Anaemia Severity Prediction From Images Using Machine Learning Models

Ziqian Xu

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2023

Abstract

The report describes the application of various machine learning techniques to predict PCV (packed cell volume) values based on color images of blood samples. The data includes various attributes - R, G, B, area and more - which are used as independent variables in the models. The study employs correlation analysis, multiple variable linear regression, K-Nearest Neighbors (KNN), Gaussian Process Regression (GPR) and Random Forest to develop and evaluate the predictive models. Results indicate that all the models achieved high accuracy, with GPR offering additional advantages such as uncertainty estimation, flexibility, and suitability for small to medium-sized datasets. Overall, the study demonstrates the potential of machine learning techniques to predict PCV values based on color images, which can have significant implications for medical diagnosis and treatment.

Research Ethics Approval

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Ziqian Xu*)

Acknowledgements

I would like to express my profound appreciation to my supervisor, Dr. Yorston, for her invaluable assistance throughout the project. I am immensely grateful for her guidance and support, which have been instrumental in the successful completion of this project.

Additionally, I extend my heartfelt thanks to Mr. Breheny for providing the project and supplying the sample images, which have been crucial in the development of the project.

Contents

Chapter 1

Introduction

The protein molecule known as haemoglobin that is carried by red blood cells is responsible for transporting oxygen through the body. Haemoglobin concentration serves as a conceptual representation of the patient's blood's ability to transport oxygen. Anaemia is a condition in which blood lacks adequate healthy red blood cells and can lead to complications such as fatigue, shortness of breath and heart failure [3]. This problem particularly affects young children, pregnant women and the elderly. According to WHO [23], 42% of children less than 5 years of age and 40% of pregnant women worldwide are anaemic.

There have been various methods to ascertain the presence of anaemia, both invasive and non-invasive in nature. Some methods even include the use of smartphone applications to analyse anaemia through human body parts such as the cuticle and conjunctiva. However, we are undertaking a distinct approach that involves capturing standardised images of blood samples using a smartphone application. Our objective is to construct machine learning models based on the data extracted from these images and determine the most optimal model for measuring anaemia.

In a second stage, a customized smartphone application could potentially be employed for detecting anemia using the ML method identified in this study. In the subsequent section, I will introduce the current techniques and subsequently conduct a thorough examination of my proposed machine learning approach.

Chapter 2

Background Research

2.1 Current popular ways to measure haemoglobin count

2.1.1 Complete Blood Count Test

Anaemia is often diagnosed by a Complete Blood Count(CBC) [14] which measures many different parts and features of your blood including red blood cells, white blood cells and haemoglobin. A health care staff takes a sample of blood by inserting a needle into a vein in arm, usually at the bend in your elbow. The blood sample is sent to a lab for analysis. A small bruise or mild soreness around the blood test site is common and can last for a few day though the patient can return to your usual activities immediately. The cost of private CBC test varies between £15 to £50 in the UK and it requires experienced staff and specialised lab equipment.

2.1.2 Existing Non-invasive Ways

Non-invasive measurement is preferable for both cleanliness and ease of use when taking measurements often. For instance, sickle cell patients frequently experience severe anaemia and require ongoing observation. However, because of the medication they get to reduce the generation of sickled cells, their veins become hard, making drawing blood challenging. To measure haemoglobin non-invasively, health providers must spend hundreds to thousands of dollars on a specialised machine [10]. The Masimo Pronto (costs around 1600 USD) is a handheld haemoglobin testing oximeter that is widely being used to monitor functional total haemoglobin. A study [20] suggested that they calculated a bias of 0.14g/dl which often leads to overestimation compared to the actual Hb value in the patients. It is worth noting that as shown by a research, the sensitivity and the specificity values of the invasive method (83.3%, 87.9%) were higher than the non-invasive method(66.7%, 77.1%). The Bland-Altman analysis was employed to assess the accuracy and bias of the non-invasive method for both genders. The results revealed that the bias of the non-invasive method for males (-0.97 g/dl) was higher compared to the bias of the invasive method for males (0.07 g/dl). It has also been known that skin pigmentation and melanin can affect a pulse oximeter's ability to accurately measure oxygen saturation. So further invasive approaches are

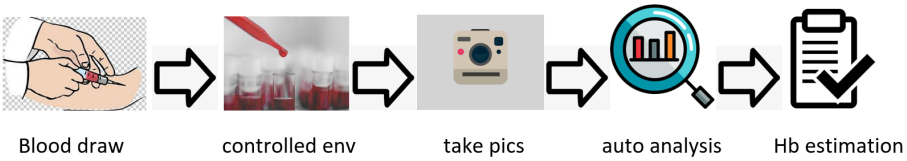

Figure 2.1: System overview of a smartphone app using chose machine learning method

often suggested to be carried out so that the result produced by the machine is less questionable.

2.2 My Project

According to statistics [21], anaemia is higher among rural residents (46.6%) than urban residents (20.1%). So there has been an unmet demand for a method to test for anaemia that is economical for rural areas with less access to CBC test. Since cost is a huge factor which is limiting modern technology in disadvantaged communities, we came up with the idea of incorporating smart phones into the diagnosis of anaemia. My task is to build a machine learning model that can be incorporated into a smartphone that is used to take pictures of blood drops with the camera. The smartphone industry has been steadily developing and growing since 2008. According to statistics [5], developing countries such as China, Vietnam and Indonesia have smartphone penetration rate of over 60%. Smartphone-based medical devices such as heart rate monitoring, sleep monitoring and a variety of telemedicine systems have also been growing. Those devices illustrate how modern smartphones are ideal choices for clinical and remote healthcare systems because of their computation and sensor capabilities. Many apps have even used the smartphone's built-in sensors alone to provide outcomes that are comparable to those of specialised equipment. Upon identifying an effective machine learning model in this project, the subsequent stage will involve the development of a smartphone application to implement the model's prediction capabilities. The application would enable the user to predict anaemia by capturing images of blood droplets using the phone's camera.

2.3 Smart-phone App to Test Canine PCV (inspiration)

The packed cell volume (PCV) is a measurement of the proportion of blood that is made up of cells [1]. PCV falls to less than normal, indicating anaemia, when your body decreases its production of red blood cells or increases its destruction of red blood cells. A scientist Craig Breheny from the University of Edinburgh investigated whether the Packed cell volume (PCV) might be extrapolated from photos taken with a smartphone and readily available equipment with a secondary aim of determining whether a controlled setting and a standardised blood volume were required to increase the accuracy of the results [4]. The controlled environment was established in order to reduce external causes of fluctuation, notably illumination. A standardized, white, paper-based box was utilized to create a consistent environment for the experiment.

Standardization was required for two factors: blood volume and the photographic environment. The box contained filter paper stained with blood to establish a uniform environment. Accordingly, images were captured under four distinct experimental conditions: dropper used outside the box with a standardised environment, pipette used outside the box with a standardised environment, dropper used inside the box with a standardised environment, and pipette used inside the box with a standardised environment.

