

# Master Thesis

## The discrimination of human and animal blood using hyperspectral imaging and chemometric methods

Gary Sean Cooney

5 January 2021

Submitted to the

Faculty of Chemistry and Mineralogy

of the

University of Leipzig

In Partial Fulfilment of the Requirements for the Degree of

Master of Science

Advanced Spectroscopy in Chemistry

Supervisors: Prof. Jörg Matysik and Dr. Claire Chalopin

## Acknowledgements

I acknowledge the support of friends and family, colleagues and associates that have helped me to realise this master thesis.

To my supervisors Dr. Claire Chalopin and Prof. Jörg Matysik. I thank you for your insight into my research topic, feedback and review of the manuscript.

To my colleagues at ICCAS, especially Hannes Köhler and Marianne Maktabi. I thank you for your patience, for your insight into new and complex topics, and your general support and camaraderie throughout the thesis.

To Dr. Carsten Babian and the Institute of Legal Medicine, Leipzig, who proposed the project collaboration, provided the HSI system, facilities, and general support. I thank you for this opportunity.

To Prof. Henschler and the Institute of Transfusion Medicine, Leipzig, who provided the human donor samples. I thank you for your assistance.

To Dr. Hirrlinger and Manuela Liebig of the MEZ, Leipzig, who provided the mouse, rat, and rabbit blood samples. I thank you for your cooperation.

To Prof. Dr. Starke and the Clinic of Hooved Animals, who provided the cow blood samples. I thank you for your cooperation.

To Dr. Fischer and Schlachthof Weißenfels, Weißenfels, who provided the pig blood samples. I thank you for your cooperation.

I would like to thank Katharina Hoffmann for her emotional support and resilience, in her actions and words, that helped inspire me to pursue this endeavour and achieve its full potential.

To my parents, who have always encouraged and supported my pursuits into further education. In particular, I am grateful for their help and support in the final stages of thesis completion.

To the ASC master cohort. I thank you for the experiences, memories, and friendships built throughout the years.

# Table of Contents

List of figures .....	4
List of tables .....	6
Abstract .....	7
1. Introduction .....	8
1.1. Blood and its components.....	10
1.1.1. Plasma .....	11
1.1.2. Red Blood Cells.....	12
1.1.3. White Blood Cells .....	12
1.2. Haematology .....	13
1.2.1. Comparative Haematology of Common Domestic Animals.....	15
1.2.2. Anticoagulants in Haematology .....	16
1.3. Methods for the discrimination of blood.....	17
1.3.1. Absorption of light by matter .....	19
1.3.2. Reflectance Spectroscopy .....	21
1.3.3. Hyperspectral Imaging .....	22
1.4. Complementary Statistical and Analytical Methods.....	24
1.4.1. Chemometrics .....	24
1.4.2. Pre-processing.....	24
1.4.3. Classification.....	25
1.4.4. Unsupervised Learning.....	25
1.4.5. Supervised Learning .....	28
1.4.6. Bayesian Optimisation.....	29
2. Materials and methods .....	30
2.1. HSI system.....	30
2.2. Human and animal dataset .....	31
2.3. Classification framework.....	33
2.3.1. Pre-processing.....	34
2.3.2. Background detection .....	35
2.3.3. Class Balancing .....	36
2.3.4. Feature Selection.....	37
2.3.5. Classification.....	38

3.	Results .....	40
3.1.	Training and Test Set spectra .....	40
3.2.	Feature Selection using NCFS algorithm .....	41
3.3.	Classification development using Bayesian Optimisation.....	44
3.4.	Automatic Background Detection of bloodstain images .....	50
3.5.	Classification of HSI images.....	54
3.6.	Reflectance Spectra Characteristics .....	57
4.	Discussion.....	62
4.1.	Data acquisition and analysis .....	62
4.2.	Class balancing and classification framework.....	63
4.3.	Pre-processing.....	64
4.4.	Neighbourhood Component Feature Selection .....	64
4.5.	Classification of HSI images.....	66
4.6.	Bloodstain spectra analysis .....	69
5.	Conclusions .....	72
	Appendix A. Hyperspectral data .....	75
	Appendix B. Haematological Parameters.....	79
	Appendix C. Machine Learning Methods .....	84
	Declaration of authorship .....	90
	Bibliography .....	91

## List of figures

Figure 1. Blood pigments with associated colour (Left to right): Haem B (deoxygenated form), Haemocyanin (oxygenated form; R = histidine), Haemerythrin (oxygenated form), and Chlorocruorin (deoxygenated form). The combination of a conjugated system (porphyrin ring of Haem B and Haemocyanin) and the oxygen-binding central atom (Iron; Fe or Copper; Cu) modifies the absorption of light within the visible spectrum.....	10
Figure 2. Typical blood composition in humans along with average haematocrit value (HCT; mL RBC/ dL blood).....	14
Figure 3. RBC morphology and common abnormalities. Anisocytosis (irregular sized RBCs) and poikilocytosis (irregular shaped RBCs) are found commonly in various blood conditions in humans. The central pallor is the concave indentation at the centre of RBCs that is lighter in colour due to the cell's relative thinness. Rouleaux (side-view) are aggregates of RBCs that form due to the large contact area facilitated by RBC shape. ....	16
Figure 4. Reflectance phenomenon. Reflection consists of two major components: Specular Reflection and Diffuse Reflection.....	21
Figure 5. Comparative bar plot of unbalanced data with respect to species and binning. ....	33
Figure 6. Classification framework for HSI data cubes of human and animal bloodstains.....	34
Figure 7. Algorithm to detect background prior to classification. ....	35
Figure 8. Figure 8. Reflectance spectra of the training dataset of a) animal and b) human blood on cotton of all ages (day 0.1 – 40) with Savitzky-Golay smoothing and SNV transformed spectra of c) animal and d) human blood. ....	40
Figure 9. Average 10-fold loss vs. incremental lambda values of NCFS models. ....	41
Figure 10. NCFS feature weights as a function of feature index (red circles) with threshold lines of 50% and 80% the maximum feature index weight. The average human blood spectrum is plotted as a secondary axis (black; x-axis wavelength [nm], y-axis SNV reflectance [a.u.]) to guide the eye.....	42
Figure 11. Confusion charts of trained k-NN (k=1) classifiers with threshold values $T = 0$ (full spectrum), $T = 0.5$ (92 wavelengths), $T = 0.55$ (68 wavelengths), $T = 0.6$ (43 wavelengths), $T = 0.65$ (34 wavelengths), $T = 0.7$ (26 wavelengths).....	43
Figure 12. Bayesian optimisation of k-NN model hyperparameters distance metric and number of neighbours as a function of estimated objective function value after 30 iterations. ....	45
Figure 13. Minimum objective values of 30 Bayesian optimisation iterations of trained k-NN models using 10-fold cross-validation.....	45
Figure 14. Minimum objective values of 30 Bayesian optimisation iterations of trained bagged trees using 10-fold cross-validation. ....	46
Figure 15. Bayesian optimisation of SVM hyperparameters kernel scale and box constraint as a function of estimated objective function value after 21 iterations.....	47
Figure 16. Minimum objective values of 21 Bayesian optimisation iterations of trained SVMs using 10-fold cross-validation. ....	47
Figure 17. Comparison of training loss, CV-loss and validation loss of Bayesian-optimised SVM iterations. Based on mean plus 1 standard error, the parameters of iteration 16 are chosen as optimum for the SVM classification model.....	49
Figure 18. Confusion chart of SVM prediction results of test data.....	50

Figure 19. RGB image of human (day 40), and cow (day 1 and day 0.1) blood stains on white cotton (top left) with coloured markers and corresponding reflectance spectra (bottom) which were used in the determination of the ABD thresholds that generate the binary background detection image (top right). Averaged spectra in order of maximum reflectance at band 1: purple = white cotton (high illumination), green = white cotton (low illumination), orange (cow blood - day 1), yellow (human blood – day 40), blue (cow blood – day 0.1), light blue (black metal), red (black cloth)..... 51

Figure 20. Steps for ABD image processing. Fresh human blood samples (t = 0.1 days) where 1 is the RGB image generated from the HSI cube, 2 is the binary ABD image where white indicates "non-background", 3 is the segmented binary image after 'opening' and border removal, and 4 is the mask of identified blood stains (blue) overlaid onto the RGB image.... 52

Figure 21. Steps for ABD image processing. Aged animal blood samples (t = 25 days) of mouse (left column), rabbit (middle column) and rat (right column), where 1 is the RGB image generated from the HSI cube, 2 is the binary ABD image where white indicates "non-background", 3 is the segmented binary image after 'opening' and border removal, and 4 is the mask of identified blood stains (blue) overlaid onto the RGB image. .... 53

Figure 22. HSI RGB images of aged human blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 2 days, 6 days, and 31 days. Yellow pixels are classified as "human" and blue as "animal". .... 54

Figure 23. HSI RGB images of aged pig blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 1 day, 5 days, and 42 days. Yellow pixels are classified as "human" and blue as "animal". .... 55

Figure 24. HSI RGB images of aged mouse blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 13 days and 21 days. Yellow pixels are classified as "human" and blue pixels are classified as "animal"..... 55

Figure 25. HSI RGB images of aged rat blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 13 days, and 25 days. Yellow pixels are classified as "human" and blue pixels are classified as "animal"..... 56

Figure 26. HSI RGB images of aged rabbit blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 1 day, and 49 days. Yellow pixels are classified as "human" and blue pixels are classified as "animal"..... 56

Figure 27. HSI RGB images of aged cow blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 2 days, 6 days, 49 days. Yellow pixels are classified as "human" and blue are classified as "animal"..... 57

Figure 28. [Top left] Average reflectance spectra with stand and deviation (shaded area) of "fresh" day 0.1 human (cyan) and animal (red), and SNV-transformed spectra [top right]. [Bottom left] Average second derivative reflectance spectra, and [Bottom right] PC scores plot of the first two principal components (explained variance PC1 = 95.3%; PC2 = 3.0%) of the SNV spectrum. .... 59

Figure 29. [Left] Average reflectance spectra and [Right] average SNV reflectance [Right] of human (red), pig (magenta), mouse (cyan), rat (green), rabbit (blue), and cow (black). The number of individuals included in the average spectrum is denoted brackets (\*) in the SNV reflectance legend..... 59

Figure 30. PC loading plot of PC1 and PC2. The wavelengths from 670-685 nm have the greatest positive PC1 loadings (+0.156) with 900-955 nm having the greatest negative PC1 loadings (-0.090)..... 60

## List of tables

Table 1. RBC morphological parameters of humans and common domestic animals. Highlighted animals are considered in this work. Note: while many animal RBCs are biconcave disks, llama RBCs are typically flat and oval-shaped. ....	15
Table 2. Human and animal dataset. ....	32
Table 3. Total dataset number of recorded HSI images and spectra.....	33
Table 4. Number of species spectrum before and after underbalancing was implemented. The class were balanced initially by age bins, followed by balancing by species. ....	37
Table 5. Training set and test set partition with balancing. ....	38
Table 6. NCFS k = 1, k-NN model F1 scores to determine the optimal number of features based on the prediction of the validation set. Model no. 3 with threshold of 0.55 and 68 wavelengths was chosen as optimal number of features. ....	44
Table 7. Bayesian optimised 3rd order polynomial SVM iterations with parameters evaluated as "best" based on objective function of 10-fold cross-validation. The iteration 16 was selected as the optimal SVM model using the mean-plus-1 standard error of the cross-validation error. ....	48
Table 8. Model statistics of SVM binary classifier. TOT = total number of observations predicted, PREV = prevalence; a measure of class distribution, ACC = accuracy, PPV = positive predictive value or precision, TPR = true positive rate or sensitivity, TNR = true negative rate or specificity.....	49
Table 9. Datasheet of 20 human donor obtained from the Institut für Transfusionmedizin, Universitätsklinikum Leipzig.....	75
Table 10. Datasheet of pig blood samples obtained from Schlachthof Weißenfels, Weißenfels.'Coagulation' signifies a separation of RBC from plasma before probe preparation. Samples were prepared and measured on the same day as slaughter. ....	75
Table 11. Datasheet of mouse blood samples obtained from the Medizinisch-Experimentelles Zentrum (MEZ), Leipzig .....	76
Table 12. Datasheet of rat blood obtained from the Medizinisch-Experimentelles Zentrum (MEZ), Leipzig .....	77
Table 13. Datasheet of rabbit blood obtained from the Medizinisch-Experimentelles Zentrum (MEZ), Leipzig .....	77
Table 14. Datasheet of cow blood samples obtained from the Klinik für Klautiere, Leipzig. Samples probes were prepared and measured on the same day as collection. ....	77
Table 15. Laboratory Variables Relevant to Hematologic Diagnosis (Normal Human Adult Values), adapted from Williams Manual of Hematology, 9ed <sup>86</sup> . ....	79
Table 16. Reference Interval for Haematologic Parameters in Diet-Restricted 7-10 Week Old CD-1 mice collected under isoflurane anaesthesia <sup>87</sup> . ....	80
Table 17. Referenced Haematological parameters of New Zealand white rabbit ( <i>Oryctolagus cuniculus</i> ). ....	81
Table 18. Reference intervals for the Domestic Pig <sup>95</sup> . ....	82
Table 19. Reference intervals for the adiva 120 from 99 Clinically Healthy Cows, 50% in First Lactation, All Milking 30-150 days, from 10 Ontario Farms <sup>106</sup> . ....	83

## Abstract

Blood is the most encountered type of biological evidence in violent crimes. It contains pertinent information to a forensic investigation, the analysis of such information leading to the difference in conviction or not in a court of law. A commonly overlooked aspect in bloodstain investigation is the origin of blood, that is, whether it is of human or non-human origin. The false presumption that blood encountered at a crime scene is human, may not be realised until after costly and time-consuming laboratory analysis is performed. Despite recent advancements in analytical and statistical methods – including the evermore frequent use of portable spectroscopic imaging systems – the main focus of these works has been the estimation of blood deposition age. This once again neglects to address the question of blood origin. In this study, the novel application of vis-NIR hyperspectral imaging (HSI) is used for the detection and discrimination of human and animal bloodstains on white cotton fabric. The HSI system is a portable, non-contact, non-destructive method for the determination of blood origin. The inclusion of such a system in a forensic investigation workflow not only removes ambiguity surrounding blood origin, but can potentially be used in tandem with blood age determination methods. Chemometric and machine learning methods are implemented to identify spectral regions of importance, and to train a support vector machine (SVM) binary classifier in the discrimination of bloodstains. On an independent test set, the SVM model achieved accuracy, precision, sensitivity, and specificity values of 96, 97, 95, and 96% respectively. Segmented images of bloodstains aged over 50 days are produced, allowing for the clear visualisation of human and animal blood. Blood components and haematological data are considered in the reasoning for the observed spectral differences between species.



# 1. Introduction

Blood is one of the most readily examined tissues of the human body, playing a key role in the diagnosis of disease, and investigative forensic science. A single drop of blood contains valuable information for use in forensic science, such as its chemical composition and morphology of the associated bloodstains on surfaces. This information, along with blood stain pattern analysis, can be utilised to reconstruct the events of a violent crime. A multidisciplinary approach encompassing natural sciences, - such as; biology chemistry, mathematics and physics, aid the investigator with the interpretation of the circumstances of the incident and qualify the information legally<sup>1</sup>. Results obtained can help differentiate whether bloodshed found at a crime scene is deliberate or accidental, or as a result of suicide or murder. In any case, a primary step upon arrival to a crime scene is to corroborate if suspect red stains are in fact blood. Presumptive tests and DNA analysis (detailed in Section 1.4) are routinely used to detect and identify suspicious drops and stains encountered at a crime scene<sup>1,2</sup>.

Blood is a common specimen collected in post-mortem forensic examinations, being the specimen of choice for detection and quantification of drugs and/or toxicants. Determination of substance concentrations in blood is useful for establishing the effect a substance might have on the victim at the time of death, or at the time of specimen sampling. This analysis can help the forensic investigation decide whether poisoning is suspected as a cause of death, or if the prolonged use of prescription medication has otherwise complicated the investigation. Nevertheless, due to the degree of decomposition and variation in substance concentrations in ante- and post-mortem blood, many other specimens are collected for toxicological investigation. These include urine, hair, gastric contents, and various organs, amongst others<sup>3</sup>.

Blood samples are an undoubtedly crucial asset to forensic investigation, especially in the case of a violent crime. However, a commonly overlooked aspect in bloodstain identification is determination of the blood stain as human or animal. Unless specific wildlife crimes are being investigated<sup>4</sup>, bloodstains encountered at a crime scene are often presumed to be of human origin. In many cases, it is only after obtaining the DNA profile with database comparison that the presumed human sample is realised to be of

animal origin. This wastes money, resources, and time. Therefore, the need for a cheap and fast method to quickly differentiate human and non-human blood has been identified. This paper addresses this need by the use of the non-contact modality hyperspectral imaging (HSI), coupled with chemometric methods to build a classification model that successfully discriminates human and animal blood stains. Even though HSI has already been proven as an effective tool in various applications of forensic science<sup>5-8</sup>, when it comes to blood and bloodstain analysis, the determination of bloodstain age is predominant<sup>9-14</sup>. While the determination of bloodstain age using HSI can objectively have greater value in a forensic investigation of a violent crime, the fact remains that the determination of blood origin is once again overlooked in the literature.

A controlled study using blood of human and 5 common animal species deposited onto white cotton was used as a preliminary investigation into the ability of visible-near infrared (VNIR) HSI to discriminate human and non-human bloodstains. Chemometric and machine learning methods were used to investigate and develop a binary human-animal classifier based on the reflectance HSI data. Image segmentation was performed on the captured RGB images using a background detection algorithm and the trained classifier, effectively visualising the discrimination of human versus animal bloodstains. The portable HSI system coupled to the human-animal classifier demonstrates the potential of this modality to significantly speed-up forensic investigation with on-site measurement capability.

To better understand the trained classifier's ability to discriminate animal and human blood, the constituents and haematology of blood is discussed in the following sections (Section 1.1). Subsequently, methods and limitations to current and nascent forensic analysis of blood is outlined in Section 1.2. The principles to reflectance spectroscopy and HSI are later detailed (Section 1.3), and the theory behind the machine learning methods used in this work is provided in Section 1.4.

## 1.1. Blood and its components

The average adult human carries between four and five litres of blood, continuously circulating through the heart, arteries, vessels, and capillaries, and interacting on the cellular level with tissues. Blood supplies nutrients and oxygen and in exchange removes cellular waste such as carbon dioxide. Several organs act to regulate the nutrient and waste content in blood. The lungs act in respiratory exchange, acquiring oxygen and releasing carbon dioxide; while the kidneys filter the blood, removing excess water and dissolved waste products. Nutrients from food make their way into the blood stream through absorption in the gastrointestinal tract, while hormones are released into the blood by endocrine system glands.

Specialised cells and fluids constitute blood. Each have specific physiological functions and provide information about an individual's state of health. Cellular composition of blood varies within the animal kingdom. Most invertebrates have relatively few blood cells compared to vertebrates, with some simple animals, such as worms or molluscs, transporting oxygen directly within the blood plasma. With greater oxygen needs, larger animals have blood pigments such as haemoglobin (iron-containing, red-coloured), haemocyanin (copper-containing, blue-coloured), chlorocruorin (iron-containing, green-coloured), or haemerythrin (iron-containing, red-coloured). Haemoglobin is an oxygen-carrying protein commonly found in vertebrates and in some invertebrates. Most vertebrates have their haemoglobin stored in erythrocytes or red blood cells (RBCs). Haemocyanin is found in some crustaceans, while chlorocruorin and haemerythrin are found in some annelids.

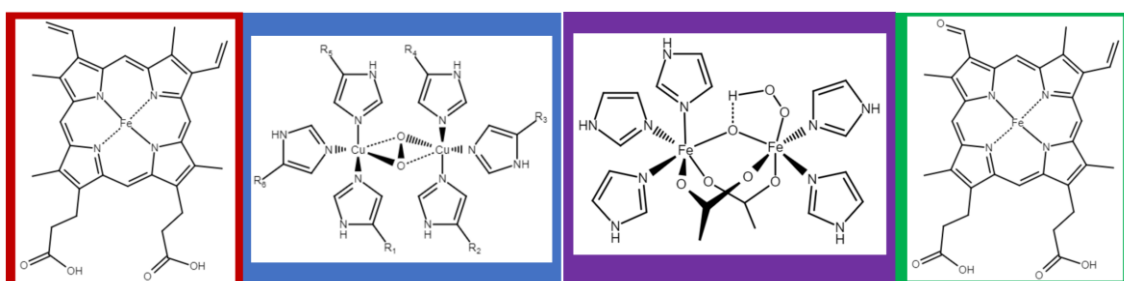


Figure 1. Blood pigments with associated colour (Left to right): Haem B (deoxygenated form), Haemocyanin (oxygenated form; R = histidine), Haemerythrin (oxygenated form), and Chlorocruorin (deoxygenated form). The combination of a conjugated system (porphyrin ring of Haem B and Haemocyanin) and the oxygen-binding central atom (Iron; Fe or Copper; Cu) modifies the absorption of light within the visible spectrum.

Due to the presence of haemoglobin, mammalian blood is a characteristic opaque red fluid. Saturated haemoglobin (oxyhaemoglobin, HbO<sub>2</sub>) and unsaturated haemoglobin (deoxyhaemoglobin, Hbb) are of slightly different shades of red, explaining the difference seen between the darker, partially deoxygenated, venous blood and the brighter, oxygenated arterial blood. The RBCs form approximately 45% of the blood volume, with white blood cells (WBCs) and platelets forming less than 1%. The remaining volume consists of plasma, a complex solution that is over 90% water.

### 1.1.1. Plasma

Plasma serves to transport blood cells, nutrients, and waste products from cellular metabolism, maintaining homeostasis of the body. As it is freely exchangeable with extracellular fluids and cells, the water within plasma also acts to maintain tissue hydration. Proteins are the most abundant plasma constituent by weight (7%), with dominant plasma protein being serum albumin (60% of all plasma proteins). Serum albumin acts to retain water in the blood via its osmotic effect, and can bind to other substances within the plasma as a nonspecific carrier protein. Other plasma proteins include globulins such as immunoglobins, which are produced in an immunogenic response to foreign antigens or substances. Cytokines are a group of small proteins that serve as chemical intracellular messengers, which play a role in the immune system response to foreign bodies as well as regulating blood cell formation (haematopoiesis). Coagulation proteins and their inhibitors are another group of important plasma proteins. Fibrinogen is converted to fibrin when blood clotting is activated, stabilising the blood clot. Phospholipids, triglycerides, free fatty acids, and cholesterol are the major fraction of total plasma lipids, with lipid concentration within plasma varying with meals. Many constituents of plasma occur in low concentrations. Such examples include glucose, a key source of energy for cells, and amino acids, a requirement for all protein synthesis throughout the body. The greatest contribution of blood plasma to the measured whole blood reflectance spectra is expected in the NIR region (700-1000 nm), where lipid<sup>15</sup> and water<sup>16</sup> are observed at ca. 900 and 950 nm respectively.

### 1.1.2. Red Blood Cells

Red blood cells (RBCs) are the most predominant cells in blood, being non-nucleated, deeply pigmented cells functioning in tissue respiration. They are highly specialised cells lacking the ability to synthesize new proteins and undergo mitosis, and are void of mitochondria and ribosomes. They contain haemoglobin, an iron-protein complex that functions as a carrier for oxygen and carbon dioxide<sup>17</sup>. RBCs take the form of biconcave disks, being 7-8  $\mu\text{m}$  in diameter in humans. This shape is both beneficial for deformation, aiding movement through the microvasculature<sup>18</sup>, and in gas transfer, where the surface-to-volume ratio is maximised in such a shape<sup>19</sup>.

### 1.1.3. White Blood Cells

White blood cells (WBCs) or leukocytes, protect the host from external pathogens. They are a heterogenous group of nucleated cells which lack haemoglobin and function to defend the body from infection by producing antibodies, and/or ingesting and destroying foreign bodies. WBCs are grouped into three main groups based on their appearance under a light microscope: lymphocytes, granulocytes, and monocytes. Each type have distinct physiological role and characteristic appearance<sup>17</sup>. Lymphocytes include B cells and T cells which recognise foreign pathogens and mediate their destruction via the production and secretion of antibodies. Granulocytes mediate body inflammation processes and can be further subdivided into neutrophils, eosinophils, and basophils. Monocytes differentiate at the site of infection into macrophages which are important antigen-presenting cells that engulf microbes via phagocytosis<sup>20</sup>.

## 1.2. Haematology

To understand the observed differences between the measured human and animal reflectance spectra the haematological parameters of each species are considered (see Appendix B).

Haematology is the study and treatment of blood and blood-related diseases. In haematology, the three main values that define the erythroid system are: the haematocrit (HCT) or packed red cell volume (PCV), the haemoglobin concentration (Hb), and the RBC count per unit volume (RCC). Haemocytometers are used in the measurement of RCC, in addition to WBC and platelet counts. The haematocrit is defined as the proportion of the blood volume occupied by RBCs, and is expressed as millilitre RBC per decilitre whole blood (mL RBC/ dL blood). It reflects the concentration of RBCs in whole blood, but not the total red cell mass. The Hb is a measure of the quantity of haemoglobin per unit volume of blood, being the measurement of choice for the physiological assessment of the erythroid system status. This is due to Hb providing a direct measurement of blood-oxygen capacity. In the determination of Hb, all predominant forms of haemoglobin present in the blood such as oxyhaemoglobin, and carboxyhaemoglobin, amongst others, are converted to the haemoglobin-cyanide by the use of Drabkin's solution<sup>21</sup>. This solution contains potassium ferricyanide ( $K_3Fe(CN)_6$ ) and potassium cyanide (KCN) which quantitatively convert all forms (except sulfhaemoglobin) to the cyanide derivative. The haemoglobin-cyanide concentration is then measured using a spectrophotometer at 540 nm and comparing with known standards<sup>22</sup>. Hb is expressed commonly as grams per decilitre of whole blood (g/dL blood). As HCT and Hb are measured based on whole blood, these parameters are therefore dependent on the plasma volume. This means that these parameters are subject to an individual's level of hydration, where levels are overestimated or underestimated in cases of extreme dehydration or overhydration respectively.

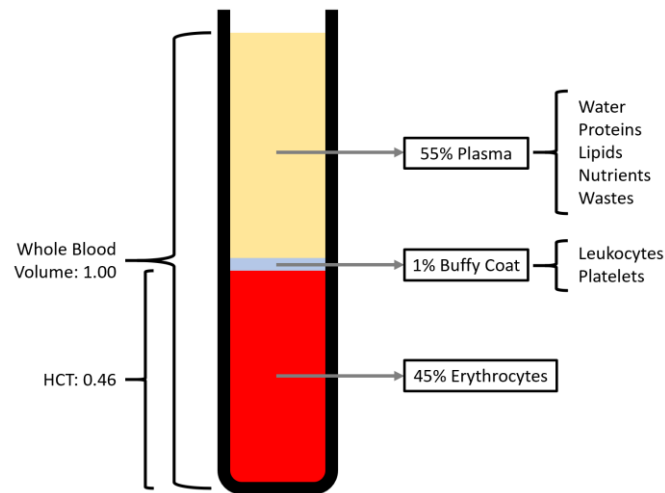


Figure 2. Typical blood composition in humans along with average haematocrit value (HCT; mL RBC/ dL blood)

In addition to HCT, Hb, and RCC measurements, three indices describe the average RBC; the mean corpuscular volume (MCV), the mean corpuscular volume haemoglobin (MCH), and the mean corpuscular haemoglobin concentration (MCHC). The MCV is the average volume of RBCs calculated as the ratio of the haematocrit and the red cell count:

$$MCV (10^{-15}/L) = \frac{HCT (L/L) \times 1000}{Red\ cell\ count (\times 10^{12}/L)}$$

The MCH is average quantity of haemoglobin within a cell and is calculated as the ratio of haemoglobin concentration to red cell count:

$$MCH (pg) = \frac{Hb (g/dL)}{Red\ cell\ count (\times 10^{12}/L)}$$

The ratio of haemoglobin to haematocrit measures the concentration of haemoglobin in the average RBC:

$$MCHC (g/dL) = \frac{Hb (g/dL)}{HCT (L/L)}$$

The MCH measures the weight of haemoglobin in the average RBC, while the MCHC indicates the haemoglobin concentration in the average RBC<sup>17</sup>. Within haematological examinations, the total number of leukocytes per unit volume (white cell count, WCC),

as well as a differential cell count that describes the proportion of each of the major cell types, is measured. Additionally, platelet and newly formed RBCs or reticulocyte counts are performed.

