Human-Centred AI Masters
September 2025
THES H6011 Research Thesis

Predicting Haemolysis in Clinical Laboratory Samples: A Human-Centred Machine Learning Approach to Operational Risk Factors

Eugene Hoe Yin Lim

A thesis submitted in the partial fulfilment for the
Masters in Science in Human-Centred Artificial Intelligence

in the
Faculty of Computing, Digital and Data
School of Enterprise Computing and Digital Transformation

Main Supervisor: Dr. Fernando Perez Tellez
Industry Partner: Public University Hospital

Declaration

I hereby certify that the material, which I now submit for assessment on the programs of study leading to the award of Master of Science, is entirely my own work and has not been taken from the work of others except to the extent that such work has been cited and acknowledged within the text of my work. No portion of the work contained in this thesis has been submitted in support of an application for another degree or qualification to this or any other institution.

Eugene Lim Hoe Yin
September 6, 2025

Abstract

Haemolysis is a common pre-analytical error that distorts biochemistry results and delays care. This thesis develops and evaluates an explainable machine-learning pipeline to predict haemolysis risk from routinely recorded clinical and logistics data at The University Hospital. Three sources were integrated: PUH chemistry records with timestamps, a courier-route reference (GP site, delivery window, transport mode, route order), and a supplementary national panel used for early baselines. Features captured operational exposure (MinutesFromGPtoLab, MinutesFromLabToResults, delivery metadata) alongside mechanistic biochemistry (Potassium, mmol/L). Following an iterative programmed, OLS/Lasso → Random Forest/XGBoost → neural networks where the problem was reframed from regression to a pragmatic binary classification for Haemolysis < 30 (Low vs High). The final model, a compact multi-layer perceptron with dropout, batch normalization, class weighting and early stopping, achieved balanced performance (accuracy = 0.602; precision High/Low = 0.609/0.596) on a stratified hold-out set. SHAP explanations consistently identified potassium, delivery window and transport mode as dominant drivers. Classical statistics corroborated these signals (Welch's t and Mann–Whitney U for potassium, $p < 0.001$; χ^2 for delivery method and window, $p \leq 0.0035$), while turnaround time showed interaction-driven rather than marginal effects. The pipeline translates hidden logistics into measurable, actionable risk, enabling prioritization of high-risk samples and targeted process checks. Limitations include class imbalance, single-site data, and missing environmental context. Future work targets multi-site validation, richer route instrumentation (temperature/vibration), calibrated decision thresholds, and causal/fairness analyses to support safe deployment as a pre-analytical quality assurance tool.

Keywords
AI, ML XAI, SHAP, MLP

Acknowledgement

I would first like to sincerely thank **Dr. Fernando Perez Tellez**, my primary supervisor, for his invaluable guidance, encouragement, and constructive feedback throughout this research. His expertise in machine learning and healthcare applications has shaped both the technical and conceptual direction of this thesis. I am also deeply grateful to the **clinical laboratory staff at The Public University Hospital**, whose collaboration, domain knowledge, and provision of data made this work possible. Their practical insights into laboratory workflows and haemolysis processes were crucial in ensuring that this project remained clinically grounded and relevant.

I would like to acknowledge my lecturers at **Technological University Dublin**, particularly within the Human-Centred Artificial Intelligence programme, for equipping me with the theoretical and methodological foundations that underpinned this research. Special thanks also go to my peers and colleagues, who provided thoughtful discussions, feedback, and encouragement during the most challenging phases of this project.

Finally, I am profoundly thankful to my family and friends for their unwavering support, patience, and belief in me throughout this journey. Their encouragement has been a constant source of motivation.