Environment	Coefficient of determination
Dropper in an uncontrolled environment	0.62
Pipette in an uncontrolled environment	0.76
Pipette in a controlled environment	0.78
Dropper in a controlled environment	0.79

Table 2.1

As shown by table 2.1, PCV predicted by the dropper in a controlled environment has the highest coefficient of determination. It produces a coefficient of determination of 0.79 against the actual measured PCV indicating that smartphone-obtained photos that are shot in a controlled environment could extrapolate an approximate PCV. The current PCV measuring method requires a number of different pieces of equipment, which limits its accessibility for both economical and logistical reasons, such as the requirement for energy. Additionally, the necessary apparatus makes it less portable and hence less useful in field settings. These restrictions would be overcome via a smartphone application. The benefit of using the traditional PCV measurement method is that blood viscosity does not affect the findings of the measurement. The pictures obtained on the smartphone, however, may be significantly impacted by the material's viscosity. The viscosity of the sample, which is primarily influenced by the packed cell volume and protein content, will determine the size of the blood drop. This inspired us to come up with a better idea of taking multiple photos with a constant time gap with the help of a timer built in the smartphoneapp so that we can take account of the viscosity.

2.4 The HemaApp

It is worth noting that there have been applications designed to predict anaemia previously with different approaches. At the University of Washington a group of researchers developed an app called HemaApp that uses the camera to find particular characteristics that indicate the level of haemoglobin [22]. There has not been much research done on using a smartphone for non-invasive blood screening before these scientists. They were the first people to develop a non-invasive method for measuring haemoglobin concentrations using an unmodified smartphone camera. Given the unavailability of non-invasive tests in underprivileged communities, this method has the extra benefit of being quickly deployable and enabling previously unfeasible treatment management choices. For example, the app can help community health professionals in screening for iron-deficient anaemia brought on by malnutrition and lessen the quantity of equipment

that healthcare providers have to come up with. The app is a solution that aims to lighten the load on these people and lowers equipment costs because smartphones are already considered basic equipment for tracking of records. People receiving treatment for anaemia who need to keep track of their status at all times benefit from the reusability of smartphones in addition to better deployment in distant places. A patient may check their haemoglobin levels at home with the use of a smartphone app and this avoids the need to invest hundreds and thousands in specialist blood testing equipment. This makes it more simple and frequent for the patient and the doctor to monitor the success of these therapies. This can aid in the early discovery of any treatment that is ineffective and causing problems. Lastly, even in a clinical situation, the ability to assess haemoglobin non-invasively is helpful for doing so more regularly.

The device enables haemoglobin assessment by analysing the blood's chromatic composition at the user's fingertip and detecting the blood's absorption characteristics at various light wavelengths. This is accomplished by lighting the fingertip with various light sources while utilising the RGB camera. They compare three distinct hardware settings that differ in the amount of hardware augmentation required: (1) white flash + infrared emitter, (2) white flash + infrared emitter + incandescent lamp and (3) white flash + custom infrared led

The results show that for each embodiment, the rank order correlation of the haemoglobin estimates made by HemaApp against the CBC's predictions were 0.69, 0.74, 0.82, respectively, with the mean error of 1.56g/dL, 1.44g/dL, 1.26g/dL. The Masimo Pronto's findings are also compared with those from the CBC, which has a rank order correlation of 0.81 and a mean error of 1.28g/dL. These figures implicate that with the addition of an incandescent light to the smartphone improves the correlation and considerably reduces the inaccuracy. The performance of HemaApp is equivalent to that of the Masimo Pronto with an additional IR source. In terms of sensitivity, the second and third embodiments goes up to 85.7% which exceed the Pronto's 69.3%. Yet the specificity is around ten percent behind the Pronto's. The analysis of HemaApp demonstrates that the smartphone-based haemoglobin system compares favourably with the Pronto predictions with some modification to the smartphone hardware. HemaApp can be used as an efficient screener to decide whether more blood testing is required, but it cannot completely replace the CBC blood test.

2.5 Conjunctiva Images

Another approach is to take images of both conjunctiva [19] from the patients using a smartphone. The palpebral conjunctiva are filled with blood vessels and contain minimal epidermis, dermis and subcutaneous fat that may obstruct light passing through. They also suffer less from physiological and environmental impacts on blood flow. More importantly, severe anaemia can be revealed by physical examination if the conjunctiva are pale. Therefore anaemia can be determined by sampling and analysing digital photographs of the palpebral conjunctiva. This non-invasive technique for determining haemoglobin content is supported by a current research [6]. They measure the haemoglobin level based on colour spectroscopy of the region of the interest, which is the conjunctival pallor that is independently extracted from the images. If the

expected haemoglobin level is less than a certain amount, anaemia is diagnosed. It is suggested by the researchers that the model they established has reached a sensitivity of 89% compared to the actual blood haemoglobin levels. The model continues to be durable under a variety of lighting conditions and device types. The estimated value of haemoglobin concentration would be generated after a quick, on-the-spot calculation that is performed using the image analysis algorithm on the smartphone application. Hence at-home anaemia diagnosis with this method is a feasible substitute for laboratory haemoglobin tests.

2.6 Cuticle Images

This group of scientists came up with another paradigm of entirely non-invasive, on-demand diagnostics [12] that may replace common blood-based laboratory testing with just a smartphone app and pictures. They take use of the finding that pallor is related to anaemia to create a technique that uses image analysis algorithms to quantitatively assess pallor in patient-sourced cuticle photographs for identifying anaemia. With this technology, a user installs an app to their smartphone, takes a picture of their fingernail beds, and the app instantly displays their Hgb level right on the screen of the device. Since melanocytes, the skin cells that produce melanin, are absent from fingernails, conjunctiva, and palmar creases, blood Hg18 serves as these anatomical features' main source of colour. Quality control software is also used to reduce common fingernail irregularities, including leukonychia and camera light reflection for the measurement of Hb levels. To validate this method, they obtained blood examples and smartphone fingernail photos from patients with anemia of various etiologies as well as healthy people to undertake a clinical assessment of this smartphone-based technology. App measured Hgb levels were measured to within 2.4g/dL, with a bias of 0.2g/dL [26] of CBC Hgb levels in 100 patients with a variety of anaemia diagnoses mixed with healthy subjects using a single smartphone image and no customised calibration step [8]. In comparison to published accuracy levels of current invasive point-of-care anaemia screening methods, this non-invasive technique provides a higher level of accuracy. The sensitivity of the test gets up to 97%, showing the potential for this test to serve as a non-invasive screening tool for anaemia using the average WHO Hgb level criterion of 12.5g/dL [24]. To sum up, the level of accuracy is significantly higher than POC screening methods still in development, such as HemaApp and conjunctival analysis, and on par with published accuracy values in POC tools. A summary of the different approaches is given in Figure 2.2.

2.7 Comparing different approaches

Item	Cost	ML Model	Blood	Staff	Equipment	Accuracy	Sanitation Risk	Controlled Env
CBC	£30	x	√	√	√	accurate	√	x
Masimo Pronto	£500	x	x	x	x	high	x	x
Conjunctiva	0	√	x	x	x	high	x	√
HemaApp	0	√	x	x	x	v.high	x	√
Cuticle	0	√	x	x	x	accurate	x	x
My App	0	√	√	x	√	?	√	√

Table 2.2

Chapter 3

Task Description

3.1 Dataset Overview

The dataset provided by Mr. Breheny, from The Royal (Dick) School of Veterinary Studies, comprises of 40 JPG images and 1 Data.xlsx file. These images as shown in Figure 3.1, have uniform dimensions and a white background, which ensures consistency in the samples.

