A probabilistic approach to quantify the strength of evidence of presence of cell types from RNA data using a multi-label method

Voerman, N.

A probabilistic approach to quantify the strength of evidence of presence of cell types from RNA data using a multi-label method

Naomi T. Voerman (s2072661)

First internal supervisor: Prof. Dr. R.D. Gill

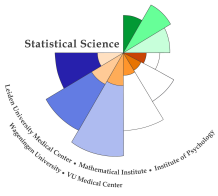
External supervisor: Dr. ir. R.J.F. Ypma

Second internal supervisor: Prof. Dr. M.A. van de Wiel

MASTER THESIS

October 17, 2019

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

In forensics it is relevant to identify the presence of one or several body fluids in a crime stain. This may be done using traditional methods however, those methods require a part of the available material, therefore leaving less residual material for the purpose of other analysis. Alternatively, one can use messenger RNA evidence: mRNA expression levels may vary among body fluids and therefore can be identified. The currently used method provides the forensic examiner with a categorical statement regarding the existence of the body fluid. However, such a method cannot express any associated uncertainty, whereas alternatively, a probabilistic method can and hence is a preferable choice. In forensic science it is common to express the level of uncertainty by means of a likelihood ratio but, due to a bad choice of statistical model or data scarcity, may be inaccurate.

This thesis first of all carries out experiments using four probabilistic classification methods, namely Multinomial Logistic Regression, Multilayer Perceptron, Extreme Gradient Boosting and a Fully connected Feed Forward model. In actual casework the crime stain often consists of multiple body fluids, which is why the classifiers are compared using synthetic representations of actual mixture samples. Multi-label approaches that enable the classifiers to express the level of uncertainty about multiple body fluids in a sample are used. The output from the logistic regression model is directly interpreted as likelihood ratio, whereas for the remaining three classifiers a post-hoc calibration step to improve the accuracy of the clasiffiers is included. Additional tests are performed to investigate how susceptible the classifiers are when the relative frequency of the body fluids in the data changes. The main focus is on two target classes, namely on saliva and a combination of vaginal mucosa and menstrual secretion, because these are most often requested to be identified in a crime stain and therefore seen as most relevant.

It is concluded that using a separate logistic regression model for each target class in combination with presence/absence data results in both accurate and reliable likelihood ratios. Results also indicate that these models are the least susceptible to a change in the frequency with which body fluids occur in the train dataset.

Furthermore, a study using an additional dataset with actual mixtures of two body fluids that are not assumed representative of forensically realistic mixtures of the same two components is done. Results show that the accuracy of the classifiers on the mixtures dataset are higher in comparison to the accuracy on the synthetic representations. This indicates that the results are overly optimistic, hereby verifying that the mixtures' cell type dataset should not be used as validation set.

A user-friendly tool is constructed that implements logistic regression to calculate the likelihood ratio from samples from actual casework. Using mRNA measurements from two cases both the practical use and the interpretability of the results are shown.

Acknowledgements

First of all I would like to thank Rolf Ypma for his remarkable and unending guidance. It is safe to say that without your help my thesis would not be where it is at right now. I would also like to thank you for joining all the meetings with other NFI staff members. I admire the way you are able to explain difficult topics in an intuitive way for everyone to understand and I have learned a great deal from that. Lastly, I would like to thank you for being understanding when I was not able to be present full time.

Secondly I would like to thank Richard Gill, who was willing to be my first supervisor even though he is already retired. It was a great pleasure to work together the past months and I learned a great deal from you during the meetings we had every 2 or 3 weeks. I am astonished about all the knowledge you carry in you and thankful for the way you answered my questions with care.

I would also like to thank Margreet van den Berge and Petra Maaskant from the Deparment of Human and Biological Traces from the NFI. Your enthusiasm was one of my main drivers for this project. Hopefully the outcome of my thesis acquires you with a suitable technique to study mRNA measurements.

Also many thanks to Marjan Sjerps and Peter Vergeer. Your critical notes about the statistical procedures were of great use.

I would like to thank Mark van de Wiel whom, even though he was my second supervisor, showed great interest in the project. Thank you for the effort you put in and the constructive criticism.

Lastly, many thanks to the colleagues from the team FBDA for providing such a pleasant working environment and the willingness from each and everyone of you to help whenever you could.

Contents

1 Introduction

1.1 Background

In forensic cases it is relevant to determine both who left a crime stain at a crime scene and what type of material was in the crime stain. Statements about the latter can be made with RNA measurements (i.e. RNA profiling). The body fluids that are forensically most interesting are blood, semen, saliva, menstrual secretion, vaginal mucosa, nasal mucosa and skin. They are often also referred to as cell types. Traditional methods to investigate the existence of cell types are microscopy, immunological, chemical and enzymatic methods, but have some drawbacks. First of all, they have to be carried out separately on each cell type which is a time consuming process. These methods also require a part of the material, leaving less material available for DNA analysis [1]. An alternative is using messenger RNA (mRNA) evidence, or mRNA-profiling. mRNA is a form of RNA which serves as messenger between two processes in which DNA is transformed into proteins. Within the DNA there are sections called genes containing instructions for making proteins. In the first process, called transcription, a gene is transferred to mRNA. When a gene (also called a marker) is used to transcribe the mRNA from the DNA, the gene is considered to be observed. Subsequently, a signal value may be observed for that marker. The second process, called translation, is when the mRNA is used to make proteins. The mRNA expression levels vary among cell types and are often specific for cell types, which is how mRNA can be used to make statements about the cell types in crime stains.

The Department of Human Biological Traces of the Netherlands Forensic Institute (NFI) currently does body fluid identification using the categorical n/2 method [2]. This method uses a RNA results table containing six categories to evaluate cell types. These categories are 'observed', 'observed and fits with', 'sporadically observed and fits with', 'sporadically observed', 'no reliable statement possible', 'non-specific due to high cDNA input' and 'not observed'. For each cell type it is determined how often a signal is present relative to the times a signal could have occurred. The latter depends on the number of RNA profiles and the number of specific markers of the cell type. For example, with four mRNA profiles and three specific markers the number of signals that can occur is twelve. It is then categorized as 'observed' when the signals are present in at least half of the possible positions. In case no signals are measured the category 'not observed' is given. When at least one signal but less than half of the signals is measured it is categorized as 'sporadically observed'. An explanation of the remaining categories can be found in Lindenbergh et al. [2]. One of the shortcomings of this method is that it is impossible to report different levels of (un)certainty since one is making a categorical statement. Also there is a so-called 'fall-of-the-cliff-effect', meaning that there is a hard cut-off between the categories 'observed' and 'sporadically observed'. When five out of twelve signals are observed the cell type is categorized as 'sporadically observed' whereas, in case of observing six signals the cell type is said to be 'observed'. In this case the strength of the evidence increases only a bit whereas, the statement changes drastically. Another shortcoming is that the same statement is reported when six out of twelve signals are observed as when all twelve signals are observed, so the method does not report a different outcome when the evidence is stronger. Another drawback is that only the cell type specific markers are selected whereas, signals of other markers may also contain relevant information [1]. So generally speaking the n/2 method ignores relevant information when identifying the body fluids in a stain and hence a method that is capable of incorporating this preferred. A probability model is able to make a probabilistic statement about the existence of cell types rather than a definite statement. Another convenience when using a probability model is that it is capable of modelling the unknown variation in the data though, it can only succeed once enough data is available. It will also use both the evidence of the markers that amplified and markers that did not amplify [1].

In former research probabilistic methods that can identify one cell type in a sample have been proposed. Dorum et al. in their study have also attempted to identify multiple (i.e. two-component) cell types. In the majority of those mixtures, the correct two cell types received the highest and second highest probability though, the second highest was a lot smaller. This is because the classifier was trained in a multi-class setting and the probabilities are calculated relative to the other probabilities. In another study, performed at the Netherlands Forensic Institute, they experimented with several probabilistic methods that were trained on a dataset with various synthetic mixtures of cell types to predict multiple cell types, using multi-label classification, on an actual mixtures dataset. They conclude that in a multi-label setting a model

is able to correctly classify both single and multiple cell type samples. Unfortunately, the results are unreliable because the mixtures dataset consists of unrealistic samples and hence their selected method presumably does not apply well to actual multiple component stains. It however remains relevant to identify a multiplex of cell types since the majority of forensic casework in which cell types of the donated material is relevant consists of mixtures [1]. The method will for example be of use in a rape case. If a male suspect claims that he is innocent of abusing the victim, mRNA could be used as evidence either against or supporting his claim. If the crime stain sampled from the victims body consists of both vaginal mucosa and semen, this may be evidence against the suspects' claim. This thesis will perform experiments with probabilistic classifiers in combination with multi-label methods using mixtures of cell types representative of actual mixtures. Note that the body fluids that the forensic scientists have to identify most often in actual casework are vaginal mucosa and saliva. Therefore, this thesis aims at producing accurate and reliable results regarding these two cell types.

1.1.1 Likelihood ratio

In the forensic field a commonly used measure that expresses the level of (un)certainty is the likelihood ratio (LR). With two hypotheses, one being "the crime stain contains cell type 1" (H_1) and the other being "the crime stain contains cell type 2" (H_2), the likelihood ratio can be calculated using the following equation:

$$\text{LR} = \frac{Pr(E|H_1)}{Pr(E|H_2)} \tag{1}$$

where E is the evidence, $Pr(E|H_1)$ is the conditional probability of the evidence under the hypothesis that the crime stain contains cell type 1 and $Pr(E|H_2)$ the conditional probability under the alternative hypothesis. The LR measures the strength of the evidence (i.e. the evidential value) for the first hypothesis compared to the second hypothesis. The higher the LR, the stronger the evidence that the crime stain contains cell type 1. The posterior probability is calculated by combining the likelihood ratio with the prior probability as in the following equation:

$$\frac{Pr(H_1|E)}{Pr(H_2|E)} = \frac{Pr(E|H_1)}{Pr(E|H_2)} \times \frac{Pr(H_1)}{Pr(H_2)} \tag{2}$$

Note however, that it is a forensic examiners' job only to calculate the LR and that it is up to a judge to define the prior. For convenience the likelihood ratio is often converted to the base 10 log likelihood ratio. The strength of evidence is then expressed on a scale which is symmetrical around zero [3].

1.1.2 Calibration

In practice, the value of the LR is handed to the court, where it is the judges' task to use the collected evidence to make their final judgement. One can image that their verdict is influenced by the value of the LR: they may be more inclined to convict when the evidential value supporting the hypothesis that the crime stain contains the cell type of interest is high. This is problematic in case the LR points in the wrong direction meaning that the evidence lends greater support to wrong hypothesis potentially leading for the judges to make a wrong decision. Hence, the forensic researcher must be confident about the correctness of the reported LR. This is when the LR is well calibrated.

The concept of calibration can be easily explained with the weather forecast example. For a sequence of days a weather forecaster predicts the probability p of it raining. The ground truth, so whether it actually rained or not, is known at the end of the day. After x days of weather forecasting, the probability of rain and the ground truth labels whether it actually rained those days are known. A way of evaluating the performance of the weather forecaster is to look at its calibration and can be measured using this knowlegde. In case rain occurs on $p * x$ of the days to which he assigns a probability of p of it raining, the forecaster is considered to be well-calibrated [4]. For example, say there are 10 days to which the same probability of rain of 0.8 is given, then it should rain 8 out of 10 days.

A probabilistic model, is well-calibrated when the output of the model yields correct LR values, meaning that the values are not too high or too low. Bad calibration indicates that a significant part of the LRs have a wrong value [5]. Generally speaking, the LRs that belong to H_1 are expected to be large (and above 1) whereas LRs that belong to H_2 are expected to be low (and below 1). Because of for example a bad choice of statistical models or data scarcity the LRs may

not be well-calibrated [6]. Also many existing machine learning models and algorithms are not optimized for obtaining accurate probabilities which is why their predictions may be miscalibrated [7]. In the former study performed at NFI the selected methods' output involved high LR values that cannot be proved to be true due to data scarcity. Therefore, the correctness of these values is questioned. Moreover, the LR is the ratio of the probability densities of the evidence given H_1 and H_2. Often there are little data points in the tails areas of the densities, so where data is expected to occur, it does not or only weakly. Therefore, the exact LR value cannot be calculated and is unknown. A validation step will be carried out to ensure that the probabilistic models in this thesis will calculate correct LR values in which 1) calibration will be measured and 2) a post-hoc calibration step will be performed.

A drawback of implementing a calibration process is that part of all the available will be used, leaving less data available for training. Besides that, one must pick a suitable calibration technique and possibly determine optimal parameter values, increasing the risk of making an error. The output for the methods for which no post-hoc calibration step is incorporated, is directly interpreted as LR value, thereby assuming that the method will produce well-calibrated LRs. However this is only true when the assumptions made by the model are met [8] which often is not the case. In this thesis the LRs resulting from so-called 'complex' methods will be transformed in a post-hoc calibration step whereas, LRs derived from 'simple' methods will not. This way the accuracy of the LRs from both methods can be compared and thereby the preferred post-hoc treatment can be determined.

1.1.3 Relative frequency of cell types in the data

The classifiers that have been experimented with require a prior probability for each cell type in order to perform its calculations and specify it by the relative frequency of the different cell types in the train data [1]. This is valid when the frequency with which the cell types occur in the data reflects the true relative occurrence. Unfortunately, it is highly unlikely that this is the case and besides that the true prior probabilities are unknown. Often a flat prior is used, i.e. assuming that the prior probability for each cell type is the same [9]. It however remains difficult to define a prior for the cell types in the data and besides that, it is not in the province of the forensic examiner to decide about its value. It is therefore,desirable for the likelihood ratio to always equal the posterior probability, irrespective of the prior value. If this is the case, than it is not necessary to specify priors, because no matter the value, the results are the same. In other words, irrespective of the relative occurrence of the different cell types in the data, the output of a given classifier is the same. In this thesis it is examined if a change in the relative frequency of the different cell types (i.e. the prior probabilities) in the train data affects the output of a given classifier.

1.2 Previous work

The first research to experiment with probabilistic models is by Zoete et al. They experimented with a naïve Bayes (NB) method based on Bayesian networks and a method based on multinomial logistic regression (MLR) [1]. The performance of both methods has been tested on several criteria. The dataset contains 158 samples and the cell types are blood, menstrual blood, saliva, semen and skin. In total 19 markers were assessed and the signal values are expressed in rfu (relative fluorescence units). These were converted into binary values, creating an presence/absence dataset: all signals with a rfu exceeding the threshold of 150rfu were converted into one and when below this value into zero.

They compared the results from the two methods and the n/2 method. Performance on the probabilistic models was measured based on the accuracy (i.e. the percentage of correct classifications on the data). They showed that both the NB method and the MLR method outperform the n/2 method and that methods discriminate well between the cell types. In their paper, they propose probabilistic methods as an alternative to the existing categorical statement methods but stress that further research is required to examine which is the best probabilistic model [9].

In another study, Dorum et al.[9] also aimed to predict the origin of a crime stain using probabilistic methods, but differentiate between the study by Zoete et al. in three ways. First the mRNA is measured differently, namely using the NGS (Next Generation Sequencing) technology. Secondly, next to evaluating the methods on the presence/absence data, they also evaluated the methods on quantitative data in the form of read counts. At last they proposed a new probabilistic method, namely Partial Least Squares Discriminant Analaysis (PLS-DA) that is able to take the quantitative information

of read counts into account. They used a different dataset to evaluate the models' performance on. The data consists of 183 samples from six different body fluids, namely blood, semen, saliva, vaginal secretion, menstrual blood and skin. In total 33 markers were assessed.

They compared the results from the two probabilistic methods from Zoete et al. with the results from PLS-DA. PLS-DA had a lower prediction error and performed better when trained on the read counts data. They however argue that more data should be gathered to build an even more robust model. They also suggest that the model should be tested on several datasets to estimate its prediction performance on real casework examples.

They also predicted the body fluids in 26 two-component mixture samples using PLS-DA model that was trained on the read count data. In all of the samples at least one of the body fluids in the mixture received a high posterior probability. For 20 samples the correct body fluids obtained the highest probabilities, however the value of the second highest probability is very small. This is mainly due to the fact that the model is trained to predict only one label (i.e. multi-class classification).

In a recent study carried out at the Netherlands Forensic Institute [10], experiments that involve multi-label classification were performed. They compared the performance in terms of classification accuracy of the n/2 method and eight probabilistic methods, namely Naive Bayes (NB), Multinomial Logistic Regression (MLR), Partial Least Squares Discriminant Analysis (PLSDA), Decision tree (DT), Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and K-nearest Neighbors (KN). Their first objective was to determine which of those methods showed the most promising results regarding the classification of different body fluids on three datasets containing samples from single cell types. The first dataset is the one used by Zoete et al. The second dataset was the dataset introduced by Dorum et al. The third dataset was created by the Department of Human Biological Traces of the NFI and consists of 866 samples from blood, menstrual secretion, saliva, nasal mucosa, vaginal mucosa, sterile semen, fertile semen, skin and penile skin and a 'blank' category. In total 17 markers were assessed. They pre-processed the datasets either by transforming the data into presence/absence data or by normalizing the signal values per sample.

They conclude that MLP and MLR outperform the n/2 method and that they achieve a higher accuracy in comparison to the other probabilistic methods. The remaining experiments were done using these two methods on both the presence/absence and the normalized data.

The second objective was to determine a probabilistic method that correctly classifies mixture samples. They introduced techniques to enable classifiers to predict more than one label. The first is to train the methods on mixture samples, rather than single cell type data. As this data was not yet available, in total 20.000 empirical mixture samples (i.e. synthetic data) were generated. The second addition is including multi-label classification methods, namely the label power-set method and the binary relevance method, to allow predicting multiple labels. They used a multiple body fluid dataset, which was also constructed by the NFI, as validation dataset. It consists of 351 samples of 7 mixtures (of two components) and is made up of the same body fluids and markers as the single cell type dataset created by the NFI. The combinations of the two methods and two multi-label methods were trained on the synthetic dataset and performance was measured on the multiple body fluid dataset. The MLP in combination with the label power-set method on the presence/absence data resulted in the highest classification accuracy. Moreover, the resulting LR values on mixture RNA data were more accurate than those of Dorum et al.

One of the shortcomings of this study is that the 'final' method is selected based on its performance on the mixture cell type dataset, even though the samples in this data are a poor representation of forensically realistic samples. So, the validity is measured in an unreliable setting [3]. An improvement would be to determine a probabilistic method based on data representing actual case data.

1.3 Outline

This thesis carries out experiments with four probabilistic classifiers in combination with multi-label methods and are trained and validated on synthetic mixture samples that are considered representative of actual mixtures. Two of the four classifiers are chosen because they showed the most promising results in the study by Scholten, namely MLP and MLR, the remaining two are Extreme Gradient Boosting and a Fully connected Feed Forward model. The first objective

is to determine which probabilistic classifier outputs the most accurate and reliable likelihood ratios for the two most relevant classes, namely vaginal mucosa and saliva. This is when the likelihood ratios are both well-calibrated and highly discriminating. Another objective is to compare the performance of classifiers from which the output is transformed in a post-hoc calibration step and a classifier from which the output is directly interpreted as likelihood ratio, thereby examining which of the two is preferred. The third objective is to determine whether, and if so to what extent, a change in the relative frequency of cell types in the train data alters the output of a given classifier. The main goal is to determine a probabilistic method that can replace the n/2 method in identifying cell types in a crime stain. Another objective is to demonstrate that the mixtures cell type dataset should not be used as validation set. The last objective is to implement the selected method into a user-friendly environment (i.e. tool built in Microsoft Excel), enabling the the Department of Human Biological Traces of the NFI to use it. The practical use of the tool is shown by adapting it to mRNA measurements from old cases and comparing the LRs to the statement that was reported based on the n/2 method.

This outline of the thesis is as follows. In section 2 the materials and methods are described. Section 2.1 goes into the datasets used for the experiments. Section 2.2 describes the way that the datasets are processed. In section 2.3 the methods implemented to perform multi-label classification, the probabilistic classifiers and the calibration technique are elaborated on. Section 2.4 defines the performance measures and illustrates the experimental setup. Section 3 shows and discusses the results. In section 4 the user-friendly tool is explained and adapted to mRNA measurements from two actual cases. This thesis ends with a discussion about the experimental results and suggestions for future work.