### 1.2.1. Comparative Haematology of Common Domestic Animals

The morphology of mature RBCs is similar in most mammalian species; they lack nuclei, are biconcave, disc-shaped cells. The major differences between animal RBCs are the size and the degree of central pallor (central zone with decreased haemoglobin due to closer apposition of membranes). As per Table 1 below, animal RBCs also vary in their degree of Rouleaux formation (aggregate stacks of RBCs), and anisocytosis (unequally sized RBCs). RBC size of laboratory animals – mice, rats, rabbits – is generally dependent on animal age, all of which have similar or slightly smaller size of RBCs compared to those found in dogs. The central pallor is also less pronounced compared to dog RBCs<sup>23</sup>.

*Table 1. RBC morphological parameters of humans and common domestic animals. Highlighted species are considered in this work. Note: while many animal RBCs are biconcave disks, llama RBCs are typically flat and oval-shaped.*

	<b>Diameter [µm]</b>	<b>Central pallor</b>	<b>Rouleaux</b>	<b>Anisocytosis</b>
<b>Human</b>	7.5	+	-	-
<b>Pig</b>	6	+	+++	+
<b>Mouse</b>	5 - 7	+	+ -	++
<b>Rat</b>	5.7 - 7	+	+ -	++
<b>Rabbit</b>	6.8	+	+ -	+
<b>Cow</b>	5.5	+ -	-	+
<b>Dog</b>	7.0	++	+	-
<b>Cat</b>	5.8	+	++	+
<b>Horse</b>	5.7	+ -	+++	-
<b>Sheep</b>	4.5	+	-	+ -
<b>Goat</b>	3.2	+ -	-	+
<b>Llama</b>	4.0 x 7.0	-	-	+ -



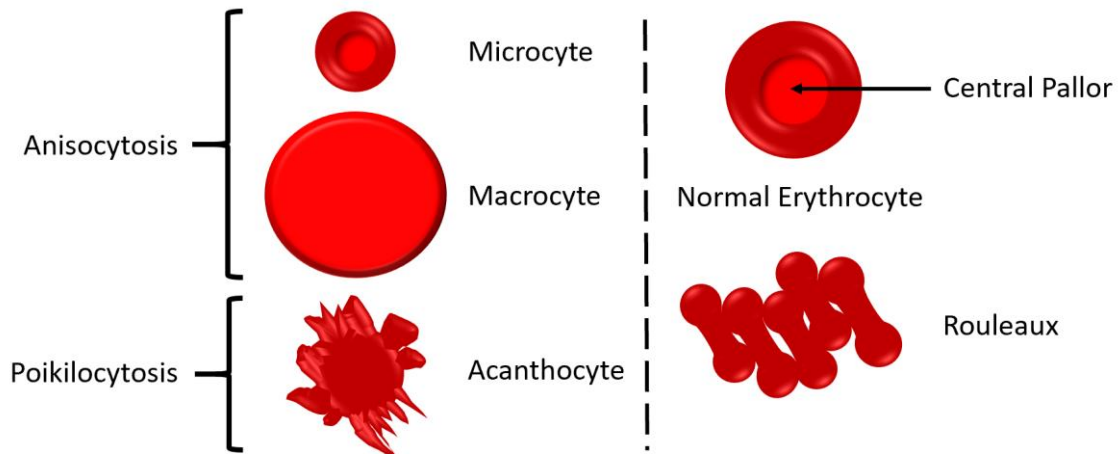


Figure 3. RBC morphology and common abnormalities. Anisocytosis (irregular sized RBCs) and poikilocytosis (irregular shaped RBCs) are found commonly in various blood conditions in humans. The central pallor is the concave indentation at the centre of RBCs that is lighter in colour due to the cell's relative thinness. Rouleaux (side-view) are aggregates of RBCs that form due to the large contact area facilitated by RBC shape.

An additional factor to consider in the discrimination of human and animal blood samples based on their reflectance spectra is their respective morphological traits. Size and degree of cell irregularity influences the relative amount of haemoglobin present within the cell. If Rouleaux formations are common, it is hypothesised that the thickness of these aggregates could induce more light scattering. As a result, the measured reflectance spectrum of the blood stains is dependent not only on the relative number of RBCs, but the abundance of the molecule haem within a RBC, and the various morphological traits that attenuate its signal<sup>24</sup>. Similarly, coagulation of whole blood would generate aggregates that would cause more scattering. For this reason, the anticoagulant EDTA was used to mediate coagulation prior to measurement.

### 1.2.2. Anticoagulants in Haematology

Anticoagulants are a group of chemical compounds that inhibit the coagulation cascade in blood and thus reduce or prevent clotting. They differ from antiplatelet drugs, which act directly on the aggregation of platelets<sup>25</sup>. The most common inhibitors of coagulation used in the analysis of blood are; ethylenediaminetetraacetic acid (EDTA) in either disodium or tripotassium salt forms, trisodium citrate, and heparin. In the case of the former two, their mechanism of action is the removal of calcium from the blood —

a clotting factor. Heparin forms a complex with plasma antithrombin III, inhibiting the formation of thrombin. EDTA contains four carboxylic acid groups and two amine groups which chelate calcium and other metal ions via their lone-pair electrons. The chelation of calcium inhibits the coagulation cascade and thus prevents the formation of blood clots; stabilising whole blood in its liquid form. EDTA is the preferred anticoagulant of blood cell counts as it has minimal effects on all blood cells with complete anticoagulation. Anticoagulated blood can be stored at 4°C for 24 hours without significant alterations to cellular morphology or cell counts<sup>25</sup>.

### 1.3. Methods for the discrimination of blood

Basic characteristics of blood found at a crime scene are determined using serological presumptive tests. The Kastle-Meyer (KM) test is a fast, cheap and efficient test used in the discrimination of blood from other substances of similar appearance. It is based on the decomposition of hydrogen peroxide into water and oxygen by the reaction with haemoglobin present in blood. This reaction, along with the indicator phenolphthalein, give a visible colour change of the medium-pink positive test result<sup>26</sup>. The main drawback to preliminary tests like the KM, and similarly the luminol test<sup>27</sup>, is the lack of a confirmatory result and their destructive nature<sup>28</sup>. The development of the Teichmann and Takayama crystal test<sup>29</sup> as a confirmatory blood test, improves ambiguity during a forensic investigation. However, these tests are destructive and cannot differentiate blood of different species. Being based on the immune reaction and subsequent generation of antibodies when foreign blood is introduced into a host, precipitin tests are often used to differentiate human from animal blood. This occurs when human blood, or any protein of human origin in a specimen sample, reacts specifically with antibodies present in the anti-human serum. The formation of a cloudy band at the interfaces of the two liquids indicates blood of human origin<sup>1,30</sup>. This test, however, has been mostly surpassed by immunoassay tests in blood identification<sup>31</sup>. Immunoassays are nonetheless prone to false-positives, where animal haemoglobin is very similar to human. Additionally, like many other presumptive and confirmatory tests, they are destructive and more often than not, require a laboratory environment to be performed.

Following the evaluation of presumptive tests, DNA analysis gives an accurate reconstruction of a person's DNA profile with a high sensitivity. The results of the DNA analysis confirm the origin of the blood, be it from the victim or the perpetrator, or possibly both<sup>2</sup>. As the analysis of samples of ambiguous origins can cost significant time and money to an investigation, DNA analysis is performed only in cases where it is absolutely necessary.

Advanced analytical methods have become increasingly predominant in forensic sciences. Mass spectrometry and chromatography techniques have become common place in toxicology labs, and spectroscopic methods such as Raman spectroscopy, UV-vis spectroscopy, and Fourier transform infrared spectroscopy (FT-IR) have proven themselves in the analysis of blood<sup>32-35</sup>. Their embrace has been partly due to the accompanying advancements in multivariate statistical analysis, leading to accurate, sensitive and reliable methods. These methods are often less destructive, or non-destructive, with less sample preparation when compared to presumptive tests<sup>28</sup>.

The most common chromophore in tissue is the iron-porphyrin complex haem<sup>36</sup>. This complex is found in oxygen carrying proteins such as haemoglobin and myoglobin, in addition to other haemoproteins such as haem peroxidase, catalases and cytochromes. In the blood, haemoglobin and its derivatives give its characteristic red pigment. When oxygenated haemoglobin (HbO<sub>2</sub>) is illuminated with white light, blue light is absorbed and red light is reflected, giving its characteristic colour. Deoxygenated haemoglobin (HHb) absorbs a higher degree of red light and thus is perceived bluer under white light. Therefore, reflectance spectroscopy can be used to provide information about haemoglobin oxygenation and by extension, concentrations<sup>37</sup>. The degree of light absorption of haemoglobin derivatives in tissue provides information about vascularity and metabolic status. Based on these principles, pulse oximetry is a simplistic version of reflectance spectroscopy. Reflectance spectroscopy has the benefits of being a non-destructive method, with modern portable spectrometers enabling fast on-scene analysis. When coupled as an imaging system – as is the case with HSI – the information procured by reflectance spectroscopy can be mapped onto a crime scene, thus aiding interpretability of forensic findings. Nevertheless, reflectance spectroscopy is limited by long initial analysis times and the level of interpretation required to build models for use

outside of the lab. As outlined in Section 1.4, advanced statistical methods are required for the study of spectroscopic and hyperspectral data sets.

### 1.3.1. Absorption of light by matter

Electromagnetic radiation interacts with matter by three mechanisms; rotational, vibrational, and electron excitation. Rotational spectroscopy deals with the measurement of transition energies between quantized rotational states of molecules typically in gas phase. This is due to intermolecular forces between molecules in liquid or solid phase preventing free rotation. It is often termed pure rotational spectroscopy — to differentiate it from ro-vibrational or vibronic spectroscopy, where both rotational and vibrational electronic states change simultaneously. Vibrational spectroscopy is the study of characteristic vibrational modes of molecules, which encompasses infrared (IR) spectroscopy and Raman spectroscopy. Absorption occurs when the frequency of radiation matches the vibrational frequency of bonds of the molecule or group of atoms within a molecule. These modes must be IR-active to be observable, that is, have a change in the electric dipole moment<sup>38</sup>. The IR-inactive modes can be observed using Raman spectroscopy, which is based on the principle of Raman scattering — the inelastic scattering of photons<sup>39</sup>. Analogous to IR spectroscopy, a molecule must have a change in its polarizability to be Raman-active.

Absorption of light in the UV, visible, and NIR ranges of the electromagnetic spectrum occur mainly by the radiation-induced transitions of single electrons from states of lower energy to ones of higher energy. Absorption occurs when the incident light frequency  $\nu$  is equal to the energy difference between the ground and excited state:

$$h\nu = E_1 - E_0 = h \frac{c_0}{\lambda}$$

where  $h$  is Planck's constant,  $c_0$  is the speed of light,  $\lambda$  is wavelength,  $E_1$  and  $E_0$  are the excited and ground states respectively. Pure electronic excitation theoretically gives sharp lines from the associated transition energy. However, electronic excitation is

always accompanied by molecular rotation and vibration, resulting in overlapping signals of different molecular vibration and rotation.

$$E = E_{ee} + E_{vib} + E_{rot}$$
$$|E_{ee}| \gg |E_{vib}| \gg |E_{rot}|$$

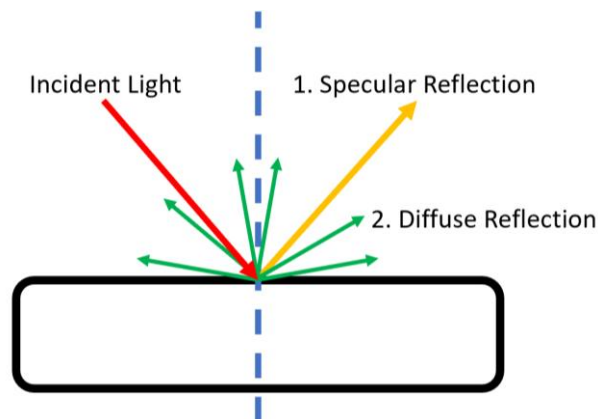
This results in the characteristic broad bands, of most liquid and solid UV-vis spectra. In multicomponent analysis, convoluted spectra can be resolved using several methods including (but not limited to) optical methods, such as the use of monochromators and narrow-band lasers<sup>40</sup>, low-temperature spectroscopy<sup>41</sup>, and computational methods such as curve-fitting and derivative spectroscopy — the latter of which is discussed in a later section.

Electrons of isolated atoms have discrete energy states, which in principle, can be calculated using the Schrödinger equation. In polyatomic molecules, the absorption of a photon is due to the excitation of a small group of atoms or specific types of electrons present. In the case of a carbonyl group (C=O), the absorption of 290 nm is normally observed. Groups with characteristic electronic transitions are termed chromophores (“colour-carrier” in Greek), and their presence in molecules and compounds often results in their observed colour<sup>42</sup>.

Molecular orbital (MO) theory is used to describe the process of electronic transition where single bonds are described by  $\sigma$  orbitals and double bonds by  $\pi$  orbitals. The inclusion of asterisk (\*) signifies antibonding nature of the respective orbital i.e.,  $\sigma^*$  and  $\pi^*$ . Electrons of n orbitals do not participate in bonding and are termed nonbonding orbitals. Electrons of  $\sigma$  orbitals typically absorb light in the far UV ( $\lambda < 180$  nm), with  $\pi$  and n transitions occurring with  $\lambda > 180$  nm. Transition metals with incomplete d subshells form various complexes with ions and polar ligands and tend to have intense colour. The transitions between the orbitals of the central ion give rise to the weak and medium-intense bands in the visible range of the spectrum. Electronic transitions between the central ion and the ligands,  $\pi \rightarrow \pi^*$  and  $n \rightarrow \pi^*$  transitions within the ligands give rise to shorter-wavelength intense bands. These are termed charge-transfer transitions and inner-ligand transitions respectively<sup>43</sup>.

### 1.3.2. Reflectance Spectroscopy

Reflectance spectroscopy is often used where the measurement of absorption via transmission-geometry is not possible due to sample thickness — as is the case in remote sensing applications<sup>44–46</sup>. The dynamic range of measurement is extremely large, ranging from UV to IR frequencies<sup>38</sup>. In the case of particulate matter, multiple scattering can amplify the contrast within weak absorption bands. A major diagnostic characteristic for determining a substance composition is the wavelength at which the reflectance band occurs, while the concentration or relative abundance can be retrieved from the intensity of spectral bands. However, in reflectance spectroscopy, both the band centre and band shape differ to that measured by transmission.



*Figure 4. Reflectance phenomenon. Reflection consists of two major components: Specular Reflection and Diffuse Reflection*

Reflection is when incident light that falls on a surface is thrown back without absorption. When the surface roughness is greater than the incident wavelength, light is scattered spherically outward in what is termed diffuse reflection. In the case of smooth surfaces such as liquids, glass or highly polished metals the light is reflected unidirectionally in the same plane and at the same angle as the incident beam. This phenomenon is termed specular reflection or regular reflection<sup>43</sup>.

Reflected radiation of diffusing media consists of two components. The first is the regular reflection at the surface where the Fresnel equations can be applied. The regular reflectivity of perpendicular incidence is given by:

$$r = \frac{[(n - 1)^2 + \kappa_0^2]}{[(n + 1)^2 + \kappa_0^2]}$$

where  $n$  is the relative refractive index of the medium and  $\kappa_0$  is the absorption index defined by the Lambert law:

$$I = I_0 \exp\left[-\frac{4\pi\kappa_0 d}{\lambda}\right] = I_0 \exp[-\alpha d]$$

With the distance  $d$ , being the distance travelled within the absorbing medium, where the radiation is reduced from  $I_0$  to  $I$ , and  $\alpha = \frac{4\pi\kappa_0}{\lambda}$  is the characteristic absorption coefficient of the substance.

The second component of reflected radiation is due to the multiple scattering of surface-penetrating radiation with individual media particles. A portion of this scattered radiation emerges back at the surface as diffuse reflection, while another portion interacts with particles giving absorption information. The radiation attenuation with distance follows the relation outlined above, however  $\alpha$  is now interpreted as the mean absorption coefficient of the sample, with the distance  $d$  being the mean penetrated layer thickness.

Both regular and diffuse reflectance are used to determine absorption properties of a medium. However, the proportion of both parts reflected is dependent on both the instrument and the absorptive properties of the medium. In the equation of regular reflectance, it is noted that the regular reflectivity increases with increasing absorption index,  $\kappa_0$ . Meanwhile, the diffuse reflectance is exponentially attenuated after a certain extent in which the diffuse component decreases with increasing absorption<sup>47</sup>.

### 1.3.3. Hyperspectral Imaging

Hyperspectral imaging (HSI) combines conventional imaging with spectroscopy, giving two-dimensions of spatial ( $x,y$ ) and one of spectral information ( $\lambda$ ). HSI was first developed for remote sensing<sup>45,48</sup>, having greater resolution than traditional broadband Landsat scanners. The power of HSI lies in the ability to obtain a continuous spectrum

for each pixel, as opposed to multispectral imaging which measures discrete spectral bands<sup>49</sup>. In addition to its original application in satellite and airborne land observation systems, HSI has found recent applications in agriculture<sup>56,51</sup>, archaeology and art conservation<sup>51,52</sup>, medicine<sup>53,54</sup>, and forensic science<sup>11,55</sup>.

In general, HSI systems feature a light source, and wavelength dispersion devices which are coupled to area detectors. Common light sources include tungsten halogen lamps. They are suitable for their stable spectrum in the visible and infrared ranges being the most common illumination source in hyperspectral reflectance imaging<sup>56</sup>. Broadband light emitting diodes (LEDs) are an up-and-coming alternative to halogen light sources. However, despite their long life, low power consumption, and small size, they provide only narrow bands of light. Additionally, NIR LEDs are more expensive than their tungsten alternative<sup>57</sup>. Lasers, being inherently monochromatic, are most commonly used in excitation-based methods such as fluorescence and Raman spectroscopy<sup>56,57</sup>.

Due to the diversity in imaging-system spectral-ranges, resolution, types of dispersion-devices and detector-arrays, there are many possibilities of classifying HSI systems. The classification of HSI systems is generally divided in terms of acquisition of spectral and spatial information. There are two conventional methods, spatial and spectral scanning, with “snapshot” methods being mainly reserved for multispectral imaging. For spatial scanning hypercube acquisition, a complete spectrum is obtained per pixel by either a point-scanning (whiskbroom) or line-scanning (pushbroom) instrument. Spectral scanning methods are also often termed “staring”, as they capture the whole 2D image scene in a single exposure and subsequently stepping through the wavelength range<sup>54</sup>. The main trade-off between acquisition techniques is the acquisition speed, resolution, and/or signal-to-noise ratio.

While HSI systems can work in a variety of wavelengths from UV, visible and IR ranges, the majority of systems operate in reflectance mode over fluorescence and transmission modes. This bias is partly due to the presumption that sample size is too thick and thus incapable of transmission measurement<sup>54,58</sup>.



## 1.4. Complementary Statistical and Analytical Methods

### 1.4.1. Chemometrics

The use of computers in chemistry dates back to the 1970s, when analytical groups used statistical and mathematical methods using mainframe computers. Svante Wold and Bruce R. Kowalski first described chemometrics in 1972, with the first description of the discipline coinciding with the foundation of the International Chemometrics Society, two years later<sup>59</sup>. The definition is as follows: “Chemometrics is the use of mathematical and statistical methods in chemistry to (1) design optimal measurement procedures and experiments, and (2) to provide maximum chemical information by the analysis of chemical data”. Today, the analytical chemist uses various software related to processing of data and/or applying mathematical methods. Besides statistical-mathematical methods, chemometrics encompasses methods of handling spectroscopic or chemical databases and artificial intelligence<sup>59</sup>. HSI data is inherently multivariate in nature and thus multivariate tools are needed to appropriately extract information from the large number of data variables in the  $x\lambda$ -space.

### 1.4.2. Pre-processing

Physical properties of the sample, such as the degree of light scattering, as well as the influence of instrumental noise, can lead to spectral variation between samples and measurements. These variances that are not caused by the sample properties directly, need to be reduced and/or eliminated by pre-processing methods. Typical pre-processing in HSI multivariate analysis includes; smoothing, normalisation, derivatives, multiplicative scatter correction (MSC), and standard normal variate (SNV). The latter two are implemented to remove non-uniform scattering interferences contributing to the observed spectrum, while smoothing reduces spectral noise, and derivatives can separate overlapping peaks, sharpening spectral features<sup>56</sup>. The selection of pre-treatment method or combination of methods is often performed via an iterative process where the best method fulfils criteria of later data treatment methods (e.g., if standardisation or normalisation of data is required) and one that produces the most robust model.

### 1.4.3. Classification

Multivariate data analysis and statistics in chemistry are used for one of two objectives: Firstly, the modelling of relationships between sets of analytical measurements and properties is the case of calibration. In typical calibration a parameter is estimated from the representative calibration coefficients. The second is the grouping and /or classifying of objects, chemicals, or compounds by means of analytical data on the basis of a property or known class membership.

### 1.4.4. Unsupervised Learning

Unsupervised classification is the grouping of objects without known membership to the particular classes. This grouping of data is performed by either the projection of high-dimensional data onto lower dimensional space or via clustering methods. Clustering aims to divide data points into groups with similar traits and assign these groups into clusters using a given “similarity” measure. Clustering methods can be subdivided into two groups; hard-clustering or soft-clustering. The data point in question can either be assigned to a cluster or not, as is the case in hard-clustering. For soft-clustering, the data point can be given a probability to which cluster it is assigned to. The most popular clustering algorithms are k-means clustering and hierarchical clustering. K-means clustering is an iteratively finds the local maxima of clustering centroids until no further improvement is achieved. Hierarchical clustering on the other hand, assigns all data to an individual cluster before merging the nearest clusters to each other in a hierarchical fashion.

The former data grouping method, data projection, is often referred to as factorial methods and can be implemented into the multivariate workflow as feature transformation. High dimension data is projected onto a line, plane, or three-dimensional coordinate system which reduces the dimensionality and can reveal grouping when the optimal projection is found. The main methods of dimension reduction applied in chemometrics are that of principle component analysis (PCA), factor analysis (FA) and singular value decomposition (SVD)<sup>60</sup>. In the following sections,

PCA and neighbourhood component feature selection (NCFS) are described in further detail.

#### 1.4.4.1. Principal Component Analysis (PCA)

Given the original multidimensional data matrix  $X$ , which consists of  $n$  rows or objects, and  $p$  columns or features,  $X$  can be projected down a  $d$ -dimensional subspace to give the object coordinates in the plane  $T$  via the projection matrix  $L^T$ .  $T$  has  $n$  rows and  $d$  columns (number of principal components) and is called the scores matrix, while  $L$  has  $d$  columns and  $p$  rows and is the loading matrix. The  $p$  rows of loading matrix are called loading vectors — while the  $d$  columns of the scores matrix are called the score vectors — with both vectors being orthogonal  $t_i^T t_j = 0$  and  $p_i^T p_j = 0$  for  $i \neq j$ . This reconstruction of the input data results in new, uncorrelated variables. PCs are determined based on the criterion of maximum variance, and thus most of the data variance is described by the first PC, next the second PC and so on and so forth. As a large proportion of the variance is described by only a few PCs, the data is often visualised by plotting PC scores against one another. The decision of the number of PCs to use is often determined by the percentage explained cumulative variance e.g., 90 %<sup>60</sup>.

#### 1.4.4.2. Neighbourhood Component feature selection (NCFS)

Neighbourhood component feature selection (NCFS) is a neighbour-based feature weighting algorithm proposed by Yang *et al.* (2012)<sup>61</sup>. It is a non-parametric method without assumptions about the data distribution, which learns the feature weights by maximising the expected classification leave-one-out accuracy with a regularisation term. The following is an adaptation from the original 2012 paper, as described by MathWorks:

Consider a classification problem with training set  $T = \{(x_i, y_i), i = 1, 2, \dots, n\}$ , where  $x_i$  is a  $d$ -dimensional feature vector,  $y_i \in \{1, \dots, C\}$  is the corresponding class labels, and  $n$  number of samples. The aim is to determine a weighting vector  $w$  that selects the

feature subset optimizing classification. The weighted distance between two samples  $x_i$  and  $x_j$  in terms of the vector  $w$  is:

$$d_w(x_i, x_j) = \sum_{r=1}^p w_r^2 |x_{ir} - x_{jr}|$$

where  $w_r$  is an associated weight of the  $r$ -th feature. Consider a classifier that randomly picks a reference point  $Ref(x)$  from  $T$ , and labels  $x$  using the label of reference point  $Ref(x)$ . The probability  $P(Ref(x) = x_j|T)$  that point  $x_j$  is picked as reference point for  $x$  is higher the closer  $x_j$  is to  $x$  as measured by  $d_w$ . Assuming  $P(Ref(x) = x_j|T) \propto k(d_w(x_i, x_j))$  where  $k(z) = \exp(-z/\sigma)$  is a kernel function with kernel width  $\sigma$ , the sum of  $P(Ref(x) = x_j|T)$  for all  $j$  must equal 1. Hence, it can be written:

$$P(Ref(x) = x_j|T) = \frac{k(d_w(x, x_j))}{\sum_{j=1}^n k(d_w(x, x_j))}$$

The probability that point  $x_j$  is picked as reference point for  $x_i$  in a leave-one-out randomised classifier is:

$$p_{ij} = P(Ref(x_i) = x_j|T^{-i}) = \frac{k(d_w(x_i, x_j))}{\sum_{j=1, j \neq i}^n k(d_w(x_i, x_j))}$$

where the label of  $x_i$  is predicted using  $T^{-i}$ , the training set without point  $(x_i, y_i)$ .

The average leave-one-out probability of correct classification is therefore,

$$p_i = \sum_{j=1, j \neq i}^n P(Ref(x_i) = x_j|T^{-i}) I(y_i = y_j) = \sum_{j=1, j \neq i}^n p_{ij} y_{ij}$$

Where  $y_{ij} = I(y_i = y_j) = \begin{cases} 1 & \text{if } y_i = y_j \\ 0 & \text{otherwise} \end{cases}$ . The average leave-one-out probability of correct classification with regularized objective function can be written as:

$$\begin{aligned} F(w) &= \frac{1}{n} \sum_{i=1}^n p_i - \lambda \sum_{r=1}^p w_r^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left[ \sum_{j=1, j \neq i}^n p_{ij} y_{ij} - \lambda \sum_{r=1}^p w_r^2 \right] \end{aligned}$$

$$= \frac{1}{n} \sum_{i=1}^n F_i(w)$$

which depends on the weight vector  $w$  and regularization parameter  $\lambda$ . The aim of NCFS is therefore to maximise  $F(w)$  with respect to  $w$ . As  $\lambda$  drives weights in  $w$  to 0, this parameter needs to be tuned for example, using cross-validation.

#### 1.4.5. Supervised Learning

Supervised learning is a group of pattern recognition methods where object class membership is known prior to training. Common methods include linear learning machines, discriminant analysis, k-nearest neighbour (k-NN), and support vector machines (SVM). Multivariate modelling methods can also be the basis for supervised methods, e.g., partial-least squares (PLS) in the form of discriminant analysis (PLS-DA). Decision trees, SVMs and k-NN are briefly described below, with the theory further detailed in Appendix C.

##### 1.4.5.1. Decision Trees

Decision trees attempt to predict a response to a dataset by following decisions from the beginning (root) down to an end (leaf) which contains the response. Decision trees are easy to interpret due to the sequential binary partitions along the data coordinates. The theory of decision trees is further detailed in Appendix C.

