# Proposal Thesis

## 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, not many of these methods have been rigorously tested or compared. 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. However, these methods have not been actually put to the test based on quantitate measures, or have been submitted to user study.

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 model-agnostic counterfactual explanations for ML models that predict the number of contributors. To achieve this, we must identify the existing techniques for generating counterfactual explanation 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 counterfactual 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. Which model-agnostic counterfactual explanation techniques exist and for which types of problems are they suitable?
4. How can existing counterfactual explanation techniques be adapted to be applied to our problem?
5. How can we evaluate the generated explanations from a machine learning perspective?
6. How can we evaluate the generated explanations from a user perspective?

**Planning**

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. The main steps are to perform literature survey to answer sub question 3, to implement any suitable techniques to the current ML model for determining the NOC, and to perform experiments with the experts to determine the value of the implemented explanations.

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.

Figure 1: Gantt-chart of general planning. On the left the main phases of the research are shown as well as any milestones in boldface.

(insert planning.pdf)

**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**.

Example:

* *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.*