Explainable artificial intelligence (XAI) for number of contributor (NOC) estimates in autosomal DNA profiles

Problem

Single source and high template DNA profiles are easy to analyse and interpret. However, crime scene samples are often mixed and of low quantity or low quality which complicates the interpretation of the DNA profiles. Mixed DNA samples may suffer from allele masking; low quality can result in degraded patterns; low quantity can result in stochastic artefacts such as heterozygote peak imbalance, drop-out, drop-in and elevated stutter. These factors, for instance, complicate the estimation of the number of contributors (NOC). Specifying the NOC is required by many of the current probabilistic genotyping software that are used to compute weights of evidence.

Artificial intelligence to aid in DNA profile interpretation

During a previous study, a machine learning model (the RFC19 model, a random forest classifier with 19 DNA profile features) was developed that takes into account various DNA profile characteristics resulting in improved NOC estimates when compared to conventional methods that are based on allele counting [1].

Research project

Although the RFC19 model outperformed allele counting methods to estimate the NOC, further improvements are regarded possible. For instance, the RFC19 machine learning model presents a prediction for the NOC and a probability for this NOC. The higher the probability, the more certain the model is about the predicted NOC which can be useful to the end user. However, to the end user it is unknown which features made the model decide on the NOC that was presented which makes it a black box model. Various explainable artificial intelligence (XAI) methods exist that enable opening black boxes [2,3] by presenting the feature importance, thus the features that positively or negatively contributed to a given prediction [4]. This could help the end-user in understanding the outcome of the model, it would also help enhancing user trust in adopting a machine learning model for NOC estimation. This XAI-NOC research project will focus on these methods to further improve the RFC19 model and our understanding of it.

Furthermore, other models or approaches can be researched to improve accuracies in NOC estimates. One option is examining a regression approach. This would have the benefit that the relationship between the classes would be clear to the algorithms. In other words, in a regression approach misclassifying a four-person mixture as a two-person mixture would yield a stronger penalty then misclassifying a four-person mixture as a three-person mixture. In the current model, the NOC is treated as a classification approach and both receive the same penalty. Furthermore, a regression approach would yield real numbers rather than integers, where the distance to the integers could be interpreted as a measure of uncertainty, similarly to how we currently use the computed probabilities.

Practical work student

The student will:

- Gain knowledge on the interpretation of mixed DNA profiles
- Gain knowledge on the various existing methods to estimate the NOC
- Gain insight in the recently developed RFC19 model
- Provide a link between data science and forensic genetics
- Research XAI approaches as a solution to the problem
- Program XAI model
- Compare performance of approaches
- Perform/help in validation of the method

Internship information:

- Background in data science or bioinformatics and affinity with data science is required for this internship project.
- The duration of the internship will be 8-9 months.
- The student will have the NFI supervision of forensic data scientists as well as forensic genetics researchers. Furthermore, this project will be performed in collaboration with a UvA data science professor.

References

[1] C.C.G. Benschop, J. van der Linden, J. Hoogenboom, R. Ypma, H. Haned. Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach. Forensic Sci. Int. Genet. 43 (2019) 102150.

[2] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti. A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018).

[3] https://christophm.github.io/interpretable-ml-book/

[4] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier. Proceeding KDD 2016, Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. August 13–17, 2016, pp. 1135–1144, , https://doi.org/10.1145/2939672.2939778.