##### 1.4.5.2. Support Vector Machines (SVM)

The separation of overlapping classes based on optimal separating hyperplanes (such as discriminant analysis) is not feasible. In classification, SVMs give linear boundaries between groups in a (usually) higher dimensional transformed  $x$ -variable space. In the transformed feature space, a maximum margin classifier is implemented with the back-transformed boundaries being non-linear. SVMs can be used for both regression and

classification problems<sup>62</sup>, and have increased in prominence in biology and chemistry<sup>63,64,65</sup>. Further information on SVMs is detailed in Appendix C.

#### 1.4.5.3. k-Nearest Neighbours (k-NN)

A simple way of classifying new data points is by using pairwise distance metrics to determine the distance between objects and assigning the new data point to the corresponding class of minimum distance. k-NN is often implemented as a reference method as it is conceptually simple, for its applicability to multiclass problems, and the fact that it does not require compact group clusters or linearly separable data. k-NN is further described in Appendix C.

#### 1.4.6. Bayesian Optimisation

In machine learning, model learning parameters and hyperparameters need to be tuned in order to obtain the optimal model that describes the data. Bayesian optimisation is an automated approach to hyperparameter tuning, in which the performance of a learning algorithm's generalization is modelled from a Gaussian process<sup>66,67</sup>. A global statistical model of the unknown objective function is iteratively developed using a posterior distribution treated as observations in a Bayesian nonlinear regression<sup>68</sup>.

Bayesian optimisation constructs a probabilistic model for the function  $f(x)$  of a bounded set  $X$  which it uses to determine where next to evaluate  $f(x)$  in  $X$ . As all the information from previous evaluations of  $f(x)$  are considered in subsequent evaluations, this requires more computation, with the benefit of requiring fewer evaluations, to obtain function minimization. This trade-off is acceptable in machine learning where training and evaluation is computationally expensive. Bayesian acquisition functions are described in Appendix C.

## 2. Materials and methods

### 2.1. HSI system

The commercial Specim IQ® (Specim, Spectral Imaging Ltd., Oulu, Finland) was used for the capturing of hyperspectral images. This system features a push-broom scanner producing hypercubes in the range of 400 – 1000 nm with a spectral resolution of 7 nm (204 spectral bands,  $\lambda$ -axis). The number of effective pixels is 512 x 512 pix (x-, y-axis) and the camera fore optic provides a field of view of 31 x 31 degrees. Therefore, at a distance of 30 cm between the camera and sample, a viewable area of 16.4 x 16.4 cm results in a spatial resolution of 0.32 mm. Illumination was achieved using two tungsten halogen broadband light sources (750 W each). The default recording mode (DRM) with simultaneous white reference method, in which the white reference panel is measured alongside the target sample, was used for data acquisition. An integration time of 10 ms was used giving a recording time of 35 ms per hypercube. This was verified using the built-in 'quick data validation', where the highest spectrum value of each pixel is visualised as a histogram with an indication of under-saturation (pixel intensity < 30 %) or over-saturation (maximum pixel intensity). The reflectance transformation is calculated from 3 measured datacubes: The raw datacube of light intensities measured, the dark frame — which is the sensor baseline signal due to the camera electronics, and the white reference. The white reference is assumed to contain only the signal from the illumination given the same measurement geometry, distance and illumination as the sample measured. The reflectance datacube is calculated using the relation:

$$R_{ij}(\lambda) = \frac{RAW_{ij} - Dark_{ij}}{White_{ij}}$$

where  $R$  is the reflectance,  $RAW$  is the raw datacube,  $Dark$  is the instrument dark frame,  $White$  is the white reference plate intensity, and  $i$  and  $j$  are horizontal and vertical pixel indices. To reduce the interference, all external light sources — including room lights — were switched-off during image recording. To avoid heat damage of the samples the halogen lights were switched-off in between measurements.

## 2.2. Human and animal dataset

Venous blood from 20 healthy human volunteers (10 male, 10 female, age  $42 \pm 16$  years) was obtained from the Institut für Transfusionmedizin, Universitätsklinikum Leipzig. The 1.7 mL blood aliquots were collected into EDTA and refrigerated to avoid coagulation prior to measurement. The samples were transported to the Institut für Rechtsmedizin, Universitätsklinikum Leipzig where the blood was then deposited onto white cotton fabric creating a spot of ca.  $5 \text{ cm}^2$  which was let dry at room temperature for 10 minutes. Samples that were stored under refrigeration were allowed to warm to room temperature prior to probe preparation in order to remove temperature dependence, if any, between samples. The hyperspectral image was recorded using the SPECIM IQ® camera under halogen light. The samples were left exposed under ambient conditions and recorded once daily for a week, and then intermittently up to 32 days.

Blood from 20 pigs was obtained from Schlachthof Weißenfels, Weißenfels and transported to the Institut für Rechtsmedizin, Universitätsklinikum Leipzig in EDTA 1.7 mL aliquots. The blood was deposited onto white cotton and let dry for 10 minutes. Varying degrees of coagulation were noticed in 11 of the aliquots which did not resuspend upon inversion. For these samples, the jelly-like mass was not deposited onto the cotton and the stains appeared lighter. As above, the HSI images were measured daily up to a week and then intermittently up to 42 days.

Venous blood from 20 cows was obtained from the Klinik für Klauentiere, Veterinärmedizinische Fakultät, Universität Leipzig. The 1.7 mL EDTA aliquots were transported to the Institut für Rechtsmedizin, Universitätsklinikum Leipzig, spotted onto white cotton and let dry as previously mentioned. RBCs could be resuspended by gentle inversion and no major coagulation was observed. HSI images were measured daily for 1 week and then intermittently up to 42 days.

Cardiac blood from 16 mice (5 female CD1/CR, 5 male CD1, 6 female Sv129), was obtained from the Medizinisch-Experimentelles Zentrum III, Universitätsklinikum Leipzig. The  $<1$  mL samples were transported to the Institut für Rechtsmedizin, Universitätsklinikum Leipzig, and spotted on to white cotton giving ca.  $3 \text{ cm}^2$  spots. HSI images were measured daily for the first week and then intermittently up to 24 days.



Venous blood from 5 rats (3 female, 2 male SPRD) and 5 rabbits (1 female, 1 male White New Zealand; 1 male, 2 female Chinchilla bastard) was obtained from Medizinisch-Experimentelles Zentrum I, Universitätsklinikum Leipzig. The 1.7 mL EDTA aliquots were transported to the Institut für Rechtsmedizin, Universitätsklinikum Leipzig and ca. 5 cm<sup>2</sup> spots were made on white cotton. HSI images were measured daily for 1 week and then intermittently up to 49 days.

The unbalanced data (Table 3) obtained from the ROIs of the corresponding HSI images was divided into bins that follow the natural exponential series  $f(x) = e^x$  where  $x = 0, 1, 2, 3$  and  $f(x) = 1, 2.718, 7.389, 20.086$ . This gives the time interval bins: 0.0-1.0, 1.0-3.0, 3.0-7.0, 7.0-20.0, and 20.0-55.0. This was in order to capture the exponential-like decrease in haemoglobin derivatives HbO<sub>2</sub>, and increase in metHb and HC with respect to degradation over time<sup>11</sup>.

Table 2. Human and animal dataset.

	<b>Number Total (Male/Female)</b>	<b>Age Mean ±Stdev</b>	<b>Age Range</b>
Human	20 (10/10)	42 ±16 years	20 – 68 years
Pig	20 (-/-)	6 months*	-
Mouse	16 (5/11)	108 days	59 – 204 days
Rat	5 (3/2)	87 days	42 - 162 days
Rabbit	5 (2/3)	1.7 ±0.9 years	0.9 – 3.2 years
Cow	20 (-/-)	-	-

Table 3. Total dataset number of recorded HSI images and spectra.

	Sample No.	Age [Days]	HSI Images	No. ROIs	Tot. Spectra
<b>Human</b>	20	0.1 – 32	21	103	231,075
<b>Animal</b>	66	0.1 – 49	59	250	454,450
Pig	20	0.1 – 42	8	83	180,000
Mouse	16	0.1 – 24	18	64	45,075
Rat	5	0.1 – 49	12	18	35,750
Rabbit	5	0.1 – 49	12	18	50,625
Cow	20	0.1 – 42	9	67	145,000

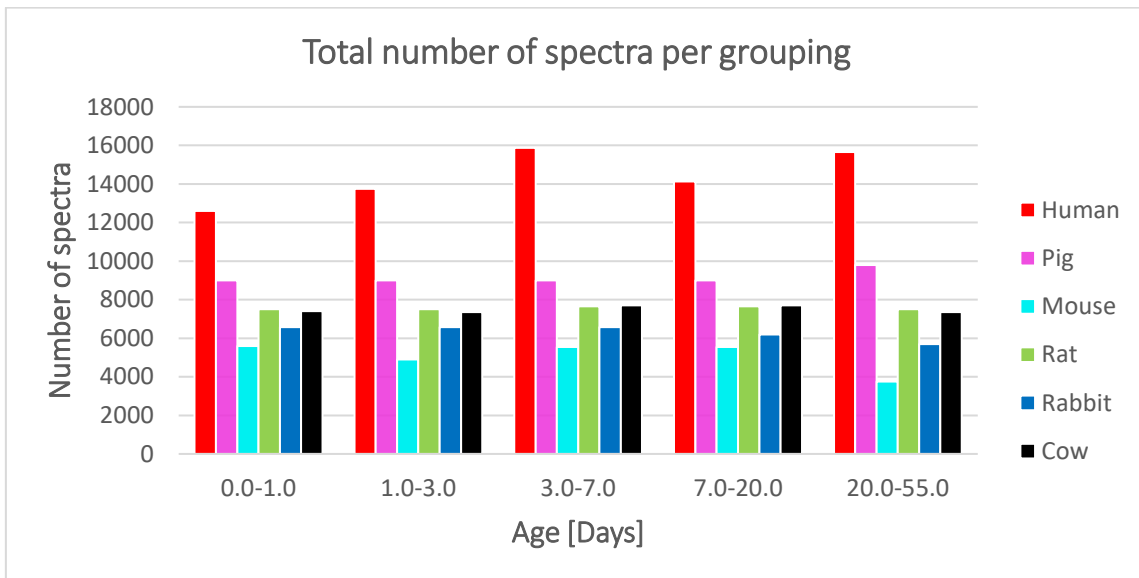


Figure 5. Comparative bar plot of unbalanced data with respect to species and binning.

### 2.3. Classification framework

The aim of the classification was the automatic identification and segmentation of blood with respect to human and animal blood using HSI. HSI data cubes were annotated using GIMP (The GIMP Development Team, 2019) to create regions of interest (ROI) of the blood samples consisting of approximately 25 x 25 pixels (625 spectra). Microsoft Excel (Microsoft Corporation 2019) was used for the documentation and analysis of results. Data balancing and data pre-processing were performed using custom scripts written in

MATLAB (version 9.8; R2020a, The MathWorks Inc.), with machine learning algorithm selection and optimisation being performed using the Statistics and Machine Learning Toolbox™, and Imaging Processing Toolbox™, both being provided by MATLAB. The data treatment is detailed in the following sections.

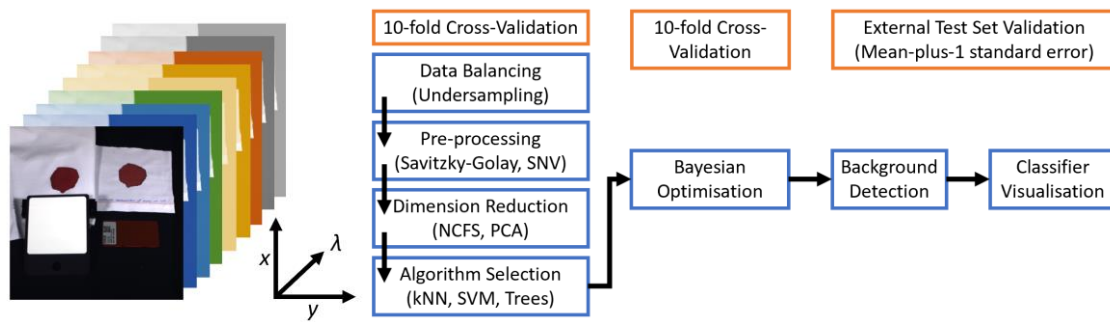


Figure 6. Classification framework for HSI data cubes of human and animal bloodstains.

### 2.3.1. Pre-processing

The spectral window of 400 – 1000 nm (204 spectral bands) was initially truncated to 435 – 965 nm (178 spectral bands) due to the low camera sensitivity and low power of the halogen light source beyond this range<sup>12</sup>. A Savitzky-Golay filter<sup>69</sup> with polynomial order of 2 and window length of 9 spectral bands was implemented to smooth the reflectance spectra giving an effective wavelength range of 445 – 955 nm (169 spectral bands). This filter uses linear least squares to fit a polynomial to successive windows of adjacent data points. Spectra were then normalised using the SNV transform<sup>70</sup>, which auto-scales the data giving a mean reflectance of zero and standard deviation of one. The SNV reflectance spectra were then used in the training of classification models, including feature selection, which is further detailed later sections.

Second-order derivative spectra were generated from the human and animal spectra using an additional Savitzky-Golay filter, in order to resolve undefined broad peaks and examine differences between datasets. Additionally, PCA was performed using the SNV spectra, and visualised via scores and loading plots of the first 3 principal components. Both the second-order derivatives and PC plots were used to investigate and rationalise the observed differences between human and animal reflectance spectra.

### 2.3.2. Background detection

Prior to image classification a decision tree algorithm for background detection was implemented. The Algorithm Background Detection<sup>71</sup> (ABG) was modified to exclude all background except for the blood stains:

```
%Input: Set of reflectance data as XxYxM matrix
%R = (x, y, m) %with X is the width, Y is the height,
%and M is the number of wavelengths of the HSI Cube
%Output: Set of pixel data as XxY matrix P = p(x,y)

function p=ABD(R)
p = zeros (512,512);
for x = 1:512
    for y = 1:512
        A = mean(R(x,y,m),3); %with 400 nm <= m <= 1000 nm
        B = mean(R(x,y,m),3); %with 510 nm <= m <= 600 nm
        C = mean(R(x,y,m),3)/B; %with 450 nm <= m <= 770 nm
        if A < 0.6 & B > 0.05 & C > 1.7
            p(x,y) = 0; %sample
        else
            p(x,y) = 1; %background
        end
    end
end
end
```

Figure 7. Algorithm to detect background prior to classification.

The algorithm parameters  $A$ ,  $B$ , and  $C$ , attempt to exclude spectra that have features that are not characteristic to blood and blood stains. Parameter  $A$  removes constant high-valued reflectance spectra, such as the white cloth deposition surface. Parameter  $B$  is based on the Q-bands of haem, which are predominant low-reflectance in spectra of blood. Parameter  $C$  attempts to capture the proportionality between the low-reflectance region of the haem Q-bands, and the high-reflectance far-red region of a typical blood reflectance spectrum. The parameter threshold values  $A < 0.6$ ,  $B > 0.05$ , and  $C > 1.7$ , were empirically determined as the optimum values for background detection.

The image morphological noise removal technique “closing” was performed on the ABD binary gradient mask by dilation using a vertical structuring element followed by a horizontal structuring element and subsequent erosion using a diamond structuring element.

### 2.3.3. Class Balancing

The species classes include different numbers of individuals and recorded HSI images which in turn lead to drastically different numbers of spectra per class (Table 3). To build an unbiased classification model, the data needs to be distributed evenly with respect to classes. This process of 'balancing' the data usually follows one of two forms of resampling: Oversampling is the addition of copies of data from the minority class(es) to artificially increase the class size to equal that of the majority class. This method is preferred for instances of little data, but can produce overfitted models due to the duplicated observations. Undersampling is the removal of observations from the majority class to decrease its size to be comparable to that of the minority class. Undersampling is generally suited for large data sets, as this method removes information that could lead to an underfitted model or a model that generalizes poorly. For this reason, undersampling was chosen as the resampling method for the balancing of the bloodstain data. This was achieved by randomly excluding spectra of a given class, so that it correlates with the number of spectra of the minority class.

In addition, to build a classification model that is independent of sample age, the data must be logically distributed to equally represent changes in the blood spectra with time. As the effect of time on blood composition is the most explored method in age determination of blood stains<sup>11,28,72</sup>, the effect of time on the absorption ratio of blood components is well documented. To capture the exponential-like decrease in HbO<sub>2</sub> and the contrary increase in metHb and HC with time, blood samples were binned into groups of 0.1 - 1 days, 1 - 3 days, 3 - 7 days, 7 - 20 days and 20 - 49 days. Within each species class, the age-group bins were first balanced using undersampling to that of the least represented bin. The animal classes were then balanced to each other, so each animal was equally represented, before balancing to the human class by the random exclusion of animal observations, maintaining age group distribution. Human observations were assigned the binary response of 1 while all animal observations were assigned the binary response of 0 thus forming the two classes.

Table 4. Number of species spectrum before and after underbalancing was implemented. The class were balanced initially by age bins, followed by balancing by species.

	<b>No.</b>	<b>Spectra</b>	<b>Balanced</b>
<b>Human</b>	20	121 535	100 135
<b>Animal</b>	66	187 750	100 135
Pig	20	59 950	20 027
Mouse	16	25 425	20 027
Rat	5	31 050	20 027
Rabbit	5	25 325	20 027
Cow	20	46 000	20 027

#### 2.3.4. Feature Selection

Neighbourhood component feature selection (NCFS)<sup>61</sup> was implemented to reduce data dimensionality and identify regions of interest in the blood spectra that contribute to the successful discrimination of human and animal blood. The stochastic gradient descent (SGD) solver algorithm, with solver-batch size of 1,000 observations, was used to estimate feature weights. The initial learning rates were tuned with a subset size of 10,000 observations. The best value for the regularization parameter  $\lambda$  that minimizes the generalisation error is expected to be a multiple of the inverse of the number of observations  $n$ . 10-fold cross-validation was thus used to tune  $\lambda$  for feature selection to find the average minimum loss value of the folds. Given the large number of observations in the training set (144,020 spectra), the expected value is  $\lambda = 6.943 \times 10^{-6}$  and therefore can be approximated as zero. Without a regularisation parameter, all features have a weight greater than 0, and therefore a feature weight threshold was implemented. Thresholds of 0.5, 0.55, 0.6, 0.65, 0.70, and 0.75 times the maximum feature weight were used to select 92, 68, 43, 34, 26, and 16 of the most important wavelengths as determined by the NCFS algorithm. These features were then used to train simple k-NN classifiers ( $k = 1$ ) of the training set, which were evaluated based on their F1 score. The F1 score is the quotient of 2 times the product of the recall and precision divided by the sum of recall and precision and arguably captures the model's performance better than the accuracy, recall and precision values individually.

### 2.3.5. Classification

The data was split into training and test sets as per Table 5 with approximately 20-25% of individuals removed from each species class to build the test set. This was further randomly divided into 66% validation and 34% test set to be used in the validation of optimised models and the final chosen model respectively.

Table 5. Training set and test set partition with balancing.

	No.	Training Set	Spectra	Test Set	Spectra
<b>Human</b>	20	15	72010	5	28125
<b>Animal</b>	66	53	72010	13	28125
Pig	20	16	14402	4	5625
Mouse	16	13	14402	3	5625
Rat	5	4	14402	1	5625
Rabbit	5	4	14402	1	5625
Cow	20	16	14402	4	5625

Five binary classification algorithms were initially tested using 10-fold cross-validation: SVM with polynomial and gaussian kernels, decision tree, bagged tree, and k-NN. The models were assessed based on their F1 scores and AUC curves, and the SVM with polynomial kernel, bagged tree, and k-NN were selected for further optimisation. Bayesian optimisation was implemented, which attempts to minimise an objective function  $f(x)$  for  $x$  by using an acquisition function  $a(x)$  to determine the next hyperparameter point for evaluation. The acquisition function “expected-improvement-per-second-plus” was used to evaluate the goodness of fit<sup>66,68</sup>.

30 iterations were used to evaluate the models with Bayesian optimisation. For the SVM the kernel scale and box constraint hyperparameters were simultaneously optimised. The distance metric and number of neighbours were optimised for the k-NN model, while the bagged decision tree was optimised based on number of learning cycles and number of leaves within the tree. The optimised models were then tested using the validation data and the F1 scores compared. The polynomial-SVM was selected and the

degree of overfitting estimated using the training error, 10-fold cross-validation error, and validation error of the 9 'best' iterations as determined by the Bayesian optimisation algorithm. The optimal SVM model was selected based on the mean-plus-1 standard error of the smallest mean CV error.



### 3. Results

#### 3.1. Training and Test Set spectra

In the age estimation of bloodstains using spectroscopic methods, the change in blood composition is the most studied method<sup>28</sup>. The  $\text{HbO}_2$  in blood is degraded to metHb and finally HC and the ratio in the values of Q-bands of the haemoglobin derivatives follows an exponential-like decrease<sup>11</sup> with respect to time. Thus, haemoglobin is at its highest concentration within the “fresh” blood samples and by extension the associated spectrum undergoes the greatest change in the first hours of exposure to the environment. This is evident in the measured HSI images (see Figure 22-27) where the drastic change in colour from bright red (almost all haemoglobin is  $\text{HbO}_2$ ) to brown (mostly metHb and HC) can be seen.

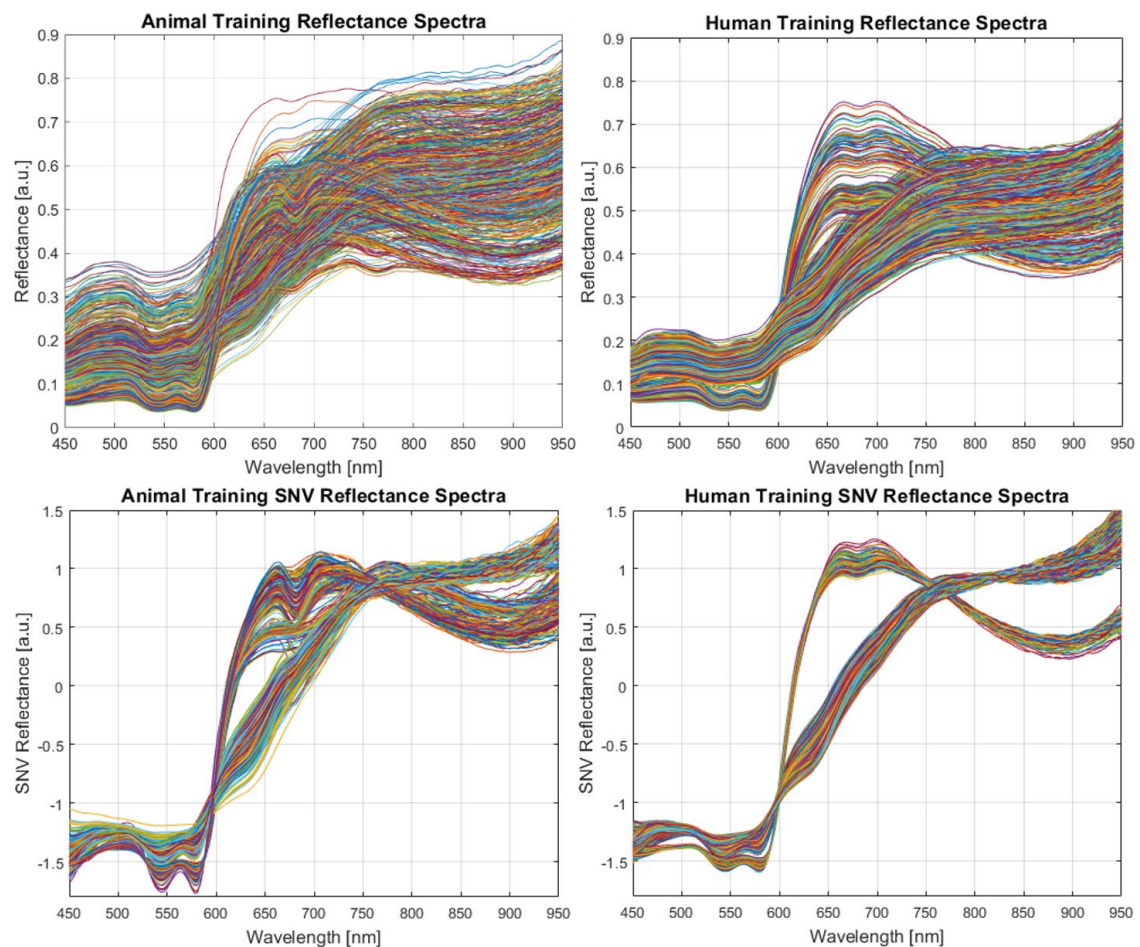


Figure 8. Reflectance spectra of the training dataset of a) animal and b) human blood on cotton of all ages (day 0.1 – 40) with Savitzky-Golay smoothing and SNV transformed spectra of c) animal and d) human blood.

Figure 8a and Figure 8b shows the reflectance spectra of animal and human blood on white cotton. The animal blood spectra contain blood spectra from pig, mouse, rat, rabbit and cow of age 0.1 days to 49 days old. The human spectrum contains spectra of training set individuals aged over a period of 32 days. The SNV transformed spectra of animal and human blood is presented in Figure 8c and Figure 8d respectively. The SNV transformation is similar to multiplicative scattering correction (MSC) in that the multiplicative interferences of light and particle scattering are corrected. After SVN transformation, two distinct groups of spectra are visible in the animal spectra and two groups in the human spectra. These correspond to “fresh” and “old” blood stains as the blood spectrum changes in composition of HbO<sub>2</sub>, MetHb and HC with respect to time.

### 3.2. Feature Selection using NCFS algorithm

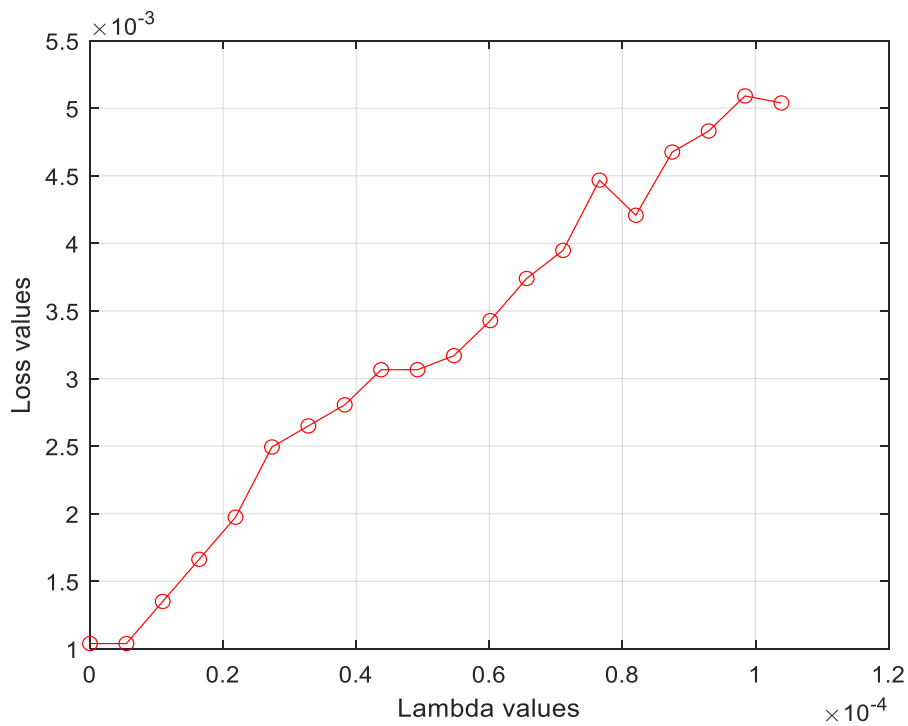


Figure 9. Average 10-fold loss vs. incremental lambda values of NCFS models.

