Additional thesis chapters

[Proposal and planning 3](#_Toc72168416)

[Problem description 3](#_Toc72168417)

[Main related works 3](#_Toc72168418)

[Contribution 4](#_Toc72168419)

[Example scenario 4](#_Toc72168420)

[Planning 4](#_Toc72168421)

[DNA Mixture interpretation 8](#_Toc72168422)

[Short Tandem Repeat (STR) profiles 8](#_Toc72168423)

[Estimating the Number of Contributors (NOC) 8](#_Toc72168424)

[*MAC-method* 10](#_Toc72168425)

[Survey on mixture interpretation and explanation types 11](#_Toc72168426)

[Set-up 11](#_Toc72168427)

[Question 1: workflow 11](#_Toc72168428)

[Analysis answers question 1 11](#_Toc72168429)

[Question 2: feature importance or counterfactual explanations 12](#_Toc72168430)

[Analysis answers question 2 12](#_Toc72168431)

[Question 3: features or raw peak information 13](#_Toc72168432)

[Analysis answers question 3 13](#_Toc72168433)

[Concluding remarks 14](#_Toc72168434)

[Experiments with various XAI techniques 15](#_Toc72168435)

[SHAP 15](#_Toc72168436)

[Anchors 15](#_Toc72168437)

[Experiment 1 15](#_Toc72168438)

[Experiment 2 16](#_Toc72168439)

[Experiment 3 17](#_Toc72168440)

[User experiment 17](#_Toc72168441)

[Conclusion 18](#_Toc72168442)

[Counterfactuals 18](#_Toc72168443)

[WhatIf 19](#_Toc72168444)

[DiCE random 19](#_Toc72168445)

[DiCE genetic / GeCo 19](#_Toc72168446)

[Multi-objective optimization 19](#_Toc72168447)

[Final user study 21](#_Toc72168448)

[Discussion 21](#_Toc72168449)

[References 22](#_Toc72168450)

# Proposal, planning

## Problem description

The NFI has previously developed a machine learning model to predict the number of contributors (NOC) to a DNA sample based on features derived from Short Tandem Repeat (STR) data [1]. This Random Forest classifier currently uses 19 statistical features of the STR data to achieve an accuracy of about 83%. However, the only output experts receive is the predicted NOC. 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]. Currently, the workflow of the experts consists of first determining the NOC based on an electropherogram profile, after which they validate their own ideas with the current NOC model. However, if these outcomes do not align, the expert has no way to determine if they made a wrong prediction or if the model did.

With the addition of 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 brings up a good argument.

We propose to develop informative explanations that can be applied to any ML model the NFI wishes to implement in the future for determining the number of contributors from STR profiles.

## 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 by the literature [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. We call these 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 class A?”* in the context of a classification problem. 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 field is new, numerous methods are being developed, yet none has particularly risen to the top as with SHAP. Similarly, these methods have hardly been submitted to user study. A recent study has classified existing techniques for generating counterfactuals according to certain properties [11]. For instance, whether they exploit parts of the underlying model, the data distribution, or how many feature changes are permitted.

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]. It seems that for this problem, counterfactual explanations could be especially valuable since the problem that experts face when determining the NOC is contrastive in nature (Why did the model predict a NOC of 5 when the expert derived a NOC of 4?). Because counterfactual methods are an active research direction, they lack a lot of practical testing, which this study could also be used for.

## Contribution

In this study, we want 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 does the NOC machine learning problem look like?
3. What purpose does an explanation of the NOC machine learning model serve?
4. Which types of local explanations could work for this problem?
5. How can local explanation techniques be adapted to be applied to this problem?
6. How can we evaluate the generated explanations from a machine learning perspective?
7. How can we evaluate the generated explanations from a user perspective?

## Example scenario

* *Expert predicts a sample to be a 4-person mixture based on feature values x=10, y=15, z=20*
* *NOC model predicts that sample to be a 5-person mixture*
* *Expert wonders why*
* *SHAP explanations shows that the feature value of x has contributed most to the prediction of a 5-person mixture by the model. The feature values of y and z have also contributed.*
* *Expert can see how these feature values are important for both a NOC of 4 or 5, but does not understand why the prediction then is not 4.*
* *Counterfactual explanations states that if feature x was 5, it would be classified as a 4-person mixture*
* *Expert determines that this lower value of x is not indicative of a 5-person mixture, and therefore assigns a NOC of 4 to this profile.*

## Planning

Originally, the thesis was planned in a linear fashion.