Table of Contents

List of Abbreviation 7

1. Introduction 9

2. Literature Review 10

3. Data Collection and Preparation 11

3.1. Data Collection 11

3.2. Ethical Approval 12

3.3. Data Cleaning 12

Figure 3.3.1. Tables showing the variables after the merging of the PUH, KNHANES and PUH route dataset. Along with its data type, units and roles and significance in the project. 13

3.4. Data Visualization (Hypothesized Key Variables) 14

Figure 3.4.1. Average Haemolysis by GP with Standard Deviation 14

Figure 3.4.2. Average haemolysis by delivery window with within-window variability. Bars show the mean HI for each delivery window 15

Figure 3.4.3. Average haemolysis by distance from PUH laboratory. Bars show mean HI for each recorded DistanceFromTUH_KM value. Whiskers are ±1 standard deviation. 16

4. AI Modelling 16

4.1. Model Selection and Development 16

Figure 4.1.1. First 5 rows from the PUH + KNHANES dataset before merging in the courier route dataset. This shows the key variable used in models initially before the final merged dataset at the beginning of this project. 16

Figure 4.1.2. Train and Test Loss for the initial baseline feedforward neural network model 17

Figure 4.1.3. Distribution of the variable “Results” that were fed into the initial baseline neural network model 18

Figure 4.1.4. Classification report for the metrics chosen for the initial baseline model 18

Figure 4.1.5. *Ordinary Least Squares (OLS) regression output with standardised coefficients and Lasso feature selection. The OLS summary provides coefficient estimates, standard errors, and metric statistics, while the standardised coefficients allows for the comparison of effect sizes across key predictors. Lasso regression (bottom) identifies the subset of features with non-zero coefficients after regularisation.* 19

Figure 4.1.6. Flowchart of how the initial statistical and analytical framework models’ function 20

Figure 4.1.7. SHAP Analysis of Random Forest Regressor Prototype 2 with weak predictive performance 21

Figure 4.1.8. SHAP Analysis of Random Forest Regressor Prototype 3 with additional categorical operation variables 22

Figure 4.1.9. Individual SHAP Analysis of XGBoost Classification of Haemolysis Severity Prototype 6 for “Mild” Haemolysis levels 23

Figure 4.1.10. Individual SHAP Analysis of XGBoost Classification of Haemolysis Severity Prototype 6 for “Moderate” Haemolysis levels 23

Figure 4.1.11. Individual SHAP Analysis of XGBoost Classification of Haemolysis Severity Prototype 6 for “Severe” Haemolysis levels 23

4.2. Final Model Selection 24

Figure 4.2.1. Methodology flowchart for the building and selection of the final MLP binary model. 24

Figure 4.2.2. Summary of the final MLP model 25

4.2.1 Hyperparameter Tuning In Final Model 25

4.2.2 Core Technologies 25

4.2.3 Model Evaluation 25

5. Results and Discussion 26

5.1. Results 26

Table 1. Overview of previous experimental prototypes for haemolysis prediction, showing model types, target variables, features and performance. 27
Table 2. Overview of experiments for haemolysis prediction, showing model types, target variables, features and performance. 28
Figure 5.1.1. SHAP Analysis of Final MLP with all categorical operation variables 29

5.2. Statistics 30
Table 3. Statistical Tests Comparing Features Between Low and High Haemolysis Groups 31

5.3. Discussion 31
5.3.1. Recap of Research Question 31
5.3.2. Summary of Findings 32
5.3.3. Interpretation of Findings 32
5.3.4. Addressing Research Questions and Objectives 33
5.3.5. Human-Centered Aspects Of The Project 34
5.3.6. Limitations 34

6. Conclusion and Future work 35

Bibliography 37

List of Abbreviation

AI — Artificial Intelligence

ML — Machine Learning

XAI — Explainable Artificial Intelligence

SHAP — SHapley Additive exPlanations

MLP — Multi-Layer Perceptron

DNN — Deep Neural Network

TL — Transfer Learning

OLS — Ordinary Least Squares (linear regression)

Lasso — Least Absolute Shrinkage and Selection Operator

RF — Random Forest

XGBoost — Extreme Gradient Boosting

SMOTE — Synthetic Minority Over-sampling Technique

GP — General Practice / General Practitioner

PUH — Public University Hospital (study site)