Figure 3.1: one of the 40 JPG files

Dog name	Packed cell volume (PCV)
A	8%
Annie	40%
Archie	44%
B	7%
Bianco	35%
C	15%
D	5%
Daisy	51%
Gracie	15%
Juno	19%
M1	15%
M2	6%
M3	7%
N1	6%

Figure 3.2: Top half of the Data.xlsx file

Each image features a circular piece of white paper in the centre, on which a blood

spot has been placed. The blood spots are canine blood and vary in color intensity and size, with some appearing lighter and larger, while others are darker and smaller. This variability is due to differences in packed cell volume (PCV) within the blood.

PCV is a measure of the proportion of red blood cells in a given volume of blood and is often used to evaluate the presence and severity of anemia. The data.xlsx file as shown in Figure 3.2 provides a table with two columns: dog name and packed cell volume, with the PCV values ranging from 5% to 60% evenly.

PCV values can vary due to several factors, including breed, age, and season. These factors can also affect PCV values in other animals, including dogs. The dataset provided by Mr. Breheny has a broad range of anemic and non-anemic states among the dogs, with PCV values ranging from 5% to 60%.

Anemia occurs when there are insufficient red blood cells to carry oxygen to the body's tissues. Low PCV values indicate anemia, as there is a lower percentage of red blood cells in the total blood volume.

Therefore, in cases of low PCV, the blood spot on the paper would appear larger in size and lighter in color due to the lower concentration of red blood cells. Conversely, higher PCV values indicate a higher concentration of red blood cells, resulting in smaller and darker blood spots on the paper.

3.2 Task Overview

My task is to analyze these images and deduce the PCV values based on the color intensity and size of the blood spots. This can be achieved by using image analysis techniques to measure the size and color intensity of each blood spot, followed by a conversion to a corresponding PCV value using a calibration curve. The resulting data can provide valuable information about the presence and severity of anemia in the blood samples.

In study [13] the authors demonstrated the usefulness of using image analysis techniques to assess the PCV values of blood samples in cats. They used a calibration curve to convert the color intensity of the blood spots to corresponding PCV values, and this approach was found to be highly accurate and reliable. Similarly, in a study [4], the authors utilized image analysis to assess the PCV values of blood samples in dogs. They found a strong correlation between the colour intensity and size of the blood spots and the corresponding PCV values. We intend to pursue these methodologies to establish whether any correlation exists within our data.

The use of image analysis techniques for assessing PCV values in animals is becoming increasingly popular due to its non-invasive and reliable nature. It is expected that these techniques will continue to be used in future research and clinical applications.

Chapter 4

Methodology

Our primary objective is to convert these images into inputs that can be utilized to construct a machine learning model. Two key factors significantly impacting the output (PCV value) are colour intensity and the area of the blood spots. We can attempt to determine the primary colour of the blood spots and subsequently detect their respective areas. Furthermore, we should assess whether data augmentation is necessary given our limited data availability.

Upon obtaining these values, we will utilize them as inputs for a machine learning model and perform model training and testing by partitioning the samples. It may also be worthwhile to standardize the data, as color and area values are expressed on differing scales. Several types of machine learning models are suitable for this case. For instance, we may initially attempt simpler models if we are confident that a linear relationship exists between the inputs and outputs.

Linear regression is a suitable model for this case because it assumes a linear relationship between the input and output variables. In this scenario, we have identified two key factors that significantly impact the output (PCV value): color intensity and area of blood spots. If we can quantify the impact of these factors on the output, we can use linear regression to develop a model that predicts the PCV value based on these input variables.

Furthermore, linear regression is a simple and interpretable model, making it an ideal starting point for modeling. It can help us understand how each input variable contributes to the output and how the inputs interact with each other. It also provides a baseline for comparison with more complex models, such as nonlinear models, that we can explore if the linear model's performance is not sufficient. Alternatively, we can employ more complex models like Gaussian Process Regression to enhance flexibility. Gaussian Process Regression is a suitable model for this case because it can model nonlinear relationships between the input and output variables.

In this scenario, we are attempting to predict PCV values based on color intensity and area of blood spots. These variables may have complex interactions and nonlinear relationships with the PCV value. Therefore, Gaussian Process Regression can provide greater flexibility in capturing these complex relationships, allowing us to make more

accurate predictions.

Furthermore, Gaussian Process Regression is a probabilistic model that can quantify the uncertainty in the predictions. This is particularly useful in medical applications where accurate and precise predictions are critical. By quantifying the uncertainty, we can assess the reliability of the model's predictions and make informed decisions accordingly.

Additionally, this issue can be perceived as a classification problem, where PCV values below 20% indicate Severe anaemia, values between 20% and 30% indicate Moderate, and those between 30% and 40% indicate Mild.

Consequently, some suitable classification machine learning models such as K Nearest Neighbours and Random Forests are worth considering. K-Nearest Neighbours (KNN) and Random Forests are suitable models for this case because they can effectively model nonlinear relationships and perform well in classification tasks. In this scenario, we are interested in predicting the severity of anemia based on the PCV value, which we classify as Severe, Moderate, or Mild based on specific PCV ranges. KNN and Random Forests are well-suited for this classification task as they can identify patterns in the data and predict the class labels based on those patterns.

KNN is a non-parametric model that identifies the k-nearest neighbours to a given data point and predicts the class label based on the majority class among those neighbours. It can handle nonlinear relationships between the input and output variables and does not assume a specific functional form for the relationship.

Random forests are known to work well with a large number of features or attributes, making them a suitable choice for this use case where we are working with numerous attributes. This is because random forests are able to handle the curse of dimensionality, which refers to the problem of having too many variables in relation to the number of observations. Random forests work by building multiple decision trees on subsets of the data and then combining their outputs through a process called ensemble learning. Each decision tree makes a prediction based on a subset of the features, which helps to reduce the risk of overfitting to the data. By combining the outputs of many decision trees, random forests can provide accurate predictions while also reducing the variance and improving the model's generalization ability. In this case, we are working with a large number of features if we can extract lots of colours from the images, which can make it difficult to find the best model. Random forests can handle this by considering a subset of features at each node of each decision tree, which reduces the impact of irrelevant or noisy features. This makes random forests more robust to overfitting and improves their ability to generalize to new data.

Chapter 5

Data Processing

5.1 Image Processing

To identify the red regions in the image, I defined two ranges of lower and upper red values in the HSV colour space [18]. This is because red hues have a circular range in the HSV colour space, which means that a single range of values would not be sufficient to capture all possible shades of red. I then applied two masks to the image using the 'inRange' function, which creates binary images with white pixels in the desired colour ranges and black for the rest. These masks are then combined using the bitwise 'and' operator to create a final mask that highlights the red regions in Figure 5.1 [9].

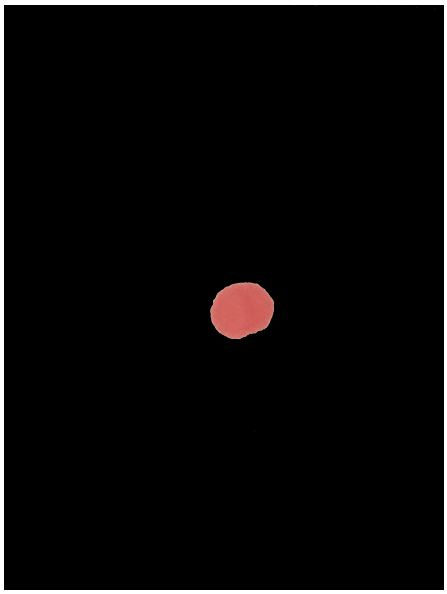

Figure 5.1: masked image

Next, I applied a 2D filter to the image using the 'filter2D' function as shown in Figure 5.2. This is a common image processing technique that can be used to smooth out noise or other unwanted variations in the image. The size of the filter, in this case, is 15x15,

which means that it considers a 15x15 pixel neighbourhood around each pixel in the image when applying the filter.