Figure 9 above shows the average loss of 10-fold cross-validation of NCFS models using the training dataset. The general trend of increasing loss value with increasing lambda value is observed. The  $\lambda$ -value with the minimum loss value is very close to 0 which

correlates to the expected ideal value of  $\lambda = 6.943 \times 10^{-6}$  for  $n = 144,020$  observations. The regularisation parameter was therefore set as zero before refitting the NCFS model to the training dataset. The lack of a value for the regularisation parameter results in no forcing of features weights to zero, and therefore by definition, all features weights are considered of some degree of importance (feature weight  $> 0$ ). This has the additional effect of illustrating the correlation between wavelengths in a given spectra as shown in below.

To identify the regions of greatest interest as determined by the NCFS algorithm in the typical blood spectrum, the average SNV reflectance human blood spectrum was plotted as a secondary axis (black) to the feature weights (red). Threshold values of 50% to 80% in increments of 5% of the feature of maximum weight (feature 80, weight = 26.12) were used to select the most important wavelengths for use in further classification. Six k-NN models where  $k = 1$  were trained using the threshold values including and the model statistics compared to a k-NN classification model without feature selection (threshold = 0, 170 wavelengths).

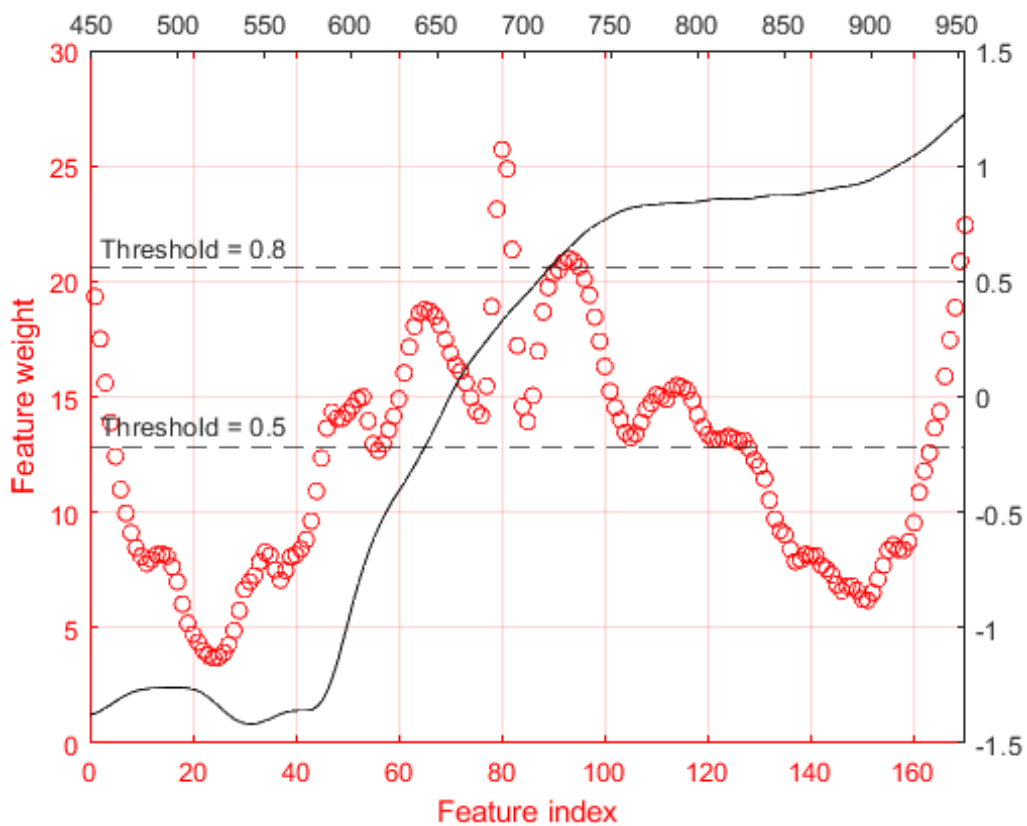


Figure 10. NCFS feature weights as a function of feature index (red circles) with threshold lines of 50% and 80% the maximum feature index weight. The average human blood spectrum is plotted as a secondary axis (black; x-axis wavelength [nm], y-axis SNV reflectance [a.u.]) to guide the eye.

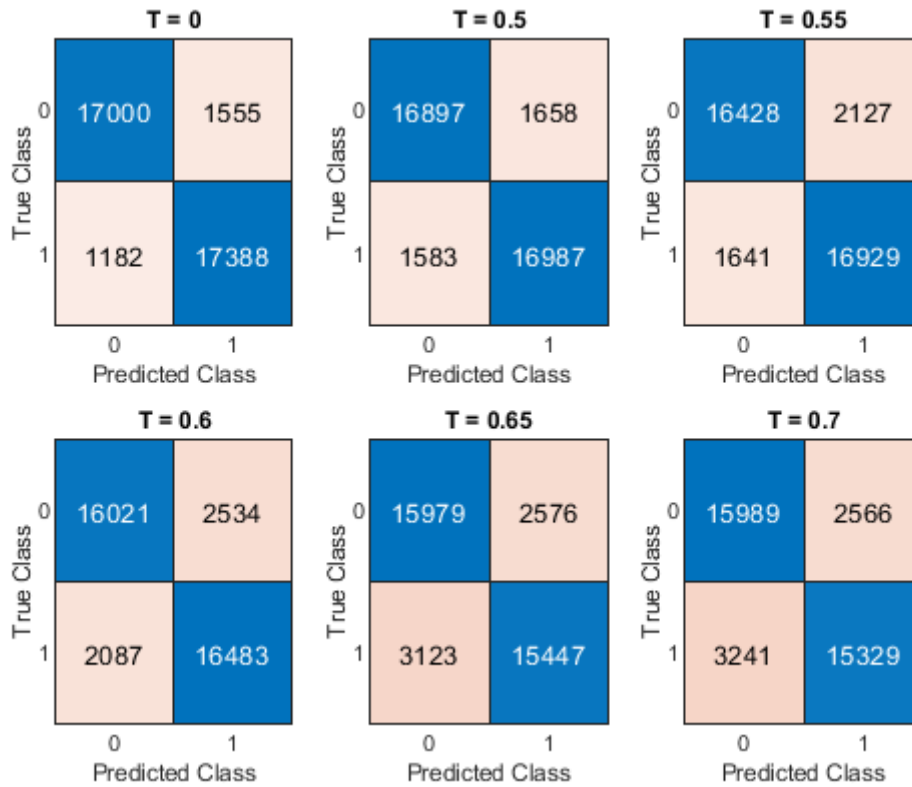


Figure 11. Confusion charts of trained  $k$ -NN ( $k=1$ ) classifiers with threshold values  $T = 0$  (full spectrum),  $T = 0.5$  (92 wavelengths),  $T = 0.55$  (68 wavelengths),  $T = 0.6$  (43 wavelengths),  $T = 0.65$  (34 wavelengths),  $T = 0.7$  (26 wavelengths).

The confusion charts for the  $k$ -NN ( $k = 1$ ) models above illustrate effect of information loss with fewer features resulting in more misclassifications generally. This is reflected in the F1 scores presented in Table 6 below. Comparing to model 1 without feature selection, there is a 2.8% decrease in F1 score with model 3 which has a feature weight threshold of 0.55. Model 4 has a relative decrease of 2.3% compared to model 3. In addition, model 3 reduces the data dimensionality by 60% from 170 features to 68 features, which is 15% greater than model 2 of 92 features. Therefore, the optimal threshold value of 0.55 with 68 features (wavelengths) was selected for use in further classification model development, as this is the best trade-off between data reduction and information retention.

Table 6. NCFS  $k = 1$ ,  $k$ -NN model F1 scores to determine the optimal number of features based on the prediction of the validation set. Model no. 3 with threshold of 0.55 and 68 wavelengths was chosen as optimal number of features.

Model No.	Threshold	No. Features	F1 score
1	0	170	0.9255
2	0.50	92	0.9125
<b>3</b>	<b>0.55</b>	<b>68</b>	<b>0.8971</b>
4	0.60	43	0.8740
5	0.65	34	0.8487
6	0.70	26	0.8463

### 3.3. Classification development using Bayesian Optimisation

Bayesian optimisation was used to optimise hyperparameters of  $k$ -NN, bagged tree, and SVM models with 10-fold cross-validation. The acquisition function “expected-improvement-per-second-plus” was used to evaluate the next hyperparameter point for evaluation. The  $k$ -NN model was optimised in terms of distance metric and the number of neighbours ranging from 2-10. This range of neighbours was chosen as  $k$ -NN models with  $k = 1$  tend to overfit, and models with large values of  $k$  are prone to underfitting. Figure 12 below of the Bayesian optimisation objective function model for 30 iterations of  $k$ -NN models shows the large differences in the estimated objective function value with respect to distance metric. This is in contrast to the number of neighbours which shows minimal variation within each distance metric. The Spearman, Mahalanobis, Jaccard, and Hamming distances give the poorest estimated objective function values of the distance metrics.

Figure 13 of the minimum objective versus the number function evaluations shows the rapid convergence of  $k$ -NN models over iterations. The best observed feasible point determined by the Bayesian optimisation was the Euclidean 5-NN model with observed objective function value of 0.0118 and estimated objective function value of 0.0122. The best estimated feasible point (according to models) was the Euclidean 9-NN model with an estimated function value of 0.0122.

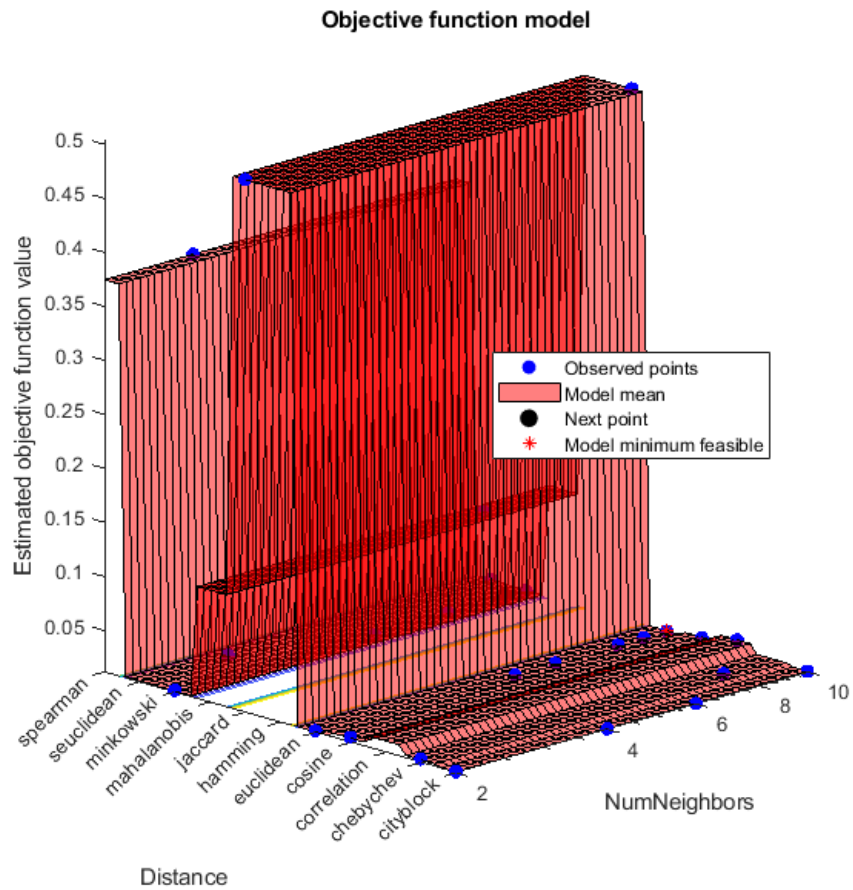


Figure 12. Bayesian optimisation of  $k$ -NN model hyperparameters distance metric and number of neighbours as a function of estimated objective function value after 30 iterations.

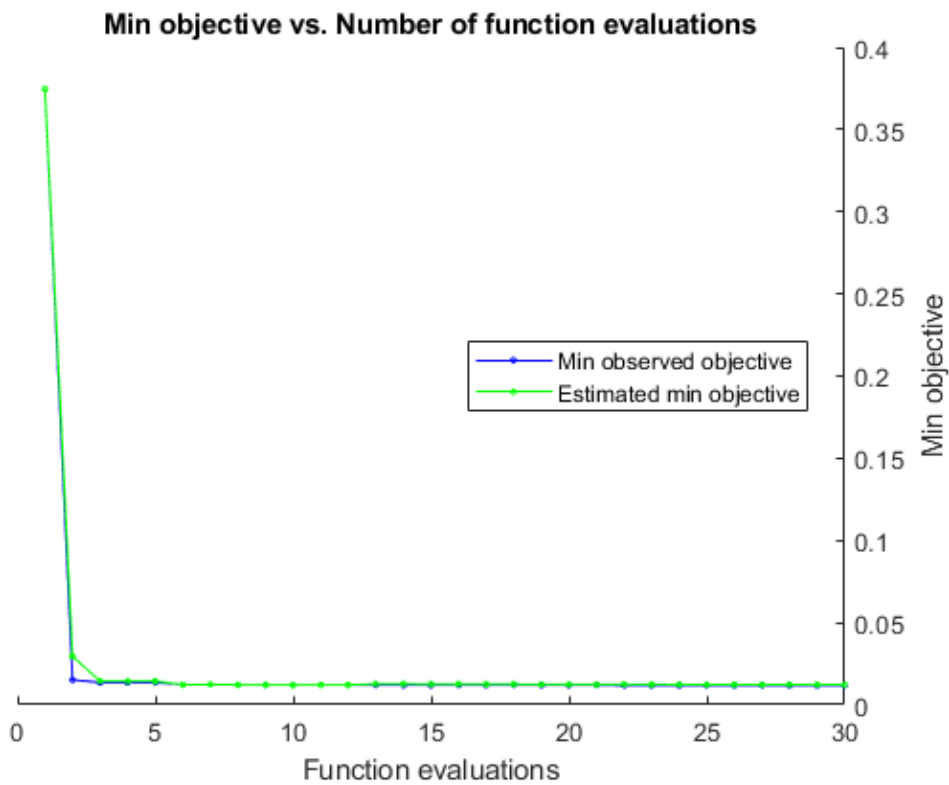


Figure 13. Minimum objective values of 30 Bayesian optimisation iterations of trained  $k$ -NN models using 10-fold cross-validation.

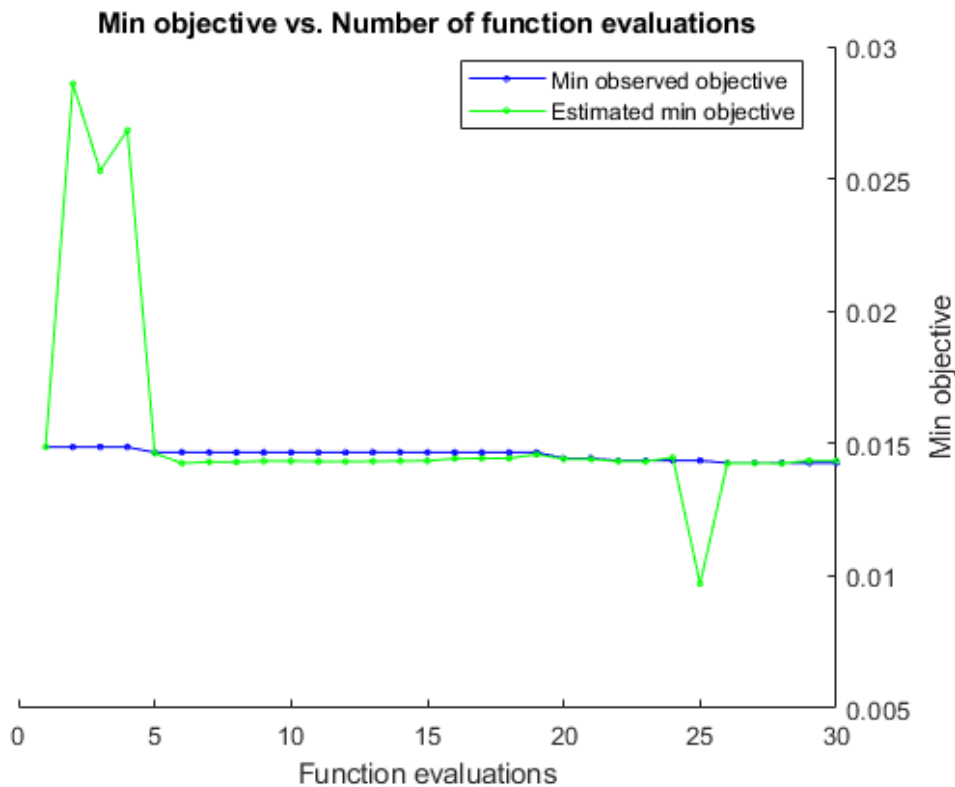


Figure 14. Minimum objective values of 30 Bayesian optimisation iterations of trained bagged trees using 10-fold cross-validation.

The bagged tree models were optimised over 30 iterations using Bayesian optimisation to evaluate the optimal number of learning cycles. The minimum objective versus number of function evaluations plot (Figure 14) shows minimal change in the minimum objective with increasing function evaluation. The best observed feasible point had 476 learning cycles with an observed function value of 0.0143 and estimated objective function value of 0.0143. The best estimated feasible point (according to models) is 0.0143.

The objective function model of the 3<sup>rd</sup>-order polynomial SVM (Figure 15) shows the Bayesian optimisation hyperspace in terms of kernel scale, box constraint and estimated objective function value after 21 iterations. The optimisation was set to run 30 iterations but was prematurely terminated for reaching a total objective function evaluation time of 150,000 seconds (41.67 hours).

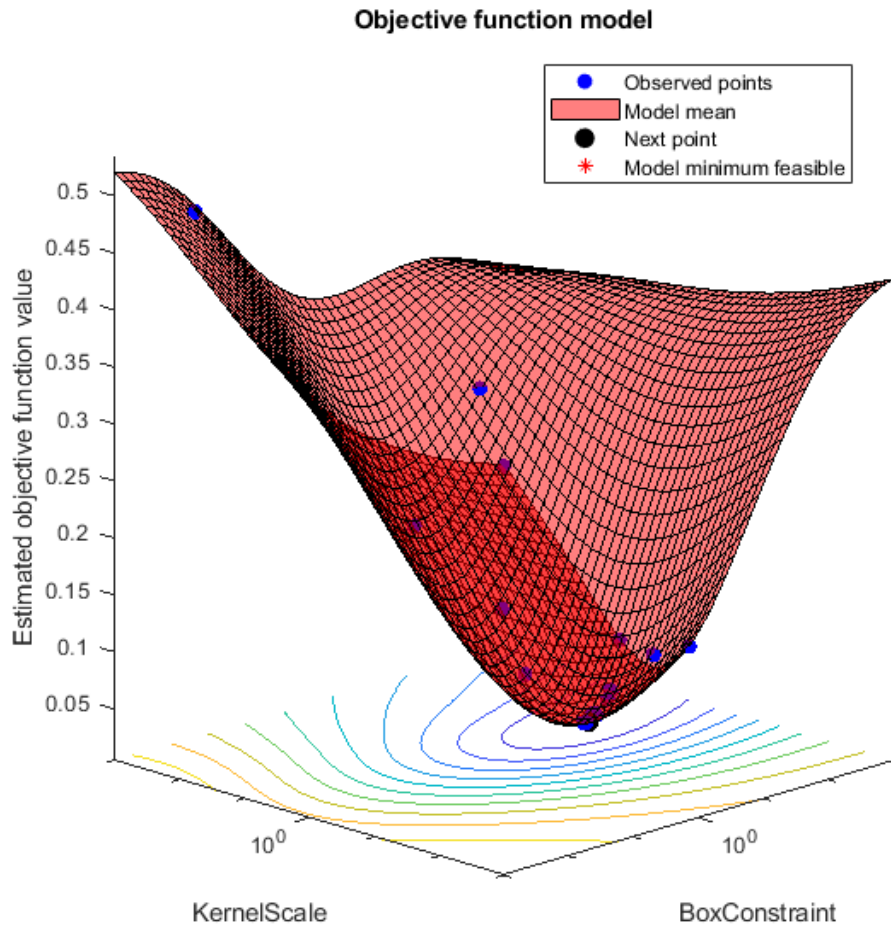


Figure 15. Bayesian optimisation of SVM hyperparameters kernel scale and box constraint as a function of estimated objective function value after 21 iterations.

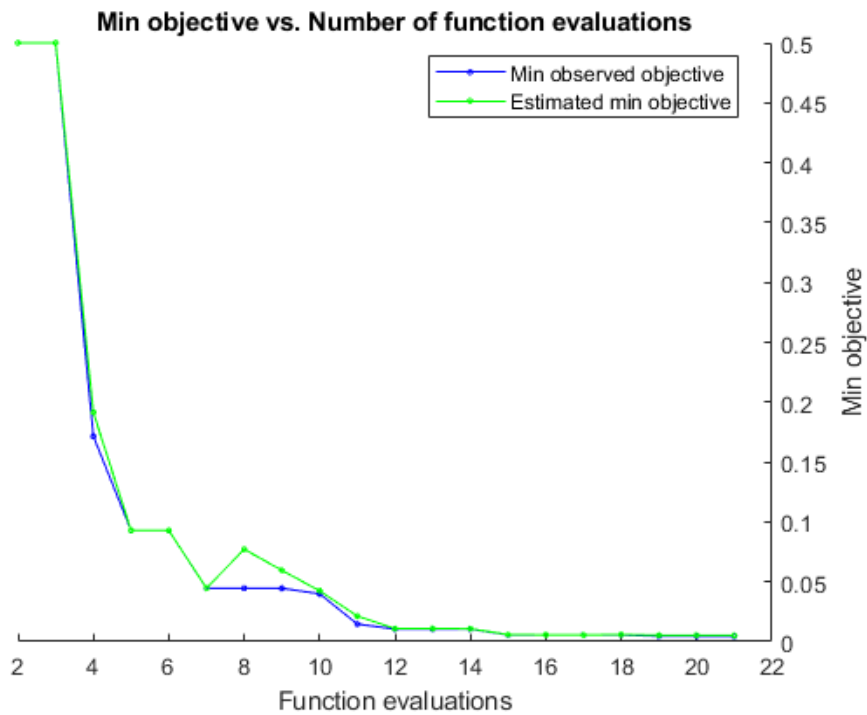


Figure 16. Minimum objective values of 21 Bayesian optimisation iterations of trained SVMs using 10-fold cross-validation.



Table 7. Bayesian optimised 3rd order polynomial SVM iterations with parameters evaluated as "best" based on objective function of 10-fold cross-validation. The iteration 16 was selected as the optimal SVM model using the mean-plus-1 standard error of the cross-validation error.

Iteration No.	Objective	Runtime (seconds)	Best so far (observed)	Best so far (estimated)	Box Constraint	Kernel Scale
2	0.4999	623.64	n/a	0.4999	0.0016	89.33
4	0.1716	464.94	0.1716	0.1915	2.8907	68.78
5	0.0929	274.83	0.0929	0.0929	20.072	21.66
7	0.0446	193.31	0.0446	0.0447	16.326	7.000
10	0.0399	318.34	0.0399	0.0242	967.70	16.03
12	0.0105	1689.7	0.0105	0.0109	162.17	4.116
<b>16</b>	<b>0.0057</b>	<b>3751.4</b>	<b>0.0057</b>	<b>0.0057</b>	<b>74.185</b>	<b>2.756</b>
19	0.0047	4023.6	0.0047	0.0054	39.554	2.239

Table 7 contains the SVM iterations of the Bayesian optimisation that were determined as "best" based on the calculated objective function for that iteration. The training time (Runtime) in seconds, and values of the optimised hyperparameters box constrain and kernel scale are also presented. Using these iteration parameters, new SVM were trained with 10-fold cross-validation, and the training loss, validation loss, and average 10-fold cross-validation loss with standard deviation, were plotted. The training loss is the misclassification error for the training set objects fitted within the 10-fold cross-validation. The 10-fold cross-validation loss is the misclassification error of the cross-validation evaluation. That is, the error of each of the 10 evaluation sets using the 9 other training sets. The mean of the 10 errors and their standard deviation are plotted. The validation loss is the misclassification error of the independent validation set not used in model training. The optimal parameters for the SVM model are determined from the mean-plus-1 standard error for the smallest mean CV error. This is determined as iteration 16 with training, validation, and 10-fold CV loss of 0.0032, 0.0441, and 0.0034 ( $\pm 6.2 \times 10^{-3}$ ) respectively.

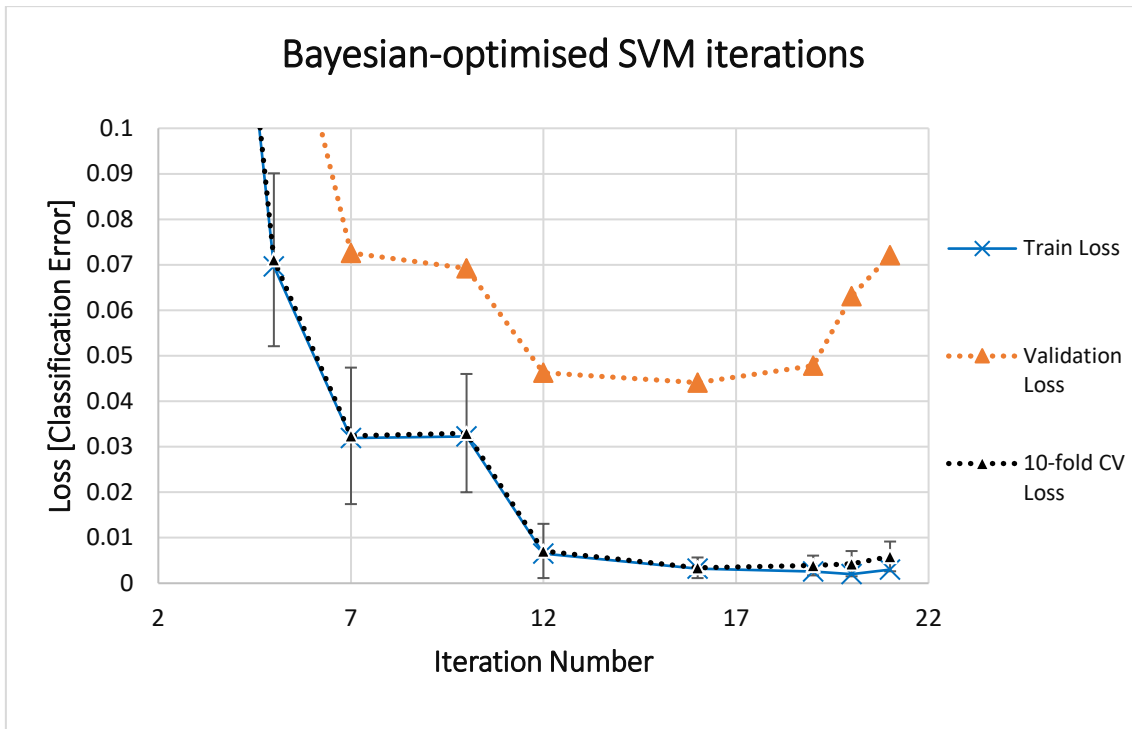


Figure 17. Comparison of training loss, CV-loss and validation loss of Bayesian-optimised SVM iterations. Based on mean plus 1 standard error, the parameters of iteration 16 are chosen as optimum for the SVM classification model.

For the final SVM binary classification model the parameters of box constraint = 74.185 and kernel scale = 2.7558 were used as per iteration 16 of the Bayesian optimisation (Table 7). The retrained model was tested with the independent test set (19,125 observations) and the model statistics calculated. Presented below are the confusion chart and model statistics of the predicted test data. This model has an accuracy and F1 score above 95% in the discrimination of human and animal blood stains up to 32 and 49 days respectively.

Table 8. Model statistics of SVM binary classifier. TOT = total number of observations predicted, PREV = prevalence; a measure of class distribution, ACC = accuracy, PPV = positive predictive value or precision, TPR = true positive rate or sensitivity, TNR = true negative rate or specificity.