KNHANES — Korea National Health and Nutrition Examination Survey

EHR — Electronic Health Record

LDL-C — Low-Density Lipoprotein Cholesterol

HI — Haemolysis Index

TAT — Turnaround Time (e.g., MinutesFromLabToResults)

ROC — Receiver Operating Characteristic

AUC / AUC-ROC — Area Under the (ROC) Curve

PR-AUC — Area Under the Precision–Recall Curve

F1 — F1-score (harmonic mean of precision and recall)

MAE — Mean Absolute Error

MSE — Mean Squared Error

RMSE — Root Mean Squared Error

R^2 — Coefficient of Determination

SD — Standard Deviation

IQR — Interquartile Range

CI — Confidence Interval

SOP — Standard Operating Procedure

QC — Quality Control

GDPR — General Data Protection Regulation (EU)

EU — European Union

1. Introduction

Every day, thousands of blood samples travel a quiet logistics network from general practices to hospital laboratories. Most arrive intact. A few do not. When red cells rupture in transit (haemolysis) the biochemical results become distorted in testing, reflex tests are triggered, and clinical decisions are delayed. In high-output laboratories, even a small fraction of haemolysed samples multiplies into repeat tests, longer stays, and higher cost. Yet while textbooks detail biological causes (extraction techniques, tourniquet time), the operational journey of a sample (who carried it, when it arrived, how long it waited) often remains invisible in data systems and, by extension, in models that aim to predict risk.

This project starts from that blind spot. Working with The University Hospital (PUH), we asked a simple but clinically vital question: Can we predict haemolysis risk from routinely recorded clinical and logistics metadata, and do so in a way clinicians can understand and act on? To answer it, we combined three data sources: (i) PUH 'routine chemistry results with time stamps; (ii) a curated courier-route reference linking samples to submitting GP sites, delivery windows, and transport modes (courier, taxi, in-house driver); and (iii) a supplementary national dataset (KNHANES) used by PUH clinicians to stabilize missing fields for exploratory baselines. From these, we engineered features that mirror real operational constraints: MinutesFromGPtoLab, MinutesFromLabToResults, Delivery Window, who delivered samples, Route_Order, IsCourierDelivery, alongside (Potassium_mmol/L).

Two design principles shaped the work. First, human-centredness. Predictive models in laboratory medicine must earn trust, not just post accuracy. We therefore raised the importance of explainability from the outset, using SHAP to show how each feature pushes an individual prediction toward "low" or "high" haemolysis risk. Second, methodological flow. Rather than fixate on a single algorithm, we treated modelling as an empirical program, we begin simple, measure, learn, and iterate. We moved from interpretable linear baselines (OLS, Lasso) to non-linear ensembles (Random Forest, XGBoost), then to a compact binary MLP once it was clear the signal was sparse, interactions mattered, and the clinically useful decision boundary was risk or no-risk under a set cap (Haemolysis < 30).

Three challenges dominated this project. Class imbalance, as the vast majority of samples are non-haemolysed and therefore necessitated undersampling and SMOTE in early prototypes, and class-weighting in later neural models. Data veracity, where inaccurate GP collection times required careful cleaning (e.g., replacing negative transit durations with GP-level medians) and a shift from raw timestamps to robust, interval-based features. Generalization posed as the risk that models learn site- or route-specific artefacts was addressed through

strict hold-out evaluation and classical statistics to verify group differences independently of any model.

The contributions are practical and methodological. We show that a small set of routinely captured variables can surface actionable levers are notably potassium level, delivery window, and transport mode that align with clinical intuition yet quantify their effect at scale. We also provide a transparent pipeline that a laboratory can adopt, from feature engineering and quality controls to an explainable classifier that flags high-risk samples before analysis, enabling queue reprioritization or targeted process checks.
In short, this project is about turning the hidden logistics of a laboratory into measurable, explainable risk while also building a model that clinicians can read, contest, and use.