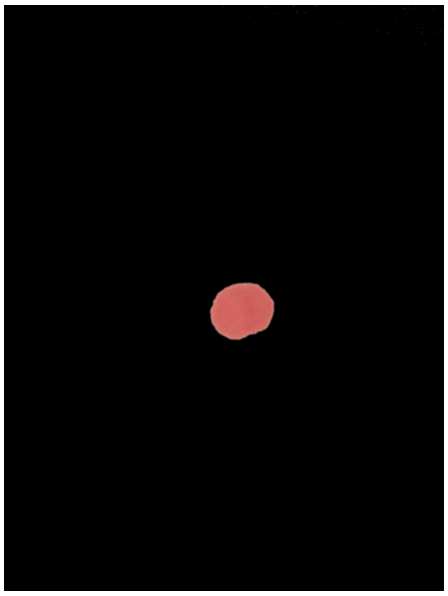

Figure 5.2: masked image after blurring

After smoothing the image, I applied the KMeans [1] clustering algorithm to the RGB values of each pixel in the image. In this case, the code clusters the RGB values of each pixel into 5 clusters. The number of pixels in each cluster is then counted and normalized to create a histogram of cluster frequencies.

One of the main advantages of using KMeans [11] for colour segmentation is its ability to handle large datasets efficiently. In the case of image processing, images typically contain millions of pixels, which can make processing them a computationally intensive task. KMeans is a fast and scalable algorithm that can handle large datasets, making it an ideal choice for this type of task. KMeans also allows for the creation of a histogram of cluster frequencies, which can be used to identify the most dominant colours in an image.

To visualize the top 5 most common colours in the image, I sorted the clusters by frequency and created coloured bars for each of the top clusters using the 'makeBar' function. These bars represent the relative frequency of each colour in the image as shown in Figure 5.3. Finally, I printed out the RGB and HSV values of each colour, along with the number of pixels that were classified as red based on the masks that were created earlier. These values are shown in Figure 5.4.

Figure 5.3: pop-out window that shows the most common colour bars

[1] sklearn.cluster.KMeans

```
Bar 1
  RGB values: (0, 0, 0)
  HSV values: (0, 0, 0)
Bar 2
  RGB values: (209, 101, 95)
  HSV values: (2, 139, 209)
Bar 3
  RGB values: (204, 117, 109)
  HSV values: (3, 119, 204)
Bar 4
  RGB values: (37, 29, 27)
  HSV values: (6, 69, 37)
Bar 5
  RGB values: (115, 79, 74)
  HSV values: (4, 91, 115)
```

Figure 5.4: printed values saved under the data folder

5.2 Basic correlation and data visualisation

Data visualization is an essential part of any data analysis project, and the choice of visualization method can have a significant impact on the insights that can be gained from the data. In this case, a violin plot was selected as the most appropriate visualization method, as it is particularly useful for displaying summary statistics and the density of each variable.

```
data.describe()
✓ 0.0s
```

	R	G	B	Output	Area
count	38.000000	38.000000	38.000000	38.0000	38.000000
mean	166.368421	40.894737	49.157895	32.0000	88943.394737
std	30.290861	34.224055	24.683430	17.6574	44533.461403
min	80.000000	2.000000	11.000000	5.0000	28781.000000
25%	147.750000	12.000000	29.250000	15.0000	54910.000000
50%	165.500000	29.500000	42.000000	37.5000	76736.500000
75%	189.000000	59.500000	62.250000	45.0000	121488.750000
max	209.000000	115.000000	106.000000	60.0000	215111.000000

Figure 5.5: description of the data that is put together

As you can see in Figure 5.5, the attributes above are in different scales. Standardizing the data with mean and standard deviation is a common pre-processing step in data analysis, as it allows variables with different units and scales to be compared more easily. By standardizing the data in this case, it was possible to observe that the mean values for each attribute were centered around 0, with a unit standard deviation on the y-scale. This made it easier to compare the variables with each other and to identify any potential outliers.

When examining the upper and lower quartiles of each attribute, it was observed that they were all within one standard deviation from the mean. This suggests that the data are not far from a normal distribution, which is an important consideration when selecting statistical tests and models for analysis.

Interestingly, all five of the violins shown in Figure 5.6 had a pear shape, with some looking similar to others and some appearing to be upside down in comparison to the others. This suggests that there may be correlations between the variables, and further

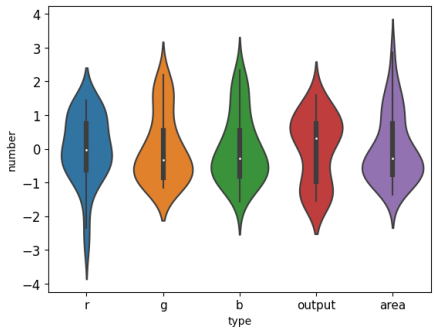

Figure 5.6: violin plot with attributes on the x axis

investigation is planned to explore this possibility. Correlation analysis can help to identify relationships between variables and to determine the strength and direction of these relationships.

	R	G	B	Output	Area
R	1.000000	0.858343	0.875924	-0.909865	0.722576
G	0.858343	1.000000	0.993321	-0.930034	0.835051
B	0.875924	0.993321	1.000000	-0.917261	0.826594
Output	-0.909865	-0.930034	-0.917261	1.000000	-0.841883
Area	0.722576	0.835051	0.826594	-0.841883	1.000000

Figure 5.7: correlation table of the attributes

The correlation table in Figure 5.7 provides us with valuable insights into the relationships between the attributes and output. It is evident that there is a significant degree of mutual correlation between the variables.

Specifically, the ‘Red’ attribute shows a strong positive correlation of over 80% with both ‘Blue’ and ‘Green’, indicating that the three variables are highly interdependent. Moreover, ‘Blue’ and ‘Green’ display an almost perfect positive correlation of nearly 100%, as they are primarily influenced by the depth of the color of the blood spots.

A deeper analysis of the correlation table reveals that ‘Red’, ‘Green’, and ‘Blue’ exhibit a highly negative correlation with the output variable ‘PCV values’, while showing a positive correlation with the ‘Area’ attribute. These results suggest that the intensity of the colour is inversely related to PCV values, with lighter colours indicating anaemia or a lack of haemoglobin.

Conversely, the ‘Area’ attribute is positively related to the color intensity, as lighter colours correspond to lower PCV values. When PCV is low, the blood viscosity decreases, leading to more dispersed blood spots (it should be noted that the images were taken after a uniform time delay following the blood drop on the paper).

In summary, the results confirm that all four attributes are correlated with PCV values, with the degree and direction of correlation varying between the attributes.

5.3 Data Augmentation

The use of image augmentation techniques to increase the size of data sets is a common approach in machine learning, especially in deep learning. As mentioned in our previous discussion, there are several image augmentation techniques available, such as geometric transformations, colour space transformations, kernel filters, random erasing, and image mixing. These techniques can increase the diversity and quantity of data, making it possible to improve the accuracy and robustness of the analysis.

There have been several studies that have explored the use of data augmentation in image analysis tasks. For example, a study [16] found that applying geometric transformations such as rotation, scaling, and cropping could improve the accuracy of deep learning models for image classification tasks. Similarly, a study [2] showed that data augmentation techniques such as flipping, rotating, and cropping could improve the performance of deep learning models for medical image analysis tasks.