At the time of writing, the main problem has been identified. The research questions are draft versions and might still evolve over the course of the thesis. In Figure 1, the next phases are planned over the course of the thesis period.

From this point forward, there will be development cycles. Each cycle starts with one or multiple goals, for which relevant literature will be consulted

The main steps are to perform a survey for questions 1, 3 and 4; analyze the literature for questions 5-8, as well as implement any suitable techniques to the current ML model. Due to the workload of the DNA experts, user studies must be kept to a minimum. Therefore, there will be a survey of the baseline for questions 1, 3 & 4, and an assessment of the final product.

There are some risks associated with this approach, for which some mitigation steps could be defined as follows:

1. There are no suitable counterfactual explanation techniques available for this type of data.
   1. Move towards other local explanation methods.
2. The suitable counterfactual 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.
   3. Implement other local explanations methods.
3. No users want to participate in the user study.
   1. Only use Corina Benschop to get at least one expert evaluation.
   2. Use feedback from colleagues at the NFI as input.

**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**.

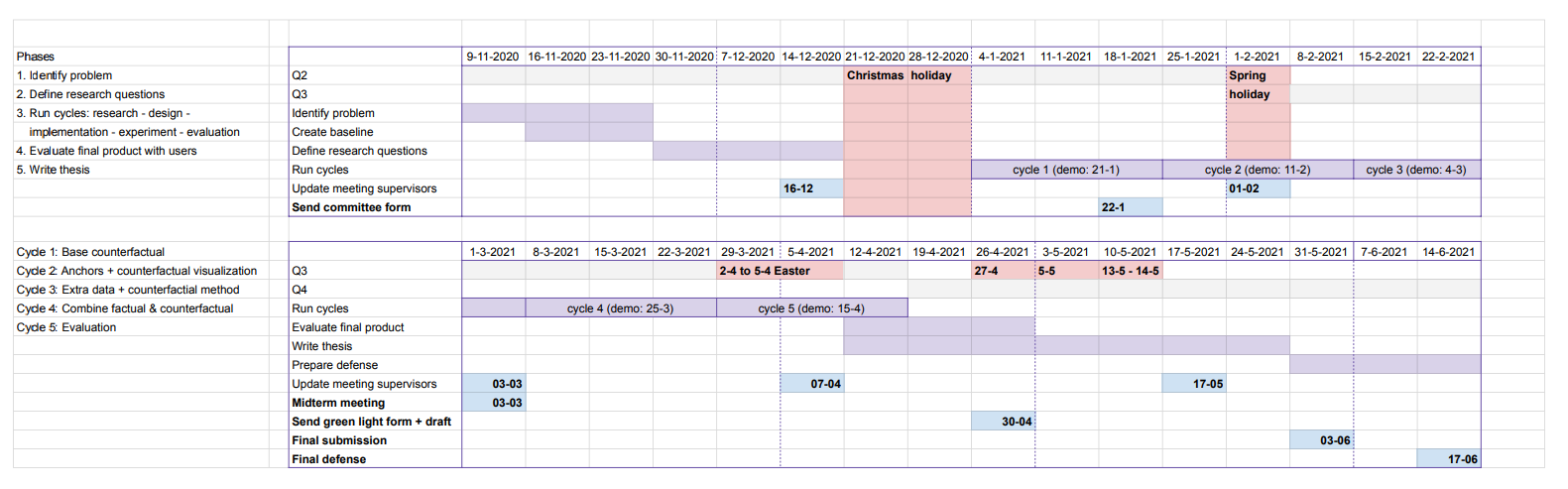


Figure 1: Gantt-chart showing the general planning of the thesis. The phases on the left show the general phases of the project; starting from identifying the problem and the research questions, then moving into cycles of development. The themes of the 5 cycles are listed below, these were adjusted during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

# 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 become 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 highlights the different methods to determine the number of contributors.