2. Literature Review

With machine learning emerging as one of the most rapidly transformative tools in clinical analysis, where the high-volume nature of data and how they are structured provide an ideal environment for computational modelling to thrive (Rabbani, 2022). Machine learning in this space had evolved greatly in the recent years where it's not able to perform beyond traditional rule-based lab systems by revealing complex, non-linear relationships that may evade human interpretation. In some recent studies, ML have been applied to perform a variety of diagnostic task, such as the estimation of cholesterol (Hwang, 2021), anemia prediction (Azarkhish, 2011), smart reflex testing (McDermott, 2020), and the optimization of routine blood test. These collection of works clearly show the evidence of both premise and challenges of implementing and integrating AI In routine clinical workflows.

(Rabbani, 2022) provide a very comprehensive review on the use and integration of ML in lab medicine, by pin-pointing the specific application in test result prediction, utilization, management, quality control and creating dynamic reference ranges in blood analysis. It was stated in the discussion of this study that Neural Networks and tree-based models such as random forest and XGBoost consistently showed superior performance across tasks compared to linear methods. However, even with the high accuracy of these models, it still struggled with overall generalization and data quality. These issues directly resonate with this current Project with PUH, where instead of finding a general blood variable, its haemolysis levels. These problems are multifaceted and are issues influenced by an array of factors such as transport, processing, etc. Therefore, it requires a very robust interpretable model for any substantial results to be expected.

When choosing the architecture, hyperparameter tuning, and modern NN techniques selection for the DNN model in this project, the study by (Hwang, 2021) played a vital role. Their application of DNN to estimate lipid profiles of patients highlights the clinical potential of ML. (Hwang, 2021) utilized DNNs to estimate low-density lipoprotein cholesterol (LDL-C) values directly from electronic health records (Similar with what this project has been doing with PUH) and demonstrated improved model performance over traditional methodologies and statistical methods such as the Friedewald equation. However, this study has shown the

fallbacks of using these techniques, which in this case is model overfitting when it is trained on population-specific datasets. They mitigated this issue with the use of transfer learning with hospital-specific data. The weights of the transferred learning model were unfortunately not published publicly. This finding is especially relevant for haemolysis prediction, where variability across general practices, couriers, and laboratories may require model retraining or fine-tuning to maintain performance across different variables.

Exploring the extent of the different advantages of nonlinear models proved to be crucial for this project. A good landmark application for the use of nonlinear models is the study done by (Azarkhish et al., 2011), where the use of artificial neural networks to predict iron deficiency (IDA) using blood parameters such as mean cell volume (MVC) and mean cell haemoglobin (MCV). This study closely relates to the current project that we are currently exploring. The study concluded that NN outperformed logistic regression and adaptive neuro fuzzy inference systems in both the diagnosis and prediction aspects of serum iron. The study showed an AUC of 0.982, which showed its robustness and sheer predictive power. Thus superior performance shows the evidence of using nonlinear models to capture hidden patterns in routine lab data and more importantly it shows the emphasis that data quality and appropriate feature selection are highly essential for predictions to be reliable. This insight, as previously said, is mirrored in haemolysis modelling, where operational features such as delivery time windows and transport distance are as critical as biochemical indicators like potassium.

Despite any sort of performance gains, the adoption of AI in any laboratory medicine are limited by the trust of the clinician. Models must not only just predict accurately but also provide transparent reasoning in their final output. Techniques such as SHAP (SHapley Additive exPlanations) proved to be successful in bridging this gap as they act as an attributing model predictions to input features (Lundberg, 2017). In both cholesterol estimation (Hwang, 2021) and reflex testing (McDermott, 2020), interpretability was crucial in the demonstrating of clinical relevance of model outputs. For haemolysis, SHAP allows laboratory staff to see, for example, how elevated potassium or certain delivery time slots increase predicted risk. This aligns with the broader human-centered AI framework that emphasizes explainability, fairness, and actionable insights (Hoche, 2025).