It's worth noting that the effectiveness of data augmentation can depend on the specific task and the type of data being analyzed. For example, some studies have shown that certain types of data augmentation may be more effective than others for certain image analysis tasks [7]. Therefore, it's important to carefully consider the type of data augmentation techniques used and evaluate their effectiveness for the specific task at hand.

However, in our case, there were specific constraints that needed to be taken into account. For example, we needed to preserve the originality of color space and blood spot size, which eliminated several augmentation techniques from consideration. Color space transformations, random erasing, and image mixing were not feasible in our case due to these constraints. We also had already applied blurring, so we could not use this technique either.

Therefore, we decided to use geometric transformations, specifically cropping as seen in Figure 6.8 and Figure 6.9. Although flipping, rotating, and zooming are common geometric transformations, they would not have had any effect on the input and output in our case. In contrast, cropping was found to be an optimal geometric transformation technique, providing a means to generate new samples without compromising the originality of the color space and blood spot size.

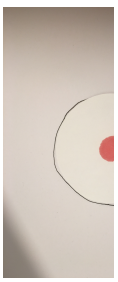

Figure 5.8: Image A cut vertically

To implement this technique, we divided the image into two halves pixel precise, first vertically and then horizontally, resulting in four new images from one. This method

Figure 5.9: Image A cut horizontally

effectively increased the sample size to five times its original amount, providing a significant increase in data size for subsequent analysis. The RGB values of half of the blood spots were tested, producing values similar to those of the original blood spots.

The use of data augmentation techniques, such as the one we employed, can improve the performance of machine learning models by increasing the diversity of the training data and reducing the risk of overfitting. This approach offers a promising solution to overcome the limitations of limited data samples, as it can help improve the accuracy and robustness of the analysis.

5.4 Augmented Data Correlation

```
data.describe()
✓ 0.1s
```

	R	G	B	Output	Area
count	190.000000	190.000000	190.000000	190.00000	190.000000
mean	167.089474	41.400000	49.673684	32.00000	88943.394737
std	29.435921	33.645019	24.324500	17.46955	44059.687699
min	80.000000	0.000000	10.000000	5.00000	28781.000000
25%	148.000000	12.000000	30.000000	15.00000	54886.000000
50%	167.000000	30.000000	43.000000	37.50000	76736.500000
75%	189.750000	61.750000	63.750000	45.00000	126515.000000
max	210.000000	115.000000	106.000000	60.00000	215111.000000

Figure 5.10: augmented data described

The augmented dataset in Figure 5.10 displayed a comparable correlation with the original table. This result can be attributed to the preservation of the area attribute in the new images, while only slightly altering the other three attributes. As a consequence, the violin plot in Figure 5.11 offers a closely similar correlation between the inputs and outputs.

Correlation analysis revealed strong relationships between the attributes and output. ‘Red’ showed a high positive correlation with ‘Blue’ and ‘Green’, while ‘Blue’ and ‘Green’ exhibited an almost perfect positive correlation. Additionally, ‘Red’, ‘Green’, and ‘Blue’ displayed a negative correlation with ‘PCV values’ and a positive correlation with ‘Area’. This suggests that lighter colors indicate low PCV values, while a larger area corresponds to lighter colors. Overall, all four attributes are correlated with ‘PCV values’ to varying degrees.

	R	G	B	Output	Area
R	1.000000	0.860508	0.873446	-0.885133	0.720649
G	0.860508	1.000000	0.992756	-0.891947	0.826763
B	0.873446	0.992756	1.000000	-0.872097	0.813368
Output	-0.885133	-0.891947	-0.872097	1.000000	-0.833747
Area	0.720649	0.826763	0.813368	-0.833747	1.000000

Figure 5.11: correlation of the augmented data

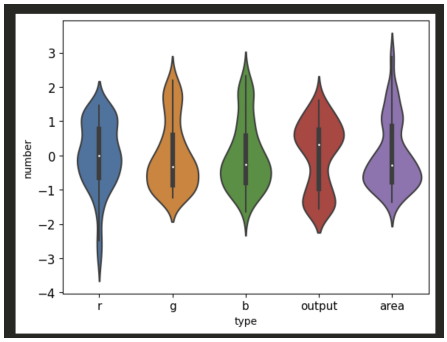

Figure 5.12: Violin plot of the augmented data

Chapter 6

Machine Learning Algorithms

In this chapter, the four main methods I used for machine learning are described along with their results: linear regression, K nearest neighbours, Gaussian process regression and random forests.

6.1 Regression Methods

6.1.1 Linear Regression

I opted to employ multiple variable linear regression [1] in order to formulate the association between four independent variables, namely R, G, B, and area of blood spot, and a dependent variable 'PCV value'. Two models were established, one encompassing all four variables and another with only three variables, and both models yielded impressive R-squared and adjusted R-squared values of approximately 0.89 as can be seen on Figures 6.1 and 6.2.

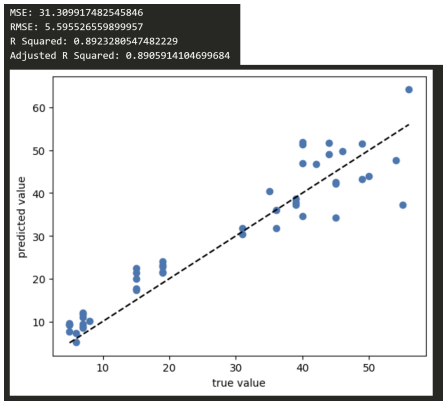

Figure 6.1: Evaluation metrics and scatter plot for 3 variable model (RGB)

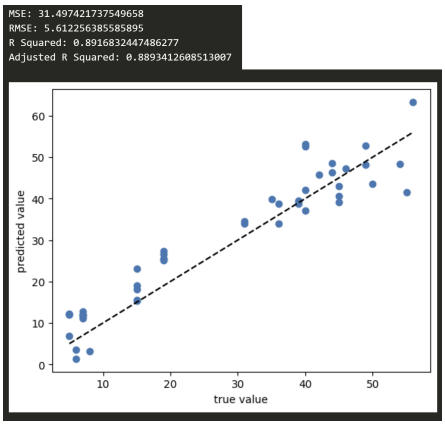

Figure 6.2: Evaluation metrics and scatter plot for 4 variable model (RGB and area)

[1] sklearn.linear_model.LinearRegression

An R-squared value of 0.89 is indicative of a strong model, as it implies that about 89% of the variance in the dependent variable can be explained by the independent variables in the model.

Notably, the adjusted R-squared value is almost identical to the R-squared value, which indicates that the inclusion of an additional independent variable (in this scenario, the 'area' variable) did not significantly enhance the model's fit. Consequently, it could be inferred that the 'area' variable might not have a highly significant relationship with the dependent variable, and could be eliminated from the model without substantial effect on its performance.

Upon observation of the plot, it can be deduced that no conspicuous outliers are present, and the regression line is well-suited to the data.

6.1.2 Gaussian Process Regression

6.1.2.1 Why Gaussian Process Regression would be useful

Gaussian process regression (GPR) is a powerful and flexible machine learning algorithm with several advantages:

1. Uncertainty Estimation: GPR not only provides point predictions of the output variable, but also estimates the uncertainty associated with each prediction. This can be particularly useful when making predictions in real-world applications, where it is important to know how confident we are in our predictions.