## 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 1, 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 x-axis shows the location of the locus on the DNA strand. 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.

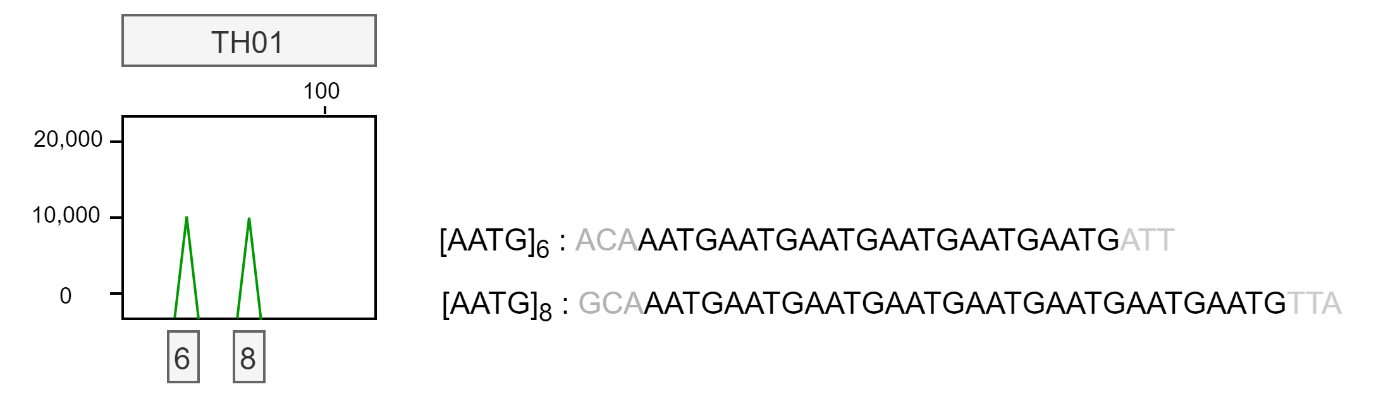


Figure 2: Simplified electropherogram result for locus TH01 showing two alleles with six and eight repeats each. The repeat sequence is shown on the right with arbitrary flanking regions.

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.

## 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 considered a mixture. The next step is to determine the number of contributors. 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]:

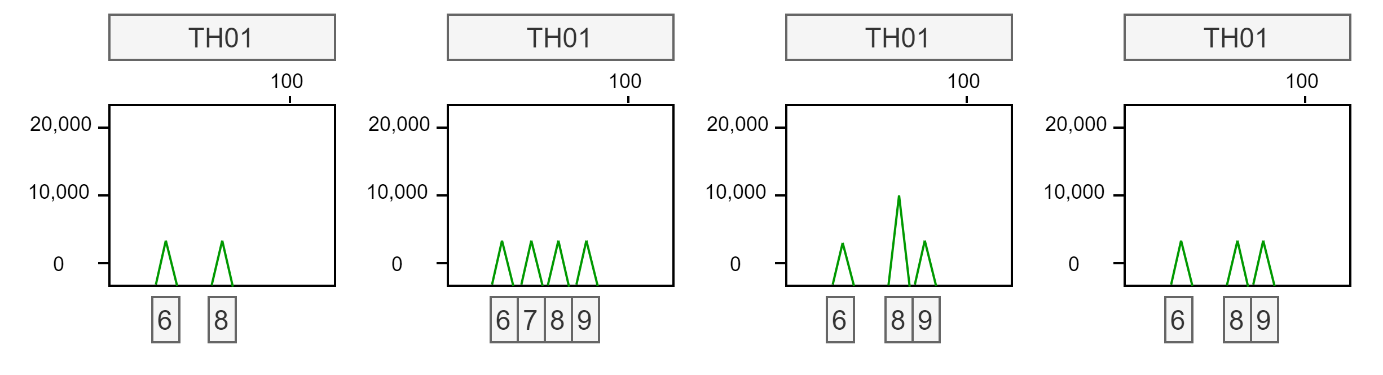


Figure 3: Four simplified electropherogram results for locus TH01. From left to right: Example of a single donor profile; Example of a 2-person mixture profile; Example of a 2-person mixture profile with allele sharing, peak 8 is twice as high compared to peak 6 and 9; Example of a 2-person mixture profile 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 of Figure 2; 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 the DNA available 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.