3. Data Collection and Preparation

3.1. Data Collection

As the data was obtained from PUH, we do not know the exact dates these results were taken. We obtained 9922 results from patients that have had their blood tested at PUH, it includes both inpatients and outpatients from all departments that have sent their blood to PUH, however we do not know which are which. Data was anonymized by the clinicians before passing the data on to us, this redacted information were crucial with the compliance to the using of healthcare data (e.g., name, address, gender, patient number).

We also collected data from the Korea National Health and Nutrition Examination Survey (KNHANES) dataset to combine it with the existing data from PUH. This was due to the fact that PUH were not able to provide a larger dataset and some data in the dataset were also missing. Therefore, the clinicians at PUH took the KNHANES dataset and merged it to fill the missing gaps in the existing dataset. The reason for using the KNHANES dataset was its similarity to our PUH dataset and how it was used for another study where they were predicting factors that affect lipid levels in blood (Lee T, 2019). We do however know that the patients who had their blood taken from this survey had blood profiles with 12 hours of fasting.

Additionally, to both these datasets a courier route reference was obtained from the clinicians which was later cross-referenced and merged again with the aforementioned PUH + KNHANES dataset. With the courier information we were able to link each patient's blood sample to its GP and courier type.

3.2. Ethical Approval

For secondary use of PUH lab records, access was governed by the EU GDPR and institutional research ethics approval. The reasoning for processing was Article 6(1)(e) by the hospital as controller, together with Article 9(2)(j) for special-category health data processed for scientific research, implemented with Article 89(1) safeguards (pseudonymisation, data minimisation, access control, retention limits, and audit) (Council, 2016)

3.3. Data Cleaning

Before modelling, rigorous data cleaning was essential to ensure reliability and validity of the analyses. Laboratory datasets, particularly those that incorporate external logistics information, are prone to inconsistencies such as missing entries, implausible timestamps, and heterogeneous variable formats. Since haemolysis is influenced by both biochemical and operational factors, the accuracy of predictor variables (e.g., potassium concentration, transport durations, delivery metadata) directly affects the interpretability and clinical utility of the models. Therefore, this stage focused on selecting clinically relevant variables, correcting errors in time-based measures, standardising data types, and addressing class imbalance. By implementing these procedures in consultation with TUH clinicians, the dataset was transformed into a robust foundation for subsequent machine learning experiments while preserving transparency and reproducibility.

Variable Name	**Type**	**Unit / Levels**	**Role**
Haemolysis	continuous	instrument index (unitless)	**Target**(regression)
HaemolysisScore	ordinal	e.g., 0–5	**Target**(classification)
Potassium_mmol/L	numeric	mmol/L	Mechanistic predictor
MinutesFromGPtoLab	numeric	minutes	Process predictor
MinutesFromLabToResults	numeric	minutes	Process predictor

GP	categorical	submitting practice/site	Context predictor
DistanceFromTUH_KM	numeric	km	Context/route predictor
Who delivered samples	categorical	courier / PTS / staff / other	Route predictor
Delivery Window	categorical	e.g., AM / Midday / PM	Temporal context
DatetimeSampleTaken	datetime	Europe/Dublin	Audit / feature
DatetimeReceivedLab	datetime	Europe/Dublin	Audit / feature
DatetimeResultReported	datetime	Europe/Dublin	Audit / feature
Est time collected from practice	numeric	minutes (est.)	Sensitivity covariate
Results	categorical/text	—	Not modelled

Figure 3.3.1. Tables showing the variables after the merging of the PUH, KNHANES and PUH route dataset. Along with its data type, units and roles and significance in the project.