2 Materials and Method

2.1 Data

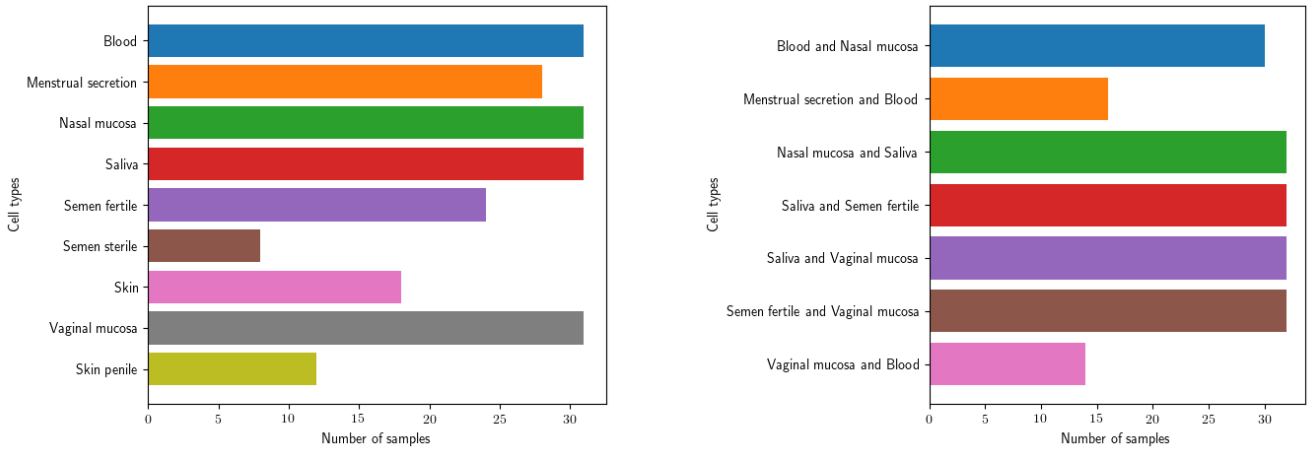
The same data that has been used by Scholten has been used in this thesis.

2.1.1 Single cell type dataset

The dataset was created by the Department of Human Biological Traces of the NFI. In total there are 866 measurements from samples including blood, menstrual secretion, saliva, nasal mucosa, vaginal mucosa, sterile semen, fertile semen, skin and penile skin. The expression levels for markers may vary a lot each time a stain is measured [1, 2] which is why it has been measured approximately four times. To get one sample for a cell type rather than four separate measurements, the average of the measurements has been calculated. Figure 2a displays the distribution of the samples per cell type. The number of samples from blood, menstrual secretion, nasal mucosa, saliva en vaginal mucosa is approximately the same, namely 30. There are less samples in the dataset of the remaining cell types. The number of samples in the dataset is 214.

2.1.2 Mixtures cell type dataset

The mixtures dataset consists of 351 measurements. Each sample has been measured approximately two times and again the average of these repeated measurements is calculated, resulting in the actual mixture samples. The mixture cell types are made up of the same cell types and markers as the single cell type dataset and there are seven different two-component mixture classes. The distribution of the samples over these classes is displayed in figure 2b. There are two times as much samples from the combination classes menstrual secretion + blood and vaginal mucosa + blood as there are samples for the other mixtures. The number of mixture sampes in the dataset is 188.



(a) Single cell type dataset.

(b) Mixture cell type dataset.

Figure 2: Distribution of samples in the two datasets created by the Department of Human biological Traces.

Since this dataset contains real mixtures of cell types it is reasonable to assume that this is an accurate validation set. Unfortunately, the samples are an inaccurate representation of forensically realistic mixture samples consisting of the same two components. Forensically realistic samples are degraded, meaning that markers attenuate regularly or contain noise. This is rarely the case in the mixtures dataset rather, for most of the samples the correct markers amplify and it barely consists of noise. Therefore, the dataset is assumed to be 'too clean' and contains easy to classify samples. Using this to evaluate the performance of a given classifier on, it will show to perform better than it actually would on forensically realistic samples. Therefore, the mixtures dataset will only be used to show that a probabilistic method will indeed perform better than on a dataset with more realistic samples.

2.1.3 Markers

For both datasets levels of expression have been measured for 15 cell type specific markers. Additionally, two housekeeping markers have been measured and they determine whether the mRNA profile is informative [2]. A sample is considered uninformative when for at least one of the housekeeping markers no signal is measured. The sample will then be disregarded. The total number of disregarded samples in the single cell type dataset is 16 and 0 in the mixture cell type dataset.

The proportion of the amplifications for the 15 relevant markers and all cell types in the single cell type dataset are shown in table 1. This proportion is calculated by the number of times the marker amplifies (i.e. the signal value is above or equal to 150) divided by the number of measurements. The column names belonging to the blue shaded cells correspond to the cell type specific markers, so the proportions therein are expected to be the largest. For menstrual secretion, there are markers for which the proportion of amplifications is higher (1, 0.496, 0.451, 0.566, 0.531) than for the menstrual specific markers (0.391 0.381, 0.558). This involves the markers for both blood and vaginal mucosa and can be explained by the fact that menstrual secretion actually a composition of several cell types under which blood and vaginal mucosa. This however does not hold the other way around: menstrual secretion does not necessarily appear in blood and/or in vaginal mucosa. The rate of amplifications for the MUC4 marker in nasal mucosa is also high (0.616) even though it is specific for vaginal mucosa. Because of this cross reaction, BPIFA1 has been added to distinguish between these cell types. The same holds for the blood specific marker CD93 that amplifies often in nasal mucosa. The remaining amplification rates in the table that are not cell type specific are considered noise. Note that this is one of the reasons why the samples in the single cell type dataset are considered to represent realistic samples. The marker STATH is specific for both nasal mucosa and saliva. Therefore, it may be hard for a given classifier to discriminate between these two classes. In this dataset, there are no markers that can identify skin or penile skin, which is why a classifier is expected to be unable to determine (penile) skin cells in a sample.

Table 1: The proportion of the amplifications for the 15 relevant markers and all cell types in the single cell type dataset using a threshold of 150.

		HBB	ALAS2	CD93	HTN3	STATH	BPIFA1	MUC4	MYOZ1	CYP2B7P1	MMP10	MMP7	MMP11	SEMG1	KLK3	PRM1
1	Blood	1	0.960	0.579	0	0	0	0	0	0	0	0	0.032	0	0	0
2	Menstrual secretion	1	0.496	0.451	0	0.009	0	0.566	0.531	0.31	0.319	0.381	0.558	0	0	0
3	Nasal mucosa	0.008	0	0.432	0.008	0.976	0.504	0.616	0.016	0.016	0	0.008	0.024	0.024	0	0
4	Saliva	0.159	0.009	0.028	0.907	0.907	0.019	0.009	0.019	0.009	0	0.009	0	0	0	0
5	Semen fertile	0.011	0.011	0	0	0	0.011	0.011	0	0	0	0	0	0.832	0.789	0.958
6	Semen sterile	0	0	0	0	0	0	0	0	0	0	0	0.031	0.875	0.656	0
7	Skin	0.264	0.014	0.111	0	0.083	0.028	0.194	0.056	0	0	0.0278	0	0	0	0
8	Vaginal mucosa	0.009	0	0.157	0	0	0	0.922	0.722	0.557	0	0.043	0.009	0	0	0
9	Skin penile	0.146	0	0.042	0	0	0	0.333	0.021	0	0.021	0.021	0.042	0	0	0.104

2.2 Data Pre-processing

2.2.1 Data split

The single cell type data is split in three parts that either are used for training, calibrating or testing. It is ensured that in each part at least one sample of each class label is present and there may not be overlapping samples in the three parts. If for example the same sample is in the train and test part, it is easier for a classifier to correctly classify this sample as it has already seen it during training, and consequently return inaccurate results. Therefore, the data is split before carrying out any of the subsequent pre-processing steps and experiments.

2.2.2 Synthetic data

The single cell type dataset consists of mRNA measurements for single cell types, which is why a given classifier is most likely to fail to predict multiple labels, if any, when only having seen this data. Therefore, a synthetic dataset with signal values for all combinations of cell types, i.e. empirical mixtures, has been created for each part after splitting the data.

	HBB	ALAS2	CD93	STATH	BPIFA1
Blood replicate 1	5715	7572	0	0	0
Blood replicate 2	6066	4547	0	0	0
Blood replicate 3	5967	5556	0	0	0
Blood replicate 4	6307	5555	0	0	0
Nasal mucosa replicate 1	0	0	0	3092	3976
Nasal mucosa replicate 2	0	0	0	2665	3824
Nasal mucosa replicate 3	0	0	0	2487	3123
Nasal mucosa replicate 4	0	0	0	3926	5565

↓ Max per replicate

	HBB	ALAS2	CD93	STATH	BPIFA1
Combination replicate 1	5715	7572	0	3092	3976
Combination replicate 2	6066	4547	0	2665	3824
Combination replicate 3	5967	5556	0	2487	3123
Combination replicate 4	6307	5555	0	3926	5565

↓ Calculate the mean

	HBB	ALAS2	CD93	STATH	BPIFA1
Empirical mixture sample	6014	5808	0	3043	4122

Figure 3: Creating one two-component empirical mixture sample.

Figure 3 illustrates how the mRNA measurements from two cell types, here blood and nasal mucosa, are combined into an empirical two-component mixture sample. Note that for simplicity only the cell type specific markers are used to illustrate the process. The continuous marker values are the expression levels. From the existing samples in the single cell type data one random sample from all the blood samples and another random sample from all nasal mucosa samples is drawn. Then, the replicates within both picked samples are shuffled by taking a random permutation of the number of replicates in the sample and are shown in the first block in the figure. In this example, both samples consist of four replicates. Starting from the first replicates from both body fluids, these are combined into the first replicate of the empirical mixture by selecting the maximum marker value. The maximum has the desirable property that the signal strength cannot be shrunken down below the threshold of 150 which as a result would decrease the signal strength. This way of combining the samples is called the or-relation and ensures that all markers that are amplified in both cell types are taken into account in the empirical mixture. The same has been done for the remaining replicates, resulting in four combined replicates. The average of these replicates results in one empirical mixture sample for blood and nasal mucosa. Note that the or-relation assumes independence between the amplification of markers. To be more concise, when replicates of different cell types are combined using the or-relation, it is assumed that no other markers than those who already are amplified, will amplify, nor will their signal strength intensify or attenuate. This is a justifiable assumption, because in reality it is unlikely that markers amplify or dissolve when combined.

The complete synthetic dataset consists of empirical samples for all the combinations of cell types (i.e. 2^K where K is the number of cell types, here 8) and are created as has been illustrated in figure 3. For each empirical mixture the same number of synthetic samples is created. Because there is randomness involved in the creation of the synthetic dataset, it will differ each time it is created even when the same samples from the single cell type dataset are used.

As has been briefly discussed, the samples in the single cell type dataset are considered to be an accurate representation of forensically realistic samples. Since the synthetic dataset is generated using these samples, the empirical mixtures therein are also considered to be representative of actual mixture samples and hence have been used to evaluate the probabilistic models on.

2.2.3 Transforming data

In former research conflicting statements regarding the type of data that results in the highest accuracy are made. Therefore, within this thesis both presence/absence data and quantitative data have been experimented with. The mixtures dataset and three synthetic datasets for training, calibrating and testing all are transformed the same way. In order to create the presence/absence dataset, the signals from each measurement are transformed into one when its value exceeds the threshold of 150 and zero otherwise. Thereafter the measurements belonging to the same sample are merged by calculating the average, which is why values between zero and one can also occur. The second way of transforming the data is by normalizing the signal values. The goal of normalization is to change the values in the dataset to a common scale, without distorting differences in ranges of values. All values have been divided by 1000, keeping a continuous scale.

In order to determine the preferred data transformation, experiments have been performed using both. Note that in order to make an honest comparison, the synthetic dataset should be generated first and transformed thereafter. However, within this thesis two synthetic datasets have been generated separately, but using the same original samples, one is converted into a presence/absence dataset and the other into a quantitative dataset. As has been explained earlier, due to randomness the synthetic dataset will differ when it is created again which is why the transformed datasets will not contain of the exact same randomly selected samples. Nonetheless, the performance on both still has been compared since they are constructed using the same original samples and thus not very susceptible to sampling variability.

2.3 Method

2.3.1 Multi-label classification

Multi-label methods to predict multiple class labels in a sample have been introduced already in the study performed by Scholten and have been implemented here rather than multi-class methods. One advantage of the multi-label approach over multi-class is that it enables a probabilistic classifier to predict more than one label simultaneously and is therefore better suited to classify mixtures. The multi-label methods that have been implemented are two problem transformation techniques, namely the label power-set method and the binary relevance method, that both transform the multi-label learning problem into one or more single-label classification problems [11].

Include prior knowledge

In some cases the forensic scientist is certain that the crime stain does or does not contain a body fluid prior to making any predictions. For example, when the sample comes from a men's genitals it is known that penile skin will be in it. In such cases there is no debate about the value of the prior probability and should be used as this will result in a more accurate LR value. Note that only in these exceptional cases the prior may be set. By doing this, one can simulate a world in which a cell type always or never exists, thereby providing a given classifier with information about the cell type. When the cell type certainly is in the crime stain the prior probability for that cell type is set to one. Consequently, the label for this cell type will always be in the list of labels that represents the class label of an empirical mixture. A prior probability of zero is set when the cell type definitely does not exist and is excluded from the list labels. Only the prior probability for penile skin is set to zero and thus excluded from all the experiments. Penile skin is thus excluded from the list of labels that the probabilistic classifier can predict with multi-label classification.

Dependence between class labels

Figure 4 shows the correlation plot for the 15 markers based on the original signal values. Blue denotes a positive correlation and red denotes a negative correlation. The brighter the color, the stronger the correlation. The markers, especially those who are cell type specific for the same cell type, are positively correlated. There also is a positive correlation between both the markers for menstrual secretion (MMP10, MMP7 and MMP11) and the markers for blood (HBB, ALAS2 and CD93) and the markers for vaginal mucosa (MUC4, MYOZ1 and CYP2B7P1). The correlation plot implies that cell types depend on one another and this should be accounted for by using a multi-label approach that is able to model these dependencies.

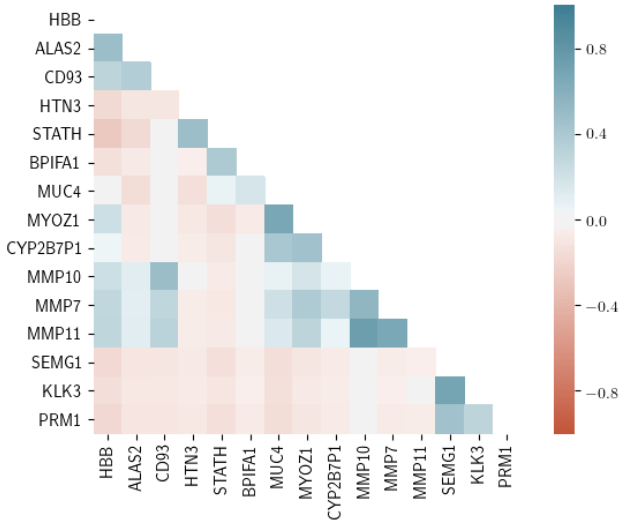


Figure 4: Correlation between the 15 relevant markers based on the original signal values.

Label powet-set method

The label power-set method transforms the multi-label problem to a multi-class problem by regarding each combination of cell types as a class. This way it directly takes into account the class label correlations [11]. Since there are $K = 8$ cell types, the number of unique combinations, called empirical labels, is 256 (2^K). Note that the majority of the empirical labels map to an empirical mixture consisting of at least two cell types, yet there are 8 labels mapping to a single cell type and one label mapping to a sample in which no measurements are present. Any given classifier will predict one of the 256 labels, namely the one for which the predicted probability is highest, when classifying a new unseen sample. This is called the majority rule. Probabilities are calculated using the softmax function:

$$Pr(c_{k^*}|x) = \frac{\mathrm{e}^{x_{k^*}}}{\sum_{j=0}^{2^K} \mathrm{e}^{x_j}} \quad \text{for } k^* = 1, 2, 3, ..., 2^K \tag{3}$$

Where c_{k^*} is the k^*-th empirical class label and x_{k^*} is any real number. The softmax function calculates the probability of each empirical class label over all possible empirical class labels and the sum of all probabilities is one. The probabilities follow a multinomial distribution.

An illustration of the process of predicting multiple labels using the label power-set method with eight samples and three cell types, here blood, nasal mucosa and saliva, is displayed in figure 5. The empirical class label of a sample is represented as hot encoded vector of length three. For example, the vector for blood and saliva is {1, 0, 1}. The number of empirical mixtures here is $2^3 = 8$. Starting off a classifier is trained on $\mathbf{X}$ and thereafter predicts the probabilities $\Pr(y^{(i)} = c_{k^*}|\mathrm{x}^{(i)})$ for i = 1,2,...,8 for $k^* = 1,2,...,8$. These probabilities are displayed in the 8x8 table in the figure. The blue shaded cells are the probabilities with the highest value and the class labels belonging to them become the class predictions, as is shown in the final table in the figure.

The main drawback of this method is that the number of empirical labels tends to become very large [11] increasing the probability of making a type I error.

Calculate Likelihood Ratio

The LR calculation differs from the way that it is calculated in equation 4 when using the label power-set method. First of all, to calculate the probability for one cell type, the 2^K probabilities are combined to become one probability. This can be done by calculating the marginal probability: the summation of all 2^K probabilities in which the cell type occurs. When interested in the probability of a mixture of cell types, one can repeat this process, but now by summing all the 2^K probabilities in which either one or both of the cell types occur. Next up the LR is calculated using the marginal probability and the following equation:

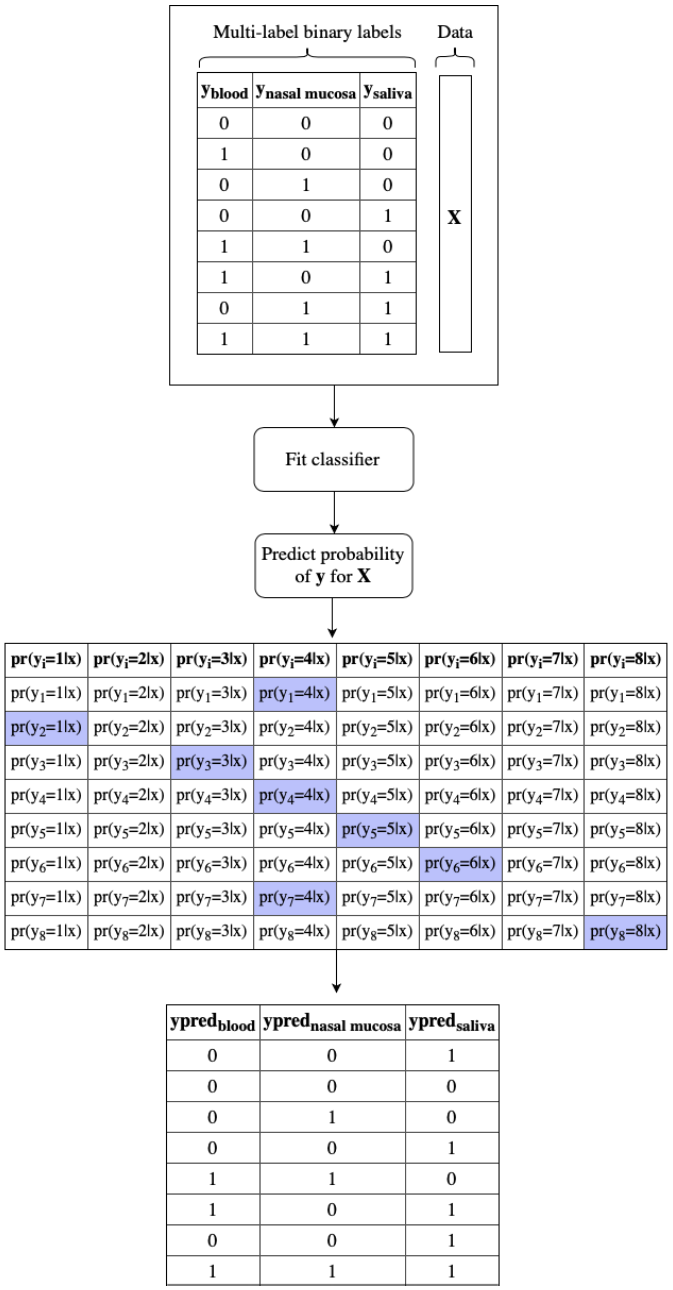
$$\text{LR} = \frac{P((\text{mixture of) cell type(s)})}{1 - P((\text{mixture of) cell type(s)})} \tag{4}$$

Independence between class labels

There also are multi-label approaches in which independence between class labels is assumed and have been implemented as well. The main advantage is that these methods are simple and, contrary to the label power-set method, less time consuming.