TOT	PREV	ACC	PPV	TPR	TNR	F1 score
19125	0.4996	0.9564	0.9671	0.9470	0.9663	0.9569

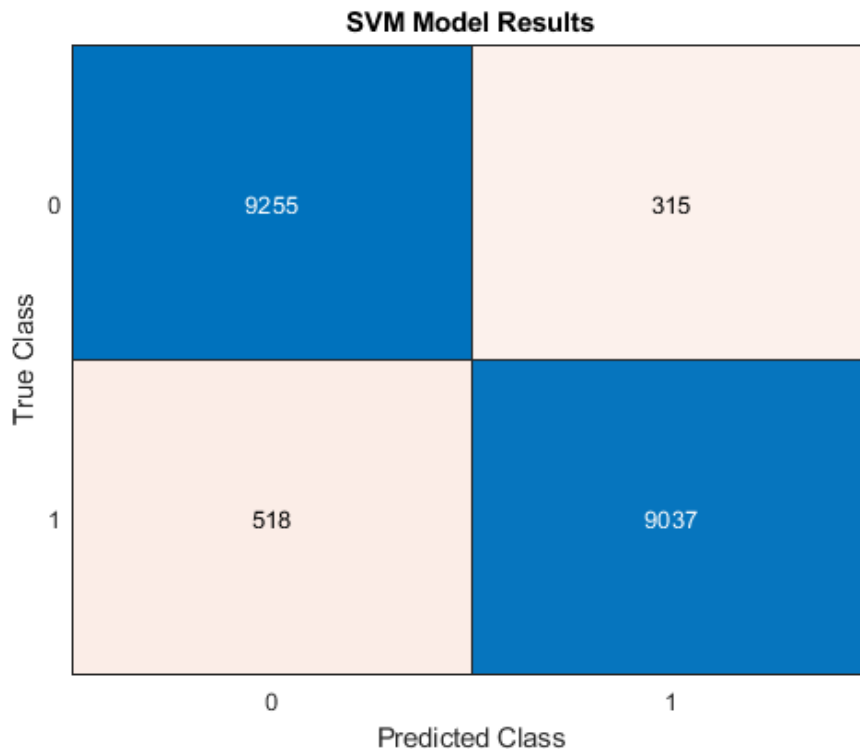


Figure 18. Confusion chart of SVM prediction results of test data.

### 3.4. Automatic Background Detection of bloodstain images

The ABD algorithm was empirically modified for the identification of animal and human blood stains on white cotton. The typical reflectance spectra of background and sample tissue were visualised using an HSI cube containing both human and animal blood at different ages (see Figure 19).

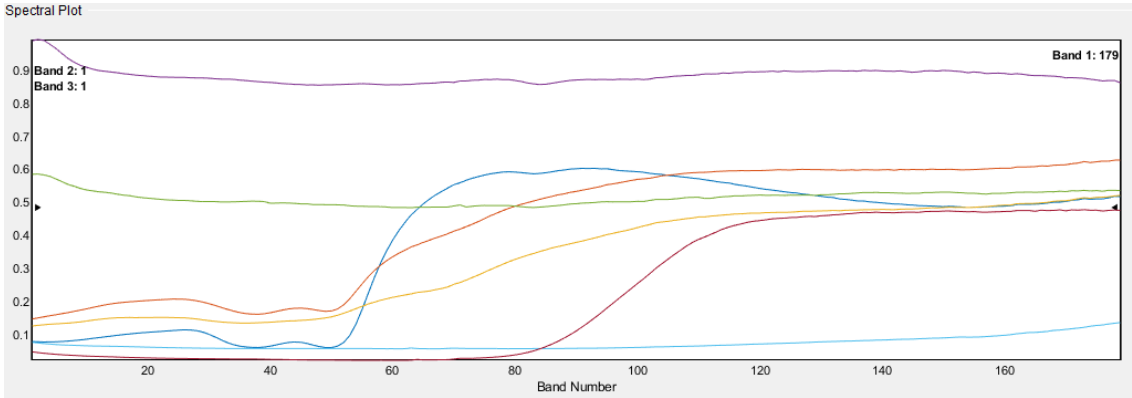
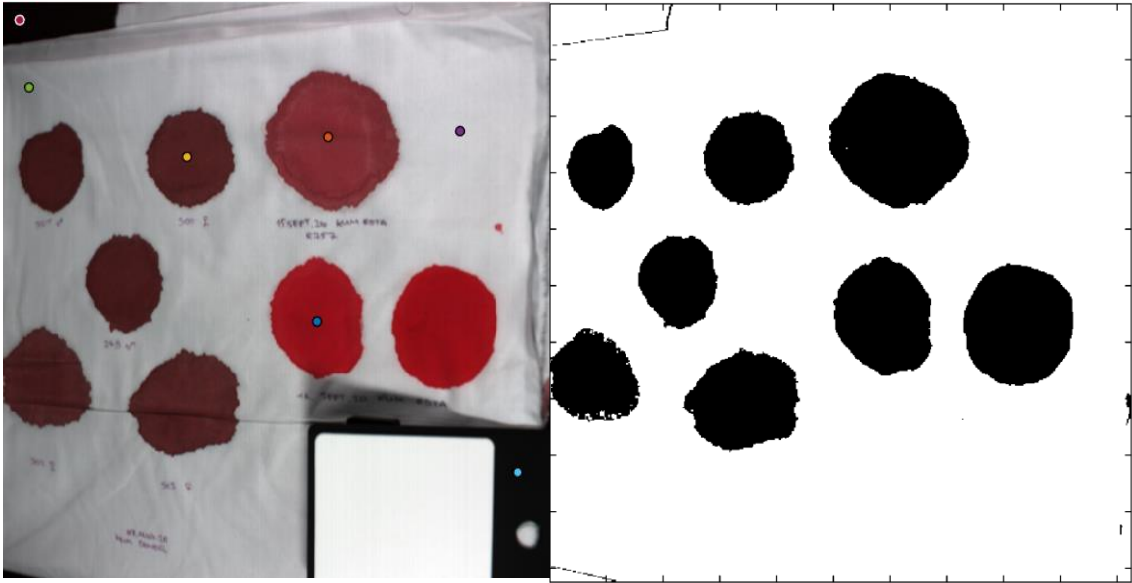
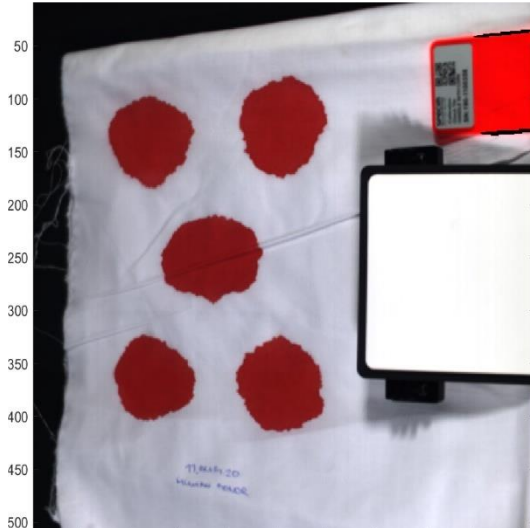


Figure 19. RGB image of human (day 40), and cow (day 1 and day 0.1) blood stains on white cotton (top left) with coloured markers and corresponding reflectance spectra (bottom) which were used in the determination of the ABD thresholds that generate the binary background detection image (top right). Averaged spectra in order of maximum reflectance at band 1: purple = white cotton (high illumination), green = white cotton (low illumination), orange (cow blood - day 1), yellow (human blood - day 40), blue (cow blood - day 0.1), light blue (black metal), red (black cloth).

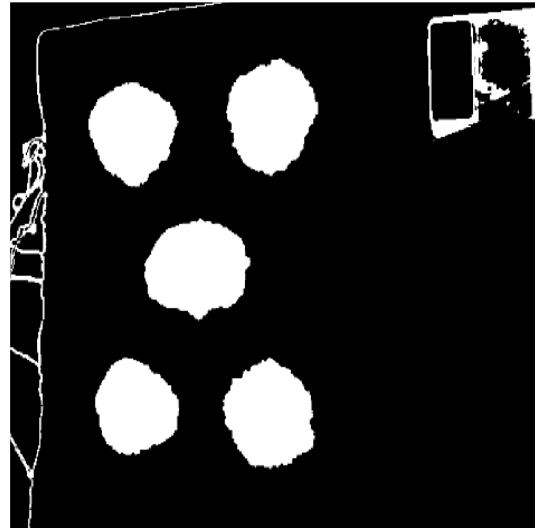
Figure 21 and Figure 22 below show the steps of the ABD algorithm processing for fresh human and aged animal blood samples. The RGB image is generated from the red, green, and blue channels of the HSI datacube and based on the reflectance values and thresholds set by the ABD algorithm. A binary image is generated of pixels identified as background (black, value = 0) and those identified as sample (white, value = 1). The binary image is then further processed with 'closing' which effectively fills holes in the binary image (see Figure 20 (b) and (c)). Groups of pixels that are identified as sample, but connect to the image border are removed, giving the segmented binary image of

the samples. When these regions are overlaid onto the original RGB image, it can be seen that the ABD method successfully identifies blood stains in a given HSI image.

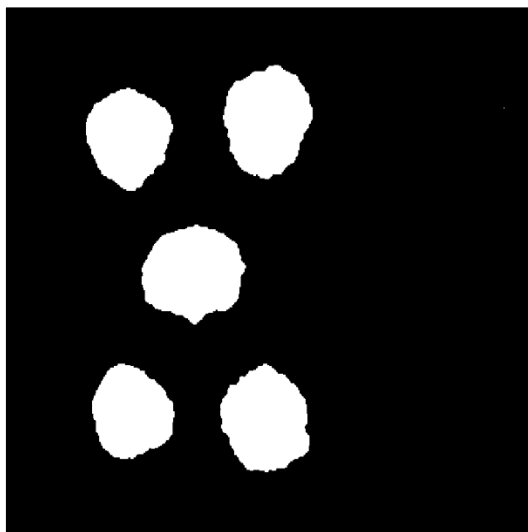
1. RGB image of day 0.1 human blood



2. ABD binary image



3. Segmented image after 'opening' and border removal

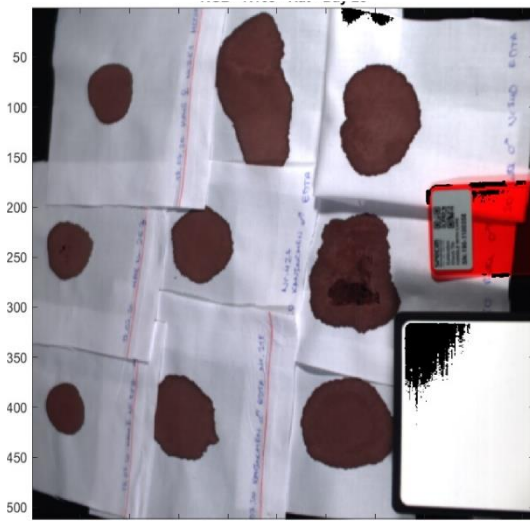


4. ABD Mask overlay on RGB image

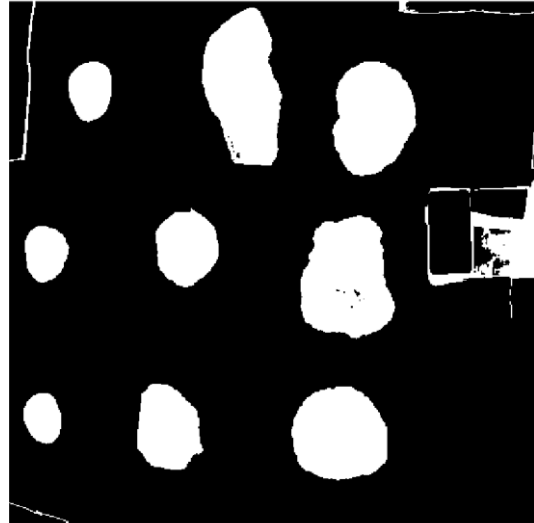


Figure 20. Steps for ABD image processing. Fresh human blood samples ( $t = 0.1$  days) where 1 is the RGB image generated from the HSI cube, 2 is the binary ABD image where white indicates "non-background", 3 is the segmented binary image after 'opening' and border removal, and 4 is the mask of identified blood stains (blue) overlaid onto the RGB image.

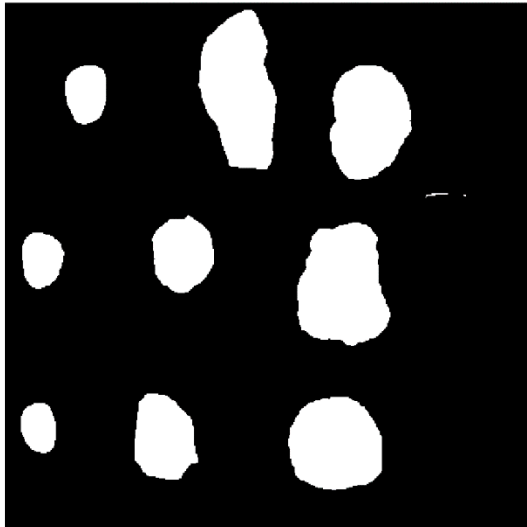
1. RGB image of day 25 animal blood



2. ABD binary image



3. Segmented image after 'opening' and border removal



4. ABD Mask overlay on RGB image

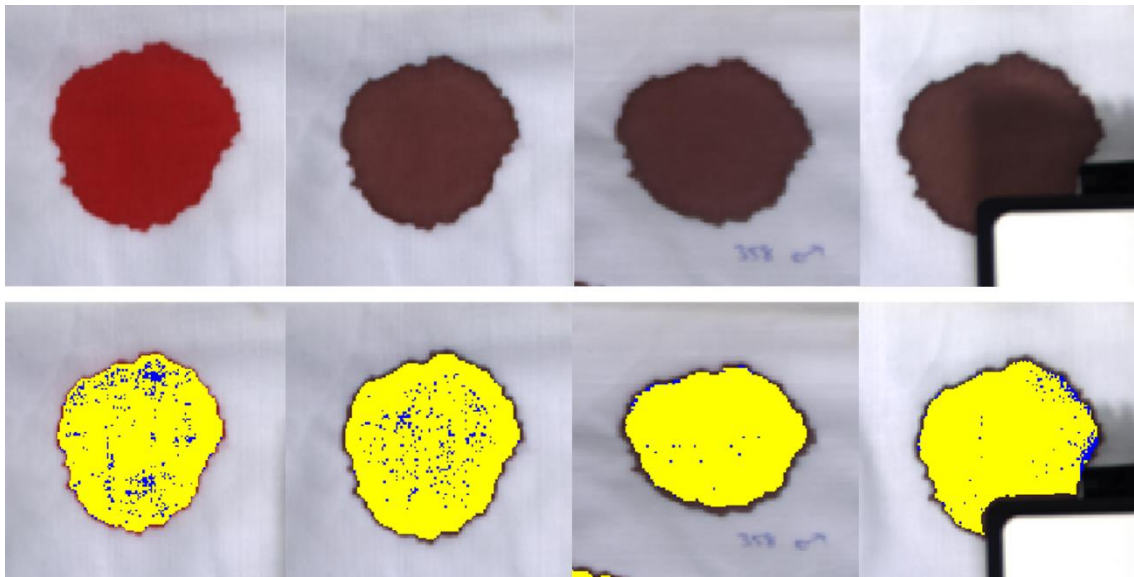


Figure 21. Steps for ABD image processing. Aged animal blood samples ( $t = 25$  days) of mouse (left column), rabbit (middle column) and rat (right column), where 1 is the RGB image generated from the HSI cube, 2 is the binary ABD image where white indicates "non-background", 3 is the segmented binary image after 'opening' and border removal, and 4 is the mask of identified blood stains (blue) overlaid onto the RGB image.

The ABD algorithm successfully identifies human and animal blood stains from background white cotton, ink, the white reference plate, and the orange calibration plate. The blood stains are also identified irrespective of sample age here being shown to identify blood samples at  $t = 0.1$  days and  $t = 25$  days.

### 3.5. Classification of HSI images

The HSI images associated with the test data set were processed and classified using the trained SVM binary classifier. 1 individual from each of the animal classes and 1 human sample is presented in Figures 22-27 below. The RGB images are presented in the first row followed by the classified images directly under. The presented images are taken from each of the classification bins to illustrate the model's performance with respect to aging. Pixels that are classified as "human" by the SVM classifier are coloured with a yellow mask, while pixels classified as "animal" are coloured with a blue mask. No further image processing was applied after SVM classification.



*Figure 22. HSI RGB images of aged human blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 2 days, 6 days, and 31 days. Yellow pixels are classified as "human" and blue as "animal".*

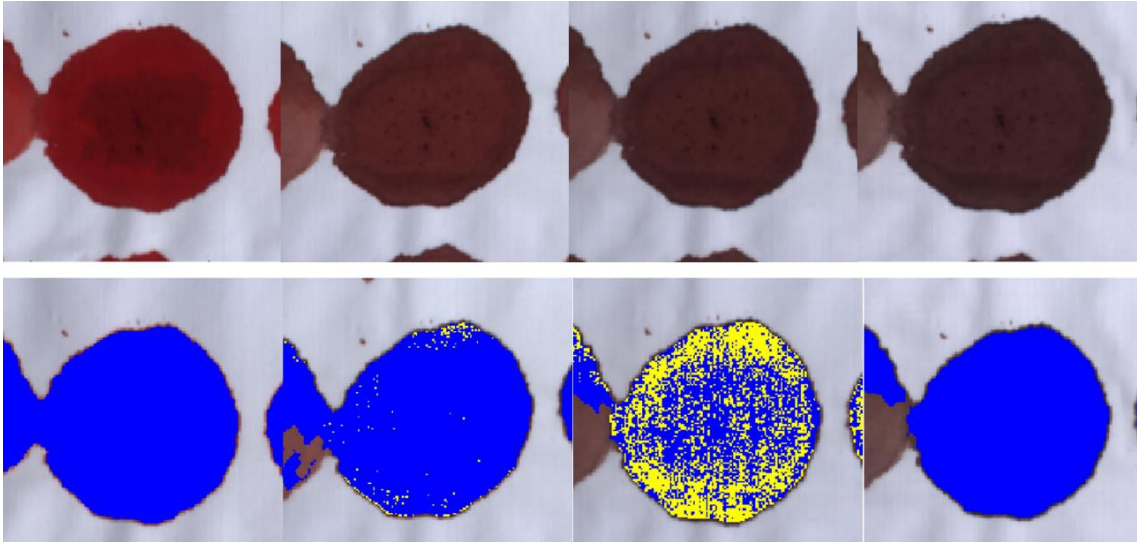


Figure 23. HSI RGB images of aged pig blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 1 day, 5 days, and 42 days. Yellow pixels are classified as “human” and blue as “animal”.

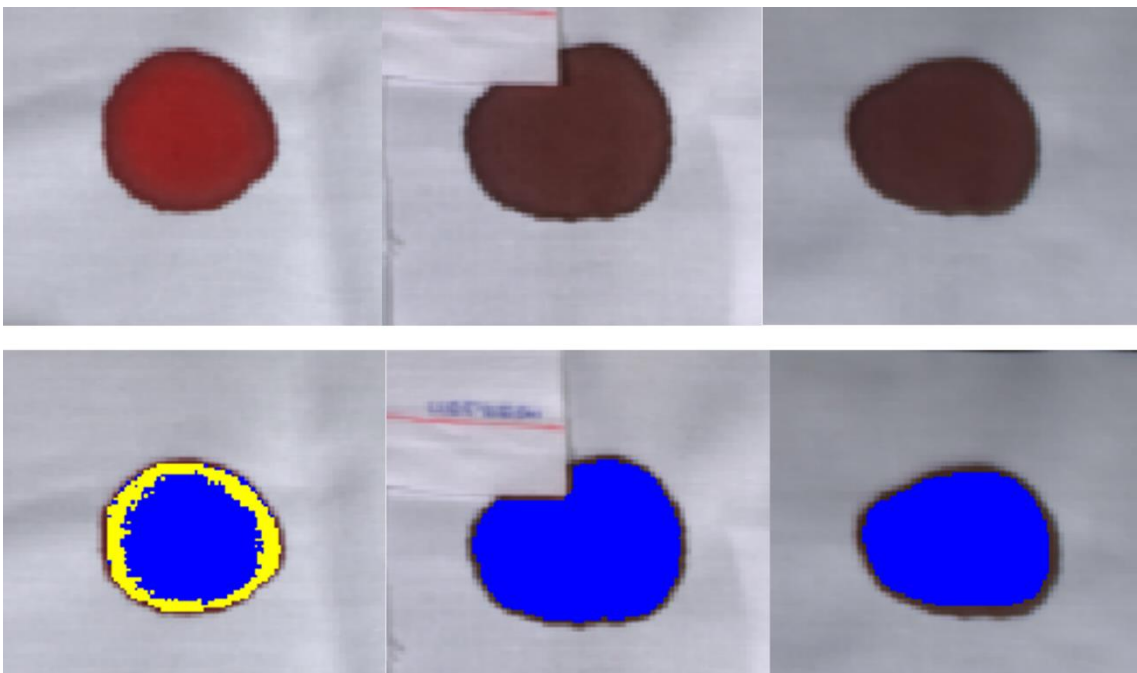


Figure 24. HSI RGB images of aged mouse blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 13 days and 21 days. Yellow pixels are classified as “human” and blue pixels are classified as “animal”.



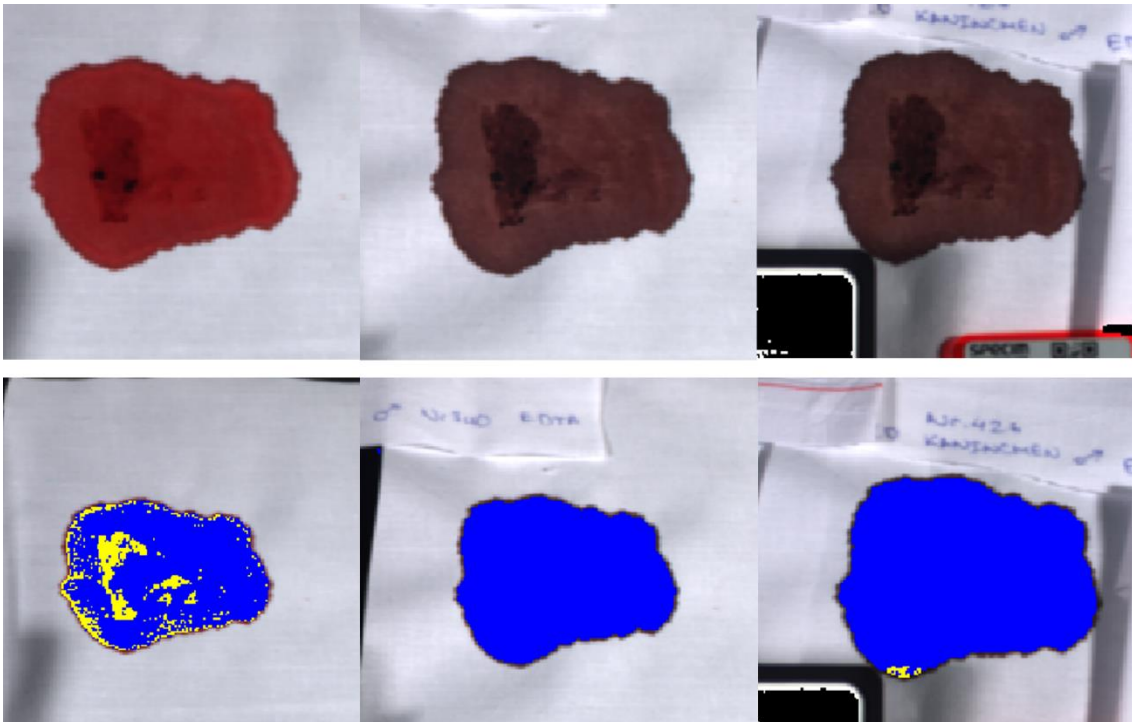


Figure 25. HSI RGB images of aged rat blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 13 days, and 25 days. Yellow pixels are classified as “human” and blue pixels are classified as “animal”.

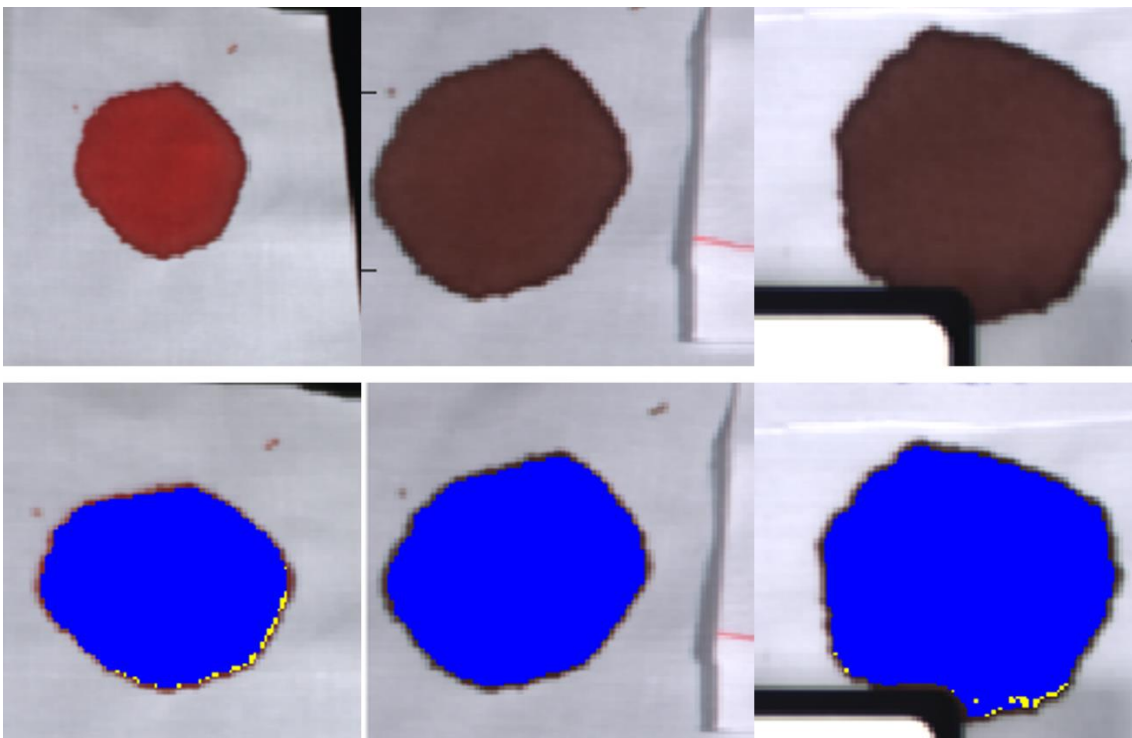


Figure 26. HSI RGB images of aged rabbit blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 1 day, and 49 days. Yellow pixels are classified as “human” and blue pixels are classified as “animal”.

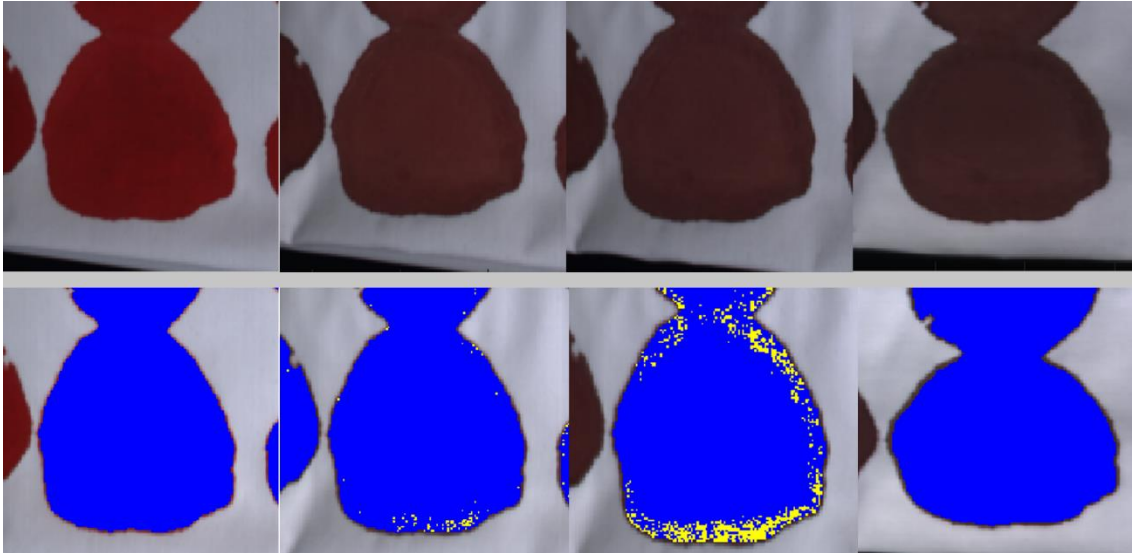


Figure 27. HSI RGB images of aged cow blood with (bottom row) and without (top row) SVM classification mask. The sample ages are (from left to right): 0.1 days, 2 days, 6 days, 49 days. Yellow pixels are classified as “human” and blue are classified as “animal”.