2. Small to Medium-Sized Data: GPR can work well with small to medium-sized datasets because it is a nonparametric algorithm that does not make any assumptions about the underlying distribution of the data. This can be useful when working with data that may not fit well with traditional parametric regression models.

3. Flexibility: GPR is a highly flexible algorithm [25] that can be customized by choosing different kernels to model the relationships between the input and output variables. This allows the model to be tailored to the specific characteristics of the data, which can improve its performance.

Overall, GPR can be a good choice for regression problems where the relationships between the input and output variables are complex [17], and where it is important to estimate the uncertainty associated with the predictions.

6.1.2.2 Implementation

For the first model [2], I only used R, G and B as attributes. After standardizing the input data, I split the data into training and testing sets.

I defined a grid of hyperparameters to search over using the GridSearchCV function from scikit-learn.

The hyperparameters included different kernels and alpha values, and I specified a range of values to search over (see Figure 6.3).

[2] sklearn.gaussian_process.GaussianProcessRegressor

```
# Define K-fold cross-validation
kfold = KFold(n_splits=5, shuffle=True, random_state=42)

# Define hyperparameters for grid search
param_grid = {'kernel': [ConstantKernel(1.0, (1e-3, 1e4)) * RBF(length_scale=1, length_scale_bounds=(1e-2, 1e2)),
                         ConstantKernel(1.0, (1e-3, 1e4)) * RBF(length_scale=10, length_scale_bounds=(1e-2, 1e2)),
                         ConstantKernel(1.0, (1e-3, 1e4)) * Matern(length_scale=1, length_scale_bounds=(1e-2, 1e2))],
              'alpha': [0.01, 0.1, 1, 10]}
# Perform grid search to optimize kernel and alpha
grid_search = GridSearchCV(gp, param_grid=param_grid, cv=kfold, scoring='r2')
grid_search.fit(X_train, y_train)
```

Figure 6.3

The GridSearchCV function performs a exhaustive search over the hyperparameters using cross-validation to evaluate the performance of each combination.

The kernel parameter specifies the covariance function to use for the GPR model. The covariance function determines the similarity between any two points in the input space, and hence governs how much the prediction at one point should be influenced by the observations at other points. The RBF (radial basis function) kernel and Matern kernel are both popular choices for GPR models.

The RBF kernel, also known as the squared exponential kernel, is a stationary kernel that assumes that nearby points are highly correlated, while points far apart are uncorrelated. The length-scale hyperparameter controls the range of correlations, and larger values of the length-scale allow for smoother functions.

The Matern kernel is a more flexible kernel that includes the RBF kernel as a special case [15]. It has a length-scale hyperparameter that controls the range of correlations, as well as a smoothness hyperparameter that governs the rate at which the correlations decay with distance.

I extracted the best kernel, alpha, and R2 score from the grid search object and built model from these parameters as seen in Figure 6.4.

6.1.2.3 Results

```
Best kernel: 1**2 * Matern(length_scale=1, nu=1.5)
Best alpha: 10
Cross-validation R2 scores: [0.63579666 0.6357968  0.90131814 0.65978126 0.6597812  0.90324146
 0.81684001 0.81683945 0.91474261 0.93959514 0.93959516 0.94570863]
Mean cross-validation R2 score: 0.8140863762873908
Testing R2 score: 0.9774326897691179
```

Figure 6.4

```
Best kernel: 1**2 * Matern(length_scale=1, nu=1.5)
Best alpha: 1
Cross-validation R2 scores: [0.74945932 0.74945948 0.95513377 0.80996053 0.8099607  0.95856583
 0.88584279 0.88584267 0.971032   0.95719034 0.95719031 0.96130928]
Mean cross-validation R2 score: 0.8875789173247952
Testing R2 score: 0.9874590108104699
```

Figure 6.5

As evidenced by the results in Figure 6.4, the cross-validation R-squared values are quite high, and the model has achieved an exceptionally high testing R-squared score.

We now include an additional attribute, namely area, in addition to R, G, and B and followed the same process of cross-validation and grid search. We observe in Figure 6.5 that the cross-validation R-squared values are even higher than before, and the testing R-squared value has also increased significantly. These are the most optimal results obtained thus far from the two GPR models examined.

In both cases, the Matern kernel had a better result than the RBF kernel, as measured by the cross-validation R2 score. This may be because the Matern kernel is more flexible and can capture a wider range of patterns in the data. Additionally, the Matern kernel has a smoothness hyperparameter that can be tuned to match the degree of smoothness in the data, which may have contributed to its superior performance.

```
[(50, 46.711377801415324), (50, 49.64748189198042), (15, 15.085198715247202), (19, 19.904952737349937),
(54, 52.17889250770028), (7, 6.868571793575931), (40, 40.45475799086336), (7, 6.710723557919742),
(25, 24.465529818082956), (7, 7.270748144666321), (49, 48.930259344448512), (39, 41.071522216187304),
(7, 6.9700457590487055), (40, 40.50820692233685), (32, 31.324389788032562), (7, 6.696203248459176),
(54, 50.66489976968363), (19, 21.470203340819218), (5, 4.985899517964938), (15, 14.935884940157555),
(45, 44.763163581679905), (60, 59.932543088283225), (56, 55.93916932321403), (39, 43.18052821056659),
(19, 17.56152632753941), (43, 45.256806648622955), (7, 6.350789777123463), (7, 6.835920156443567),
(55, 48.07078513089357), (15, 14.976326392223214), (40, 46.12122262379768), (19, 17.625517630730542),
(40, 40.45475799086336), (6, 6.170038925208685), (40, 40.66429722139645), (40, 39.987487375355386),
(15, 16.167663506047745), (15, 14.837004090951197)]
```

Figure 6.6

As seen in Figure 6.6, the results are mostly really close to the real value. Only two out of the 38 test samples are more than 5% away from the real value.

6.2 Classification Methods

6.2.1 K-Nearest Neighbours

I developed a K-Nearest Neighbors (KNN) model [3] with the independent variables of R, G, B, and area after standardizing the data. I chose not to specify any parameters, allowing the model to use the default values. Upon analysis, the accuracy of this KNN model was found to be 73.68% (see Figure 6.8).

To improve the accuracy, I attempted to optimize the K value (see Figure 6.7) using a plot to visualize the relationship between K and the training and testing accuracy. I found that when K equaled 5, the training accuracy remained high while the testing accuracy was the highest as shown in Figure 6.9.