Binary relevance method

The binary relevance method is also known as the one-vs-rest strategy and converts the multi-label problem into several binary classification problems [11] and therefore cannot account for correlations. In each binary classification problem the class labels of the dataset are assigned in the following way: all samples from the positive class get label one and all samples from the negative class get label zero. After a given binary classifier is fitted to the data it will predict the probability of it being the positive and the negative class using equation 3 for $K = 1$. By combining the predicted class labels (this is when the predicted probability for the positive class is above 0.5) from all binary classifiers, the set of class labels can be constructed.



y_{blood}	$y_{nasal\ mucosa}$	y_{saliva}
0	0	0
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	1	1

X

$pr(y_i=1\|x)$	$pr(y_i=2\|x)$	$pr(y_i=3\|x)$	$pr(y_i=4\|x)$	$pr(y_i=5\|x)$	$pr(y_i=6\|x)$	$pr(y_i=7\|x)$	$pr(y_i=8\|x)$
$pr(y_1=1\|x)$	$pr(y_1=2\|x)$	$pr(y_1=3\|x)$	$pr(y_1=4\|x)$	$pr(y_1=5\|x)$	$pr(y_1=6\|x)$	$pr(y_1=7\|x)$	$pr(y_1=8\|x)$
$pr(y_2=1\|x)$	$pr(y_2=2\|x)$	$pr(y_2=3\|x)$	$pr(y_2=4\|x)$	$pr(y_2=5\|x)$	$pr(y_2=6\|x)$	$pr(y_2=7\|x)$	$pr(y_2=8\|x)$
$pr(y_3=1\|x)$	$pr(y_3=2\|x)$	$pr(y_3=3\|x)$	$pr(y_3=4\|x)$	$pr(y_3=5\|x)$	$pr(y_3=6\|x)$	$pr(y_3=7\|x)$	$pr(y_3=8\|x)$
$pr(y_4=1\|x)$	$pr(y_4=2\|x)$	$pr(y_4=3\|x)$	$pr(y_4=4\|x)$	$pr(y_4=5\|x)$	$pr(y_4=6\|x)$	$pr(y_4=7\|x)$	$pr(y_4=8\|x)$
$pr(y_5=1\|x)$	$pr(y_5=2\|x)$	$pr(y_5=3\|x)$	$pr(y_5=4\|x)$	$pr(y_5=5\|x)$	$pr(y_5=6\|x)$	$pr(y_5=7\|x)$	$pr(y_5=8\|x)$
$pr(y_6=1\|x)$	$pr(y_6=2\|x)$	$pr(y_6=3\|x)$	$pr(y_6=4\|x)$	$pr(y_6=5\|x)$	$pr(y_6=6\|x)$	$pr(y_6=7\|x)$	$pr(y_6=8\|x)$
$pr(y_7=1\|x)$	$pr(y_7=2\|x)$	$pr(y_7=3\|x)$	$pr(y_7=4\|x)$	$pr(y_7=5\|x)$	$pr(y_7=6\|x)$	$pr(y_7=7\|x)$	$pr(y_7=8\|x)$
$pr(y_8=1\|x)$	$pr(y_8=2\|x)$	$pr(y_8=3\|x)$	$pr(y_8=4\|x)$	$pr(y_8=5\|x)$	$pr(y_8=6\|x)$	$pr(y_8=7\|x)$	$pr(y_8=8\|x)$

$ypred_{blood}$	$ypred_{nasal\ mucosa}$	$ypred_{saliva}$
0	0	1
0	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	0	1
1	1	1

Figure 5: Predicting multiple labels using the label power-set method.

An illustration of this process using eight samples and three cell types is shown in figure 6. First the positive and negative class labels for *blood* are determined and are in the red shaded column. A binary classifier is fitted on **X** and thereafter used to predict probability of blood is for $x^{(i)}$ for i = 1,2,...,8. The blue shaded cells contain the highest probabilities and when $\Pr(y^{(i)} = 1|x^{(i)})$ is highest, the sample is considered to contain blood. If $\Pr(y^{(i)} = 0|x^{(i)})$ is the higher of the two, it will not get the label blood. All predicted labels for blood are shown in the first column of the final table. This same process has been carried out separately for nasal mucosa and saliva and finally the predicted labels for

the three classes are collected in the last table. For example, the first sample (on the first row) is predicted to be *saliva* and the fifth sample is predicted to be a mixture of *blood* and *nasal mucosa*.

An advantage of this approach is its interpretability. Since each class is represented by one classifier only, it is possible to gain knowledge about the class.

Sigmoid activation function

Another way of predicting probabilities of (multiple) class labels without modelling dependencies is using the sigmoid activation function. This is especially useful, and in this thesis only made feasible, for artificial neural networks. The sigmoid function shown in equation 5 maps the values from the output layer in the range (0, 1) and hence will return a probability rather than any real number.

$$P(c_k|x) = \frac{1}{1 + \exp(-x_k)} \quad \text{for } k = 1, 2, ..., K \qquad (5)$$

Where c_k is the k-th class label and x_k is a real number. The sigmoid function calculates the probability of each class separately from all the other classes and the sum of all probabilities does not necessarily have to be equal to one. The probabilities follow a bernoulli distribution.

Calculate Likelihood Ratio

Using the binary relevance method, the predicted probability for a cell type resulting from the binary classifier may be directly used to calculate the LR with using equation 4. To predict the probability of a mixture of cell types, the correct class labels have to be assigned to the samples first: all samples containing either one or both cell types get label one and the remaining samples get label zero. This way, after fitting a binary classifier, it will be able to predict the probability of the mixture class (i.e. the positive class). The LR again can be calculated using equation 4.

When the probabilistic classifier is an artificial neural network, the number of cell types equal the number of nodes in the output layer. When predicting the probability for one cell type, the number of nodes is one. The sigmoid activation function ensures that the output is a probability that can be used to calculate the LR with using equation 4. To predict a mixture of cell types, the network is trained on the train dataset from which the class labels are set to one if either one or both cell types are in the sample and zero otherwise. The value resulting from the one output node in the final layer is transformed using the sigmoid activation function and represents the predicted probability for the mixture class. The LR can be calculated as has discussed before.

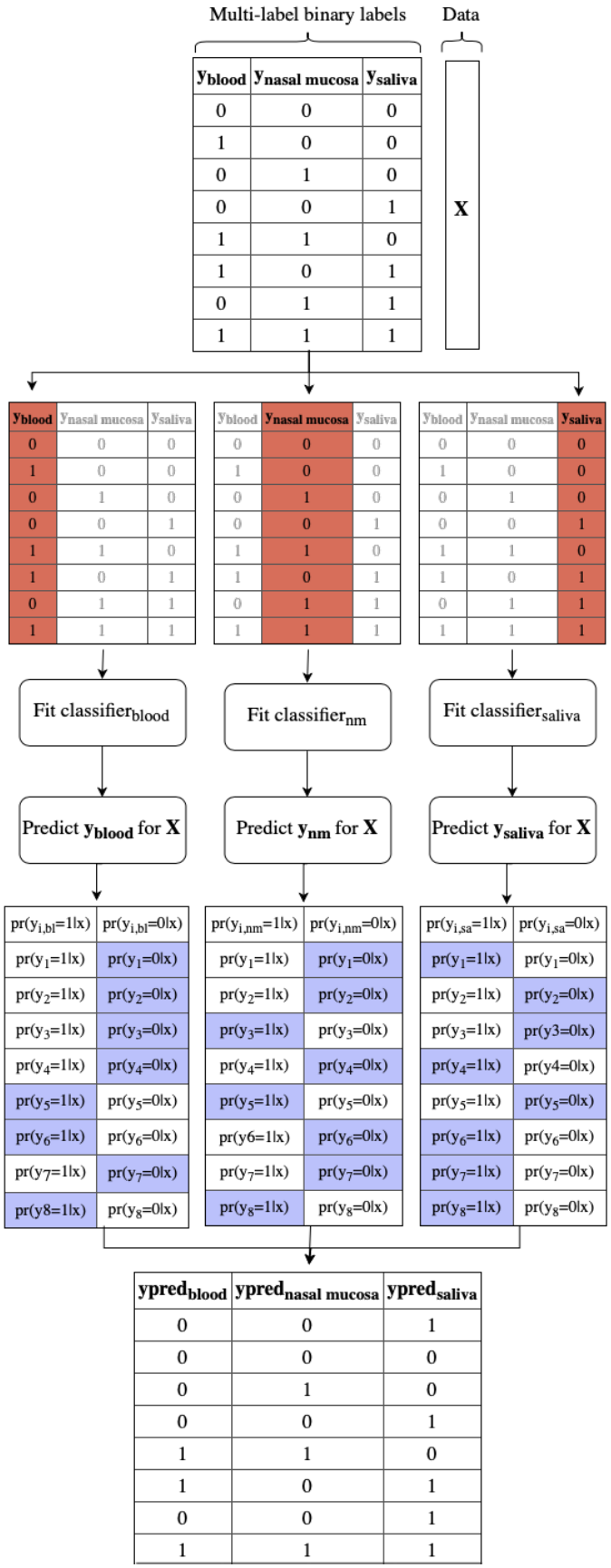


Figure 6: Predicting multiple labels using the binary relevance method.

2.3.2 Probabilistic classifiers

There are four different probabilistic classifiers that have been experimented with. These are (Multinomial) Logistic Regression (MLR), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB) and a Fully connected Feed Forward model (FFF). The first two classifiers are the same as the classifiers implemented and tested by Scholten, and are selected out of the eight methods that they experimented with because they showed the most promising results. XGB, FFF and MLP are known to be well performing classifiers in terms of classification accuracy [12]. They however also are known not to be optimized for obtaining accurate probabilities and produce a set of ill-calibrated LRs [13]. Therefore, the methods will benefit from a post-hoc calibration step that will transform the set into a well-calibrated one. An advantage of MLR is that it is a white-box model and the model coefficients are thus interpretable. MLR is not very flexible (when only the original covariates are inlcuded), meaning that it cannot vary as much with the train data, but therefore is less susceptible to overfitting [12]. The MLP and FFF classifiers are black-box models meaning that they do not allow for an interpretation of their model parameters. It is expected that their discriminating power is better than that of white-box models [12] and they are more susceptible to overfitting because they or more flexible. The classifiers have been implemented in Python using either the Scikit-learn module, the XGBoost library or the Keras API.

Each probabilistic classifier has been carried out in combination with both a multi-label approach accounting for dependence between class labels and another multi-label approach that ignores dependencies between class labels.

(Multinomial) Logistic Regression

Logistic regression is a statistical model that models a binary dependent variable. It transforms a linear equation using the logit function (equation 6) into a set of probabilities that sum to one. The optimal parameter values, or rather the coefficients and intercept, are obtained using maximum likelihood estimation [1, 12].

$$\text{logit}(Pr) = \log\left(\frac{Pr}{1-Pr}\right) = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + ... + \beta_p x_p^{(i)} \tag{6}$$

Under the assumption of dependence between the class labels the label power-set method in combination with a generalization of logistic regression to multiple classes, namely multinomial logistic regression, have been implemented. The L2 penalty has been added to prevent the model from overfitting. Once the parameters ($\boldsymbol{\beta}_{k^*}$ for k^* in 1,2,...,2^K) are obtained using maximum likelihood estimation, these can be used to predict 2^K probabilities (that add up to one) for an unseen sample $x^{(i)}$ as in equations 7.

$$\begin{aligned}
Pr(y^{(i)} = 1) &= \frac{\mathrm{e}^{\boldsymbol{\beta}_1 x^{(i)}}}{1 + \sum_{k=1}^{2^K-1} \mathrm{e}^{\boldsymbol{\beta}_k x^{(i)}}} \\
Pr(y^{(i)} = 2) &= \frac{\mathrm{e}^{\boldsymbol{\beta}_2 x^{(i)}}}{1 + \sum_{k=1}^{2^K-1} \mathrm{e}^{\boldsymbol{\beta}_k x^{(i)}}} \\
&\vdots \\
Pr(y^{(i)} = 2^K - 1) &= \frac{\mathrm{e}^{\boldsymbol{\beta}_{2^K-1} x^{(i)}}}{1 + \sum_{k=1}^{2^K-1} \mathrm{e}^{\boldsymbol{\beta}_k x^{(i)}}} \\
Pr(y^{(i)} = 2^K) &= \frac{1}{1 + \sum_{k=1}^{2^K-1} \mathrm{e}^{\boldsymbol{\beta}_k x^{(i)}}}
\end{aligned} \tag{7}$$

Under the assumption of independence between the class labels the binary relevance method has been used. After training a logistic regression model on the dataset, the parameter values ($\boldsymbol{\beta}$) are obtained. Again, the L2 penalty has been added to prevent the model from overfitting. A binary logistic regression model predicts the probability of the positive class ($Pr(y^{(i)} = 1)$) for an unseen sample $x^{(i)}$ as in equation 8 and the probability for the negative class can be calculated by $1 - Pr(y^{(i)} = 1)$.

$$Pr(y^{(i)} = 1) = \frac{1}{1 + \mathrm{e}^{-(\beta_{t0} + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + ... + \beta_p x_p^{(i)})}} \tag{8}$$

Multilayer Perceptron

A Multilayer Perceptron (MLP) is an artificial neural network that is built up of three layers: an input layer, a hidden layer and an output layer. The input layer contains nodes that combine the input values of the 15 input features (i.e. markers) with a set of weights and biases by taking the product. The sum of the products are then passed through a node's activation function and sends this to the next layer. So, each layer's output is the subsequent layers input. The hidden layer consists of 100 nodes encoding the values from the former layer. The number of nodes in the output layer equal the number of class labels and the produced output variables are the predicted probabilities for the class labels. The model is trained, or in other words the parameters are optimized, by minimizing a loss function [12]. The binary cross-entropy loss in equation 9 is an appropriate loss function for binary classification problems and will also be used for multi-class classification in combination with the softmax function [14].

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} \cdot \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \cdot \log(1 - \hat{y}^{(i)}) \tag{9}$$

The training process can be described as follows: the MLP starts from random weights and biases that map the input to a set of predictions. Then the loss, or error, is calculated using equation 9. With backpropagation the weights and biases are updated: it backpropagates information about the error in reverse through the network so that it can alter the parameters. This is an iterative process and continues until the maximum number of iterations, here 500, is reached or the error is below the threshold value. The adam optimization function, that helps to minimize the loss function, has been used.

With the use of the label power-set method dependence between class labels is accounted for. Subsequently, the number of nodes in the final layer is 2^K and the softmax function is used to transform the values into a range from 0 to 1. The sum of the probabilities resulting from all those nodes is equal to one. Once the model is trained, an unseen sample $x^{(i)}$ can be 'fed' to the network that will use the parameters (weights and biases) and the information from the sample to predict 2^K probabilities for all empirical class labels.

Under the independence assumption the sigmoid function has been used as activation function in the final layer. The number of output nodes is equal to the number of classes (may also be mixtures of classes) of which one wants to predict a probability for. Note that the MLP without the hidden layer and using the sigmoid activation function in the output layer is exactly the same as logistic regression [12].

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an implementation of the gradient boosted trees algorithm. This is a way of ensemble learning that combines the estimate of a set of weaker trees (i.e. their predictions have a high bias and the predictive power is only somewhat better than random guessing) to make predictions. Trees are built sequentially such that each subsequent tree aims to reduce the error of the previous tree. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new trees. Note that the type of loss function depends on the problem. Here, the number of weaker trees, that when combined predict the class probability, is 100 and the maximum number of leafs in each tree is 3.

Under the assumption of dependence between class labels the label power-set method has been implemented. The loss function that is minimized during training is the softmax objective function. Once the model is trained, for each empirical class label k^* a sequence of 100 weaker trees is obtained, resulting in 2^K sets of 100 weaker trees. These sets can be used to predict the probabilities of the 2^K empirical class labels with.

Under the dependence assumption the binary relevance method has been implemented. The loss function that is minimized during training is the logistic objective function and is the same as equation 9. After training a XGBoost classifier on the dataset, a sequence of 100 weaker trees is obtained and are used to predict the probability of the positive class $Pr(y^{(i)} = 1)$ for sample $x^{(i)}$ and calculate the probability of the negative class by $1 - Pr(y^{(i)} = 1)$.

Fully connected Feed Forward model

The Fully connected Feed Forward model is also an artificial neural network and is equivalent to the Multilayer Perceptron, but differs in the number of layers in the network. Namely, it is built up of four layers: one input layer, two hidden layers and an output layer. On the first layer dropout is implemented, in which at a rate of 0.05 nodes are dropped during the training process. This is a way to prevent the model from overfitting. The input layer consists of 15 nodes, the first hidden layer consists of 20 nodes and the second hidden layer consists of 80 nodes. The number of nodes in the output layer equal the number of class labels and the produced output variables are the predicted probabilities for the class labels. Both the training process and the way that the input values, weights and biases are used to calculate the output of the network are the same as for the Multilayer Perceptron. Note that MLP only supports the binary cross-entropy loss function, whereas both the binary cross-entropy loss and the categorical cross-entropy loss function (equation 10) are supported by FFF. The latter can be used in multi-class classification problems where only one result can be correct, whereas binary cross-entropy can be used in multi-label problems.

$$L(y, \hat{y}) = -\sum_{i=j}^{M}\sum_{i=1}^{N} y^{(ij)} \cdot \log(\hat{y}^{(ij)}) \tag{10}$$

The FFF has been trained for 30 epochs. An epoch is one forward pass and one backward pass of all training examples. This is the same as one iteration if the batch size is equal to the size of the training set, which here is the case. The adam optimization function has been used.

Under the independence assumption the label power-set method has been implemented together with the categorical cross-entropy loss function and softmax activation function in the output layer. The number of nodes in the output layer is 2^K and the softmax function is used to transform the output of the network with into the range (0, 1) and thereafter referred to as probabilities. The sum of these probabilities sum up to one. Once the model is trained, the probability of 2^K empirical class labels can be predicted.

Under the dependence assumption the sigmoid function has been used as activation function in the final layer and the loss function during training is binary cross-entropy function. The number of nodes in the final layer is equal to the number of classes (may also be mixtures of classes) of which one wants to calculate the probability for.

2.3.3 Calibrating the scores

One of the aims of this thesis is to compare the performance of probabilistic classifiers, namely MLP, XGB and FFF, from which the derived LRs are transformed in a post-hoc calibration step with the performance of a probabilistic classifier, namely MLR, from which the resulting LRs are not transformed. Post-hoc model calibration is a way of transforming a set of ill-calibrated LRs into a presumably well-calibrated one [7]. Ill-calibrated LRs are referred to as scores (s) rather than LRs. After calibrating the scores they may be interpreted as LRs as they then are more realistic LR values. The synthetic calibration dataset has been used in the building process of the calibration model.

Calibration techniques are designed for two-class problems and when expanding it to multiple classes, the results are not likely to be accurate [15]. So, to enable calibrating the scores for multiple classes, a calibration model has been built for each class, thereby making it a set of two-class problems. The scores belonging to a class are transformed using the calibration model of that same class.