The misclassification observed in the human blood samples are randomly distributed, being observed across all samples ages. Such misclassification is evident in the animal bloodstains, in particular day 5 pig blood (Figure 23.c), day 0.1 mouse (Figure 24.a), day 0.1 rat (Figure 25.a), and day 6 cow (Figure 27.c). These misclassifications are attributed to inhomogeneities in the bloodstains, where darker regions evident in the RGB image tend to be misclassified. No trend in the visual misclassification with increasing bloodstain age was observed, and it is reasoned the binning of classes as per Section 2.2 negates the influence of sample age on the classification of bloodstains.

### 3.6. Reflectance Spectra Characteristics

The human and animal binary datasets of “fresh” day 0.1 bloodstains were plotted as average reflectance (Figure 28 Top left), average SNV-transformed reflectance (Figure 28 Top right), and second derivative (Figure 28 Bottom left) with standard deviations as shaded area. Second-order derivative spectra were generated in order to resolve the undefined broad peaks, and to investigate differences between datasets. The animal spectra (red) have an overall greater standard deviation than the human (cyan), which

is to be expected given the greater total number of individuals in this class (20 human versus 66 animals), and by extension a greater number of spectra.

The average reflectance and SNV reflectance spectra have typical profiles of blood reflectance spectra with the (low reflectance) double peaks of the haemoglobin Q-bands centred around 530 and 580 nm. A notable peak at ca. 670 nm is more pronounced in the animal reflectance spectrum than seen in human. In blood spectra, this is typically associated with the haemoglobin derivative, deoxyhaemoglobin<sup>73</sup>. This difference is accentuated in the SNV-transformed spectrum, which is considered significant given no overlap of the standard deviation curves. Additionally, the NIR region from ca. 760-910 nm differs between human and animal SNV spectra, equating to compositional differences between human and animal blood. The 2<sup>nd</sup> derivative spectrum (Figure 28. *Bottom left*) elucidates 3 major peaks (negative peaks in reflectance appear as positive peaks in the 2<sup>nd</sup> derivative). These are situated at 530, 580, and 670 nm corresponding to the haemoglobin  $\alpha$  and  $\beta$  Q-bands, and deoxyhaemoglobin. Less-pronounced peaks are centred around 630, 650, 725, 745 nm, with oscillations into the NIR portion, including the suspected lipid peak at 900 nm<sup>15</sup>. The derivative spectrum reinforces the assignment of the observed SNV peaks and resolves the likely contribution of lipids to the NIR region.

Figure 29 presents the average reflectance (*Left*) and SNV reflectance (*Right*) spectra for the “fresh” day 0.1 images of the individual animal groups; pig, mouse, rat, rabbit and cow along with the human data. After SNV transformation, the greatest differences between spectra can be seen from 600 nm onwards. It is noted that pig, cow, and rabbit have a more pronounced 670 nm peak than that of rat, mouse, and especially human.

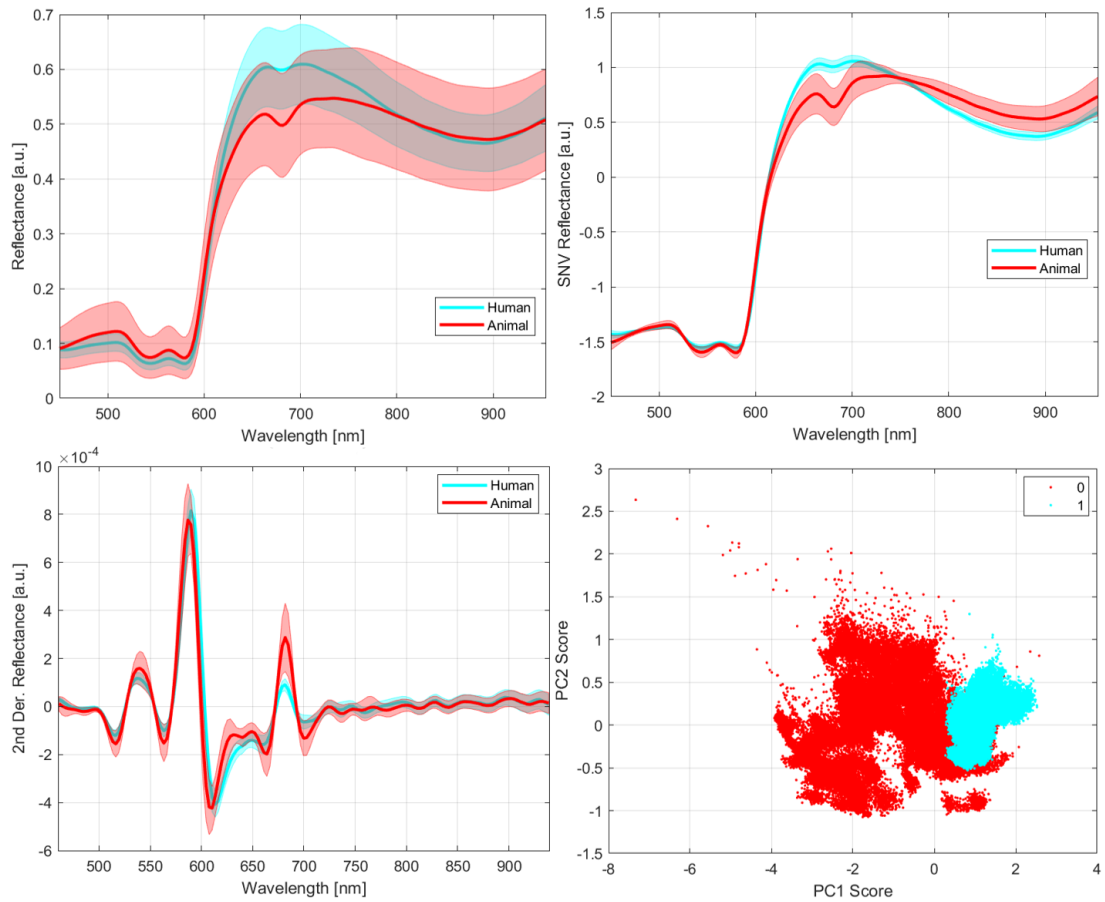


Figure 28. [Top left] Average reflectance spectra with stand and deviation (shaded area) of “fresh” day 0.1 human (cyan) and animal (red), and SNV-transformed spectra [top right]. [Bottom left] Average second derivative reflectance spectra, and [Bottom right] PC scores plot of the first two principal components (explained variance PC1 = 95.3%; PC2 = 3.0%) of the SNV spectrum.

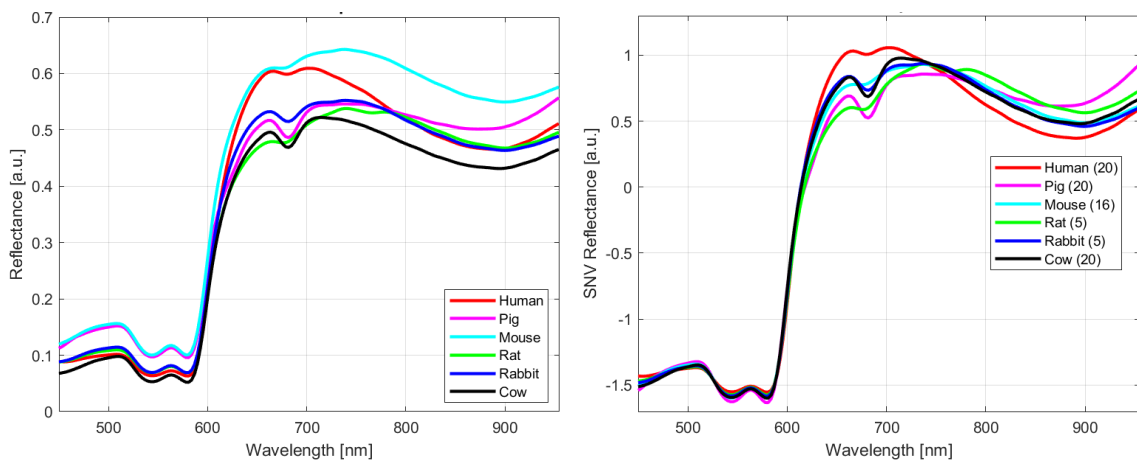


Figure 29. [Left] Average reflectance spectra and [Right] average SNV reflectance [Right] of human (red), pig (magenta), mouse (cyan), rat (green), rabbit (blue), and cow (black). The number of individuals included in the average spectrum is denoted brackets (\*) in the SNV reflectance legend.

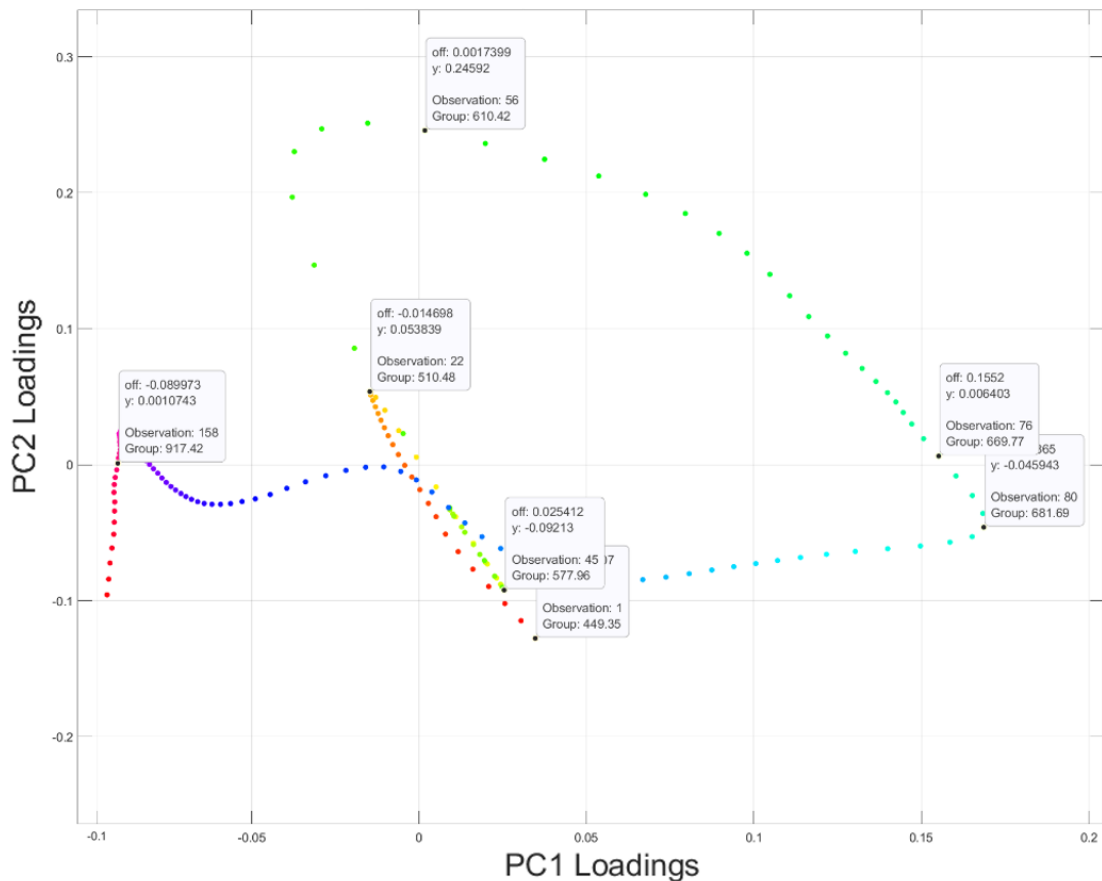


Figure 30. PC loading plot of PC1 and PC2. The wavelengths from 670-685 nm have the greatest positive PC1 loadings (+0.156) with 900-955 nm having the greatest negative PC1 loadings (-0.090). The wavelengths centred around 610 nm and to a lesser extent 510 nm, have the greatest positive PC2 loadings (+0.246 and +0.054) with 449 and 578 nm having the greatest negative PC2 loadings (-0.092).

PCA was performed on the SNV-transformed reflectance spectra and the first two principal components plotted (Figure 28 *Bottom Right*). Two clusters of data corresponding to human (cyan, 1) and animal (red, 0) can be seen with greatest separation occurring along the PC1 (95% explained variance) score axis. This signifies the theoretical separation of these groups based on the first two principal components. The variable contribution to the PCs was further investigated using the corresponding PCs loadings plot.

The PC loading plot of the SNV reflectance data is shown in Figure 30 above. A positive loading indicates that a principal component score and a given variable are positively correlated, while the negative loadings indicate negative correlation for a given variable and principal component. The largest loadings thus indicate the variables that have the greatest effect on the principal component. For PC1, the regions 670-685 nm and 900-

955 nm have the greatest positive and negative correlation respectively. The latter region is associated with lipid and water absorbances<sup>15,16</sup>, and the former with deoxyhaemoglobin which has a peak centred around 660 nm<sup>73</sup>. This agrees with the wavelength regions weighted by the NCFS algorithm in Section 3.2, and suggests the proportionality of oxy- to deoxyhaemoglobin in animal and human blood can be used in their discrimination. The wavelength values 610 and 449 nm have the greatest respective positive and negative correlation with PC2.

## 4. Discussion

### 4.1. Data acquisition and analysis

The collection of blood samples and initial spotting was carried out over a period of 61 days (17 July – 15 September 2020, see Appendix A). The number of samples varied between individuals of a species. For example; the 20 human samples were collected in batches of 5 individuals taken between 13 days from first to last sample, while all 20 pig samples obtained from the slaughterhouse were spotted on cotton and measured simultaneously on the same day as slaughter. The rat and rabbit samples had the longest time between the first and last sample collection of 48 days. The reason for this long acquisition phase is multifaceted: sample collection was often at the discretion of the provider and limited by the associated collection procedure. The mice, rat, rabbit, and pig blood samples were all obtained shortly after euthanasia, which for the latter, was timed per the slaughterhouse weekly working schedule. The former three cases, were dependent on the coordinating research institute's own experimental endeavours, as the blood was obtained as an animal "side-product". The human and cow blood samples for this work were taken in addition to routine blood donation and veterinary inspection respectively. For both sample groups, the complete cohort of samples was obtained in less than two weeks. The acquisition phase was also lengthened due to the need for establishment of collaborative connections between the parties involved, as well as the consideration of ethics and approval therein of the proposed study.

Despite the ill-defined acquisition time period, sample storage and preparation measures were considered to provide close to "fresh" blood sample measurements. Each blood sample was collected into anticoagulant-containing aliquots, and where necessary, refrigerated until preparation. To remove temperature dependence between samples, blood aliquots stored under refrigeration were allowed to warm to room temperature prior to probe preparation. EDTA<sup>25</sup> was chosen as the anticoagulant of choice mainly due to its lack of absorbance in the 400-1000 nm operating window of the HSI system. It has the added benefit of not significantly altering blood component morphology or counts, and therefore assumed to have little to no effect on the observed reflectance spectrum of blood.

Prior to exposure to HSI illumination and subsequent measurement, each blood sample was carefully blotted onto the white cotton surface and allowed to dry-in over a minimum period of 10 minutes. This allowed for the blood to be effectively absorbed onto the cotton fabric and reduced pooling of blood that would otherwise affect the reflectance measurement. Additionally, in the context of forensic analysis, blood would rarely be encountered by the forensics team as being deposited “fresh” at a crime scene and bloodstains would have likely surpassed the time of absorption onto surfaces and fabrics before analysis. To prevent the premature degradation of blood, the halogen light sources were only switched on for the duration of the HSI measurement. This was approximately 1-minute total from centring the sample in the image frame and image focusing, to total hypercube recording time.

Once images of the initial “fresh” blood stains were recorded, the samples were left exposed to ambient temperature and allowed to undergo degradative aging. Blood samples prepared on a single day, such as the pig blood, are expected to be free of any error, with respect to rate of degradation per individual during the aging process primarily due to environmental conditions. This is in contrast to the rat and rabbit samples (see Appendix A), in which the first three and last two individuals differs by over a month. As ambient temperatures in mid-summer can differ significantly compared to late-summer in the not climate-controlled room of the HSI system set-up, this could be relevant to the rate of degradation. Therefore, it can be reasoned that blood samples potentially did not age at equivalent rates between individuals of a given species and ultimately between species groups. Moreover, this could make the comparison of all samples by relevant age questionable, hence sample ages are taken as a guide to the expected extent of degradation. This observation was an additional reason for the binning of HSI images by age before training the classification models.

## 4.2. Class balancing and classification framework

Given the high number spectra, the undersampling balancing method was chosen to balance the data per bin and per species. The low number of individuals in the rat group and the average deposited blood volume, makes the rat group the minority group to



which the data was balanced to. The pig, mouse, rabbit, rat, and cow were all balanced to 1/5 of the total number of human spectra, to give 1:1 balancing between the “human” and “animal” classes of the binary classification. Based on total number of spectra, this balancing method should give a classification model that is without bias between the two binary classes, and ensures one particular animal species is not over or underrepresented within the animal class. The results of Figures 22-27 show little-to-no age bias, or bias towards species in the misclassification of pixels. This point is further explored in Section 4.5.

### 4.3. Pre-processing

The reflectance spectra were initially pre-processed by truncation removing noise at the extremities caused by the low camera sensitivity and low power halogen light source. A Savitzky-Golay filter of polynomial order 2 and a window of 9 spectral band was empirically determined based on the degree of smoothing in the NIR range from 700-1000 nm and against the trade-off of the sharpness of the haemoglobin Q-bands in the visible region. The effective wavelength range was thus 445-955 nm. While this method eliminated significant noise in the spectra inherent to the HSI system electronics, the Soret band<sup>74</sup> of *ca.* 400-420 nm is not observed. This band has been previously shown to correlate the degree of blue shift with blood age<sup>75</sup>. However, this peak has not been widely studied in mammalian species and its usefulness in the HSI classification of human and animal bloodstains is unknown.

### 4.4. Neighbourhood Component Feature Selection

In the building of the binary classification model several methods of feature selection or feature reduction were tested using a basic 1-NN classifier with the training dataset. These methods include derivatisation, SNV transform, PCA with SNV-transformed data, PCA with second derivative, and various feature selection methods (Chi-squared, minimum redundancy maximum relevance (MRMR), and NCFS). The best method was determined as SNV-transformed NCFS. The NCFS method was then optimised by 10-fold

cross-validation to determine the optimal  $\lambda$ -value for feature weighting. As demonstrated in Figure 9, there is a general increasing trend in the 10-fold loss with increasing  $\lambda$ -value, and therefore the  $\lambda$ -value of zero was taken. Given the highly-correlated nature of individual wavelength values in a given reflectance measurement, the negative impact of excluding wavelengths via a larger regularisation parameter is expected. This in turn gives every wavelength a feature weight greater than 1. This signifies each wavelength is important to some degree in the differentiation between human and animal blood spectra. For this reason, instead of viewing features as either “important” (feature weight  $> 0$ ) or “not important” (feature weight = 0), the features can be ranked in terms of importance based on their respective percentage weight of the maximum weighted feature. Figure 10 shows for a given feature/wavelength there is an associated weight as determined by the NCFS algorithm.

The red circles of NCFS feature values form a feature weight-index curve that illustrates the correlation between wavelengths and their respective importance in distinguishing human and animal blood reflectance spectra. It can be concluded that the wavelength bands of most importance (in descending order of feature weight) are centred around: 680, 955, 725, 445, 645, 775, and 600 nm. These peaks can then be used to infer the contributions of blood components to the observed reflectance spectra and give an indication of the potential underlying differences between human and animal blood. The importance of the extrema (445 and 995 nm) could be a result of the SNV transform applied to the spectra for normalisation and standardisation. This transform rescales the reflectance values to have values between 1 and -1 which are centred around the mean of 0. This has the effect of distorting the spectra with respect to the spectrum maximum. This is seen in Figure 28 where the human and animal reflectance spectra without SNV are overlapping irrespective of sample age, while two (or more) groups of spectra are seen in both human and animal training SNV-transformed spectra correlating to “old” and “fresh” blood samples. In the averaged reflectance spectra of the individual species (see Figure 28), different values for the extrema are also a noteworthy characteristic of each species’ spectrum.

As blood spectra from all samples are included in the NCFS calculation, it can be assumed that the feature weights are independent of blood age and thus spectral differences

attributed to the varying composition. This is evident in the alpha and beta Q-bands at *ca.* 550 and 575 nm which are a result of oxy- and deoxyhaemoglobin<sup>76</sup> derivatives, having feature weights below 50% of the maximum weighted feature by the NCFS algorithm. That being said, the region between 600 and 750 nm is often attributed to the derivatives HHb (broad peak 600-700, sharp peak at 760 nm) and metHb at 630 nm.

#### 4.5. Classification of HSI images

In building the machine learning models, a train/test split method was implemented as model test validation. 10-fold cross-validation was also implemented within the NCFS optimisation and validation of the Bayesian-optimised SVM iterations. The mean plus 1 standard error was used in determining the optimal model which is the least likely to experience overfitting, and this model tested on the independent test set. An alternative validation method to the train/test split validation is the leave-one-out cross-validation (LOOCV). This method is a special form of cross-validation, where the complete dataset is used, and the number of validation folds is equal to the number of observations in the data set.  $n-1$  folds are used as a training set for the selected learning algorithm with a single observation being used as the test set. This process is then repeated for each fold of the dataset. Despite LOOCV being a very robust method for testing models, it is also computationally expensive. In addition, it can be assumed that the variation between individual spectra of a given bloodstain is less than that between different bloodstains. A modified form of LOOCV is the leave-one-patient-out cross-validation; where the data is divided into folds based on the number of individuals in the dataset and the complete data from each individual is used in turn as test set. Nevertheless, this approach would require significantly more calculation and would not be expected to improve the binary classification, considering the relatively small differences between 20 bloodstains (or less) per species.

The optimised SVM classifier has good statistical values for the independent test dataset. All major predictors including accuracy, precision, sensitivity, specificity, and F1 score are all above 95% (round up) as per Table 8. The model's predictive power in the discrimination of human and animal blood is evident in the correctly classified images

of Figures 22-27. The bloodstains in general across all species are correctly classified, where the human sample is classified yellow signifying the “human” classification label, and pig, mouse, rat, rabbit, and cow blood are all classified blue as “animal”.

There is no apparent trend in the observed misclassification of pixels with respect to aging. For example, the human day 0.1 (Figure 22a) and day 2 (Figure 22b) have seemingly more misclassifications (blue pixels) comparatively to the day 6 (Figure 22c) and day 31 (Figure 22d). This is in contrast to the pig blood which has a high misclassification (yellow pixels) for the day 5 (Figure 23c) image versus the images of the fresher (Figure 23a and 23b) and older blood (Figure 23d).

Although not very apparent in the human classification images, there is often a ring-like misclassification on the surrounding edges, or of areas of darker red seen in the RGB images. This is most significant in Figure 24a (mouse blood, day 0.1), Figure 25a (rat blood, day 0.1), and Figure 27c (cow blood, day 6), albeit present to a lesser extent in other images. This misclassification on the periphery of the blood stain is most likely the result of capillary flow<sup>77</sup>, where the edges of the blood stain appear darker due to the higher concentration of haemoglobin and its derivatives. This has been observed in blood stain age studies that used cotton<sup>78</sup> as a deposition surface. Nevertheless, for the reflectance measurements in this study the region of interest was centred in the middle of the blood stain for each sample. This method should still account for inhomogeneities in the blood stains. Such inhomogeneities are apparent in the pig and rat blood images (Figure 23 and Figure 25). These darkened areas, or particles, are believed to be coagulated RBCs which formed irrespective of EDTA use. Another factor to this is the fragility of pig RBCs which readily undergo haemolysis with improper handling<sup>79</sup>. A jelly-like substance of coagulated material was observed in 11 out of the 20 pig samples, which was not spotted onto the cotton fabric. These samples also gave lighter coloured blood stains, which is likely due to the reduced number of suspended RBCs. This coagulation would have a similar effect as the dark peripheral rings from capillary flow, in that the tighter packed RBCs would absorb more light and light scattering effects would increase with sample thickness. Despite this, the observation of dark areas within the blood stain does not always result in misclassification. This fact can be seen in the

apparently uniform human blood (Figure 22a) which has notable misclassified pixels, while the cow blood (Figure 27a) is without.

Given the differences between individual species spectra (see Figure 29), it is reasoned that the classification methods outlined herein could be expanded for the discrimination between different animal species. This is supported additionally by the identified important features from the NCFS algorithm, and PCA. The parameters of the trained SVM could be used as a starting point for the new multi-species classifier, which could also include a broader animal dataset. This point is extended to the current binary-classifier, which lacks common domestic animals such as cats, dogs, sheep, goat, and horse, – and less common – fish, reptiles, and birds. The development of such a classifier would have additional applications in wildlife crime, where identification of species on-site would significantly reduce investigation times.

The developed classifier could be modified, or coupled to a separate algorithm for the simultaneous determination of bloodstain age. In the reconstruction of a crime, one of the primary goals of the investigator is the determination of the time when a crime is committed. The establishment of the age of a bloodstain could prove the difference between a conviction or not in a court of law. For these reasons, there has been significant interest in the research of deposited blood age<sup>80</sup>. Raman spectroscopy<sup>81</sup>, IR spectroscopy<sup>82,83</sup>, and reflectance spectroscopy<sup>78,84</sup> have all been used in blood age determination. In particular, vis-NIR HSI similar to that used in this paper, has been successfully used for the age estimation of bloodstains<sup>10-12</sup>. The methods for age determination outlined in these papers could be adapted and incorporated into a workflow that includes the presented SVM human-animal classifier. For example, the methods of Cadd *et al.* (2018)<sup>9</sup> are based on the absorption of haemoglobin and its derivatives between 400 and 680 nm, which has little overlap with the NCFS region (ca. 620-800 nm) selected in this work. These two methods could be therefore used in tandem to achieve both species determination and age estimation from one HSI system.

## 4.6. Bloodstain spectra analysis

To better understand the SVM classification, the spectral differences between the animal and human spectra were investigated. The binary classes of human and animal spectra containing spectra from all animal species, were plotted as average reflectance with standard deviation as an indication of variance within the dataset (Figure 28). The data from the “fresh” or day 0.1 HSI measurements were exclusively used to examine the differences without the composition changes associated with aging. Both human (cyan) and animal (red) reflectance spectra have the characteristic double peaks of the haemoglobin Q-peaks at *ca.* 540 nm for the  $\alpha$ -band and *ca.* 580 nm for the  $\beta$ -band. Given the conserved nature of haemoglobin within the RBCs of mammals, this is to be expected. The deviation from the mean, in both human and animal reflectance spectra, is most likely due to inhomogeneities in the blood samples as well as subtle differences between measurement conditions including instrumental noise. However, in spite of this, the region between 620 and 750 nm differs between human and animal average reflectance spectra. A peak centred around 670 nm is evident, which is more pronounced in the animal spectra than the human.