[3] sklearn.neighbors.KNeighborsClassifier

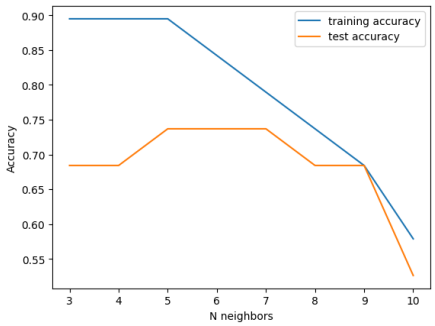

Figure 6.7

To further enhance the model's accuracy, I conducted a hyperparameter grid search with various algorithms, metrics, power, and weights. This analysis resulted in a slight increase in accuracy to 78.95%, with one additional correct prediction (see Figure 6.9).

```
Number of correctly predicted results: 14
Number of incorrectly predicted results: 5
Accuracy: 73.68%
Precision73.68%
Recall:73.68%
F1 score : 0.7368421052631579
Cohens Kappa coefficient: 0.6184738955823292
Classification report for predicted data:
              precision    recall  f1-score   support

        mild       0.43      0.75      0.55         4
    moderate       0.00      0.00      0.00         1
  no_anaemia       0.83      0.62      0.71         8
      severe       1.00      1.00      1.00         6

    accuracy                           0.74        19
   macro avg       0.57      0.59      0.56        19
weighted avg       0.76      0.74      0.73        19
```

Figure 6.8: Baseline KNN model results

```
Number of correctly predicted results: 15
Number of incorrectly predicted results: 4
Accuracy: 78.95%
Precision78.95%
Recall:78.95%
F1 score : 0.7894736842105263
Cohens Kappa coefficient: 0.689795918367347
Classification report for predicted data:
              precision    recall  f1-score   support

        mild       0.50      0.75      0.60         4
    moderate       0.00      0.00      0.00         1
  no_anaemia       0.86      0.75      0.80         8
      severe       1.00      1.00      1.00         6

    accuracy                           0.79        19
   macro avg       0.59      0.62      0.60        19
weighted avg       0.78      0.79      0.78        19
```

Figure 6.9: Optimised KNN model-results

6.2.2 Random Forest

6.2.2.1 High dimensions and random forest

Random Forest is a widely used algorithm for high-dimensional datasets because of its ability to handle many input features without overfitting.

We obtained not just one set of RGB values, but four sets as seen in Figure 6.10, resulting in a total of 13 variables. This data came from the initial most common colour detection and we can leverage this and incorporating these variables while creating the models.

Random Forest is a good choice for a 13-dimensional dataset because it can determine the importance of each feature, allowing for feature selection or reduction. This is especially useful when dealing with high-dimensional datasets where some features may be irrelevant or redundant.

```
Bar 1
  RGB values: (0, 0, 0)
  HSV values: (0, 0, 0)
Bar 2
  RGB values: (209, 101, 95)
  HSV values: (2, 139, 209)
Bar 3
  RGB values: (204, 117, 109)
  HSV values: (3, 119, 204)
Bar 4
  RGB values: (37, 29, 27)
  HSV values: (6, 69, 37)
Bar 5
  RGB values: (115, 79, 74)
  HSV values: (4, 91, 115)
```

Figure 6.10

6.2.2.2 Why we are using PCA

Principal Component Analysis is a powerful tool for reducing the dimensionality of high-dimensional data, improving the generalization of machine learning models, and aiding in data visualization.

It is a technique used to reduce the dimensionality of high-dimensional data by identifying a smaller number of key variables, known as principal components, that capture most of the information in the original data. PCA works by finding linear combinations of the original variables that explain the maximum amount of variation in the data. These linear combinations are the principal components, which are orthogonal (perpendicular) to each other and ordered by the amount of variance they explain. It is useful in our case because high-dimensional data (we have 13 dimensions) can lead to overfitting, which occurs when a model is too complex and fits the noise in the data rather than the underlying patterns.

By reducing the number of variables, PCA can help to prevent overfitting and improve the generalization of machine learning models, which in our case is random forests.

6.2.2.3 Implementation and results

Once again, I split the dataset into training and testing sets and standardized it. After building a baseline random forest model [4] with default parameters, I obtained a recall score of 0.833 with only two minor errors as seen in Figure 6.11.

[4] sklearn.ensemble.RandomForestClassifier

'Baseline Random Forest recall score'

0.8333333333333334

	predicted_no	predicted_mild	predicted_severe
actually_no	1	1	0
actually_mild	1	3	0
actually_severe	0	0	6

Figure 6.11

I then attempted to use PCA [5] to obtain more valuable variables and see if it could increase the accuracy of the random forest model (Figure 6.12 and 6.13). I created a bar plot showing the importance of each feature and another plot(Figure 6.14 and 6.15) that is cumulative, showing that the top 7 features can explain 99.8% of the variance.

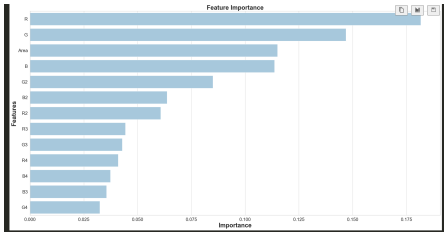

Figure 6.12

	Features	Gini-Importance
0	R	0.181664
1	G	0.146829
2	Area	0.115038
3	B	0.113662
4	G2	0.084978
5	B2	0.063700
6	R2	0.060681
7	R3	0.044302
8	G3	0.042863
9	R4	0.040961
10	B4	0.037367
11	B3	0.035556
12	G4	0.032398

Figure 6.13

[5] sklearn.decomposition.PCA

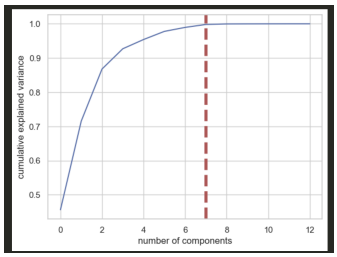

Figure 6.14

	Cumulative Variance Ratio	Explained Variance Ratio
0	0.456160	0.456160
1	0.715227	0.259067
2	0.867644	0.152417
3	0.926974	0.059329
4	0.953986	0.027012
5	0.977779	0.023793
6	0.989471	0.011692
7	0.998055	0.008584
8	0.999620	0.001565
9	0.999812	0.000192
10	0.999914	0.000102
11	0.999987	0.000073
12	1.000000	0.000013

Figure 6.15

I proceeded to build another random forest model with these 7 features (R, G, Area, B, G2, B2 and R2) and used grid search to obtain the best n estimator, max features, max depth, min samples split, and min samples leaf.

However, the results did not turn out as expected as it was worse than the baseline random forest model, with a recall score of 0.667 as shown in Figure 6.16. There was one major mistake where no anemia was diagnosed as severe anemia.

```
'PCA & Hyperparameter Random Forest recall score'

0.6666666666666666
```

	predicted_no	predicted_mild	predicted_severe
actually_no	1	1	0
actually_mild	1	2	1
actually_severe	1	0	5

Figure 6.16

This could result from the most important features that contribute to the target variable not being well captured by the principal components identified by PCA. This can result in a loss of discriminatory power and predictive accuracy.

Furthermore, some algorithms, such as decision trees and random forests, are capable of handling high-dimensional data directly, without the need for dimensionality reduction. In such cases, applying PCA may not result in any significant improvement in performance, and may even reduce accuracy, as important features may be lost during the PCA process.

Chapter 7

Conclusion

	RGB	RGB and Area
Linear Regression	0.891	889
Gaussian Process Regression	0.977	0.987

Table 7.1: Results from the regression methods

	Baseline Model	Optimised Model
K Nearest Neighbours	73.7%	79.0%
Random Forest	83.3%	66.7%

Table 7.2: Results from the classification methods

The K-nearest neighbors (KNN) algorithm and Random Forests yielded satisfactory accuracies. The linear regression results were also acceptable, yet GPR generated remarkably high R squared value. This is due to its suitability for small datasets and its ability to provide estimations of uncertainty for the predictions. In the future, we intend to integrate the GPR (RGB and Area) model into the application we are developing as it gives us a numerical result that is very close to the real value.

Chapter 8

Future Work

The primary objective of the proposed mobile application is to provide users with a simple and easy-to-use platform to predict the severity of anaemia based on a photo of a blood sample. To accomplish this goal, we plan to design a mobile app that will employ cutting-edge technologies such as React Native for the frontend, Flask for the backend, and MongoDB for the database.