Kernel density estimation

The ideal calibration of the score would be to map it to the ratio of true distributions under H_1 and H_2 [13] evaluated in point s:

$$s \rightarrow \log \frac{P(s|H_1)}{P(s|H_2)} \tag{11}$$

where H_1 is the hypothesis stating "the sample contains cell type k" and H_2 states "the sample does not contain cell type k". However, the true score distributions $P(s|H_1)$ and $P(s|H_2)$ are unknown. One fortunately can estimate a

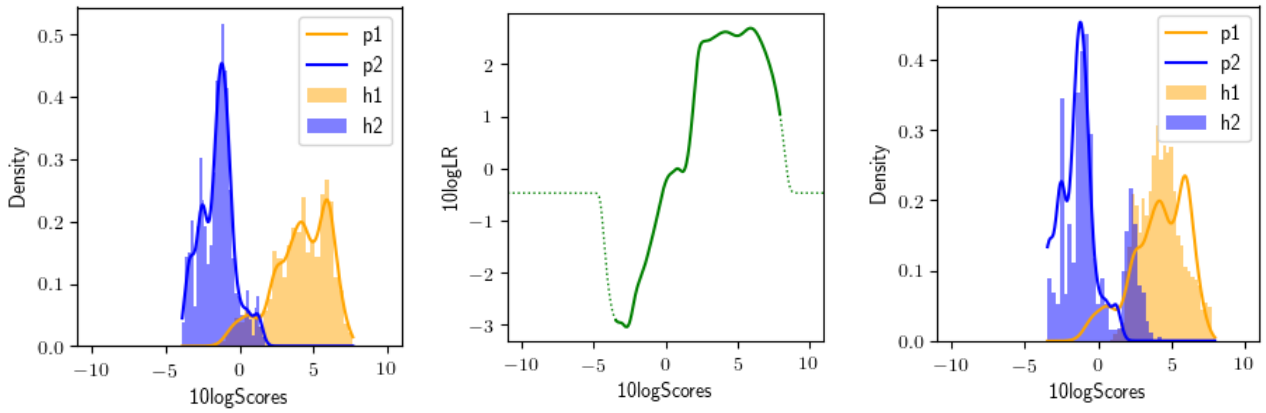
density model that fits the distribution for the scores under both hypotheses in case the ground truth labels are known. Moreover, the distributions of the scores may be predicted from the log10-transformed scores [13, 16] and are referred to as $\hat{P}(s|H_1)$ and $\hat{P}(s|H_2)$. These estimates have been obtained using kernel density estimation.

Kernel density estimation (KDE) is a non parametric way to estimate a probability density function $\rho(y)$ [17]. The density estimate at a point y within a group of points $\mathbf{x}$ is given by:

$$\rho_K(y) = \sum_{q=1}^{Q} K\left(\frac{(y - x^{(q)})}{h}\right) \tag{12}$$

where h is the bandwidth and K is the kernel. Mathematically, a kernel is a positive function $K(x;h)$ which is controlled by the bandwidth parameter h. The bandwidth acts as smoothing parameter, controlling the trade-off between the bias and variance. A large bandwidth leads to a very smooth (i.e. high bias) density distribution. A small bandwidth leads to an unsmooth (i.e. high variance) density distribution [17] that possibly overfits the calibration data. The process of estimating the probability density function using KDE is described as follows: on every data point $x^{(q)}$ in $\mathbf{x}$ it places the kernel function with bandwidth h. The average of all the placed probability masses $K(u;h)$ (equation 12) results in the final estimate [17].

In this thesis the gaussian kernel has been used and the bandwidth has been determined with the Silverman bandwidth. This is a way to estimate the optimal bandwidth parameter using Silverman's rule of thumb [17]. Figure 7 visualizes a part of the calibration model building process. In Figure 7a the two actual distributions of the log10(scores) under H_1 and H_2 and the estimated probability density estimates, p1 and p2, are displayed. The mapping function, being the ratio of the two density estimates, is shown in figure 7b. The log10(scores) are on the x-axis and the values to which the scores are mapped are on the y-axis, for example a 10log(score) of 5 is mapped to circa 2.5.



(a) Distributions of the scores derived from the calibration dataset and the probability density estimates plotted thereover.

(b) Mapping function.

(c) Distributions of the scores derived from the test dataset and the probability density estimates plotted thereover.

Figure 7: Visualization of part of the calibration model building process using kernel density estimation.

There are some drawbacks of building a calibration model using KDE. Since there are often little to no data points in the tail areas of both distributions of the scores, the density estimate is based on extrapolation and hence only weakly or not at all supported by the data [18]. The dotted line in figure 7b is the mapping function based on the extrapolation and that is for all log10(scores) below -4 and above 7.5. It should also be noted that the line is not monotonically increasing, so the order of the log10(scores) is not preserved. This could result in mapping a high log10(score), of say 8, to a lower log10(LR) than that of log10(score) below 8. If this is the case, the strength of the evidence of the transformed score is falsely shrunken too much. Another drawback is that KDE is susceptible to overfitting, meaning that it will fit the dataset rather than the underlying distribution. Due to sampling variability, the distributions of the scores from the calibration and test dataset are likely to differ. If the probability density functions perfectly fit the distributions of the

scores derived from the calibration set, but do not fit the distributions of the scores derived from the test set, the mapping function is overfitted. This is clearly shown in figure 7c where the density estimates are plotted over the distribution of the scores from the test set. As a result, the transformation from scores will not lead to a set of well-calibrated LRs. For example, a log10(score) of 2.5 remains roughly 2.5 after the transformation. Under H_2 this LR lends support to the wrong proposition and thus is not optimally calibrated.

2.3.4 Test the susceptibility to a change in the relative frequency of the cell types in the data

Another objective in this thesis is to determine if, and if so, to what extent the prior probabilities affect the output of a probabilistic classifier. These priors are implicitly incorporated as the relative frequency of the cell types in the train data. Therefore, a convenient way of assessing the affect of the priors is by altering these relative frequencies and conclude whether the output of that same classifier is changed. Initially a given classifier is trained and tested on a uniform dataset in which the number of samples for all cell types is equal (i.e. flat prior). In a non-uniform dataset the relative frequency of the cell types differ, so for example the relative frequency of cell type 1 may be 10 times that of the others. Uniform and non-uniform datasets thus have been used to assess the susceptibility to a change in the relative frequency of the cell types and an illustration showing how is displayed in figure 8. Starting with two train datasets, one being uniform and the other non-uniform, a given classifier is separately fitted on both. Both versions of the classifier are then used to calculate the LRs with from a uniform test dataset, resulting in two sets of LRs, namely $\text{LRs}_{\text{uniform}}$ and $\text{LRs}_{\text{non-uniform}}$.

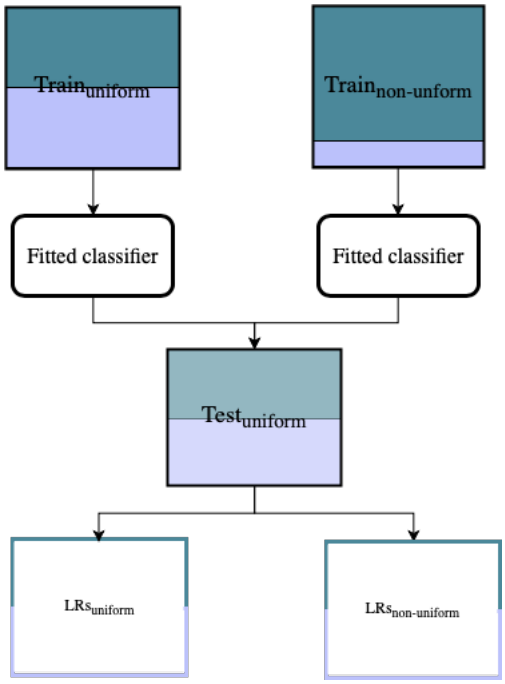


Figure 8: Illustration of the process to retrieve two sets of LRs from a given classifier that has been trained on both a uniform and non-unform train datasets.

This process has been carried out six times using a different non-uniform train dataset each time. The uniform train dataset is used as baseline and subsequently all six sets of $\text{LRs}_{\text{non-uniform}}$ are compared with $\text{LRs}_{\text{uniform}}$. In order to make an honest comparison it is necessary that the size of all train datasets is the same. Otherwise, the classifier that is trained on the larger of two datasets will presumably perform better as it could learn from more samples thereby having more information about the underlying distribution. The non-uniform datasets are constructed as follows: in three of them the relative frequency for either blood, vaginal mucosa or skin is 10 times that of the others and in the remaining three datasets the relative frequency for either of those cell types is 10 times less than that of the others.

2.4 Measure performance

2.4.1 Target classes

As has been mentioned in the introduction, there are two body fluids that forensic examiners have to identify most often in cases, namely vaginal mucosa and saliva, and are therefore mainly focused on. However, there also exist vaginal cells in menstrual secretion, so to consider all the information about vaginal mucosa, menstrual secretion should also be taken into account. Therefore in this thesis a combination class of vaginal mucosa + menstrual secretion is used in stead of vaginal mucosa. Henceforth, saliva and the combination will be referred to as target class saliva and target class vaginal/menstrual respectively. Performance of the probabilistic classifiers is measured based on both.

2.4.2 Performance metrics

The metrics with which the performance of the classifiers has been measured are based on likelihood ratios [3]. Desired properties are to have good discrimination between the classes and good calibration [19]. Disrimination represents the capability of a classifier to distinguish amongst different class labels [20]. The better the discriminating power, the less decision errors will be made. Note however that good discriminating power does not necessarily mean that the actual LR values are correct [4]. It is a desired property, but should not be primarily focused on. Good calibration is when the LR value is not too high or too low, so the better the calibration the more correct the LR values are. The accuracy measures both properties [20] and hence has been used as primary performance metric. This section also discusses the way to inspect the susceptibility of a given classifier to a change relative frequency of cell types in the data and a bootstrap procedure to assess the reliability of the LRs. Note that the performance metrics are intended to evaluate one class and have been calculated for both target classes respectively.

Accuracy

The log-likelihood-ratio cost (C_{llr}) is a measure of accuracy based on both the LRs and the classifier. Moreover, it is a gradient metric, meaning that it will take account the value of the LR [3]. Generally speaking, it is the cost of decisions based on the strictly proper scoring rules (spsr) that stems from the Bayesian framework. Strictly proper scoring rules may be seen as loss functions that assign a penalty to a given value of a predicted posterior probability depending on the true value of the ground-truth label [19]. Ideally, the spsr assigns a lower penalty to larger posterior probabilities when H_1 is true and to lower posterior probabilities under H_2. The logarithmic scoring rules have this property and their functions are shown in equation 13. These are the same as the optimization objective function for logistic regression [21].

$$\begin{aligned} &-\log_2(P) && \text{if } H_1 \text{ is true} \\ &-\log_2(1-P) && \text{if } H_2 \text{ is true} \end{aligned} \tag{13}$$

where P represents the posterior probability. Note however that the spsr framework cannot directly be applied to forensic science, because it is based on measuring the performance of posterior probabilities [4] and the forensic scientist only calculates the LR. Therefore, spsr has been extended to the forensic field by computing the cost for a wide range of prior probabilities [6]. A summarizing measure is at prior log-odds of 0, or at uniform priors, and is called the log-likelihood-ratio cost (C_{llr}) and can be calculated using equation 14 [3].

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_1}\sum_{i=1}^{N_1}\log_2\left(1+\frac{1}{LR_{1_i}}\right) + \frac{1}{N_2}\sum_{j=1}^{N_2}\log_2\left(1+LR_{2_j}\right)\right) \tag{14}$$

where N_1 and N_2 are the number of LRs under H_1 and H_2, and LR_1 and LR_2 are the likelihood ratios that are known to belong to either of both classes. Since the C_{llr} is a function of cost, a lower value is preferred and indicates a higher accuracy. Moreover, the extent to which the C_{llr} is less than one is a measure of validity of a classifier [3]. A C_{llr} of one or above one indicates that the classifier is badly calibrated and can better not be used. Moreover, this means that the classifier performs worse than a neutral system for which the LR values always equals 1 [19, 22].

Discrimination

Discrimination can be assessed by visualizing the LRs, or rather the log10(LRs), in a histogram. Moreover, it is useful to draw histograms separately for $\log10(LRs_1)$ and $\log10(LRs_2)$. The degree of overlap between these two histograms is a measure of the discriminating power. In case of complete separation between the histograms, meaning that there is no overlap, the LR values are said to be perfectly discriminated. This is when all $\log10(LRs_1) > 0$ and all $\log10(LRs_2) < 0$ in case the theshold is set to 0 [4].

Another way of inspecting the discrimination power is by assessing the Receiver Operation Characteristic (ROC) curve or rather the Area Under the ROC Curve (AUC) [19]. The plot showing the ROC curve has the false positive rate (fpr), that is the proportion of negative samples that are mistakenly considered as positive with respect to all positive samples, displayed on the x-axis. One the y-axis the true positive rate (tpr) is displayed. This is the proportion of the positive samples that are correctly considered positive with respect to all positive samples. The fpr and tpr vary together as the threshold that determines when a value is mistakenly missclassified or correctly classified varies. The ROC curve illustrates how they vary for a given classifier. Once the ROC curve is known, the area under this curve (AUC) can be determined and serves as numerical measure for discrimination. Generally speaking, the higher the AUC the better any classifier is at correctly classifying the samples and thus higher the discriminating power. When AUC equals 1 there is perfect discrimination. An AUC of 0 means that all the wrong labels are predicted. In case the AUC is 0.5, this means the model cannot distinguish between classes.

Calibration

A way of inspecting the calibration is based on the Pool Adjacent Violators (PAV) transformation [16]. This is an algorithm that transforms a set of LRs into well-calibrated LRs. By plotting the 10log(LRs) against the PAV-transformed log10(LRs) together with a diagonal line, the deviation from that line can be assessed. One can conclude by looking at the deviation if the original LRs are well- or ill-calibrated [22]: the smaller the deviation from the line is, the better calibrated the log10(LRs) are.

Susceptibility to a change in the relative frequency of cell types

To determine the susceptibility of a given probabilistic classifier to a change the relative frequency of cell types in the data, a visual aid is used. By plotting the $\log10(LRs_{\text{uniform}})$ against the $\log10(LRs_{\text{non-uniform}})$ together with a diagonal line in a scatterplot, the deviation from that line can be assessed. A larger deviation from the line indicates that there are more differences between the sets of LRs. From a large deviation, one can infer that a given classifier is susceptible to a change in the relative frequency of cell types in the data, whereas when the scatterpoints (almost) perfectly lie on the diagonal, the opposite conclusion may be drawn. Note that there has not been made use of a numerical measure that expresses to deviation from the diagonal. One must infer, by looking at the scatterplots whether the classifiers' output is affected by the relative frequency. Conclusions are drawn in a comparative manner, for example one may conclude that classifier 1 is less susceptible than classifier 2 when there is less deviation from the diagonal line.

Reliability of the LRs

The reliability of the LRs of the selected probabilistic classifier have also been studied. Reliability can also be understood as precision: the extent to which the output of a given classifier is consistent [3, 20]. A bootstrap procedure is performed, which is a resampling method that will enable to construct a confidence interval using bootstrap samples for the LRs. To preserve the same number of samples from both classes in the train data, bootstrap samples for the positive and negative class have been created separately. First the positive samples from the train dataset were resampled with replacement and thereafter the negative samples from the train dataset were resampled with replacement. The samples from both classes are combined into a bootstrap sample. In total 2000 bootstrap samples were generated, each of which the classifier was trained on separately and used to calculate the LRs from the test set with. The 95% confidence interval for each LR is constructed using the percentile method by taking the 25th and 975th largest of the 2000 replicates. With the

lowerbound and upperbound from an bootstrap interval defined as $vlowerbound^{(i)}$ and $vupperbound^{(i)}$ for sample i, one can state the following: with 95% confidence $LR^{(i)}$ is between $vlowerbound^{(i)}$ and $vupperbound^{(i)}$. The 'true' LRs and the confidence intervals are plotted to assess the overall reliability: the wider the intervals are, the less consistent the output of a given classifier is and thus the less reliable the LRs are.

2.4.3 Procedure and sensitivity analysis

Experiments have been carried out using two transformations of a given dataset, two multi-label approaches and four probabilistic classifiers. The main goal is to compare the results from all sixteen combinations, that from now on will be referred to as methods, simultaneously. The procedure in case a post-hoc calibration step is added is illustrated in figure 9.

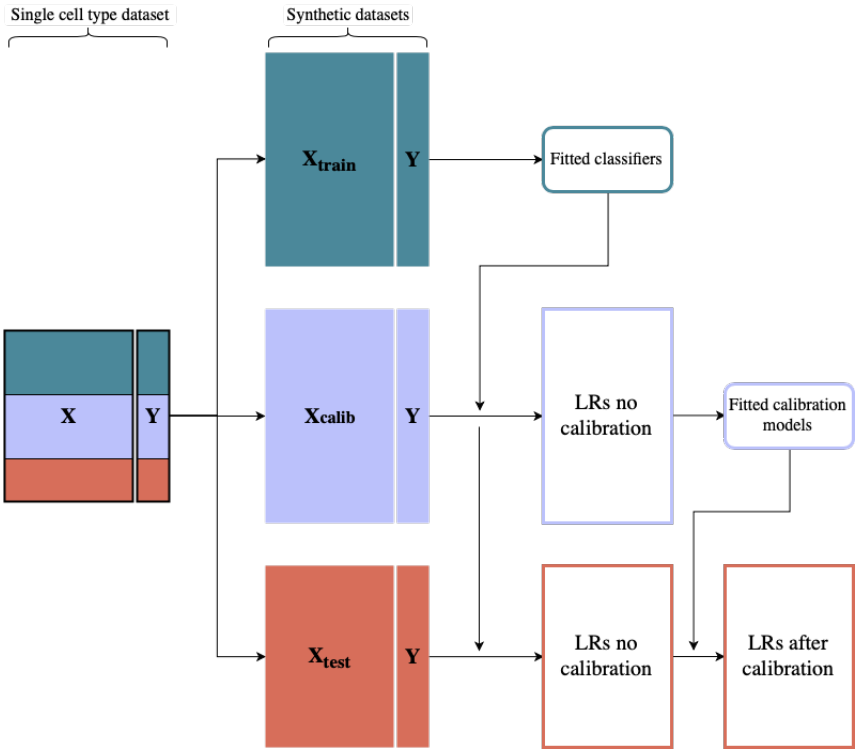


Figure 9: Illustration of the procedure.

Starting off, the single cell type dataset is split in three parts: 40% is used to create the synthetic train dataset, another 40% to create the synthetic calibration dataset and the last 20% is used for generating the synthetic test dataset. The synthetic sets are then converted into presence/absence datasets from which the process, that is about to be described, will follow. After that process ends, the synthetic datasets are generated again, but now are transformed into quantitative datasets from which the same process follows. This process is described in the following way: each of the four probabilistic classifiers in combination with one multi-label approach is fitted on the train data and thereupon used to calculate the $LRs_{\text{no calibration}}$ with from the calibration data. Those LR values are used to build the calibration models. The fitted probabilistic classifiers are also used to calculate the scores (and not yet LRs) with from the test data and are transformed into LRs using the calibration models. Note that in case no calibration has been implemented, as for MLR, the train and the calibration dataset together are used to train the classifier with and the calibration steps are skipped. Moving on this same process is repeated but using the four classifiers in combination with the other multi-label method. In the end there are sixteen sets of LRs derived from the test set and can be used to evaluate the performance of all sixteen methods with. The results of all methods derived from the mixtures cell type dataset are also calculated separately.

A sensitivity analysis has been performed in order to study the uncertainty in the output, and subsequently also in the calculated performance metrics, from the methods. This is a process of recalculating the output under alternative assumptions and in this thesis understood as repeating the procedure that has just been described multiple (30) times.

