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

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

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

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 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 the counterfactual Anchor from Table 3, 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.

### 4.2.4 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, it is defined 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 3, 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.
* Changing the input instance to match the counterfactual Anchor creates an instance that is unrealistic and/or does not match the perturbed data.
* The sampled data does not represent the dataset accurately and therefore the specific values of the Anchors do not hold for the dataset.

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. This means that major contributing features could be excluded from the explanation.

### 4.2.5 User interpretation

We presented the picture in Figure 9 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 premise of this explanation does not hold.

The one positive remark about Anchors was the fact that it shows a range. One expert mentioned that in the example of Figure 9, if more alleles were 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.

### 4.2.6 Conclusion

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

* 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. 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 that are 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 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 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 7.

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).

|  |  |  |
| --- | --- | --- |
| **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) |

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. Presenting a counterfactual profile that has similar quirks to the input, even though those do not fit the average, is preferred. The counterintuitive differences must be solved from a different perspective.

### 4.3.3 Multi-objective counterfactuals

For solving multi-objective optimization problems, there exist several approaches all with their own ad- and disadvantages. We considered two different 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 seen in Table 9. Therefore, 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.

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

|  |  |  |
| --- | --- | --- |
|  | **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 |

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.

### 4.3.5 Future feature engineering

Many of the DNA-experts expressed that they did not like the current features, they are too complicated, they do not represent information that the experts think relates to the number of contributors, and they are incomplete. From these brainstorming sessions, we gained a lot of information about the experts’ opinions of these features. From working with the features, we also uncovered some insights related to their use from a machine learning perspective.

For future feature engineering, the following changes are recommended from both a machine learning, and the DNA-experts’ perspective:

1. Remove features that are redundant. For example, MinNOC*locus* encodes the same information as AlleleCount*locus*. MinNOC*locus* is equal to AlleleCount*locus* divided by two, rounded up. 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 loci 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.

1. Make profile features complete. For instance, the 19 features now includes 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 viability. Most of these were brought up by the users.

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 are variable 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.

# 5. Final user study

It was important to do a soft evaluation get the DNA-experts’ opinion of the final explanation. There were 8 responses in total, of which 7 were useable.

## Set-up

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.

As the features are not yet understandable, we did not base our evaluation on how well users can predict the NOC with the help of the explanation. Instead, the evaluation is focused on whether the explanation can help the user understand based on what information the model makes predictions. This could provide evidence that could help users be convinced of a certain prediction, or make them wonder if the model made a mistake. The two main questions of the evaluation are therefore:

1. Can the explanation take away some doubt in the user when the model seems certain in its prediction of the NOC?
2. Can the explanation introduce some doubt in the user when the model seems uncertain in its prediction of the NOC?

For these two questions, we used two different profiles that correspond to the described situations. Profile “1\_6B.Trace#01” is fairly certain for the model while users have struggled determine the NOC, and profile “2A3.3” was misclassified by the model. For comparison, we used two state-of-the-art techniques that are fit for each type of explanation. For question 1, we compared our visualization against a SHAP force plot, while for question 2 a counterfactual table was used for comparison.

We also wanted to get the users’ opinion of the visualization:

1. Do users find the compound visualization nice to use?

This question is also posed in comparison to the SHAP force plot and counterfactual table.

## Profile 1\_6B.Trace#01

This profile was chosen because it was difficult in an old training; 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.

The explanation shows that it is indeed on the edge between 3 and 4, but it shows that quite a lot of change is required to reach 4.

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?
4. Do you think the prediction of 3 (3.22) is correct?
5. 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)?
6. Please select all number(s) of contributors you would consider when you can consult the explanation above (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.
7. Do you think the prediction of 3 (3.22) is correct?
8. Can you explain why you have answered questions 5 and 6 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)?

## Profile 2A3.3

This 3-person mixture profile was predicted by the model to have 4 contributors. Because the output value of 3.53, it also seems that it is on the edge between 3 and 4.

The explanation

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.
4. Do you think the prediction of 4 (3.53) is correct?
5. 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)?
6. Please select all number(s) of contributors you would consider when you can consult the explanation above (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.
7. Do you think the prediction of 4 (3.53) is correct?
8. Can you explain why you have answered questions 5 and 6 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)?

We also found it important to see if users gained some insight into how the model makes predictions, and when it might be wrong. Do users let this influence their decision or not?

With the help of the visualization, you do grasp the image of a profile better. In this way, a large amount of data is available, in the simplest representation. It will take some getting used to.

Instead, there are some ways to circumvent this issue. For instance, by providing an example that mainly focusses on the well-known features such as the MAC and TAC.

The explanation can create some insights on how the model operates. Some decision are quite easy for the model; if the explanation shows that, and the

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DiCE** | **WhatIf** | **ReCo** |
| Creates counterfactuals from | sampling | training data | training data + filter |
| Minimizes | distance to profile +  no. of features changed | distance to profile | distance to profile +  no. of features changed |

Show DiCE result as experiment?

**Brainstorm sesh: 23feb\_BiS presentatie**

## Discussion user evaluation

Over the entire duration of the thesis, we have tried to engage as many end-users as possible. However, there were several factors that made this task a lot more difficult.

First of all, the Covid-19 pandemic made it so that no on-site activities could be organized. It is more difficult to engage users in a brainstorming session when on a video call, since people can only talk one at a time, making fluid conversation is more difficult. This resulted in 1 or 2 people mainly contributing, and others listening in. In a video call, people also get distracted as they are not in the same room as the activity. We would have preferred to plan brainstorming sessions and the final evaluation in a more controlled environment. 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.

Secondly, the DNA-experts at the NFI are under quite some work-related pressure. Any time they participated in a survey or a discussion, that would take time from their normal activities. Therefore, there was limited response. Out of approximately 35 workers, 12 responded to the initial survey, 6 participated in a brainstorming session, and 7 usable responses were collected in the final survey. This introduces a significant representation bias in the obtained results.

Because of the limited response, uncontrolled environment, and various sources of bias, we decided to only consider the brainstorming session and surveys as mainly subjective opinions. We did not perform any statistical analysis because those results would be unreliable and biased. The results were used for guidance of direction, and a general collection of users’ sentiments.

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 users lose the total picture.

### Reflection on methodology

Despite confirming with supervisors from the NFI and TU Delft that a linear strategy would suffice, I 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 no hinder progress as much.

With similar flexibility, I adapted the main research question to better fit the needs of users. Where I originally wanted to only focus on the counterfactuals, the users made me realize that a combination with feature importance is more valuable.

One aspect I would 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 agreed 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 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-expert’s management as well so that this could be planned into their schedule. Within Covid restrictions, a specific moment could still have been planned, but it would lose some the controlled setting.

Perhaps a more informal survey would have been sufficient for the final evaluation, as with such a small group of responders and the uncontrolled environment, the results could not be statistically evaluated. There was too much bias and too many distracting factors.

What I did well was to be aware of this risk from the start, and plan the user evaluations in advance.

# 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.