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 [14]:

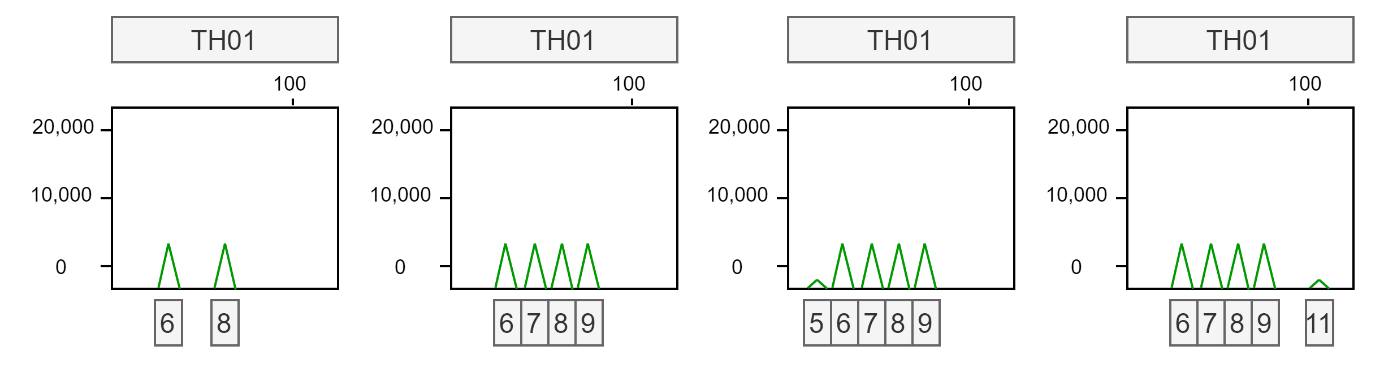


Figure 4: Four simplified electropherogram results for locus TH01. From left to right: Example of a single donor profile; Example of a 2-person mixture profile; Example of a 2-person mixture profile with a stutter peak at allele 5 caused by the folding of a STR with 6 repeats; Example of a 2-person mixture 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 3, this is
* 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 3, 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 due to a low-quantity donor.

### *MAC-method*

*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 in general 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].*

## 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.

### 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 it lines up with the studied literature. 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 23 loci per profile. With this question, we intended to find any preference regarding the data.

### 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.

### 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.**

### 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.

### Analysis answers question 2

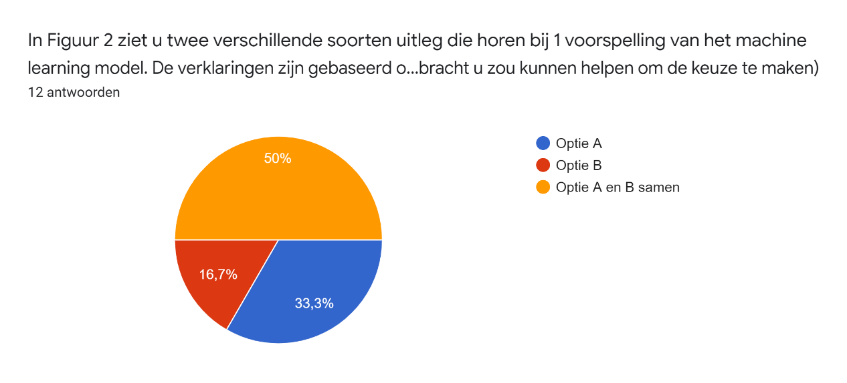


Figure 6: Pie chart of answers to question 2

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.