After seeking domain experience and insights from the clinicians at PUH, it was decided that not all columns were relevant if we were trying to predict features that affect Haemolysis levels. That would include almost all other columns in the initial PUH dataset (e.g. protein, glucose, iron, etc.) DateSampleTaken, TimeSampleTaken, DateReceivedLab, DateResultReported, TimeResultReported, GP, DistanceFromTUH_KM, Potassium_mmol/L, Haemolysis. Therefore, only these variables were chosen from the KNHANES dataset as well. Experts from the clinical analysis department have also provided insights regarding which variables that they know directly corelates with haemolysis levels, which in this case are potassium levels. Therefore, potassium was kept while other blood components were left out. The dataset had been checked and cleaned of null and missing values by the hospital beforehand, which was later double checked and proved to be true.

Additionally, it was also concluded that the exact datetime of sample taken from the GPs and sample received by the lab would not be relevant as the arbitrary duration it spent in transit would be much more relevant to the project. Therefore, we have calculated the actual minutes from GP to lab and minutes from lab to results to replace the datetime variables in the dataset. We were informed by the clinicians at PUH that with the nature of the couriers that they were utilizing at the hospital, there were no way to confirm if the datetime of the sample taken was accurate at all. This was further confirmed when we found entries with durations from GP to lab being 0 or sometimes even a negative value, which is not physically feasible. To correct this, we have replaced all of these values with their respective GP median duration. Duration from lab to results were confirmed to be highly accurate by the clinicians as they were done inhouse and were logged by their very own clinicians.

We also corrected variables and type. Non-numeric key variables in Potassium_mmol/L, Haemolysis, and distance were changed to numeric. We enforced timestamp homogeneity, excluding rows with negative intervals (26 with receipt before collection; 3 with report before receipt). The routes reference was left-joined by GP to further explain lab events with DistanceFromTUH_KM, Who delivered samples, and Delivery Window, delivery-mode labels were standardised. Haemolysis was modelled as continuous and, for classification, as HaemolysisScore (0–5) and, in sensitivity analyses, three bands for Haemolysis < 30. Categorical features were one-hot encoded. Continuous predictors were standardised for linear/MLP models (not for tree-based models). Class imbalance was addressed with class-weights and SMOTE where applicable. These were used due to class imbalance, especially for haemolysis levels, where haemolysis levels ranges from 0 – 788 and 95% of the entries were < 90 and out of that 95%, 85% were < 30.

3.4. Data Visualization (Hypothesized Key Variables)

A quick overview on the key variables that were hypothesized to be highly relevant by clinical professionals.