The SNV-transform reflectance spectrum accentuates the subtle differences between the human and animal average spectra. For the most part, the spectra are near identical up to *ca.* 620 nm where the human spectrum has a higher SNV reflectance value than the animal spectrum. The standard deviation shaded regions do not overlap from *ca.* 640-690 nm signifying a definite characteristic between the two datasets. The region from *ca.* 760-925 nm also deviates between the SNV spectra. Upon inspection of the second derivative spectrum (Figure 28, *Bottom left*), the first region *ca.* 670 nm correlates with that described in the SNV spectrum and is attributed to HHb. The pure spectrum of HHb<sup>24</sup> features an additional peak centred around 760 nm, which is not evident from the reflectance nor SNV-transformed spectrum. When considering the individual species' reflectance spectra of Figure 29, it is noteworthy that this second HHb peak features in the rat spectrum (green). If the peak at 670 nm was due to Hbb, it would be expected that the second peak at 760 nm would likewise be visible in all species' spectra. This discrepancy could be due to; rat blood having higher relative concentration of HHb, unknown components in other species' blood obscuring the latter HHb peak,

additional scattering phenomenon as a result of differences in RBC morphology and haematology between species. It could also be that the peak at 670 nm is of different origin other than HHb. As the spectra are representative of “fresh” day 0.1 blood sample, the effects of haemoglobin degradation to MetHb (630 nm absorbance) or HC should be minimal and can be ruled out as a major cause for the differences observed between spectra.

The unique structure and composition of blood, results in special scattering and absorption, that is influenced by osmolarity, haemolysis, and haematocrit<sup>24</sup>. This is, in part, due to the optical path of a photon within RBCs being increased due to multiple reflections at its internal boundaries of the cell. Thus, haemolysed RBCs have reduced scattering properties, resulting in the overall decrease in absorption coefficient. Additionally, the RBC refractive index increases with higher RBC haemoglobin concentration in the hyper-osmolar state, with higher number of internal reflections increasing the total absorption coefficient. The absorption coefficient of isotonic blood increases linearly with HCT up to 45%, where RBC aggregates are suspected to cause increased scattering, leading to a non-linear dependence above 45%<sup>24</sup>. In pig blood, where RBCs are crenelated and fragile (see Appendix B), osmolarity and haemolysis are of particular interest. The pig RBC fragility was evident in the obtained blood samples – many coagulated despite the use of EDTA (see Appendix A). The coagulation could also be due to the high tendency of pig RBCs to form Rouleaux (Table 1). For the reasons outlined above, this explains the overall higher reflectance of pig blood compared to human, rat, rabbit, and cow (Figure 29, *Left*). The bloodstains generated from samples with evident coagulation, were also pale-red in appearance compared to non-coagulated samples – which is a result of the reduced RBCs. The typical HCT values for all species – except for cow (21-30%) – are within a comparable range (40-53%).

Due to the aforementioned internal reflection, animal RBCs that are larger have greater absorption coefficient due to increased scattering. Cow RBCs are also amongst the smallest of the studied species at 5.5  $\mu\text{m}$ , having overall lower Hb (8.4-12 g/dL blood), signifying a reduced oxygen capacity. Despite this, the averaged cow reflectance spectrum (Figure 29) shows the highest overall reflectance. Mice and rat RBCs are comparable in size (5-7  $\mu\text{m}$ ), being not much smaller than the average human RBC (7.5

$\mu\text{m}$ ). While Hb values are also diminished in rodents compared to humans, this is accounted for by the almost double RCC. Despite this, the haemoglobin concentration in the average RBC, or MCHC value is 28-32 g/dL RBCs, compared to 33-36 g/dL RBCs in humans. Rodent blood would therefore be expected to absorb light to a lesser extent than human blood. Upon inspection of the reflectance spectra however, it can be seen that mouse has an overall greater reflectance than human, and rat. This could be due to the mouse blood acquired being cardiac blood, as opposed to venous blood of the other species. In any case, mouse, rat, – and by extension rabbit blood – have an expected lower oxygenation capacity and higher HHb concentration compared to human. This is seen in the reflectance spectra of the lesser slope at 600-660 nm and the double HHb peaks centred around 670 nm and 760 nm.



## 5. Conclusions

The novel application of vis-NIR HSI was successfully used in the detection and discrimination of human and animal bloodstains on white cotton. The classification of blood was based on 68 features (wavelengths) of the SNV-transformed reflectance spectrum, primarily in the range 600-800 nm as determined by NCFs. These features were used to train several classifiers, where the optimised polynomial-SVM achieved an F1 score of 95.7% when classifying the independent test dataset. In the processing of hyperspectral images, bloodstains were detected using the background detection algorithm ABD, and pixels identified as human or animal blood were coloured yellow and blue respectively. This enhanced the visualisation of the SVM classifier results. The bloodstains were aged over a period of 50 days, with successful classification being achieved independent of bloodstain age.

The main cause for misclassification of pixels was due to capillary flow, and bloodstain inhomogeneities as a result of particulate matter i.e., coagulation. The aforementioned capillary flow is characteristic of the deposition fabric cotton, which would be expected to be minimal on other deposition surfaces. Coagulation is believed to have been a result of species blood properties, – such as the fragility associated with pig RBCs – or the tendency for the formation of Rouleaux. In these cases, aggregated matter was nevertheless observed, despite use of the EDTA as anticoagulant. This demonstrates a potential weakness in the classifier, as blood encountered at a crime scene features no anticoagulant and thus the formation of aggregates are likely. The enhanced scattering properties of thicker blood stains could result in more misclassification, and potentially failure in the detection of bloodstains using the ABD method. It is therefore suggested, that further studies are required into the effects of bloodstain thickness on the ability to detect and discriminate blood stains.

The discrimination of blood stains is hypothesised to be based on compositional differences between species blood – in particular deoxyhaemoglobin content, and morphophysiological differences between RBCs. To validate this hypothesis, further studies into the effect of morphophysiological differences of animal RBCs on the measured vis-NIR reflectance spectrum are required.

The deposition surface of white cotton was selected due to its high reflectivity, which aids the visualisation of bloodstains. It is spectroscopically featureless, meaning there was no contribution to the measured vis-NIR reflectance of blood. Cotton is also forensically relevant, being the predominant fabric in domestic clothing. Future extensions of this work should be performed using a variety of domestic fabrics e.g., wool, leather, cellulose/viscose, linen, and synthetics (polyester). These would be textually distinct to cotton, being different physically in the absorption of blood into the fabric with different degrees of pooling, or wicking of blood after initial deposition. As vis-NIR reflectance is used, the effect of fabric colour on the ability to discriminate human and animal blood should be investigated. As has already been reported in the age-determination of bloodstains using spectroscopic methods, darker-coloured materials typically reduce the performance of the classifier. This is important in forensic investigations where blood can be deposited onto a variety of coloured materials in various environments. Following fabrics, materials such as wood, stone, ceramics, plastics, and metals need to be investigated. This would give conclusive insight into the classifiers' ability to function in any given condition encountered at a crime scene.

To better establish the effectiveness and reliability of the discrimination of animal and human blood on white cotton, large-scale blind tests are required. The inclusion of more species such as; cat, dog, horse, chicken, sheep, goat, fish, and reptiles into the animal repertoire, would increase the confidence in the classifier's ability to truly discriminate human blood from that of any animal. In addition, a multiclass model to discriminate between animal species is a possible future extension of this work. Not only would this bring greater applicability of the method to wildlife forensics, but the analysis of the classifier would bring greater understanding of the differences between human and animal blood reflectance spectra. Environmental variables, such as humidity and temperature should be further explored, as these could influence the degradation of bloodstains. Following from this, limits of detection with respect to age and dilution should be investigated.

The featured HSI system has significant benefits in the discrimination of human and non-human bloodstains. Its portable, non-contact, and non-destructive nature lends itself to forensic investigation, having the capability of on-site detection and evaluation of

forensic evidence. This, in turn, reduces time and money that would be otherwise lost to lengthy laboratory analysis procedures. Such a system could be expanded to include blood age-determination methods as previously outlined, and has the potential to become the workhorse of forensic investigation. The vis-NIR reflectance analysis of many forensically-relevant substances could have future applications in research areas of toxicology and post-mortem investigation, and other types of crimes including (but not limited to) fraud, forgery, and arson. Ultimately, when fully developed, this HSI methodology could be implemented directly at the crime scene, for the discrimination of human and non-human bloodstains. This would significantly benefit forensic casework, especially in the initial stages.

## Appendix A. Hyperspectral data

Table 9. Datasheet of 20 human donor obtained from the Institut für Transfusionsmedizin, Universitätsklinikum Leipzig. Samples were prepared and measured on the same day as collection.

Number	Gender	Age [years]	Collection Date
1	Female	68	05 August 2020
2	Male	56	05 August 2020
3	Male	61	05 August 2020
4	Male	26	05 August 2020
5	Female	59	05 August 2020
6	Male	20	07 August 2020
7	Female	56	07 August 2020
8	Male	42	07 August 2020
9	Female	21	07 August 2020
10	Female	31	07 August 2020
11	Female	25	11 August 2020
12	Male	50	11 August 2020
13	Male	61	11 August 2020
14	Female	48	11 August 2020
15	Female	52	11 August 2020
16	Male	20	17 August 2020
17	Female	28	17 August 2020
18	Female	25	17 August 2020
19	Male	40	17 August 2020
20	Male	50	17 August 2020

Table 10. Datasheet of pig blood samples obtained from Schlachthof Weißenfels, Weißenfels. 'Coagulation' signifies a separation of RBC from plasma before probe preparation. Samples were prepared and measured on the same day as slaughter.

Number	Gender	Age	Other	Slaughter Date
1	-	-	Coagulation	09 September 2020
2	-	-	Coagulation	09 September 2020
3	-	-	-	09 September 2020
4	-	-	Coagulation	09 September 2020
5	-	-	Coagulation	09 September 2020
6	-	-	Coagulation	09 September 2020
7	-	-	-	09 September 2020

8	-	-	-	09 September 2020
9	-	-	-	09 September 2020
10	-	-	Coagulation	09 September 2020
11	-	-	-	09 September 2020
12	-	-	-	09 September 2020
13	-	-	-	09 September 2020
14	-	-	-	09 September 2020
15	-	-	-	09 September 2020
16	-	-	Coagulation	09 September 2020
17	-	-	Coagulation	09 September 2020
18	-	-	Coagulation	09 September 2020
19	-	-	Coagulation	09 September 2020
20	-	-	Coagulation	09 September 2020

*Table 11. Datasheet of mouse blood samples obtained from the Medizinisch-Experimentelles Zentrum (MEZ), Leipzig. Sample probes were prepared and measured on the same day as collection.*

Number	Strain	Gender	Age [Days]	Collection Date
1	CD1/CR	Female	107	17 July 2020
2	CD1/CR	Female	107	17 July 2020
3	CD1/CR	Female	107	17 July 2020
4	CD1/CR	Female	107	17 July 2020
5	CD1/CR	Female	107	17 July 2020
6	CD1	Male	46	30 July 2020
7	CD1	Male	46	30 July 2020
8	CD1	Male	46	30 July 2020
9	CD1	Male	46	30 July 2020
10	CD1	Male	46	30 July 2020
11	Sv12952	Female	90	18 August 2020
12	Sv12952	Female	90	18 August 2020
13	Sv12952	Female	90	18 August 2020
14	Sv12952	Female	204	18 August 2020
15	Sv12952	Female	90	18 August 2020
16	Sv12952	Female	191	18 August 2020

Table 12. Datasheet of rat blood obtained from the Medizinisch-Experimentelles Zentrum (MEZ), Leipzig. Sample probes were prepared and measured on the same day as collection.

Number	Strain	Gender	Age [Days]	Collection Date
1	SPRD	Male	72	17 July 2020
2	SPRD	Male	72	17 July 2020
3	SPRD	Male	72	17 July 2020
4	SPRD	Female	183	2 September 2020
5	SPRD	Female	183	2 September 2020

Table 13. Datasheet of rabbit blood obtained from the Medizinisch-Experimentelles Zentrum (MEZ), Leipzig. Sample probes were prepared and measured on the same day as collection.

Number	Strain	Gender	Age [Year, Month]	Collection Date
1	White New Zealander	Female	1 yr 8 m	17 July 2020
2	Chinchilla Bastard	Male	3 yr 2 m	17 July 2020
3	White New Zealander	Female	1 yr 8 m	17 July 2020
4	Chinchilla Bastard	Female	11 m	2 September 2020
5	Chinchilla Bastard	Female	10 m	2 September 2020

Table 14. Datasheet of cow blood samples obtained from the Klinik für Klauentiere, Leipzig. Samples probes were prepared and measured on the same day as collection.

Number	Gender	Age	Other	Collection Date
1	-	-	Separation	11 September 2020
2	-	-	Separation	11 September 2020
3	-	-	Separation	11 September 2020
4	-	-	Separation	11 September 2020
5	-	-	Separation	11 September 2020
6	-	-	Separation	11 September 2020
7	-	-	-	11 September 2020
8	-	-	-	11 September 2020
9	-	-	-	11 September 2020
10	-	-	Coagulation	11 September 2020
11	-	-	-	11 September 2020
12	-	-	-	14 September 2020
13	-	-	-	14 September 2020
14	-	-	-	15 September 2020
15	-	-	-	15 September 2020
16	-	-	Coagulation	15 September 2020

17	-	-	Coagulation	15 September 2020
18	-	-	Coagulation	15 September 2020
19	-	-	Coagulation	15 September 2020
20	-	-	Coagulation	15 September 2020

## Appendix B. Haematological Parameters

Table 15. Laboratory Variables Relevant to Hematologic Diagnosis (Normal Human Adult Values), adapted from Williams Manual of Hematology, 9ed<sup>85</sup>.

Variable (Common Abbreviation)	Male	Female	Units
Haematocrit (HCT) <u>or</u> Packed Cell Volume (PCV)	42-51	36-46	% <u>or</u> mL RBC/dL blood
Haemoglobin (Hb)	14-18	12-15	g/dl blood
Red cell count (RCC)	4.5-6.0	4.1-5.1	10 <sup>6</sup> /mL
Mean cell volume (MCV)	80-96	79-94	fL/cell
Mean cell haemoglobin (MCH)	27-33		pg/cell
Mean cell haemoglobin concentration (MCHC)	33-36		g/dL red cells
Red cell distribution width (RDW)	<15		%
Reticulocyte count	0.5-1.5		% of red cells
Reticulocyte haemoglobin (CHR)	27-33		pg/cell
Total blood volume (TBV)	65-85	55-75	mL/kg
Plasma volume (PV)	39-44		mL/kg
Red cell mass (RCM)	25-35		mL/kg
Platelet count	175-450		10 <sup>3</sup> /μL
White cell count (WBC, WCC)	4.8-10.8		10 <sup>3</sup> /μL
Absolute monocyte count	0.3-0.8		10 <sup>3</sup> /μL
Absolute neutrophil count	1.8-7.7		10 <sup>3</sup> /μL
Absolute lymphocyte count	1.0-4.8		10 <sup>3</sup> /μL

Haematology of Laboratory Mice and Rats (*Mus musculus* and *Rattus norvegicus*)<sup>86</sup>

As with most other species, blood cell counts in rodents are generally higher in peripheral veins than in central or cardiac blood<sup>87</sup>. Due to the demanding nature of venepuncture in rodents, a blood volume of 5.5 mL/kg body weight can be safely collected from live rats at various sites but collection is typically a terminal procedure in mice. There are several preclinical factors that can affect haematological results in rodents including (but not limited to); sex, age, diet, fasting status, collection site, anticoagulant used, and stress induced by prior handling. For example, older rats and mice have higher RBC counts and less reticulocytes than younger animals.



Table 16. Reference Interval for Haematologic Parameters in Diet-Restricted 7-10 Week Old CD-1 mice collected under isoflurane anaesthesia<sup>86</sup>.

Variable (Common Abbreviation)	Male	Female	Units
Haematocrit (HCT) <u>or</u> Packed Cell Volume (PCV)	42.7-52.9	43.2-56.3	% <u>or</u> mL RBC/dL blood
Haemoglobin (Hb)	12.6-16.3	13.2-16.4	g/dL blood
Red cell count (RCC)	7.82-10.11	7.9-10.12	10 <sup>6</sup> /mL
Mean cell volume (MCV)	47.6-56.2	48.8-58.9	fL/cell
Mean cell haemoglobin (MCH)	14.7-16.8	15-16.7	pg/cell
Mean cell haemoglobin concentration (MCHC)	28.7-32.1	27.9-33.2	g/dL red cells
Red cell distribution width (RDW)	11.6-13.5	11.7-14.8	%
Absolute Reticulocyte	202.9-388.4	150-477	10 <sup>6</sup> /mL
Total blood volume (TBV)	6.3-8.0		mL/kg
Plasma volume (PV)	39-44		mL/kg
Platelet count	1121-1752	630-1559	10 <sup>3</sup> /μL
White cell count (WCC)	0.47-5.16	0.25-5.18	10 <sup>3</sup> /μL
Absolute monocyte count	0-0.08	0-0.09	10 <sup>3</sup> /μL
Absolute neutrophil count	0.29-1.3	0.02-1.12	10 <sup>3</sup> /μL
Absolute lymphocyte count	0.49-3.92	0.23-4.51	10 <sup>3</sup> /μL

Mature RBCs in mice are round, anucleate, biconcave disks with central pallor. In adult mice, RBCs have a mean diameter between 5 and 7 μm with a thickness of 2.1-2.13 μm and cell volumes of 40-50 fL. Rat RBCs are also anucleate biconcave disks with a mean diameter ranging from 5.7-7 μm. Due to the higher concentration of reticulocytes, mice and rats have greater anisocytosis (unequal sized RBCs) and polychromasia (abnormally high number of immature RBCs) compared to non-rodent species. The estimated lifespan of RBCs in rats is between 56 and 69 days<sup>88</sup>, while in mice it is between 41 and 52 days<sup>86</sup>.

#### Haematology of Laboratory Rabbits (*Oryctolagus cuniculus*)

Blood collection is usually performed at the marginal ear vein of rabbits, with other collection sites such as the jugular vein and direct cardiac puncture being reserved for terminal exsanguination procedures<sup>89</sup>. The total blood volume in rabbits is typically 53.8 ± 5.2 mL/kg<sup>90,91</sup> with the maximum safe volume of blood collection being 7.7 mL/kg body

weight<sup>92</sup>. The haematological values for New Zealand white rabbits are presented in Table 17 below.

Table 17. Referenced Haematological parameters of New Zealand white rabbit (*Oryctolagus cuniculus*).

Variable (Common Abbreviation)	Male	Female	Units
Haematocrit (HCT) <u>or</u> Packed Cell Volume (PCV)	40.4 ±3.05	37.8 ±2.31	%
Haemoglobin (Hb)	13.7 ±1.0	12.8 ±0.78	g/dl blood
Red cell count (RCC)	6.75 ±0.533	6.22 ± 0.484	10 <sup>6</sup> /mL
Mean cell volume (MCV)	59.9 ±2.78	60.9 ±2.4	fL/cell
Mean cell haemoglobin (MCH)	20.4 ±0.97	20.8 ±0.93	pg/cell
Mean cell haemoglobin concentration (MCHC)	34.0 ± 0.52	34.1 ± 0.61	g/dL red cells
Total blood volume (TBV)	53.8 ±5.2		mL/kg
White cell count (WCC)	9.5 ±2.07	8.4 ±2.24	10 <sup>3</sup> /μL
Monocyte count	1 ±1.1		%
Heterophil count	32 ±10.95	34 ±15	%
Lymphocyte count	62 ±13.2	61 ±11.3	%

Rabbit RBCs are biconcave disks with an average diameter of 6.7-6.9 μm and a thickness between 2.15-2.4 μm. Anisocytosis is not uncommon in rabbits with polychromasia being observed in 1-2% of RBCs<sup>89</sup>. The rabbit RBC lifespan ranges between 45 and 70 days<sup>93</sup>.

### Haematology of Swine

Due to the low intrinsic value of individual animals, difficulties in blood collection, and the wide range of reported haematological parameter values, routine haematologic tests are not frequently performed in pigs<sup>94</sup>. Porcine RBCs are relatively fragile, with improper handling or excess turbulence often resulting in haemolysis<sup>79</sup>. Despite poor accessibility of veins, blood can be collected from the external jugular or anterior vena cava in sufficient quantities. The interpretation of porcine haematology data requires the consideration of variables such as sex, breed, diet, age, and management practice<sup>95-97</sup> as the range of haematological values are wide (see below).

Table 18. Reference intervals for the Domestic Pig<sup>94</sup>.

Variable (Common Abbreviations)	Range	Average	Units
Haematocrit (HCT) <u>or</u> Packed Cell Volume (PCV)	32-50	42.0	%
Haemoglobin (Hb)	10.0-16.0	13.0	g/dL blood
Red cell count (RCC)	5.0-8.0	6.5	10 <sup>6</sup> /mL
Mean cell volume (MCV)	50-68	60	fL/cell
Mean cell haemoglobin (MCH)	17.0-21	19.0	pg/cell
Mean cell haemoglobin concentration (MCHC)	30.0-34.0	32.0	%
Reticulocyte	0.0-1.0		%
Platelet count	5.2 ±1.95		10 <sup>5</sup> /μL
White cell count (WCC)	11-22	16	10 <sup>3</sup> /μL
Monocyte count	2-10	5.0	%
Neutrophil count	28-47	37.0	%
Lymphocyte count	39-62	53.0	%

The porcine RBC is on average 6.0 μm in diameter, with artefactual crenation (abnormal notched surface due to osmotic water loss) and formation of rouleaux being common. The osmotic resistance has found to be pH-, temperature-, and time-dependent<sup>98,99</sup>. Anisocytosis is predominant in younger pigs and still present to a lesser extent in adult pigs. The RBC lifespan is typically 86 ±11.5 days.

### Haematology of Cattle

The mature RBC of adult bovine has a relatively long lifespan of 130 days<sup>100</sup>. RBCs are biconcave with minimal central pallor, and have a diameter of 5-6 μm. The erythrocyte shape is uniform in adults with poikilocytosis (abnormal shaped RBCs) being found in otherwise apparently healthy calves. This could be due to the unique haemoglobin molecules found in ruminant animals; its association with the RBCs, or otherwise<sup>101</sup>. The RBC of ruminants, including cattle, is unique with respect to the phospholipid composition of the cellular membrane. Sphingomyelin (SM) is found in higher concentration compared to phosphatidylcholine (PC) in ruminants<sup>102</sup>. It is believed this is due to evolutionary pressure of coexisting ruminal ciliates (protozoa) which have a similar membrane make-up to RBCs and the resultant formation of anti-PC antibodies<sup>103</sup>.

There is a large amount of haemoglobin polymorphism in ruminants with respect to; breed, species, and individual development from embryo to adult<sup>104</sup>. The greatest polymorphism is observed in the protein  $\beta$ -chain.

*Table 19. Reference intervals for the adiva 120 from 99 Clinically Healthy Cows, 50% in First Lactation, All Milking 30-150 days, from 10 Ontario Farms<sup>105</sup>.*

<b>Variable (Common Abbreviation)</b>	<b>Range</b>	<b>Units</b>
Haematocrit (HCT) <u>or</u> Packed Cell Volume (PCV)	21-30	%
Haemoglobin (Hb)	8.4-12.0	g/dL blood
Red cell count (RCC)	4.9-7.5	$10^6$ /mL
Mean cell volume (MCV)	36-50	fL/cell
Mean cell haemoglobin (MCH)	14-19	pg/cell
Mean cell haemoglobin concentration (MCHC)	38-43	%
Red cell distribution width (RDW)	16-20	%
Reticulocyte	0.0-1.0	%
Platelet count	1.6-6.5	$10^5$ / $\mu$ L
White cell count (WCC)	5.1-13.3	$10^3$ / $\mu$ L
Monocyte count	0.1-0.7	$10^3$ / $\mu$ L
Neutrophil count	1.7-6.0	$10^3$ / $\mu$ L
Lymphocyte count	1.8-8.1	$10^3$ / $\mu$ L

The haematological reference intervals in cattle are broad, mainly due to the lack of consideration of; animal age, sex, physiological state, or form of restraint when sampling occurs. Differences in breeds have been reported in beef cattle compared to dairy cattle, the latter of which have lower RBC values. Similarly, bulls have greater RBC counts compared to cows, with lactating cows having lower RBC, WBC and plasma protein counts than non-lactating<sup>105,106</sup>. Diet has a significant influence on bovine plasma colouration, ranging from dark yellow to colourless depending on plant chromogens present.

## Appendix C. Machine Learning Methods

### Decision Trees

A regression tree contains numeric responses while a classification tree gives nominal responses e.g., “true” or “false”, or other categorical class labels. In building the tree, the goal is to split the x-variables along the coordinates into regions, such that a given measure of misclassification is as small as possible. Given a set of training data  $(x_1, \dots, x_n)$  with responses  $(y_1, \dots, y_n)$  where  $y_i$  has a discrete value for k-groups, the positive count of group membership of object  $x_i$  in region  $R_l$  is given by the index function  $I(y_i = j)$  with a result of 1 if  $y = j$  and is 0 otherwise. The relative frequency  $p_{lj}$  of the  $j$ -th group in the  $l$ th region is therefore given by:

$$p_{lj} = \frac{1}{n_l} \sum_{x_i \in R_l} I(y_i = j)$$

The relative frequencies of each group in region  $R_l$  can therefore be computed by varying  $j$ , and the objects in region  $R_l$  are resultantly classified to the group  $j(l)$ , with the largest relative frequency i.e., the majority class. The aim is then to minimise misclassification by a chosen measure of quantifying misclassification in region  $R_l$  of tree  $T$ . The choice of criterion depends on the data set, with the main measures being:

Misclassification Error:  $\frac{1}{n_l} \sum_{x_i \in R_l} I(y_i \neq j(l)) = 1 - p_{l,j(l)}$

Gini Index:  $\sum_{j=1}^k p_{lj}(1 - p_{lj})$

Cross-Entropy or Deviance:  $-\sum_{j=1}^k p_{lj} \log(p_{lj})$

The misclassification error is the fraction of objects that do not belong to the majority class, while the Gini Index is the sum of products of relative frequencies of one class with the relative frequencies of all other classes. The cross-entropy is similar to the Gini Index in principle. Based on one of these criteria, the split point  $s$  with smallest measure of misclassification is selected. Both the optimal split point and the best split variable can be found by scanning through all x-variables of the dataset. The first branch of the tree is then given for the variable with the best split-point. This procedure is then repeated in both of the regions arising from the first split and the classification grows like a tree

resulting in smaller and smaller regions and measure of error. To avoid overfitting of the data, i.e., a separate region for each object with an error measure = 0, cross-validation methods are implemented to control tree-size with a complexity parameter (CP). This criterion uses one of the aforementioned misclassification criterions  $Q_l(T)$  and the parameter  $\alpha \geq 0$  to control tree size:

$$CP_\alpha(T) = \sum_{l=1}^{|T|} n_l Q_l(T) + \alpha |T|$$

where  $|T|$  is tree size and large values of  $\alpha$  penalizes large trees.

Decision trees are limited by their binary hierarchical structure, in which small changes in the data generates a slightly different initial split, that cascades through the tree and can result in entirely different subsequent splits<sup>60,107</sup>. “Bagging” is a procedure that uses the averaging of many trees to reduce this instability<sup>108</sup>.