### 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).

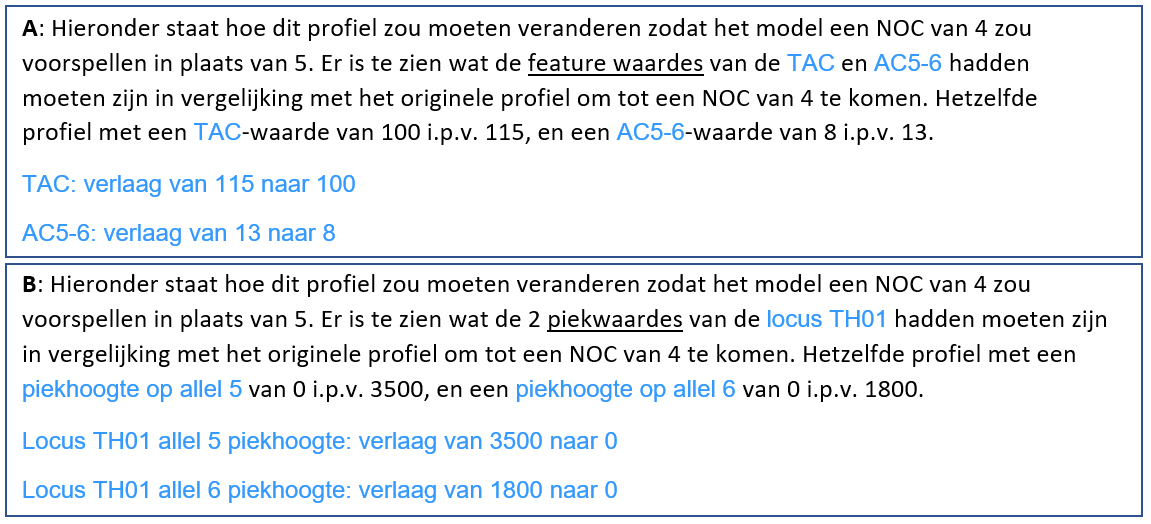


Figure 7: Question 3 with option A showing an explanation based on features and option B showing an explanation using peak height information.

### Analysis answers question 3

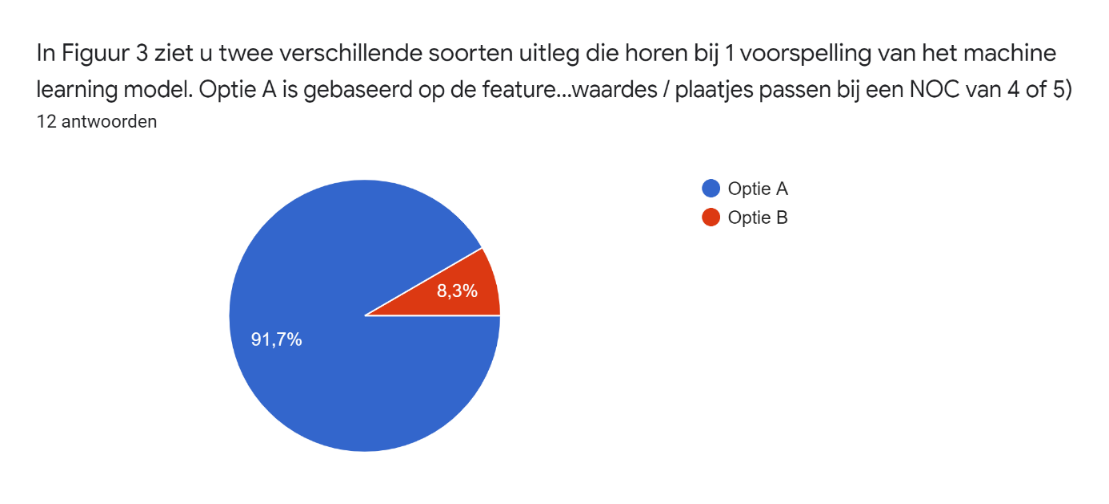


Figure 8: Pie chart of answers for question 3

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.