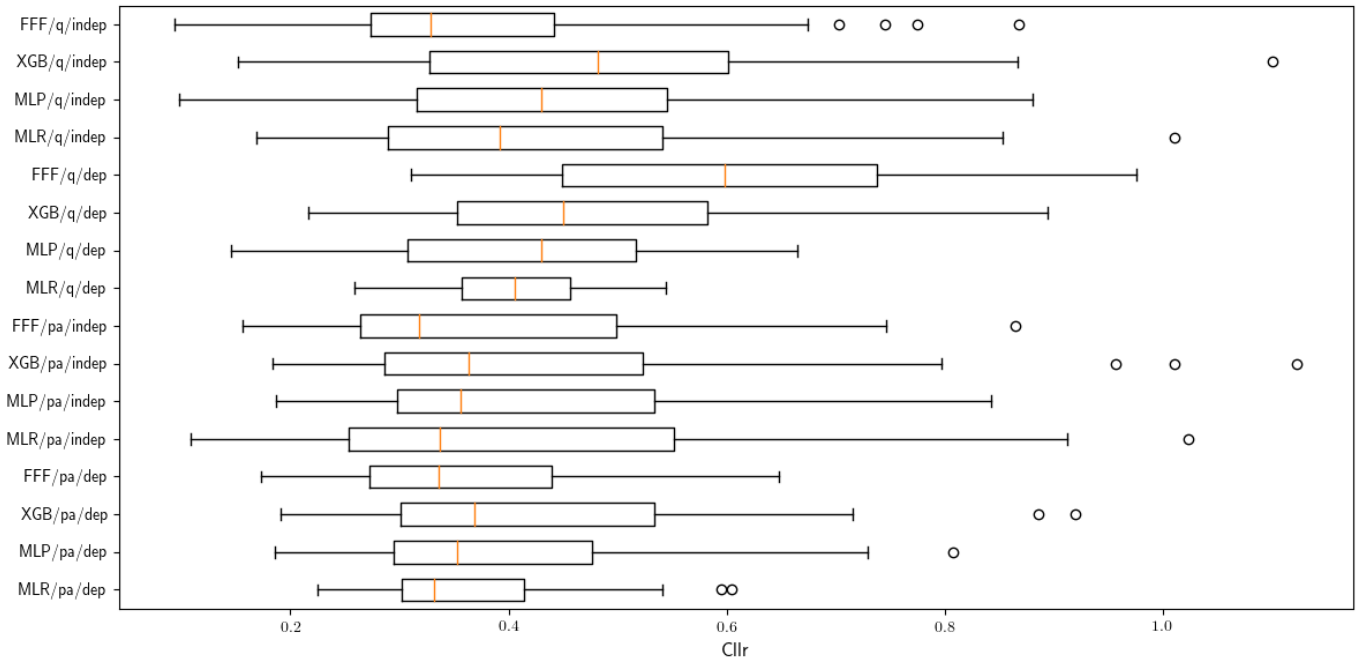
The output is expected to vary because of sampling variability. In each reiteration 8448 empirical multiple body fluid samples are generated for the train dataset, as well as for the calibration dataset. The number of generated samples for the test set is 5632.

3 Results

3.1 Synthetic test data

3.1.1 Comparison of the methods

The C_{llr}'s resulting from the 30 runs for target class vaginal/menstrual are displayed in figure 10. Each label on the y-axis represents one of the methods and consists of three elements: the name of the classifier, the type of data transformation and the assumption about the class labels (i.e. multi-label approach). For example, *MLP/pa/indep* refers to combination of the Multilayer Perceptron, the presence/absence data and the multi-label approach that ignores dependencies between class labels. The median is a measure of the center of the C_{llr}'s and in the figure is marked by the band inside the boxes of the boxplots. The width of the boxes and the length of the whiskers show the spread of the C_{llr}'s. Values outside the whiskers are outliers.



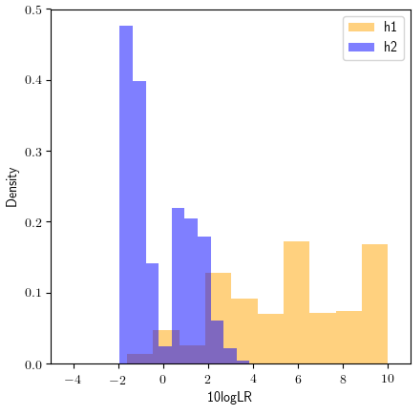
pa = presence/absence data, q = quantitative data,
indep = independence between class labels, dep = dependence between class labels

Figure 10: C_{llr}'s for target class vaginal/menstrual resulting from the 30 runs on the synthetic test data.

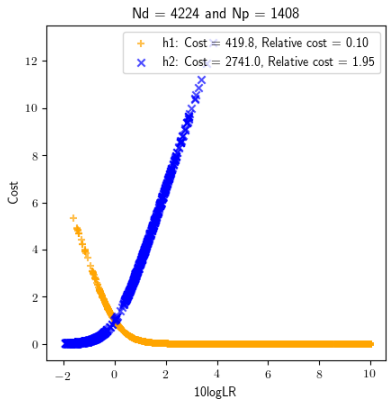
Figure 10 shows that the boxplots overlap, meaning that there is no significant difference between the accuracy of the methods. Moreover, there is no method that clearly outperforms the rest in terms of calculating accurate LR values. Most of the C_{llr}'s in the figure are below 1, meaning that the methods are more useful than a neutral system (always returning a LR of 1). It also shows that for some of the methods the median is significantly higher in comparison to that of other methods, meaning that their output is generally less accurate. This for example is the case for the method *FFF/q/dep* where the median of all the C_{llr}'s is roughly 0.6 and for the method *XGB/q/indep* where the median is roughly 0.5. Furthermore, the results show that using the presence/absence data generally leads to a higher accuracy in comparison to using the quantitative data.

There are some methods for which the C_{llr} in at least one run exceeds 1. To get a better understanding why this is the case for *MLR/pa/indep*, the way that the C_{llr} is calculated for in this particular run is displayed in figure 11. Moreover, in figure 11a the histograms of the LRs under H_1 and H_2 are plotted on a log-10 scale. It becomes clear that part of the log10(LRs) under H_2 point in the wrong direction (i.e. are above zero) as well as some log10(LRs) under H_1 (i.e. are below zero), and therefore will be penalized. Figure 11b shows the assigned penalty (or cost) to each value in the set.

The log10(LRs) under H_2 supporting the wrong propositions are most heavily penalized (i.e. penalty goes up to 12). The total cost under H_2 is roughly 6 times that of the total cost under H_1, so the reason why the C_{llr} in this run exceeds one is because many of the non-target class samples were misclassified. One could further explore why the method failed to classify them correctly by inspecting the samples, but that is outside the scope of this thesis.



(a) Separate histograms for log10(LRs) under H_1 and H_2.

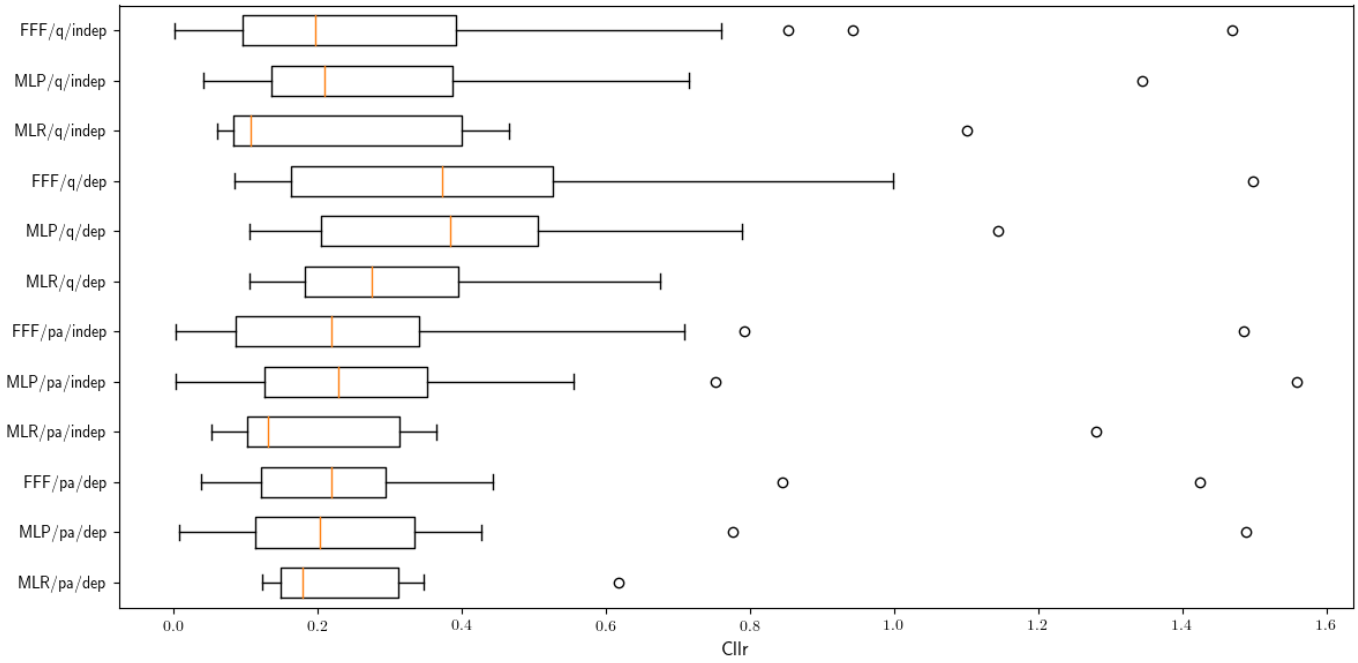


(b) Costs per log10(LR) from the histograms.

Figure 11: The distributions of the log10(LRs) calculated with the method *MLR/pa/indep* for one run and the penalty given to those values.

The C_{llr}'s resulting from the 30 runs for target class saliva are shown in figure 12. Note that the results from the four XGB methods are excluded because of some extremely large outliers (C_{llr} of circa 8) that made the results from the remaining methods difficult to study. The fact that these outliers are this extreme indicates that XGB is subject to overfitting and therefore delivers inaccurate LRs that lend strong support to the wrong hypothesis. Furthermore, note that for the majority of the methods in at least one run the C_{llr} exceeds 1. This demonstrates that in one of the 30 runs, the test dataset consisted of samples that were hard to classify and led those methods to give misleading results. Here the boxes from the boxplots overlap as well, indicating no significant difference between the accuracy of the methods, however there are two methods from which the median is significantly lower than that of the other methods. This is for *MLR/q/indep* and *MLR/pa/indep* and demonstrates that their accuracy generally is higher than the accuracy of the other methods.

Furthermore, observe there is a large spread in the C_{llr}'s. This is caused by sampling variability: each run the train, calibration en test dataset consist of different empirical mixtures because they are generated with randomly selected measurements of cell types.

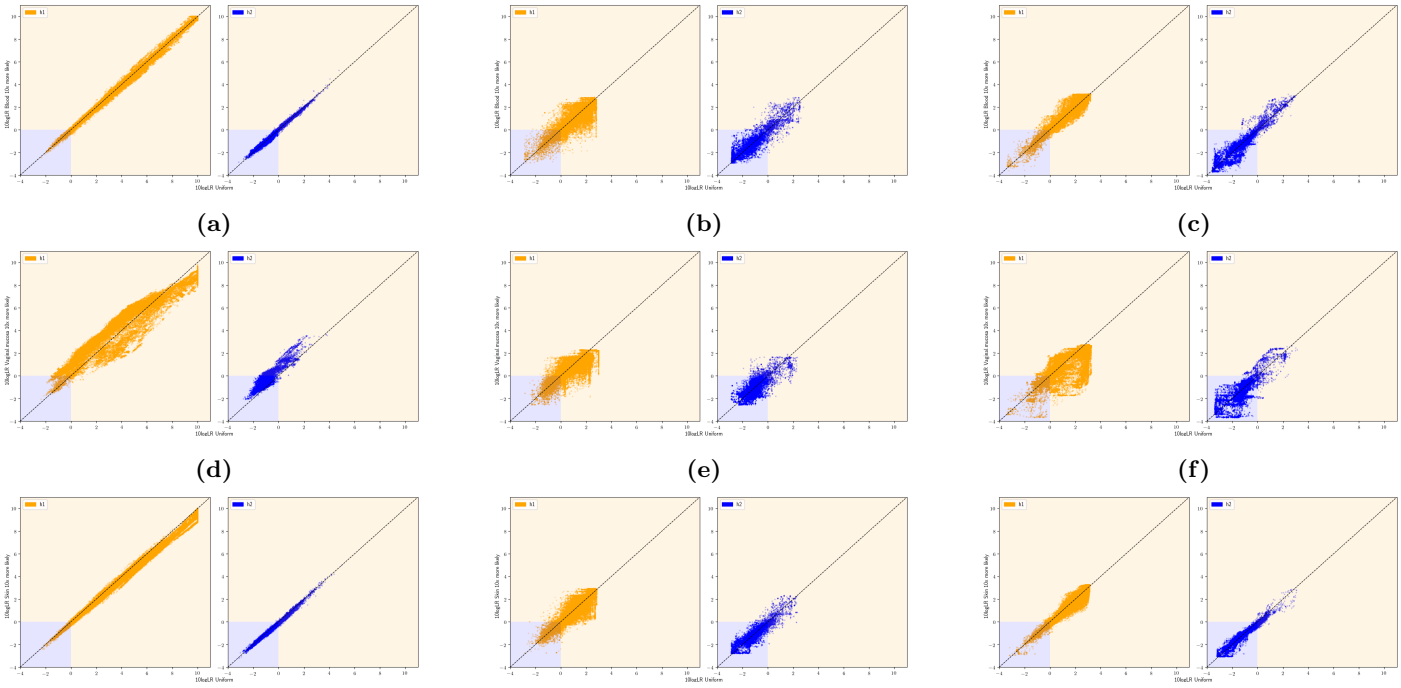


pa = presence/absence data, q = quantitative data,
indep = independence between class labels, dep = dependence between class labels

Figure 12: C_{llr}'s for target class saliva resulting from the 30 runs on the synthetic test data.

The susceptibility to a change in the relative frequency of the cell types in the data is assessed for three methods, namely *MLR/pa/indep*, *MLP/q/dep* and *FFF/pa/indep*. This selection of methods enables one to make two important comparisons. The first is to make a comparison between methods of which the derived LRs are transformed in a post-hoc calibration step to methods from which the derived LRs received no post-hoc treatment. Secondly, one can compare the susceptibility of *MLP/q/dep*, also known as the most promising method from the study by Scholten, to the other methods. The scatterplots plotting the $\text{LRs}_{\text{uniform}}$ against the sets of $\text{LRs}_{\text{non-uniform}}$ for target class vaginal/menstrual are shown in figure 13 and for target class saliva are shown in figure 14 respectively. The three sets of $\text{LRs}_{\text{non-uniform}}$ stem from the train data wherein the relative frequency of a given cell type is 10 times higher. Because it helps visualizations, the LRs under H_1 and H_2 are plotted separately. Note that the sets of LR values are collected from 10 additional runs.

Both figures show that the pairs of uniform and non-uniform LRs calculated with the MLR method are more alike in comparison to those of the other two methods. Moreover, the scatter points lie closer together on the diagonal under both H_1 and H_2. Observe however in figure 13d that the majority of the scatters points appear above the diagonal: $\text{LR}_{\text{non-uniform}}$ is higher than $\text{LR}_{\text{uniform}}$. Here the non-uniform dataset consists of relatively more vaginal mucosa samples and the target class vaginal/menstrual is assessed. So, a MLR model trained on this non-uniform dataset therefore learned more about vaginal mucosa and hence is more confident (i.e. higher LR) to classify a sample with ground-truth label vaginal mucosa. Another observation is that the pairs of uniform and non-uniform LRs calculated with the FFF method are more alike in comparison to those from the MLP method and hence less susceptible to a change in the relative occurrence of cell types. The fact that there is no clear pattern in the scatterplots for MLP may be a result of overfitting which will be checked later in this section. The scatterplots plotting the $\text{LRs}_{\text{uniform}}$ against the sets of $\text{LRs}_{\text{non-uniform}}$ that are retrieved from the methods trained on data for which the relative frequency of the three cell types is 10 times lower are shown in Appendix A.

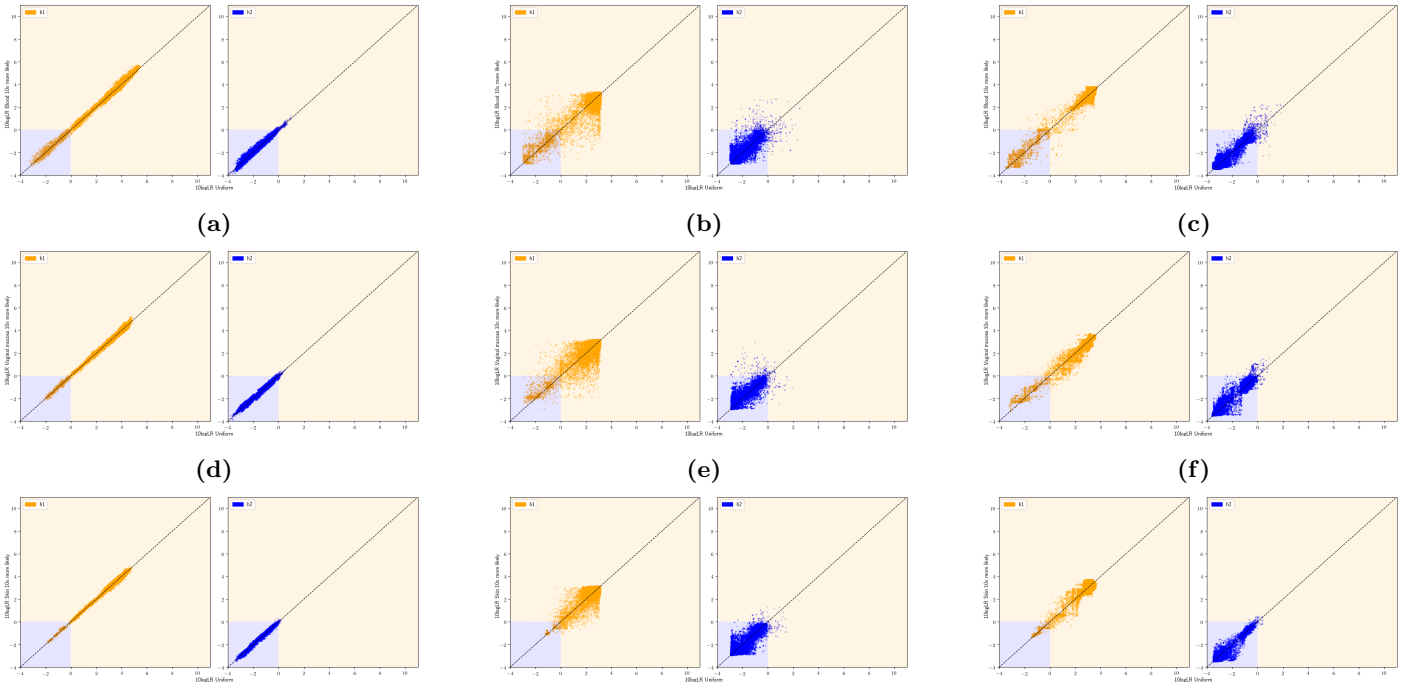


(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) *MLR/pa/indep* **(h)** *MLP/q/dep* **(i)** *FFF/pa/indep*

Figure 13: All pairs of $\text{LRs}_{\text{uniform}}$ and $\text{LRs}_{\text{non-uniform}}$ from 10 runs for the target class vaginal/menstrual. The results are from the three non-uniform datasets in which the relative frequency of the cell types blood (first row), nasal mucosa (second row) and skin (third row) is 10 times that of the others.