## Support Vector Machines (SVMs)

Discriminant analysis and linear learning machines strictly classify an object by which side of the separation hyperplane it lies. In the case of overlapping groups, one can allow for some objects to lie on the incorrect side of the separation margin. Given a hyperplane

$$f(x) = x^T w + w_0$$

where  $w$  is a weighting vector and  $w_0$  is the offset (cf. DA), decision rules can then be defined as:

$$G(x) = \text{sign}(x^T w + w_0)$$

A function  $f(x) = x^T w + w_0$  with  $y_i f(x_i) > 0$  can be found for all  $i$  given the classification vector  $y$  which lies in the interval  $[-1, +1]$ . For the training points for class -1 and 1, a hyperplane with the biggest margin between the two can be calculated. This forms the optimisation problem:

$$\min \left( \frac{1}{2} \|w\|^2 \right) \text{ subject to } y_i(x_i^T w + w_0) \geq 1, i = 1, \dots, n$$

In the case of overlapping classes, a hyperplane can still be defined which allows some points to lie on the wrong side of the margin. The slack variable  $\xi$  can then be defined to modify the optimisation problem constraints:

$$\min \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right\} \text{ subject to}$$

$$y_i(x_i^T w + w_0) \geq 1 - \xi_i \text{ for all } i, \xi_i \geq 0$$

where the maximisation of the margin and the penalty for samples on the wrong side of the margin, are contained in the first and second terms respectively. Lagrangian theory can be used to solve this optimisation problem of quadratic function minimization with linear constraints. The weight vector of the decision function can therefore be given by:

$$\hat{w} = \sum_{i=1}^n \hat{\alpha}_i y_i x_i$$

which is the linear combination of the Lagrange multiplier and the training data. The training vectors at the class boundary or margin errors have a nonzero Lagrange multiplier alpha. These are termed support vectors and determine the boundary decision function.

Basis expansions such as polynomials or splines can be used to transform the x-variables to their corresponding basis functions  $h(x)$  to enlarge the feature space:

$$f(x) = h(x)^T w + w_0$$

These basis expansions effectively translate non-linear boundaries in the original dataspace to better separated linear boundaries in the enlarged space. The particular transformation information  $h(x)$  is not strictly required when the input features are represented in the optimization problem as their inner products. Therefore, only the kernel function  $K(x_i, x_j)$  is required giving the nonlinear decision function:

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x, x_i) + w_0$$

with typical SVM kernel functions being:

Gaussian:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Sigmoidal:

$$K(x_i, x_j) = \tanh(a_1 x_i^T x_j + a_2)$$

Polynomial:

$$K(x_i, x_j) = (a_1 x_i^T x_j + a_2)^p$$

with parameters  $a_1$  and  $a_2$  of polynomial order  $p$ .

## k-Nearest Neighbours (k-NN)

Given a data matrix  $X$  of  $m \times n$  (1-by- $n$ ) row vectors  $x_1, x_2, \dots, x_m$  and a data matrix  $Y$  of  $m' \times n$  (1-by- $n$ ) row vectors  $y_1, y_2, \dots, y_{m'}$ , the distances between the vectors  $x_s$  and  $y_t$  are:

$$\text{Minkowski distance: } d_{st} = \sqrt[p]{\sum_{j=1}^n |x_{sj} - y_{tj}|^p}$$

where the special case  $p = 1$  gives the City block distance,  $p = 2$  gives the Euclidean distance, and  $p = \infty$  gives the Chebychev distance. The Mahalanobis, Cosine, Correlation, Hamming, Jaccard, and Spearman distance metrics are other metrics used and will be described as they occur in text.

Fix and Hodges first described the k-nearest neighbour (k-NN) method for classification in 1951. Compared to other methods, k-NN is a nonparametric classification method, as it works in the local of the considered data point to be classified without model fitting. A distance metric to determine the neighbourhood along with the closest  $k$ -points are used to estimate group membership of new objects. As a distance metric is used, input data should be auto-scaled (variable mean = 0 and variance = 1), with cross-validation being implemented to determine neighbourhood size  $k$ .



Given an object  $x$ , a distance metric (e.g., Euclidean distance) is used to determine the number of neighbours  $k$  by calculating the distance between object  $x$  and all other objects of the training data. The closest  $k$  nearest neighbours are denoted  $x_1, \dots, x_k$ , and using the known class memberships  $y(x_1), \dots, y(x_k)$  of the neighbours, the predicted class membership  $\hat{y}(x)$  is obtained as the most frequent occurring class amongst the neighbours. Hence, the parameter  $k$  strongly influences the decision boundary between groups. For instance, small values of  $k$  often result in overfitting as isolated regions in the dataspace are observed as separate from the main data. For  $k = 1$  (1-NN), the predicted class membership will always be that of the nearest neighbour. If too large a  $k$ -value is used then underfitting can occur, with the decision boundary becoming ‘blurred’ between groups.

$k$ -NN is often implemented as a reference method as it is conceptually simple, for its applicability to multiclass problems, and the fact that it does not require compact group clusters or linearly separable data. While no training of classifiers are needed, the complete — or a representative set of the data — is needed when classifying new objects. Due to the calculation of very many distances, larger data sets and many variables may be time consuming.

## Bayesian Optimisation

To select the hyperparameter value for subsequent iterations, an acquisition function which helps in determining the next evaluation point is used. The “expected-improvement” (EI) acquisition function ignores values that would increase the objective function while evaluating the expected degree of improvement. EI is given by:

$$EI(x, Q) = E_Q[\max(0, \mu_Q(x_{best}) - f(x))]$$

where  $x_{best}$  is defined as the lowest posterior mean location, and  $\mu_Q(x_{best})$  is the lowest posterior mean value.

The evaluation of each function can require different lengths of time, especially considering the training of a SVM which can dwell over certain regions of the dataspace.

In this case, the “per-second” (pS) time-weighting can be introduced into the acquisition function, giving a cost-weighted expected improvement. During evaluations, a second Bayesian model is maintained to evaluate the function time as a function of position  $x$ , giving:

$$EI_{pS}(x) = \frac{EI_Q(x)}{\mu_S(x)}$$

where  $\mu_S(x)$  is the posterior mean of the timing Bayesian model.

The acquisition function should also consider exploitation (regions where the objective function is deemed “low”) and exploration (regions where uncertainty is high) as a trade-off. “Plus” is a modification of an acquisition function which estimates the overexploitation of a region. Given that:

$$\sigma_Q^2(x) = \sigma_F^2(x) + \sigma^2$$

where  $\sigma_F(x)$  is the standard deviation of the posterior objective function at  $x$  and  $\sigma$  is the posterior standard deviation of the additive noise. If  $t_\sigma$  is now defined as the chosen exploration ratio, then the point  $x$  in the next iteration is deemed as overexploiting if the condition is satisfied:

$$\sigma_F(x) < t_\sigma \sigma$$

In this case the acquisition function’s kernel function is modified and the variance of  $\sigma_Q$  is increased, as suggested by Bull<sup>109</sup>. The new point generated is again tested for overexploitation before accepting the new  $x$  as the next evaluation point.

## Declaration of authorship


I do solemnly declare that I have written the presented research thesis by myself without undue help from a second person others and without using such tools other than that specified.

Where I have used thoughts from external sources, directly or indirectly, published or unpublished, this is always clearly attributed. In the selection and evaluation of research materials, I have received support services from following individuals/institutions: Dr. Claire Chalopin and the IMI group, iCCAS, University Leipzig; Prof. Matysik, Institute of Analytical Chemistry, University Leipzig; Dr. Babian, Institute for Legal Medicine, University Clinic Leipzig; Institute of Transfusion Medicine, University Clinic Leipzig; MEZ, Medical Faculty Leipzig, University Clinic Leipzig; Hooved Animal Clinic, Veterinary Medicine Faculty Leipzig, University Leipzig; Weißenfels Schlachthof, Weißenfels.

The presented intellectual work of this research thesis is my own. In particular, I have not taken any help of any qualified consultant.

I have not directly nor indirectly received any monetary benefit from third parties in connection to this research thesis. In the situation this has been the case, I have received monetary benefits from the following persons or institutions: \_\_\_\_\_n/a\_\_\_\_\_ I declare that no conflict of interest occurs due to these benefits.

Furthermore, I certify that this research thesis or any part of it has not been previously submitted for a degree or any other qualification at the University of Leipzig or any other institution in Germany or abroad.

Date: 5 JAN 2021 Signature: 

## Bibliography

1. Pokupcic, K. Blood as an Important Tool in Criminal Investigation. *J. Forensic Sci. Crim. Investig.* (2017). doi:10.19080/jfsci.2017.03.555608
2. Karger, B., Rand, S., Fracasso, T. & Pfeiffer, H. Bloodstain pattern analysis-Casework experience. *Forensic Sci. Int.* (2008). doi:10.1016/j.forsciint.2008.07.010
3. Dinis-Oliveira, R. J. *et al.* Collection of biological samples in forensic toxicology. *Toxicology Mechanisms and Methods* (2010). doi:10.3109/15376516.2010.497976
4. Linacre, A. & Tobe, S. S. An overview to the investigative approach to species testing in wildlife forensic science. *Investigative Genetics* (2011). doi:10.1186/2041-2223-2-2
5. Corrêa, R. S. *et al.* Soil forensics: How far can soil clay analysis distinguish between soil vestiges? *Sci. Justice* **58**, 138–144 (2018).
6. Leblanc, G., Kalacska, M. & Soffer, R. Detection of single graves by airborne hyperspectral imaging. *Forensic Sci. Int.* **245**, 17–23 (2014).
7. de la Ossa, M. F., Amigo, J. M. & García-Ruiz, C. Detection of residues from explosive manipulation by near infrared hyperspectral imaging: A promising forensic tool. *Forensic Sci. Int.* **242**, 228–235 (2014).
8. Edelman, G. J., Gaston, E., van Leeuwen, T. G., Cullen, P. J. & Aalders, M. C. G. Hyperspectral imaging for non-contact analysis of forensic traces. *Forensic Science International* **223**, 28–39 (2012).
9. Cadd, S. *et al.* The non-contact detection and identification of blood stained fingerprints using visible wavelength reflectance hyperspectral imaging: Part 1. *Sci. Justice* **56**, (2016).
10. Li, B., Beveridge, P., O'Hare, W. T. & Islam, M. The age estimation of blood stains up to 30days old using visible wavelength hyperspectral image analysis and linear discriminant analysis. *Sci. Justice* **53**, 270–277 (2013).
11. Cadd, S., Li, B., Beveridge, P., O'Hare, W. T. & Islam, M. Age determination of blood-stained fingerprints using visible wavelength reflectance hyperspectral imaging. *J. Imaging* **4**, (2018).
12. Edelman, G., van Leeuwen, T. G. & Aalders, M. C. G. Hyperspectral imaging for the age estimation of blood stains at the crime scene. *Forensic Sci. Int.* **223**, 72–77 (2012).
13. Li, B., Beveridge, P., O'Hare, W. T. & Islam, M. The application of visible wavelength reflectance hyperspectral imaging for the detection and identification of blood stains. *Sci. Justice* **54**, 432–438 (2014).
14. Majda, A., Wietecha-Posłuszny, R., Mendys, A., Wójtowicz, A. & Łydzba-Kopczyńska, B. Hyperspectral imaging and multivariate analysis in the dried blood spots investigations. *Appl. Phys. A Mater. Sci. Process.* **124**, 1–8 (2018).
15. van Veen, R. L. P., Sterenborg, H. j. c. m., Pifferi, A., Torricelli, A. & Cubeddu, R. Determination of VIS- NIR absorption coefficients of mammalian fat, with time- and spatially resolved diffuse reflectance and transmission spectroscopy - OSA Technical Digest. in *Biomedical Topical Meeting* (2004).

16. Hale, G. M. & Querry, M. R. Optical Constants of Water in the 200-nm to 200- $\mu$ m Wavelength Region. *Appl. Opt.* (1973). doi:10.1364/ao.12.000555
17. Lee, R. G. & Wintrobe, M. M. *Wintrobe's Clinical Hematology. Volume 1.* (Philadelphia; London: Lea & Febiger, 1993, 1993).
18. Schmid-Schoenbein, H. Erythrocyte rheology and the optimization of mass-transport in the microcirculation. *Blood Cells* 285–306 (1975).
19. Lenard, J. G. A note on the shape of the red cell. *Bull Math Biol* 55 (1974).
20. Rédei, G. P. White Blood Cell. in *Encyclopedia of Genetics, Genomics, Proteomics and Informatics* (2008). doi:10.1007/978-1-4020-6754-9\_18156
21. Drabkin, D. L. & Austin, J. H. Spectrophotometric constraints for common haemoglobin derivatives in human, dog and rabbit blood. *J Bio Chem* 719 (1932).
22. Eilers, R. J. Notification of final adoption of an international method and standard solution for hemoglobinometry. Specifications for preparations of standard solution. *Am J Clin Pathol* 212 (1967).
23. Reagan, W. J., Rovira, A. R. I. & DeNicola, D. B. *Veterinary Hematology: Atlas of Common Domestic and Non-domestic Species.* (John Wiley & Sons, 2019).
24. Roggan, A., Friebel, M., Dörschel, K., Hahn, A. & Müller, G. Optical Properties of Circulating Human Blood in the Wavelength Range 400–2500 nm. *J. Biomed. Opt.* (1999). doi:10.1117/1.429919
25. Banfi, G., Salvagno, G. L. & Lippi, G. The role of ethylenediamine tetraacetic acid (EDTA) as in vitro anticoagulant for diagnostic purposes. *Clinical Chemistry and Laboratory Medicine* (2007). doi:10.1515/CCLM.2007.110
26. Fukushima, A. R. *et al.* Actual trends in the use of the kastle-meyer test: applications in different species and verification of the limit of detection of sensitivity and vestigiality. *J. Dairy, Vet. Anim. Res.* (2019). doi:10.15406/jdvar.2019.08.00261
27. Quickenden, T. I. & Creamer, J. I. A study of common interferences with the forensic luminol test for blood. *Luminescence* (2001). doi:10.1002/bio.657
28. Sharma, V. & Kumar, R. Trends of chemometrics in bloodstain investigations. *TrAC - Trends in Analytical Chemistry* **107**, 181–195 (2018).
29. Greaves, A. V. The use of takayama's solution in the identification of blood stain. *Br. Med. J.* (1932). doi:10.1136/bmj.1.3724.932
30. Tomar, S., Mishra, M. K., Kesharwani, L. & Saran, V. Determining the Sensitivity of Dried Human Blood Stain by Precipitin Test at Different Time Intervals and On Different Surfaces. (2015).
31. Horjan, I., Barbaric, L. & Mrcic, G. Applicability of three commercially available kits for forensic identification of blood stains. *J. Forensic Leg. Med.* (2016). doi:10.1016/j.jflm.2015.11.021
32. Yang, H., Zhou, B., Deng, H., Prinz, M. & Siegel, D. Body fluid identification by mass spectrometry. *Int. J. Legal Med.* (2013). doi:10.1007/s00414-013-0848-1
33. Andrasko, J. The Estimation of Age of Bloodstains by HPLC Analysis. *J. Forensic Sci.* (1997). doi:10.1520/jfs14171j

34. Virkler, K. & Lednev, I. K. Raman spectroscopic signature of blood and its potential application to forensic body fluid identification. *Anal. Bioanal. Chem.* (2010). doi:10.1007/s00216-009-3207-9
35. Elkins, K. M. Rapid presumptive 'fingerprinting' of body fluids and materials by atr ft-ir spectroscopy. *J. Forensic Sci.* (2011). doi:10.1111/j.1556-4029.2011.01870.x
36. Poulos, T. L. Heme enzyme structure and function. *Chemical Reviews* (2014). doi:10.1021/cr400415k
37. Wallace, M. B., Wax, A., Roberts, D. N. & Graf, R. N. Reflectance Spectroscopy. *Gastrointest. Endosc. Clin. N. Am.* **19**, 233–242 (2009).
38. Hapke, B. *Theory of reflectance and emittance spectroscopy, Second Edition* (2012). doi:10.1017/CBO9781139025683
39. Smith, E. & Dent, G. *Modern Raman Spectroscopy - A Practical Approach. Modern Raman Spectroscopy - A Practical Approach* (2005). doi:10.1002/0470011831
40. TKACHENKO, N. V. Ultra-fine spectrum resolution. in *Optical Spectroscopy* (2006). doi:10.1016/b978-044452126-2/50037-x
41. Iqbal, M. S. & Rashid, F. Simple system for low-temperature spectral measurements in the UV-Vis range. *Appl. Spectrosc.* (1989). doi:10.1366/0003702894203516
42. Atkins, P. & de Paula, J. 13A Electronic Spectra. in *Atkin's Physical Chemistry 1005* (Oxford University Press, 2014).
43. Talsky, G. *Derivative Spectrophotometry*. (VCH Verlagsgesellschaft mbH, 1994). doi:10.1002/3527601570
44. Goetz, A. F. H., Vane, G., Solomon, J. E. & Rock, B. N. Imaging spectrometry for earth remote sensing. *Science (80-. )*. (1985). doi:10.1126/science.228.4704.1147
45. Li, Z. *et al.* A review on the geological applications of hyperspectral remote sensing technology. *Work. Hyperspectral Image Signal Process. Evol. Remote Sens.* 3–6 (2012). doi:10.1109/WHISPERS.2012.6874235
46. Pringle, J. K. *et al.* The use of geoscience methods for terrestrial forensic searches. *Earth-Science Rev.* **114**, 108–123 (2012).
47. Modern Aspects of Reflectance Spectroscopy. (1968). doi:10.1007/978-1-4684-7182-3
48. Goetz, A. F. H. *et al.* Imaging Spectrometry for Earth Remote Sensing Published by : American Association for the Advancement of Science Stable **228**, 1147–1153 (1985).
49. Hagen, N. & Kudenov, M. W. Review of snapshot spectral imaging technologies. *Opt. Eng.* **52**, 090901 (2013).
50. Govender, M., Chetty, K. & Bulcock, H. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water SA* (2007). doi:10.4314/wsa.v33i2.49049
51. Fischer, C. & Kakoulli, I. Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications. *Stud. Conserv.* (2006). doi:10.1179/sic.2006.51.supplement-1.3
52. Liang, H. Advances in multispectral and hyperspectral imaging for archaeology and art conservation. *Appl. Phys. A Mater. Sci. Process.* (2012). doi:10.1007/s00339-011-6689-1

53. Carrasco, O., Gomez, R. B., Chainani, A. & Roper, W. E. Hyperspectral imaging applied to medical diagnoses and food safety. in *Geo-Spatial and Temporal Image and Data Exploitation III* (2003). doi:10.1117/12.502589
54. Lu, G. & Fei, B. Medical hyperspectral imaging: a review. *J. Biomed. Opt.* **19**, 010901 (2014).
55. Schuler, R. L., Kish, P. E. & Plese, C. A. Preliminary Observations on the Ability of Hyperspectral Imaging to Provide Detection and Visualization of Bloodstain Patterns on Black Fabrics. *J. Forensic Sci.* (2012). doi:10.1111/j.1556-4029.2012.02171.x
56. Kamruzzaman, M. & Sun, D. W. *Introduction to Hyperspectral Imaging Technology. Computer Vision Technology for Food Quality Evaluation: Second Edition* (Elsevier Inc., 2016). doi:10.1016/B978-0-12-802232-0.00005-0
57. Qin, J., Chao, K., Kim, M. S., Lu, R. & Burks, T. F. Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering* (2013). doi:10.1016/j.jfoodeng.2013.04.001
58. Gendrin, C., Roggo, Y. & Collet, C. Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review. *Journal of Pharmaceutical and Biomedical Analysis* (2008). doi:10.1016/j.jpba.2008.08.014
59. Otto, M. What is Chemometrics? in *Chemometrics* (Wiley-VCH Verlag GmVH & Co. KGaA, 2016). doi:10.1002/9783527699377.ch1
60. Varmuza, K. & Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics* (2016). doi:10.1201/9781420059496
61. Yang, W., Wang, K. & Zuo, W. Neighborhood component feature selection for high-dimensional data. *J. Comput.* **7**, 162–168 (2012).
62. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* (1995). doi:10.1023/A:1022627411411
63. Brereton, R. G. *Applied Chemometrics for Scientists* (2007). doi:10.1002/9780470057780
64. Schein, C. H., Ivanciuc, O. & Braun, W. Bioinformatics Approaches to Classifying Allergens and Predicting Cross-Reactivity. *Immunology and Allergy Clinics of North America* (2007). doi:10.1016/j.iac.2006.11.005
65. Thissen, U., Pepers, M., Üstün, B., Melssen, W. J. & Buydens, L. M. C. Comparing support vector machines to PLS for spectral regression applications. *Chemom. Intell. Lab. Syst.* (2004). doi:10.1016/j.chemolab.2004.01.002
66. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. in *Advances in Neural Information Processing Systems* (2012).
67. Rasmussen, C. E. Gaussian Processes in machine learning. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* (2004). doi:10.1007/978-3-540-28650-9\_4
68. Gelbart, M. A., Snoek, J. & Adams, R. P. Bayesian optimization with unknown constraints. in *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014* (2014).

69. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
70. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **43**, 772–777 (1989).
71. Maktabi, M. *et al.* Classification of hyperspectral endocrine tissue images using support vector machines. *Int. J. Med. Robot. Comput. Assist. Surg.* **16**, 1–10 (2020).
72. Bremmer, R. H., Nadort, A., van Leeuwen, T. G., van Gemert, M. J. C. & Aalders, M. C. G. Age estimation of blood stains by hemoglobin derivative determination using reflectance spectroscopy. *Forensic Sci. Int.* (2011). doi:10.1016/j.forsciint.2010.07.034
73. Takatani, S. & Graham, M. D. Theoretical Analysis of Diffuse Reflectance from a Two-Layer Tissue Model. *IEEE Trans. Biomed. Eng.* (1979). doi:10.1109/TBME.1979.326455
74. Cadd, S. *et al.* The non-contact detection and identification of blood stained fingerprints using visible wavelength hyperspectral imaging: Part II effectiveness on a range of substrates. *Sci. Justice* **56**, 191–200 (2016).
75. Hanson, E. K. & Ballantyne, J. A blue spectral shift of the hemoglobin soret band correlates with the age (time since deposition) of dried bloodstains. *PLoS One* (2010). doi:10.1371/journal.pone.0012830
76. Kyriacou, P., Budidha, K. & Abay, T. Y. Optical techniques for blood and tissue oxygenation. *Encycl. Biomed. Eng.* **1–3**, 461–472 (2018).
77. Deegan, R. D. *et al.* Capillary flow as the cause of ring stains from dried liquid drops. *Nature* (1997). doi:10.1038/39827
78. Bremmer, R. H., Nadort, A., van Leeuwen, T. G., van Gemert, M. J. C. & Aalders, M. C. G. Age estimation of blood stains by hemoglobin derivative determination using reflectance spectroscopy. *Forensic Sci. Int.* **206**, 166–171 (2011).
79. Fontaine, M., Hamelin, N. & Martineau, G. P. Effect of time, storage conditions and mailing on the stability of porcine blood values. *Med Vet Quebec* **60**, 222–227 (1987).
80. Zadora, G. & Menzyk, A. In the pursuit of the holy grail of forensic science – Spectroscopic studies on the estimation of time since deposition of bloodstains. *TrAC Trends Anal. Chem.* **105**, 137–165 (2018).
81. Doty, K. C., McLaughlin, G. & Lednev, I. K. A Raman “spectroscopic clock” for bloodstain age determination: the first week after deposition. *Anal. Bioanal. Chem.* **408**, (2016).
82. Edelman, G., Manti, V., van Ruth, S. M., van Leeuwen, T. & Aalders, M. Identification and age estimation of blood stains on colored backgrounds by near infrared spectroscopy. *Forensic Sci. Int.* **220**, 239–244 (2012).
83. Lin, H. *et al.* Species identification of bloodstains by ATR-FTIR spectroscopy: the effects of bloodstain age and the deposition environment. *Int. J. Legal Med.* **132**, (2018).
84. Edelman, G. J., Roos, M., Bolck, A. & Aalders, M. C. Practical Implementation of Blood Stain Age Estimation Using Spectroscopy. *IEEE J. Sel. Top. Quantum Electron.* (2016). doi:10.1109/JSTQE.2016.2536655
85. Lichtman, M. A. *et al.* *Williams Manual of Hematology*. (McGraw-Hill Education, 2019).



86. Bolliger, A. P. & Everds, N. E. Haematology of Laboratory Animals. in *Schalm's Veterinary Hematology* 1232 (John Wiley & Sons, 2011).
87. Nemzek, J. A., Bolgos, G. L., Williams, B. A. & Remick, D. G. Differences in normal values for murine white blood cell counts and other hematological parameters based on sampling site. *Inflamm. Res.* **50**, 523–527 (2001).
88. Derelanko, M. J. *et al.* Toxicity of Cyclohexanone Oxime. *Toxicol. Sci.* **5**, 128–136 (1985).
89. Zimmerman, K. L., Moore, D. M. & Smith, S. A. Haematology of Laboratory Animals (*Oryctolagus cuniculus*). in *Schalm's Veterinary Hematology* 1232 (John Wiley & Sons, 2011).
90. Marshal, K. L. Rabbit Hematology. *Vet Clin Am Exot. Anim Pr.* 551–567 (2008).
91. Melby, E. C. & Altman, N. H. *CRC Handbook of Laboratory Animal Science*. (CRC Press, 1976).
92. Mittruka, B. M. & Rawnsley, H. M. *Clinical Biochemical and Hematological Reference Values in Normal Experimental Animals and Normal Humans*. (Masson, 1981).
93. Vaha, J. Red Cell Lifespan. in *Red Blood Cells of Domestic Mammals* (eds. Agar, N. S. & Board, P. G.) 420 (Elsevier, 1983).
94. Thorn, C. E. Haematology of the Pig. in *Schalm's Veterinary Hematology* 1232 (John Wiley & Sons, 2011).
95. Dubreuil, P., Farmer, C., Couture, Y. & Al., E. Hematological and biochemical changes following an acute stress in control and somatostatin-immunized pigs. *Can J anim Scie* **73**, 241–252 (1993).
96. Elbers, A. R. W., Counotte, G. H. M. & Tielen, M. J. M. Hematological and Clinicochemical blood profiles in slaughter pigs. *Vet Q* **14**, 57–62 (1992).
97. Yeh, S. H., Tai, J. J. L., Chang, H. L. & Al., E. Blood Profile of Lanyu pigs in Taiwan. *J Taiwan Livest. Res* **27**, 187–195 (1994).
98. Oyewale, J. O. Changes in osmotic resistance of erythrocytes of cattle, pigs, rats and rabbits during variation in temperature and pH. *J vet Med A* **39**, 98–104 (1992).
99. Oyewale, J. O. Effect of storage on the osmotic fragility of mammalian erythrocytes. *J Vet Med A* **40**, 258–264 (1993).
100. Vacha, J. Red Cell life span. in *Red Blood Cells of Domestic Mammals* (eds. Ager, N. S. & Boards, P. G.) 67–132 (Elsevier, 1983).
101. Okabe, J., Tajmia, S., Yamato, O. & Al., E. Haemoglobin types, erythrocyte membrane skeleton and plasma iron concentration in calves with poikilocytosis. *J Vet Med Sci* **58**, 629–634 (1996).
102. Florin-Christensen, J., Suarez, C. E., Florin-Christensen, M. & Al., E. A unique phospholipid organization in bovine erythrocyte membranes. *Proc Nat Acad Sci USA* **98**, 7736–7741 (2001).
103. Giminez, G., Florin-Christensen, M., Belaunzaran, M. L. & Al., E. Evidence for a relationship between bovine erythrocyte lipid membrane peculiarities and immune pressure from ruminal ciliates. *Vet Immunol Immunopathol* **119**, 171–179 (2007).

104. Kitchen, H. & Brett, I. Embryonic and fetal Hb in animals. *Ann NY Acad Sci* **241**, 653–671 (1974).
105. Wood, D. & Quiroz-Rocha, G. F. Normal Haematology of Cattle. in *Schalm's Veterinary Hematology* 1232 (John Wiley & Sons, 2011).
106. Monke, D. R., Kociba, G. J., DeJarnette, M. & Al., E. Reference values for selected hematological and biochemical variables in Holstein bulls of various ages. *Am J Vet Res* **59**, 1386–1391 (1998).
107. Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and Regression Trees. *Biometrics* (1984). doi:10.2307/2530946
108. Hastie, T., Tibshirani, R., Friedman, J. H. & MyLibrary. The elements of statistical learning data mining, inference, and prediction : with 200 full-color illustrations. *Springer Ser. Stat.* (2001).
109. Bull, A. D. Convergence rates of efficient global optimization algorithms. *J. Mach. Learn. Res.* (2011).