## 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.

# Experiments with various XAI techniques

## SHAP

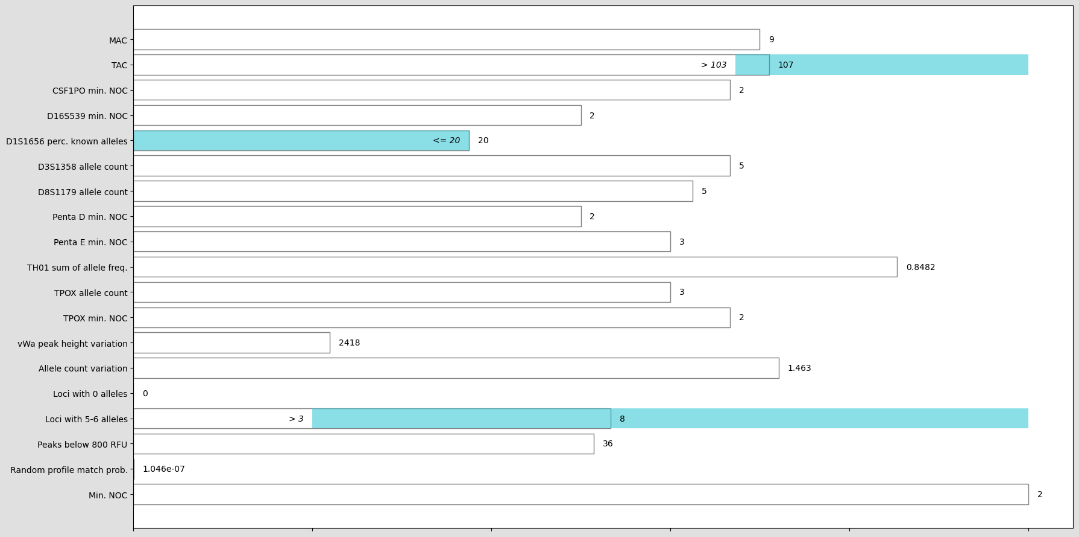
<How works for multi-class/regression>

* Interpretation by experts

It is an issue that feature importance methods split the impact on the model over correlated features. The result of this issue is that the importance value for correlated contributing features is underestimated, in contrast to if their importance was left undivided. However, since main goal of these values is to give an impression of the contributing factors to a prediction, the exact values are not a priority. This part of the explanation is to observe a general sense of which features contributed to the prediction in which direction. For this purpose, we deem SHAP adequate.

## 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 [21]. 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. The definition of Anchors also seemed to lend itself to create sparse counterfactuals from. *If the Anchor shows between which feature values the current prediction holds, if we go outside those ranges the prediction will probably switch.* For that reason, we experimented with some examples.

Experiment 1

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.

|  |  |
| --- | --- |
| 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 apply to 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 apply to original data with a probability of 0.02* |

Changing the input features that do not match the counterfactual Anchor so that they do:

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

Prediction: **2.0** with probability 0.67

This seems to work well, the prediction changed after only adjusting two feature values. 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.

For example:

|  |  |
| --- | --- |
| 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 apply to 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 apply to original data with a probability of 0.01.* |

Changing the input features that do not match the counterfactual Anchor so that they do:

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

Prediction: **1.0** with probability 0.87

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

### Experiment 2

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.

|  |  |  |
| --- | --- | --- |
|  | 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: **2.0** with probability 0.48

However, it apparently matters *how much* we change it.

|  |  |  |
| --- | --- | --- |
|  | 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** with probability 0.59

Perhaps this has to do with how anchors are *binned*. 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.

### Experiment 3

Considering the following Anchor:

|  |
| --- |
| Anchor |
| *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 apply to original data with a probability of 0.01.* |

Setting all other feature values to 0 results in a **prediction of 1.0 with a probability of 1.0**.

### User experiment

### Conclusion

An anchor A is a set of predicates that returns true for an instance x if all the predicates hold.

e.g. if x = “This movie is not bad.”, f(x) = Positive, A(x) = 1 where A = (“not”, “bad”)

D(·|A) is the conditional distribution where rule A applies.

A is an anchor if .

Where A(x) = 1 means that the anchor A holds for instance x. And the anchor A is a sufficient condition for f(x) with high () probability, for a sample z from D(z|a), for a certain prediction f(x) = f(z).

D is defined by a validation dataset (training set?). By fixing A, then sampling the rest of the row, D(z|A) is defined.

“anchors are by construction faithful, adapting their coverage to the model’s behavior and making their boundaries clear.”

Formally, we define the coverage of an anchor as the probability that it applies to samples from D, i.e. cov(A) = ED(z) [A(z)].”.