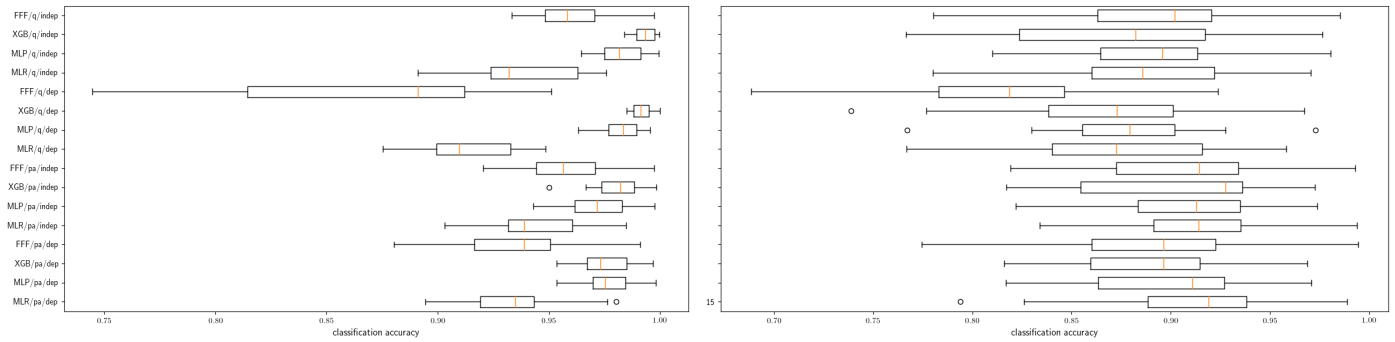
(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) *MLR/pa/indep* **(h)** *MLP/q/dep* **(i)** *FFF/pa/indep*

Figure 14: All pairs of $\text{LRs}_{\text{uniform}}$ and $\text{LRs}_{\text{non-uniform}}$ from 10 runs for the target class saliva. The results are from the three non-uniform datasets in which the relative frequency of the cell types blood (first row), nasal mucosa (second row) and skin (third row) is 10 times that of the others.

One can assess whether the probabilistic methods that have been experimented with are subject to overfitting using the classification accuracies on the train and test dataset. Note that the classification accuracy is the fraction of correct class label assignments (here for target class vaginal/menstrual). Overfitting is determined by comparing the train and test classification accuracy: when the train accuracy is high, but the test accuracy is low, the classifier is fitted to the train samples too well and will therefore fail to correctly classify the samples in the test set. Figure 15a shows that the median of the test classification accuracy for all the methods consisting of XGB and MLP are centered around 0.95. On the other hand, one can inspect from figure 15b that the median for the test classification accuracy for these same methods is centered around 0.9. So, XGB and MLP indeed are subject to overfitting. All four MLR methods however are not subject to overfitting: the median for the train classification accuracy and test classification accuracy are similar and centered around 0.92. The same holds for the methods including FFF.



(a) Train classification accuracy. **(b)** Test classification accuracy.

pa = presence/absence data, q = quantitative data,
indep = independence between class labels, dep = dependence between class labels.

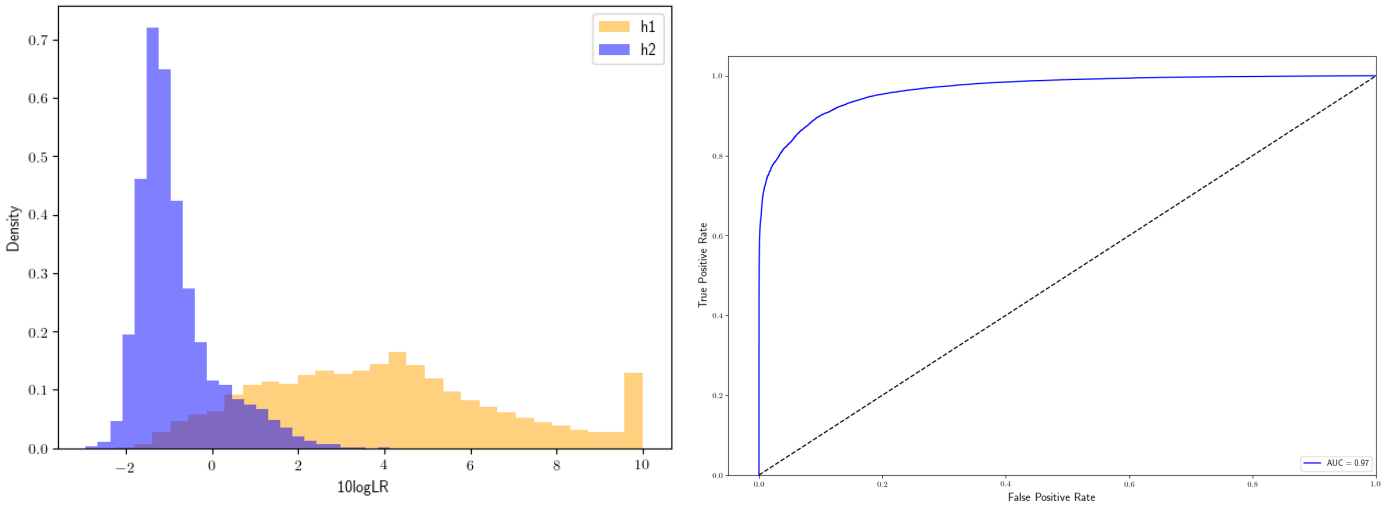
Figure 15: Classification accuracies of target class vaginal/menstrual from 30 runs on the synthetic test data set.

3.1.2 Discrimination and calibration

Additionally, the discriminating power and calibration on target class vaginal/menstrual were assessed using the LR values resulting from 30 runs for the method *MLR/pa/indep*. Figure 16a shows the LRs on a log-10 scale under H_1 and H_2 respectively. First of all there is some degree of overlap between the two distributions, indicating that a part of log10(LRs) under H_1 support the wrong proposition (i.e. are below 0) and a part of log10(LRs) under H_2 support the wrong proposition (i.e. are above 0). This tells us that there is no complete separation between the LR values belonging to the target class and the LR values belonging to the non-target class. To gain further knowledge about the degree of discriminating power, one is referred to the AUC. The ROC curve is displayed in fig 16b and the Area Under the Curve of the ROC is equal to 0.97. This implies that the probabilistic classifier can distinguish well between the target class and non-target class, or to be more precise: there is 97% chance that model will be able to distinguish between the classes. Furthermore, notice that the ROC curve is more in the upper left corner which means that there are more combinations of tp-rates and fp-rates in which the tpr is high. So, regardless of the value of the threshold, the classifier will be able to nearly always correctly classify samples from the target class. Also notice that LRs under H_1 can become as high as 1E10.

Figure 17 shows the PAV transformation of the LRs. The x-axis represents the pre-PAV calibrated LRs and the y-axis is the optimal transformation of those LRs following the Pool Adjacent Violators algorithm. The green solid line is the mapping function, mapping the pre-PAV calibrated LRs to the post-PAV calibrated LRs. The log10(LR) values under H_1 and H_2 are inserted at the bottom of the plot. First of all note that the line revolves around, but does not perfectly lie on the diagonal. Moreover, the deviation underneath the diagonal enlarges for decreasing log10(LRs) and the deviation above the diagonal enlarges for increasing log10(LRs). Van Es et al. in their paper declare that "*... for a well-calibrated system, the largest deviations from the line y=x are observed when the data is scare.*" [16]. This is what may be observed here as well: the largest deviations are around -2 where there is little data under H_1 and around 4 where there is little

data under H_2.



(a) Histogram of the log10(LRs). **(b)** ROC curve.

Figure 16: Visualizations that help assess the discriminating power of *MLR/pa/indep*

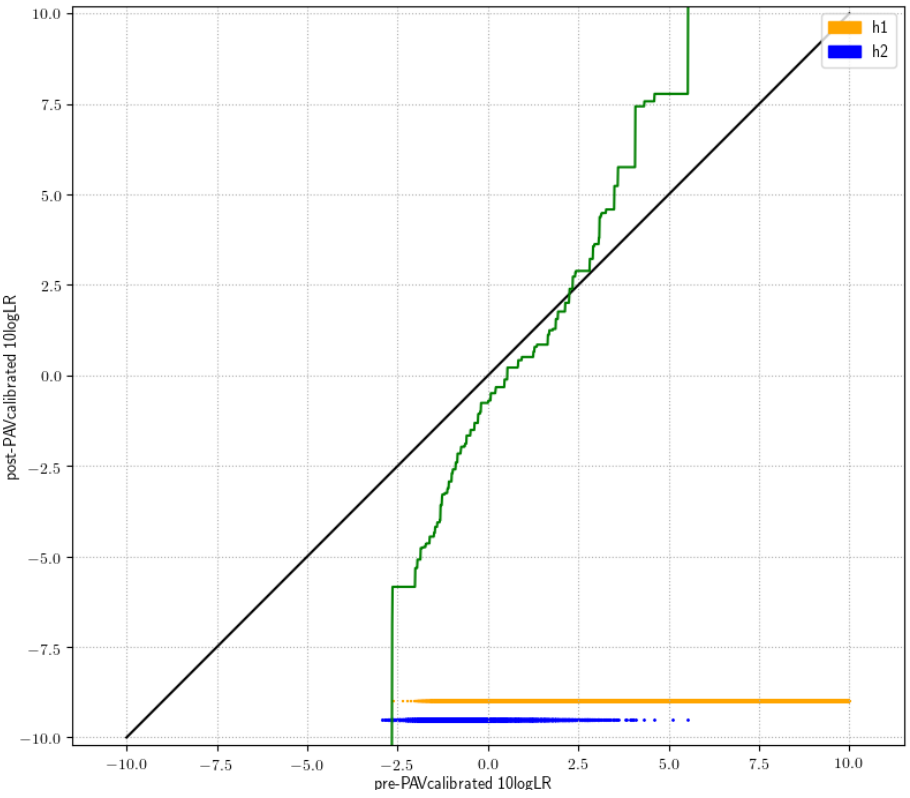


Figure 17: PAV transform of the log10(LRs) from the 30 runs.

3.1.3 Reliability of the LRs

The bootstrap intervals together with the 'true' LRs from the test data (black solid line) using the method *MLR/pa/indep* are displayed in figure 18. Note that the LRs from one out of the 30 runs have been been used to calculate the confidence interval with. First of all observe that the width of the interval depends on the underlying LR and hence is not fixed. This may be explained by the fact that for lower log10(LRs), the estimation of the density of the LRs given H_1 and H_2 is well supported by the data (see figure 16a). The confidence interval widens from log10(LRs) above 8 and is most narrow for log10(LRs) below -1. The widest interval roughly can be defined as (log10(LR)-0.6, log10(LR)+0.6).

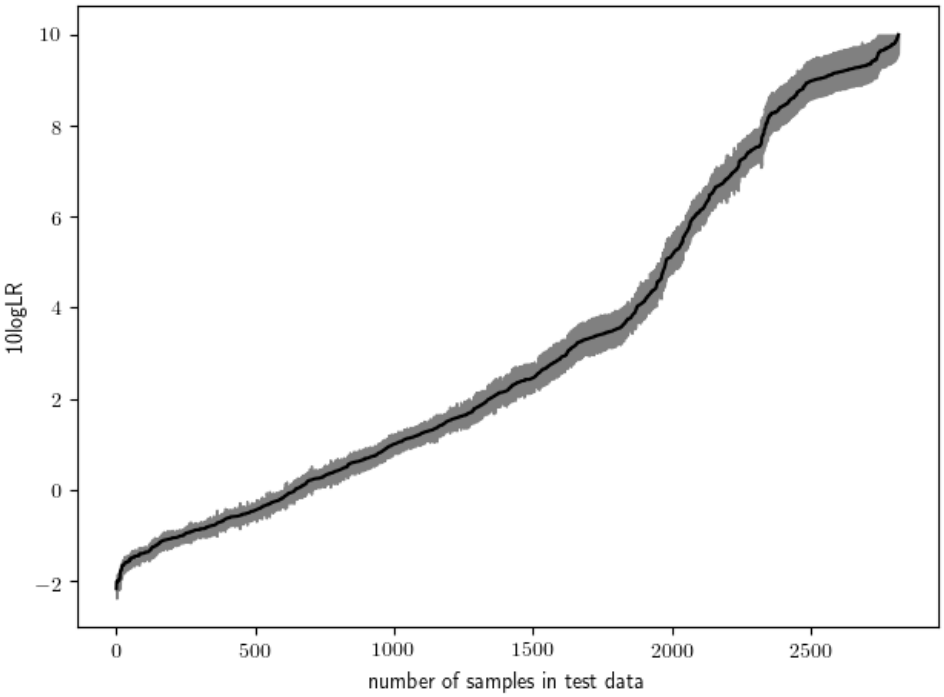


Figure 18: Plot showing the log10(LRs) and the 95% bootstrap confidence intervals.

3.2 Mixture cell type data

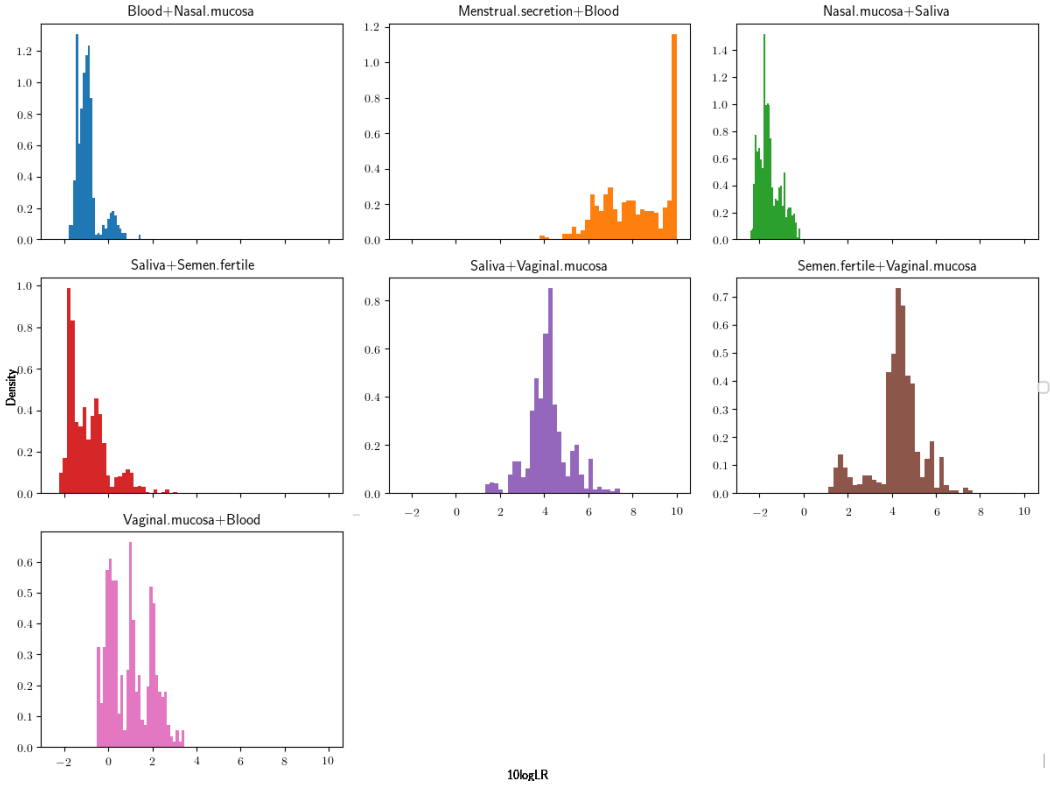
The LRs derived from the mixtures dataset using the method *MLR/pa/indep* on both target classes are illustrated in figure 24. The histograms are plotted for each unique mixture class on a log-10 scale. When the mixture includes the cell type(s) of a target class, the LR should be above one (or zero on the log scale) because they correspond to the actual cell type in the sample. The LRs should be below one (or zero on the log scale) otherwise. Then the LRs derived from the samples are correct and lend support to the correct hypothesis.

Figure 19a displays the LRs for target class vaginal/menstrual. The majority of the LR values that are derived from the mixtures are accurate. Moreover, all the LRs derived from the samples from the mixtures classes menstrual secretion + blood, saliva + vaginal mucosa, nasal mucosa + saliva and semen fertile + vaginal mucosa support the true proposition to a high degree. The highest values (up to 1E10) are obtained for menstrual secretion+blood. This is because blood is a component of menstrual secretion. In the remaining three classes a small fraction of the LRs derived from the samples support the wrong proposition. However, this is only to a small degree.

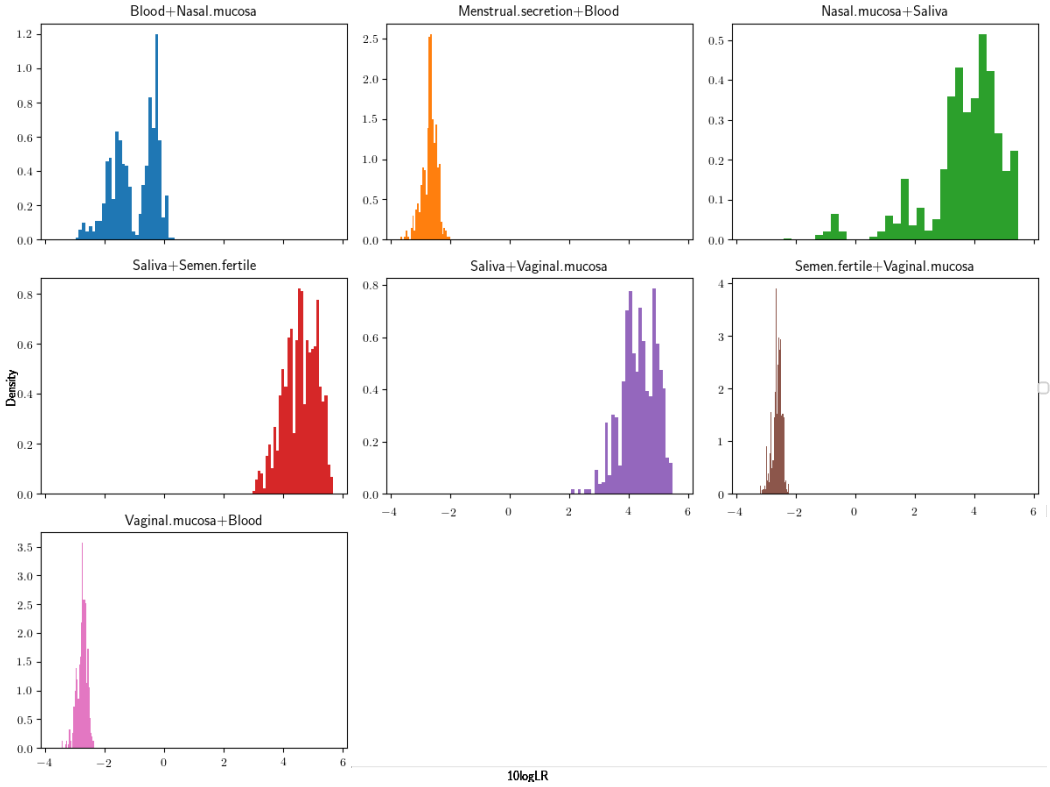
The LRs for saliva are displayed in figure 19b and the histograms clearly show that the values for the samples that contain saliva are higher than zero and for the samples that do not contain saliva are below zero. There is however a small fraction of samples from nasal mucosa + saliva for which the log10(LR) is below zero. This may be explained by the fact that nasal mucosa and saliva have one overlapping marker and one specific marker each. If in the mixture sample the saliva specific marker does not appear but the overlapping does, the classifier will be uncertain about the existence

of saliva and return a low LR.

In Appendix B the distribution of the 30 C_{llr}'s from all methods for both target classes are shown. The most important observation is that the C_{llr}'s are much lower than those resulting from the synthetic test dataset, implying that the accuracy of the classifier appears to be better than it truly is.



(a) Target class vaginal/menstrual.



(b) Target class saliva.