React Native is a great choice for this mobile application. It allows for the rapid development of high-quality mobile applications that can be used on both Android and iOS platforms. This means that we can create a single codebase that works on multiple platforms, reducing development time and costs.

The frontend of the mobile application will feature a streamlined and user-friendly interface that includes a login/signup screen, home screen, camera screen, and results screen. The login/signup screen will enable users to sign up for a new account or log in to an existing account. The home screen will serve as the central dashboard for the app, where users can view their past results and access the camera screen. The camera screen will allow users to capture a photo of a blood sample or select a pre-existing photo from their library. The results screen will display the predicted severity of anaemia based on the user's input.

The backend of the application will be developed using Flask, which will handle requests from the frontend and serve the machine learning model.

Flask is a lightweight and flexible web framework that is well-suited for building RESTful APIs, which makes it a great choice for serving machine learning models through an API.

Once the user selects a photo, the app will send a request to the backend, which will preprocess the photo and pass it on to the machine learning model for prediction. The machine learning model will generate a prediction based on the colour intensity and area of the blood spots in the image, which will be sent back to the frontend and displayed to the user.

The machine learning model will be trained using the methodology outlined in the previous section. Once the model is trained, it will be saved and loaded into the backend

of the app. When the user takes a photo or selects a photo from their library, the photo will be preprocessed to extract the necessary data, which will then be passed to the machine learning model for prediction. The model will generate a prediction, which will be returned to the frontend and displayed to the user.

MongoDB is a NoSQL database that is well-suited for storing and querying large volumes of unstructured or semi-structured data, making it an excellent choice for our anaemia severity prediction app.

In our case, we will be storing user information such as email addresses and encrypted passwords, as well as previous results generated by the machine learning model. MongoDB's flexibility in handling complex and dynamic data structures means we can easily store and retrieve this information.

Additionally, MongoDB is a scalable database, which means that as our user base grows, we can easily expand our database to handle the increased load without having to worry about complex schema migrations or downtime.

Another advantage of using MongoDB is its ability to handle geospatial data, which may be useful for future expansion of the app. For example, if we wanted to track the distribution of anaemia cases across different regions, we could easily store and query this data using MongoDB's built-in geospatial features.

When the user logs in, their previous results will be retrieved from the database and displayed on the home screen. This will allow users to easily track their progress over time and see how their anaemia severity comes out.

Bibliography

[1] Pcv. https://labtestsonline.org.uk/tests/pcv. Accessed: 2022-10-29.

[2] A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[3] WebMD Editorial Contributors. Anemia. https://www.webmd.com/a-to-z-guides/understanding-anemia-basics. Accessed: 2022-10-18.

[4] Adam G Gow Craig R Breheny, Steven E Kinsey. The use of smartphone-obtained images to extrapolate canine packed cell volume. https://pubmed.ncbi.nlm.nih.gov/32543750/.

[5] Statista Research Department. Penetration rate of smartphones in selected countries 2021. https://www.statista.com/statistics/539395/smartphone-penetration-worldwide-by-country/. Accessed: 2022-10-26.

[6] Sagnik Ghosal, Debanjan Das, Venkanna Udutalapally, Asoke K. Talukder, and Sudip Misra. shemo: Smartphone spectroscopy for blood hemoglobin level monitoring in smart anemia-care. *IEEE Sensors Journal*, 21(6):8520–8529, 2021.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[8] Ashwini Kalantri, Mandar Karambelkar, Rajnish Joshi, Shriprakash Kalantri, and Ulhas Jajoo. Accuracy and reliability of pallor for detecting anaemia: A hospital-based diagnostic accuracy study. *PLOS ONE*, 5:1–6, 01 2010.

[9] Jiss Kuruvilla, Dhanya Sukumaran, Anjali Sankar, and Siji P Joy. A review on image processing and image segmentation. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 198–203, 2016.

[10] Jennifer Langston. Hemaapp screens for anemia, blood conditions without needle sticks. https://www.washington.edu/news/2016/09/07/hemaapp-screens-for-anemia-blood-conditions-without-needle-sticks/. Accessed: 2022-10-21.

[11] M. Luo, Yu-Fei Ma, and Hong-Jiang Zhang. A spatial constrained k-means approach to image segmentation. In *Fourth International Conference on Information,*

Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, volume 2, pages 738–742 vol.2, 2003.

[12] Robert G. Mannino, David R. Myers, Erika A. Tyburski, Christina Caruso, Jeanne Boudreaux, Traci Leong, Gari D. Clifford, and Wilbur A. Lam. Smartphone app for non-invasive detection of anemia using only patient-sourced photos. *Nature Communications*, 9, 2018.

[13] John W McMurdy, Gregory D Jay, Selim Suner, and Gregory Crawford. Noninvasive Optical, Electrical, and Acoustic Methods of Total Hemoglobin Determination. *Clinical Chemistry*, 54(2):264–272, 02 2008.

[14] Medlineplus. Complete blood count (cbc). https://medlineplus.gov/lab-tests/complete-blood-count-cbc/. Accessed: 2022-10-20.

[15] Duy Nguyen-Tuong, Matthias Seeger, and Jan Peters. Model learning with local gaussian process regression. *Advanced Robotics*, 23(15):2015–2034, 2009.

[16] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *ArXiv*, abs/1712.04621, 2017.

[17] Carl Edward Rasmussen. *Gaussian Processes in Machine Learning*, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[18] Stephanie Robertson, Hossein Azizpour, Kevin Smith, and Johan Hartman. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, 194:19–35, 2018. In-Depth Review: Diagnostic Medical Imaging.

[19] Selim Suner, James Rayner, Ibrahim U. Ozturan, Geoffrey Hogan, Caroline P. Meehan, Alison B. Chambers, Janette Baird, and Gregory D. Jay. Prediction of anemia and estimation of hemoglobin concentration using a smartphone camera. *PLOS ONE*, 16(7):1–16, 07 2021.

[20] J Cardiothorac Surg. Accuracy of the masimo pronto-7® system in patients with left ventricular assist device. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3776432/. Accessed: 2022-10-21.

[21] Tinsae Shemelise Tesfaye, Fasil Tessema, and Habtemu Jarso. Prevalence of anemia and associated factors among "apparently healthy" urban and rural residents in ethiopia: A comparative cross-sectional study. *Journal of Blood Medicine*, 11:89–96, 2020. PMID: 32210654.

[22] Edward Jay Wang, William Li, Doug Hawkins, Terry Gernsheimer, Colette Norby-Slycord, and Shwetak N. Patel. Hemaapp: Noninvasive blood screening of hemoglobin using smartphone cameras. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 593–604, New York, NY, USA, 2016. ACM.

[23] WHO. Anaemia. https://www.who.int/health-topics/anaemia. Accessed: 2022-10-20.

[24] WHO. Worldwide prevalence of anaemia 1993-2005 : Who global database on anaemia. / edited by bruno de benoist, erin mclean, ines egli and mary cogswell. https://apps.who.int/iris/handle/10665/43894. Accessed: 2022-10-29.

[25] C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. pages 599–621, 1998.

[26] Joel A. Wolf, Jacqueline F. Moreau, Oleg Akilov, Timothy Patton, III English, Joseph C., Jonhan Ho, and Laura K. Ferris. Diagnostic Inaccuracy of Smartphone Applications for Melanoma Detection. *JAMA Dermatology*, 149(4):422–426, 04 2013.