The probability that they apply to the sampled instances, not the actual data. Therefore, it is very dependent on this sampled neighborhood. As the random sampling does not respect any feature correlations, the neighborhood is not quite realistic.

Moreover, the process of adding rules to the Anchor is a stochastic process. This means that the features that are picked to be a part of the Anchor are not necessarily the ones that are the most influential or the most important. This also means that if the same profile is explained multiple times in a row, different Anchors will be shown.

As such, the Anchors are not very useful:

* The rules included do not present the most important feature values that led to the current prediction.
* Their ranges do not hold for real data due to the sampling process not handling correlated features and the neighborhood being limited.
* The definition of the neighborhood for which the Anchors hold, cannot be communicated clearly to users.

## 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. For example, in one of the instances it showed to *increase* the TAC to *decrease* the NOC. This is counterintuitive since more alleles should correspond to more contributors. One way to mitigate that is to show multiple profiles, but people do not want to have to extrapolate which information is relevant themselves. Similarly, by showing the distribution of multiple profiles you can get the same effect. However, distributions can be confusing to people that are not used to them. They can also occlude local effects that are not visible from a global perspective.

Though a matching tolerance for numerical features sounds like a good idea for enhancing sparsity in counterfactuals [22], we argue that this might overlook the exact values on which a model makes a decision. This could thus lead to missing the target prediction.

### WhatIf

### DiCE random

### DiCE genetic / GeCo

### Multi-objective optimization

For solving multi-objective optimization problems, there exist several approaches all with their own ad- and disadvantages.

The first paper which generated multi-objective counterfactuals, used an implementation of finding the non-dominated instances [23]. What this entails is that for each of these instances in the set, there exists no better alternative; there cannot be an improvement for one objective, without decreasing the score for another objective. This is also called a Pareto or optimal set. Besides this optimality guarantee, another advantage is that scores with different scales and distributions can be compared without any normalization. One disadvantage is the computational effort, since many comparisons need to be made in order to identify the non-dominated set [24, 25]. Another issue that is particular to this implementation is that we wanted to show only one instance, and the non-dominated strategy produces a set.

Some studies have also considered the number of feature differences alongside the distance [23, 26, 27]. This can be implicitly incorporated by starting from the input instance and adjusting one feature at a time until the prediction switches [28, 29] **TODO: check refs here**. Others have implemented the number of feature differences as a constraint [22, 27, 30]. The difficulty with the latter is to define beforehand how many differences are allowed or even plausible. This can vary strongly between various inputs. Another approach is a post-hoc filtering step, where generated counterfactuals are greedily restored to the input instance until the target switches [26].

weighted sum: difficult to balance the two objectives properly; how much increase in distance, would you consider equal to an added different feature. The mean and variance of the two scores are also not equal; even though both scores lie between 0-1, the distance method... However the decision maker, in order to choose the coefficients, must have a clear perception of how this choice influence optimal points.

Its main disadvantage is the difficulty to determine the appropriate weight coefficients to be used when enough information about the problem is not available (this is an important concern, particularly in real-world applications). Also, a proper scaling of the objectives requires a considerable amount of extra knowledge about the problem. To obtain this information could be a very expensive process

# Final user study

## Discussion

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 where there to support that person, the outcome might have been different.

Secondly, the experts at BiS 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. The results were used for guidance of direction, and a general collection of users’ sentiments.

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

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

23. Dandl, S., et al. *Multi-Objective Counterfactual Explanations*. in *Parallel Problem Solving from Nature – PPSN XVI*. 2020. Cham: Springer International Publishing.

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

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

26. Mothilal, R.K., A. Sharma, and C. Tan. *Explaining machine learning classifiers through diverse counterfactual explanations*. 2020.

27. Grath, R.M., et al., *Interpretable Credit Application Predictions With Counterfactual Explanations.* ArXiv, 2018. **abs/1811.05245**.

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

29. White, A. and A. Garcez. *Measurable Counterfactual Local Explanations for Any Classifier*. in *ECAI*. 2020.

30. Moore, J., N. Hammerla, and C. Watkins, *Explaining deep learning models with constrained adversarial examples*. 2019. p. 43-56.