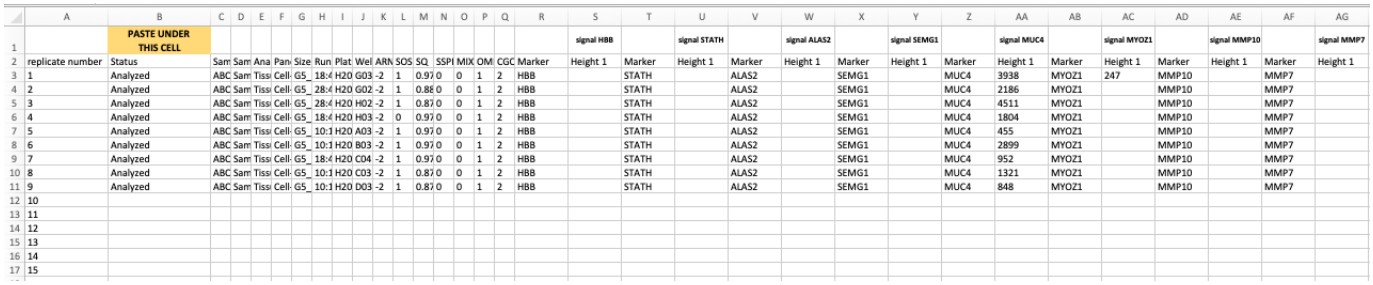
Figure 19: Histograms of the log10(LRs) derived from the mixture cell type dataset.

4 Application to data from actual cases

4.1 User friendly tool in Microsoft Excel

In order for the Department of Human Biological Traces to adapt the proposed method to actual cases, an userfriendly tool has been constructed with the use of Microsoft Excel. The main usage of the tool is to calculate the LR from data from actual cases and to compare it to the result from the n/2 method. There are two features that the tool is required to have. First of all, it should be able to use the mRNA measurements to calculate the LR for the correct class. Secondly, it should clearly show the influence of each marker on the result. This will enable the forensic examiners to both understand the result from the method and to perform their own experiments (for example to assess the change of the result in case more/less markers amplify).

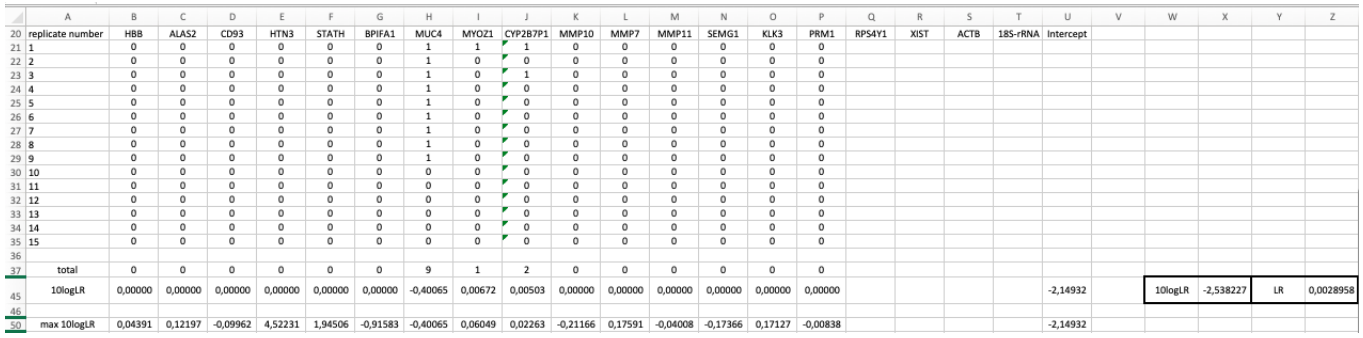
In total six logistic regression models have been fitted separately and their resulting coefficients have been processed into the tool. These are for the combinations of three classes and two prior probabilities for penile skin. The classes being saliva, the combination class of vaginal mucosa and/or menstrual secretion and vaginal mucosa. In the combination class the information about vaginal cells in menstrual secretion is taken into account when calculating the LR, whereas for the class vaginal mucosa all the method does not consider menstrual secretion. Per case it is up to the forensic examiner to decide about the appropriate model: when he or she is certain that menstrual secretion does not exist in the stain, the logistic regression model for vaginal mucosa should be picked. When the forensic examiner is certain that the stain contains penile skin, a model that also considers penile skin should be used (i.e. prior probability is 1).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1		PASTE UNDER THIS CELL																	signal HBB		signal STATH		signal ALAS2		signal SEMG1		signal MUC4		signal MYOZ1		signal MMP10		signal MMP7
2	replicate number	Status	Sam	Sam	Ana	Pan	Size	Run	Plat	Wel	ARN	SOS	SQ	SSPI	MIX	OM	CGC	Marker	Height 1	Marker	Height 1	Marker	Height 1	Marker	Height 1	Marker	Height 1	Marker	Height 1	Marker	Height 1	Marker	Height 1	Marker	Height 1
3	1	Analyzed	ABC	Sam	Tissi	Cell	G5_	18:4	H20	G03	-2	1	0.97	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	3938	MYOZ1	247	MMP10		MMP7	
4	2	Analyzed	ABC	Sam	Tissi	Cell	G5_	28:4	H20	G02	-2	1	0.88	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	2186	MYOZ1		MMP10		MMP7	
5	3	Analyzed	ABC	Sam	Tissi	Cell	G5_	28:4	H20	H02	-2	1	0.87	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	4511	MYOZ1		MMP10		MMP7	
6	4	Analyzed	ABC	Sam	Tissi	Cell	G5_	18:4	H20	H03	-2	0	0.97	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	1804	MYOZ1		MMP10		MMP7	
7	5	Analyzed	ABC	Sam	Tissi	Cell	G5_	10:1	H20	A03	-2	1	0.97	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	455	MYOZ1		MMP10		MMP7	
8	6	Analyzed	ABC	Sam	Tissi	Cell	G5_	10:1	H20	B03	-2	1	0.97	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	2899	MYOZ1		MMP10		MMP7	
9	7	Analyzed	ABC	Sam	Tissi	Cell	G5_	18:4	H20	C04	-2	1	0.97	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	952	MYOZ1		MMP10		MMP7	
10	8	Analyzed	ABC	Sam	Tissi	Cell	G5_	10:1	H20	C03	-2	1	0.87	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	1321	MYOZ1		MMP10		MMP7	
11	9	Analyzed	ABC	Sam	Tissi	Cell	G5_	10:1	H20	D03	-2	1	0.87	0	0	1	2	HBB		STATH		ALAS2		SEMG1		MUC4	848	MYOZ1		MMP10		MMP7	
12	10																																
13	11																																
14	12																																
15	13																																
16	14																																
17	15																																

Figure 20: First part of the tool: 9 example mRNA measurements inserted in the tool constructed in Excel for one model

Within the tool one solely has to insert the mRNA measurements from an external textfile. Figure 20 shows the part of the Excel tool for a given model in which nine example mRNA measurements are entered. The columns above which it says '**signal *marker name***' must contain the signal value(s), if any, for that marker. The number of replicates that the tool can process is 15.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
20	replicate number	HBB	ALAS2	CD93	HTN3	STATH	BPIFA1	MUC4	MYOZ1	CYP2B7P1	MMP10	MMP7	MMP11	SEMG1	KLK3	PRM1	RPS4Y1	XIST	ACTB	18S-rRNA	Intercept					
21	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0										
22	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
23	3	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0										
24	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
25	5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
26	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
27	7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
28	8	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
29	9	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0										
30	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
31	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
32	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
33	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
34	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
35	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
36																										
37	total	0	0	0	0	0	0	9	1	2	0	0	0	0	0	0										
45	10logLR	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	-0,40065	0,00672	0,00503	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000					-2,14932		10logLR	-2,538227	LR	0,0028958
46																										
50	max 10logLR	0,04391	0,12197	-0,09962	4,52231	1,94506	-0,91583	-0,40065	0,06049	0,02263	-0,21166	0,17591	-0,04008	-0,17366	0,17127	-0,00838					-2,14932					

Figure 21: Second part of the tool in Excel: replace the signal value in case equal or above 150 and calculate the log10(LR) using the coefficients from the Logistic Regression model

First the signal values are converted into either a one when it equals or exceeds the threshold of 150 and into zero otherwise. Figure 21 shows this modification for the nine example measurements in the tool. In the row stating 'total'

the amount of replicates in which a marker is observed is displayed. In the row below that it shows contribution of each marker to the resulting log10(LR) (i.e. the sum of each element is equal to the log10(LR)). The final row saying 'max log10(LR)' shows the influence per marker and can be used to interpret the results. If the model here would be the logistic regression model for vaginal mucosa, then the maximum log10(LR) for HBB can be interpreted as follows: in case HBB is observed in all mRNA measurements, the log10(LR) of vaginal mucosa increases with 0.044. The LR is also calculated by $10^{\log_{10}(LR)}$. In Appendix C the interpretable coefficients for all six logistic regression models are shown.

4.2 Results

The tool has been used to calculate the LRs from data from two actual cases. For both cases, the requested class is vaginal mucosa and the stain has been measured three times (i.e. three replicates). There is no information about whether the stains contain penile skin and/or menstrual secretion. In order to evaluate vaginal mucosa using the n/2 method, one must count the number of times the three specific markers are observed (i.e. signal value above 150) and divide by the number of possible positions, which is nine. Table 2 shows the results from the n/2 method both numerically and as a verbal statement and the LRs resulting from all six logistic regression models.

Table 2: Results from the n/2 method and the six Logistic Regression models for the mRNA measurements from two cases

Case number	Cell type of interest	n/2 method	n/2 method verbal	LR (log10(LR)) from Logistic Regression					
				Vaginal mucosa		Vaginal mucosa + Menstrual secretion		Saliva	
				NP	P	NP	P	NP	P
1	Vaginal mucosa	6/9	observed	500,927 (2,700)	142,372 (2,153)	99,222 (1,997)	27,281 (1,436)	0,002 (-2,647)	0,003 (-2,515)
2	Vaginal mucosa	3/9	sporadically observed	2,020 (0,305)	0,523 (-0,281)	0,511 (-0,291)	0,110 (-0,959)	0,002 (-2,705)	0,003 (-2,550)

Since 6 out of 9 vaginal mucosa markers are counted for case 1, the n/2 method categorizes vaginal mucosa as 'observed'. The LRs from the logistic regressions models for both vaginal mucosa and vaginal mucosa + menstrual secretion are above 1 meaning that they lend support to the hypothesis that states that vaginal mucosa is present in the stain. The evidential strength however varies for the four methods and the largest LR (500.927) results from the logistic regression model that does not consider penile skin for vaginal mucosa. Here it shows that logistic regression and the n/2 method would both report that vaginal mucosa is in the stain, but the logistic expresses this with a level of uncertainty which is more useful than the categorical statement. On the other hand, the LRs from the model for saliva are below 1 meaning that they lend evidence to the hypothesis that saliva does not exist in the stain.

In the second case, 3 out of 9 vaginal mucosa markers are counted and the n/2 method therefore categorizes vaginal mucosa as 'sporadically observed'. Only the LR from the vaginal mucosa logistic regression model that does not consider penile skin is above 1. The LR from the three models for vaginal mucosa and vaginal mucosa + menstrual secretion are below 1, thereby lending support for the hypothesis that vaginal mucosa does not exist in the stain. Unfortunately, because there is no (prior) information about whether the stains contain penile skin and/or menstrual secretion, it is not known which of the four methods is the appropriate method to evaluate the cell type with. If the prior probability for penile skin and menstrual secretion is zero, than the appropriate model to use would be the vaginal mucosa logistic regression model that does not consider penile skin from which the derived LR is 2.020. In that case the probabilistic method would report evidence supporting the hypothesis that vaginal mucosa is in the stain, whereas the n/2 method would not. This shows the advantage of a probabilistic classifier over the categorical method. Furthermore, the three models for vaginal mucosa and vaginal mucosa + menstrual secretion would report that there is weak or limited evidence that vaginal mucosa is in the stain. However, it is still more useful than the results from the n/2 method.

5 Discussion

In the introduction the five main objectives of this thesis were introduced, the first three being: determine a probabilistic classifier from which the output is both accurate and reliable, introduce a calibration technique and compare the performance of classifiers from which output is transformed in a post-hoc calibration step to a classifier from which output is directly interpreted as likelihood ratio, and examine the susceptibility of the classifiers to a change in the relative frequency of cell types in the train data. The main objective was to propose a classifier as alternative for the currently used categorical method.

The results on the synthetic test data show that *MLR/pa/indep* is preferred over the other methods. In other words, a separate logistic regression model for each target class together with the presence/absence data shows the most promising results on the two classes that are of most interest in forensic casework. First of all, MLR is the least susceptible to a change in the relative frequency of different cell types in the train data in comparison to two other methods. So, the likelihood ratios were affected the least for a different prior probability of the cell types. The results also show that the likelihood ratios for both target classes are accurate although, the accuracy varies among different compositions of samples in the train and test dataset. Additional experiments to assess the discriminating power and calibration regarding the target class vaginal/menstrual demonstrate that MLR distinguishes well between the target and the non-target class (AUC=0.97) and returns well-calibrated LR values. Furthermore, the 95% bootstrap intervals showed that the likelihood ratios of the MLR method are reliable. These results lend support that a logistic regression model using presence/absence data is sufficient to be applied in practice to analyse mRNA measurements. Note however that the LRs with this method can become as high as 1E10 and it is up to the forensic examiner to decide whether he or she trusts that this value, or any value for that matter, is correct.

Another objective was to compare the output from logistic regression with the calibrated output from the three remaining probabilistic classifiers. The experimental results show that calibrating the output is not necessarily beneficial: the post-hoc calibration likelihood ratios are not more accurate. Additionally, the MLP method and FFF method are more susceptible to a change in the relative frequency of the cell types in comparison to the MLR method. Note however that it was not determined what the exact cause of this result was, therefore being unable to conclude which post-hoc treatment of the LRs is preferred.

The fourth objective was to demonstrate that the mixtures cell type dataset should not be used as validation set. The logistic regression results show that the majority of the LRs derived from samples from the target class are high and the LRs derived from samples from the non-target class are low. Moreover, the accuracy of the LRs derived from the mixtures dataset is higher in comparison to the accuracy of the LRs from the synthetic test dataset. This holds for both target classes. Validating the classifiers on the mixtures dataset thus leds to overly optimistic conclusions about the performance of a classifier.

Finally, a tool was created and can be used by the Department of Human Biological Traces of the NFI. The relevance of the tool is showed by applying it on real data. Moreover, the results from the logistic regression models resulted in reporting the same conclusion regarding the existence of vaginal mucosa as the n/2 method in case 6 out of 9 markers were counted. Furthermore, in case 3 out of 9 markers were measured, the proposed method showed its main advantage, namely it reported a likelihood ratio whereas the n/2 method was not able to provide the forensic examiner with any evidence or uncertainty.

5.1 Future work

Two of the probabilistic classifiers that have been experiment with, namely MLP and XGB, have been implemented using default settings for the model parameters. Moreover, no hyper parameter optimization was performed to obtain a set of optimal parameters. As a result, both classifiers were not tuned to optimally solve the problem and were subject to overfitting. There are many ways to prevent a model from overfitting, for example by adding a dropout layer in an artifical neural network. The results indeed show that the Fully connected Feed Forward method, to which a dropout layer was added, was less subject to overfitting. Moreover, the FFF method was the only competitive method to the

MLR method as it showed promising results. In future research one could experiment with probabilistic classifiers such as FFF and XGB that are optimized to solve the problem, and study whether these are preferred over logistic regression.

A limitation of calibration techniques in general is that they are designed for two-class problems and not for multi-class problems. Here, a way to use a given calibration technique in a multi-class settings was introduced. However, there is no theory to justify the correctness of this way of using these techniques. Furthermore, the calibration technique that has been implemented here also has some limitations. First of all, kernel density estimation is controlled by a bandwidth parameter and the optimal bandwidth should be determined which unfortunately is not straightforward. When a suboptimal bandwidth is chosen, kernel density estimation can be subject to overfitting. The resulting calibration model will then not be capable of transforming the scores into well-calibrated LRs. In this thesis no other calibration techniques were implemented, nor has the used calibration technique been optimized by determining the optimal bandwidth parameter. In the future it could be beneficial to considering several techniques and determine which is 'best'. That way one can fully utilize the advantages of calibration.

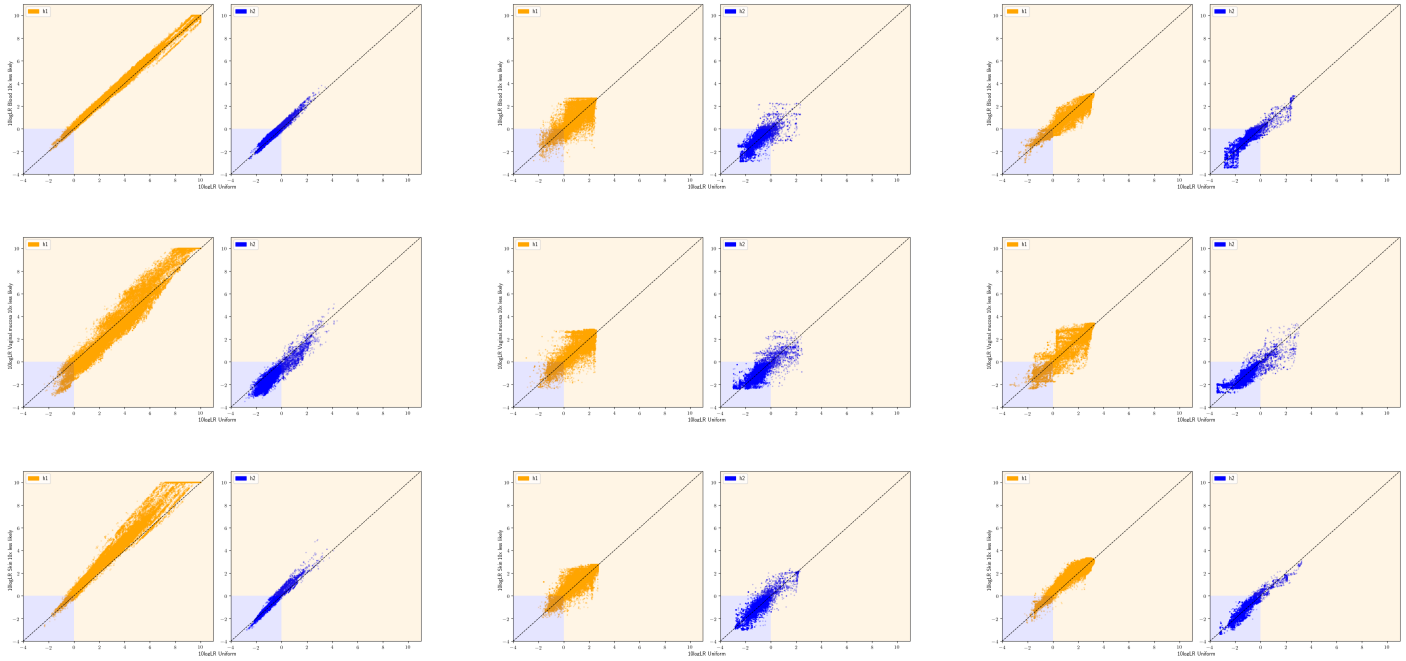
In this thesis it was assumed that the fraction of cell types in a mixture sample is equal. Moreover, equal weights were given to the cell types in the data in the way that the synthetic datasets were created: by selecting the same number of replicates for each cell type. However, in reality mixtures often are made up of different fractions of cell types. For example, a mixture sample could be 90% of cell type 1 and 10% of cell type 2. When the true fractions are known a practical solution to incorporate this information is in the creation of the synthetic dataset.

Before the proposed method can be applied in practice, one should consider to do a validation step and some adjustments. The method is chosen based on its performance on two target classes and no other classes have been taken into account. However, there are more body fluids that forensic examiners are requested to identify in actual casework, an example being exhaled blood that contains nasal mucosa, saliva and blood. Hence, the method should also be able to return accurate and reliable LRs when assessing alternative cell types. Therefore, one should examine whether the method is capable of doing this. Furthermore, as has been mentioned before, the LRs can become as high as 1E10. Even though a higher LR value lends greater support to the hypothesis stating that the cell type is in the sample, which seems desirable, it may not be desirable to report this value, because there are not enough data points to actually calculate whether this LR is correct (i.e. the exact value of the LR is not known). In case of misleading evidence, the judges could possibly make a wrong decision, because they are misinformed. An alternative would be to determine a ceiling and report that in stead of the LR when it is above this ceiling. This will shrink the evidential value, but will prevent a judge from being misinformed. At last, a user-friendly tool that allows the forensic examiner to calculate the LR for all forensically relevant body fluids and also gives them the option to incorporate their prior knowledge about the existence of one or several body fluid(s), should be build.

The method currently can only distinguish between the cell types it considers. This becomes a problem when the crime stain contains a cell type that the method has not seen before. It will give wrong results and the forensic examiner should be aware of this.

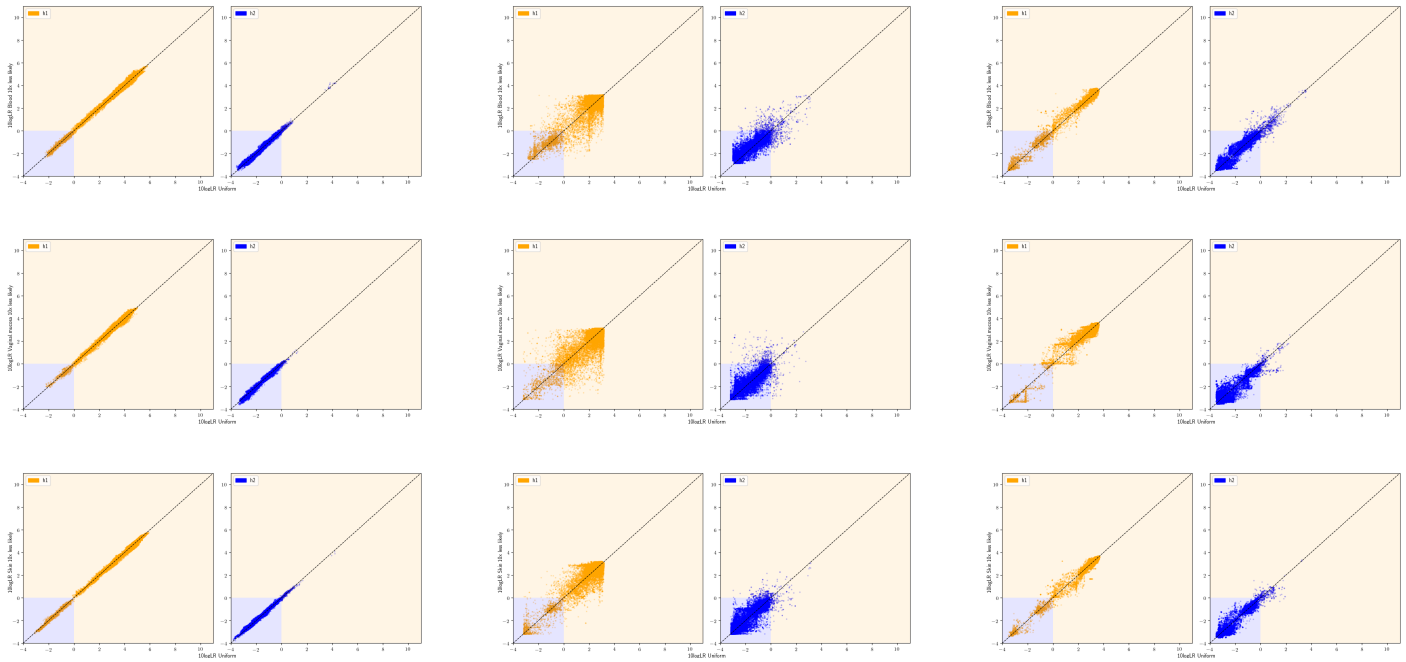
Appendices

A Scatterplots of the uniform LRs and three non-uniform sets of LRs where the frequency of cell types is lower



(a) *MLR/pa/indep* **(b)** *MLP/q/dep* **(c)** *FFF/pa/indep*

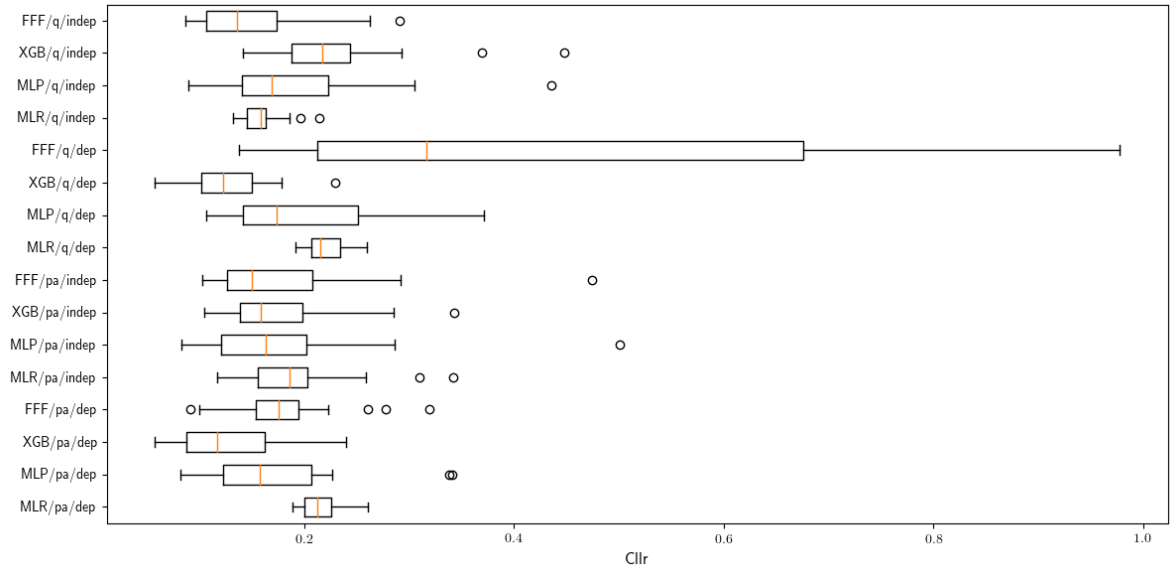
Figure 22: All pairs of $\text{LRs}_{\text{uniform}}$ and $\text{LRs}_{\text{non-uniform}}$ from 10 runs for the target class vaginal mucosa and/or menstrual secretion. The results are from the three non-uniform datasets in which the relative frequency of the cell types blood (first row), nasal mucosa (second row) and skin (third row) is 10 times less than that of the others.



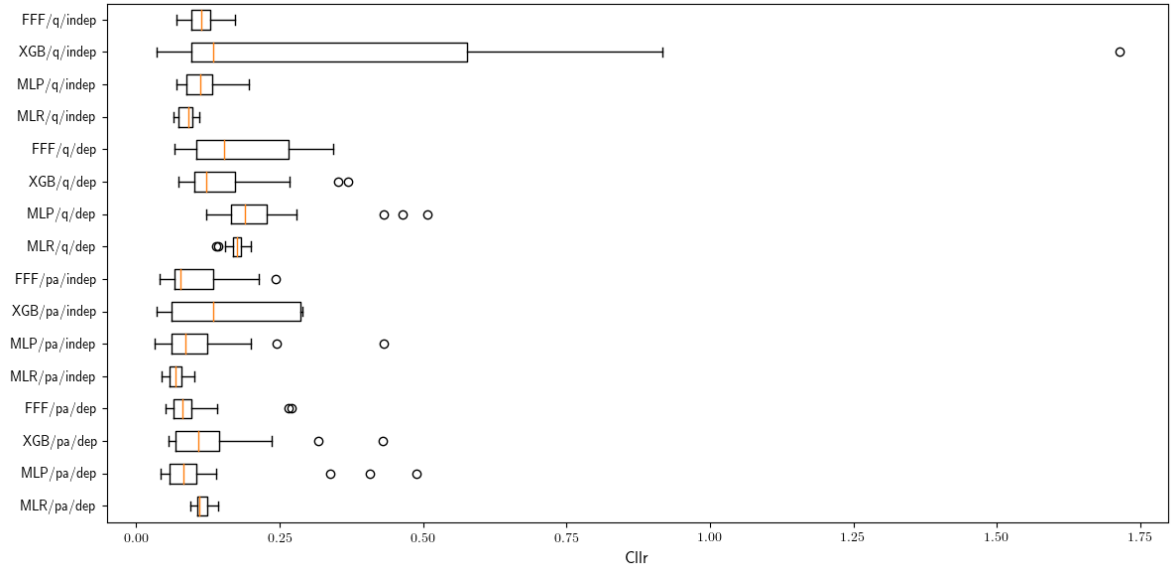
(a) MLR presence/absence indep **(b)** MLP quantitative dep **(c)** FFF presence/absence indep

Figure 23: All pairs of $\text{LRs}_{\text{uniform}}$ and $\text{LRs}_{\text{non-uniform}}$ from 10 runs for the target class saliva. The results are from the three non-uniform datasets in which the relative frequency of the cell types blood (first row), nasal mucosa (second row) and skin (third row) is 10 times less that of the others.

B The accuracy of the LRs derived from the mixtures cell type data



(a) Vaginal mucosa and/or menstrual secretion

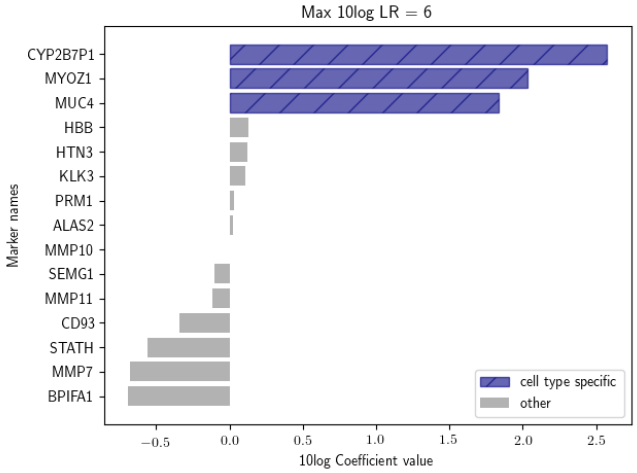


(b) Saliva

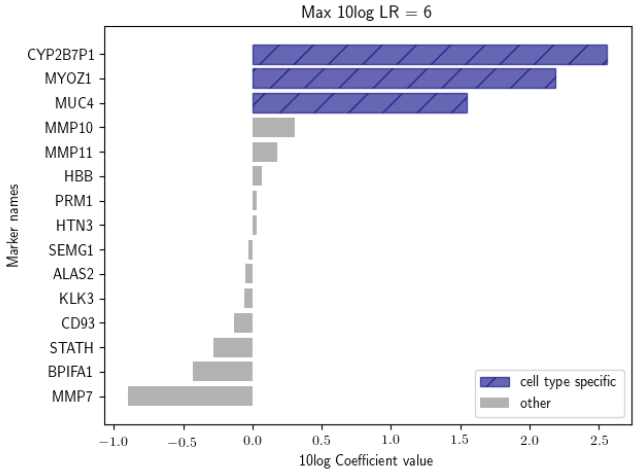
pa = presence/absence data, q = quantitative data,
indep = independence between class labels, dep = dependence between class labels

Figure 24: C_{llr}'s resulting from the 30 runs on the mixtures cell type dataset.

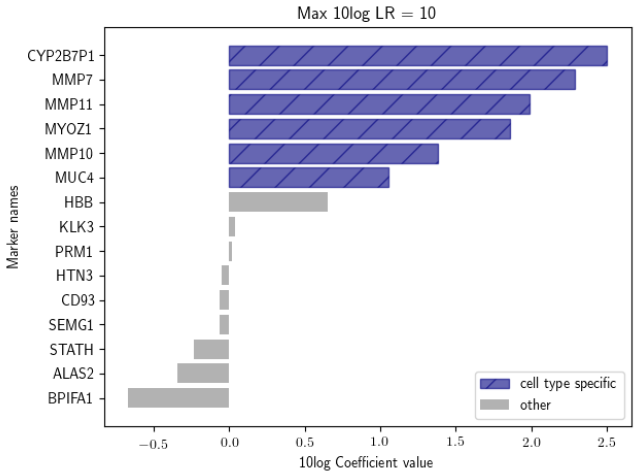
C Coefficient interpretation for the six logistic regression models



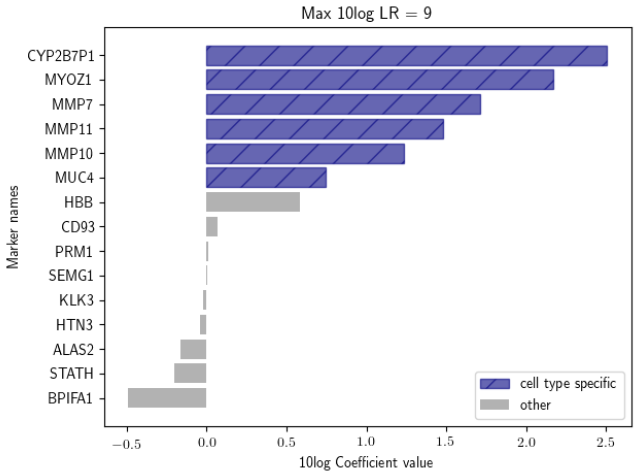
(a) Vaginal mucosa (NP)



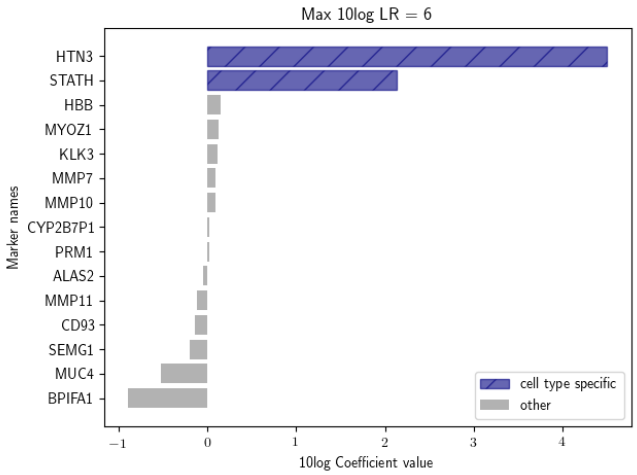
(b) Vaginal mucosa (P)



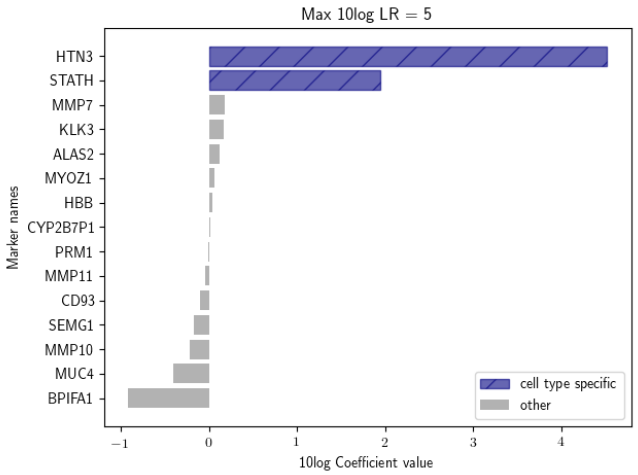
(c) Vaginal mucosa and/or Menstrual secretion (NP)



(d) Vaginal mucosa and/or Menstrual secretion (NP)



(e) Saliva (NP)



(f) Saliva (P)

Figure 25: 10log Coefficient values for all six Logistic Regression models

References

[1] Jacob de Zoete, James Curran, and Marjan Sjerps. "A probabilistic approach for the interpretation of RNA profiles as cell type evidence". In: *Forensic Science International: Genetics* 20 (2016), pp. 30–44. ISSN: 1872-4973. DOI: https://doi.org/10.1016/j.fsigen.2015.09.007. URL: http://www.sciencedirect.com/science/article/pii/S1872497315300697.

[2] Alexander Lindenbergh, Petra Maaskant, and Titia Sijen. "Implementation of RNA profiling in forensic casework". In: *Forensic Science International: Genetics* 7.1 (2013), pp. 159–166. ISSN: 1872-4973. DOI: https://doi.org/10.1016/j.fsigen.2012.09.003. URL: http://www.sciencedirect.com/science/article/pii/S1872497312001986.

[3] Geoffrey Stewart Morrison. "Measuring the validity and reliability of forensic likelihood-ratio systems". In: *Science Justice* 51.3 (2011), pp. 91–98. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2011.03.002. URL: http://www.sciencedirect.com/science/article/pii/S1355030611000256.

[4] Grzegorz Zadora et al. *Statistical analysis in forensic science. Evidential value of multivariate physicochemical data.* John Wiley & Sons, 2014.

[5] David Van Leeuwen and Niko Brummer. "The distribution of calibrated likelihood-ratios in speaker recognition". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Apr. 2013).

[6] Daniel Ramos and Joaquin Gonzalez-Rodriguez. "Reliable support: Measuring calibration of likelihood ratios". In: *Forensic science international* 230 (May 2013). DOI: 10.1016/j.forsciint.2013.04.014.

[7] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* 2015 (Apr. 2015), pp. 2901–2907.

[8] Meelis Kull, Telmo Silva Filho, and Peter Flach. "Beyond Sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration". In: *Electronic Journal of Statistics* 11 (Jan. 2017), pp. 5052–5080. DOI: 10.1214/17-EJS1338SI.

[9] Guro Dørum et al. "Predicting the origin of stains from next generation sequencing mRNA data". In: *Forensic Science International: Genetics* 34 (2018), pp. 37–48. ISSN: 1872-4973. DOI: https://doi.org/10.1016/j.fsigen.2018.01.001. URL: http://www.sciencedirect.com/science/article/pii/S1872497318300061.

[10] Valerie A.C. Scholten. "Probabilistic approaches for body fluid identification using RNA-data". MA thesis. the Netherlands: Radboud University, Oct. 2018.

[11] Gjorgji Madjarov et al. "An extensive experimental comparison of methods for multi-label learning". In: *Pattern Recognition* 45.9 (2012). Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), pp. 3084–3104. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2012.03.004. URL: http://www.sciencedirect.com/science/article/pii/S0031320312001203.

[12] "Logistic regression and artificial neural network classification models: a methodology review". In: *Journal of Biomedical Informatics* 35.5 (2002), pp. 352–359. ISSN: 1532-0464. DOI: https://doi.org/10.1016/S1532-0464(03)00034-0.

[13] Niko Brümmer and Daniel Garcia-Romero. "Generative Modelling for Unsupervised Score Calibration". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Nov. 2013). DOI: 10.1109/ICASSP.2014.6853884.

[14] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python ". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[15] Bianca Zadrozny and Charles Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2002). DOI: 10.1145/775047.775151.

[16] Andrew van Es et al. "Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis". In: *Science Justice* 57.3 (2017), pp. 181–192. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2017.03.002. URL: http://www.sciencedirect.com/science/article/pii/S1355030617300266.

[17] Berwin A. Turlach. "Bandwidth Selection in Kernel Density Estimation: A Review". In: *CORE and Institut de Statistique*.

[18] Peter Vergeer et al. "Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating?" In: *Science Justice* 56.6 (2016), pp. 482–491. ISSN: 1355-0306. DOI: https://doi.org/10.1016/j.scijus.2016.06.003. URL: http://www.sciencedirect.com/science/article/pii/S1355030616300363.

[19] Daniel Ramos et al. "Deconstructing Cross-Entropy for Probabilistic Binary Classifiers". In: *Entropy (ISSN 1099-4300)* 20.3 (Mar. 2018), p. 208. DOI: http://dx.doi.org/10.3390/e20030208.

[20] Didier Meuwly, Daniel Ramos, and Rudolf Haraksim. "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation". In: *Forensic Science International* 276 (2017), pp. 142–153. ISSN: 0379-0738. DOI: https://doi.org/10.1016/j.forsciint.2016.03.048. URL: http://www.sciencedirect.com/science/article/pii/S0379073816301359.

[21] Niko Brümmer and Johan du Preez. "Application-independent evaluation of speaker detection". In: *Computer Speech Language* 20.2 (2006). Odyssey 2004: The speaker and Language Recognition Workshop, pp. 230–275. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2005.08.001. URL: http://www.sciencedirect.com/science/article/pii/S0885230805000483.

[22] Yara van Schaik. "Measuring calibration of likelihood ratio systems". MA thesis. the Netherlands: University of Amsterdam, Nov. 2018.