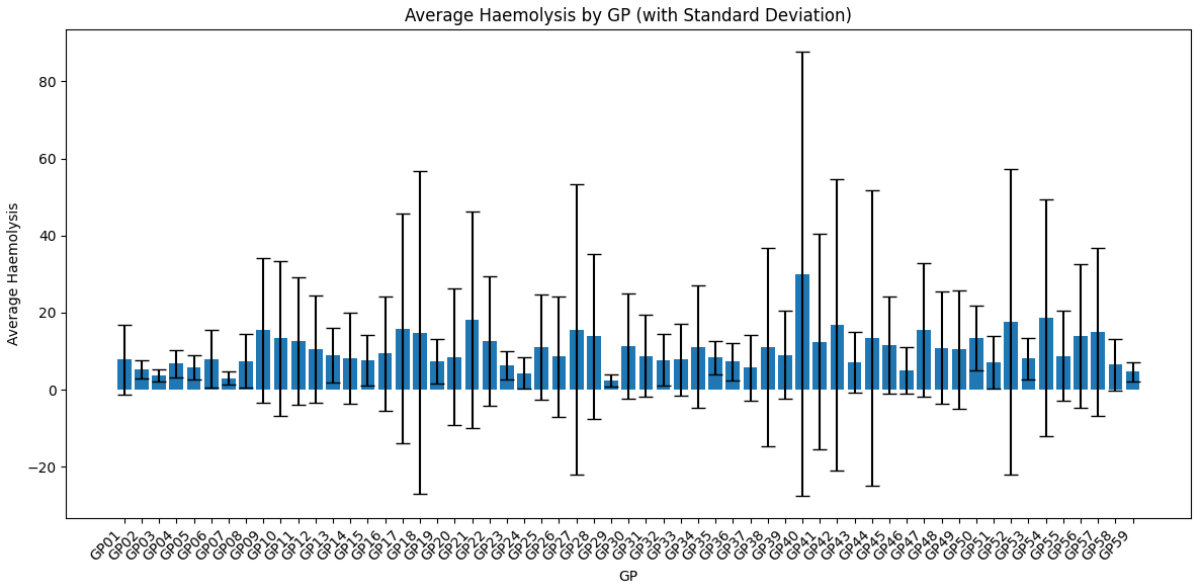

Figure 3.4.1. Average Haemolysis by GP with Standard Deviation

With a quick glance into figure 3.4.1, shows the average haemolysis by GP with within-practice variability. Bars show mean HI per GP, error bars are ±1 standard deviation to display dispersion in each practice's samples. The distribution is right-skewed.

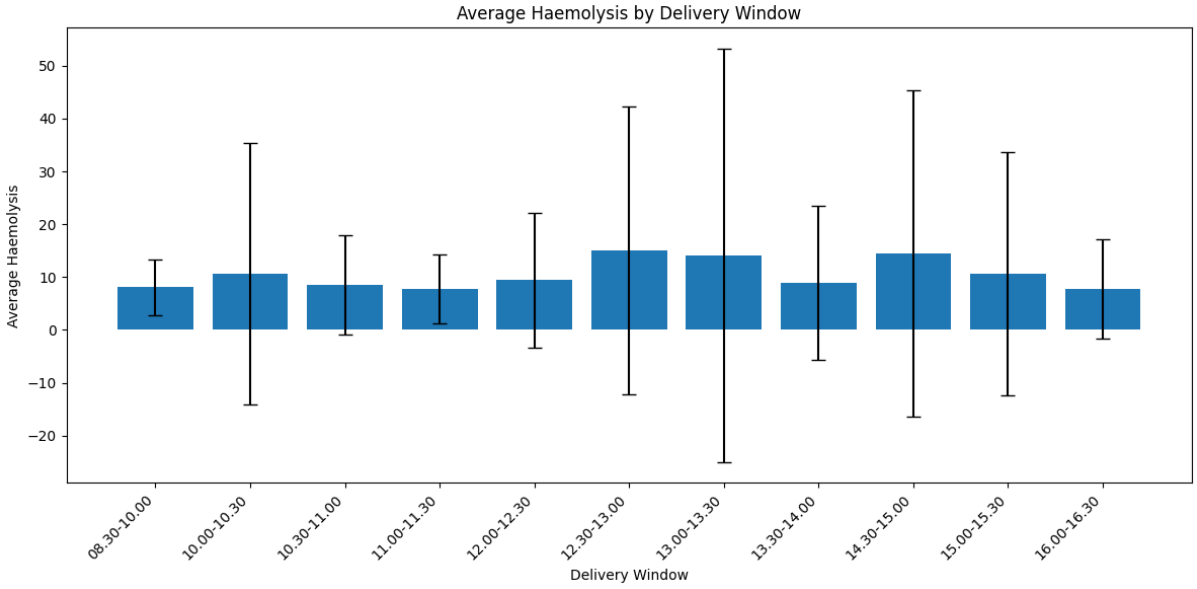

Figure 3.4.2. Average haemolysis by delivery window with within-window variability. Bars show the mean HI for each delivery window

Early windows (e.g., 08:30–10:00, 10:00–10:30) sit near the average with modest spread. Mid-day/early-afternoon windows (about 12:30–15:30) display higher means and much larger variation, consistent with a heavier right tail. Late windows (about 16:00–16:30) appear closer to baseline again. The pattern (Figure 3.4.2) suggests time-of-day operational effects (e.g., batching, queueing, ambient temperature, courier schedules) that may influence haemolysis levels.

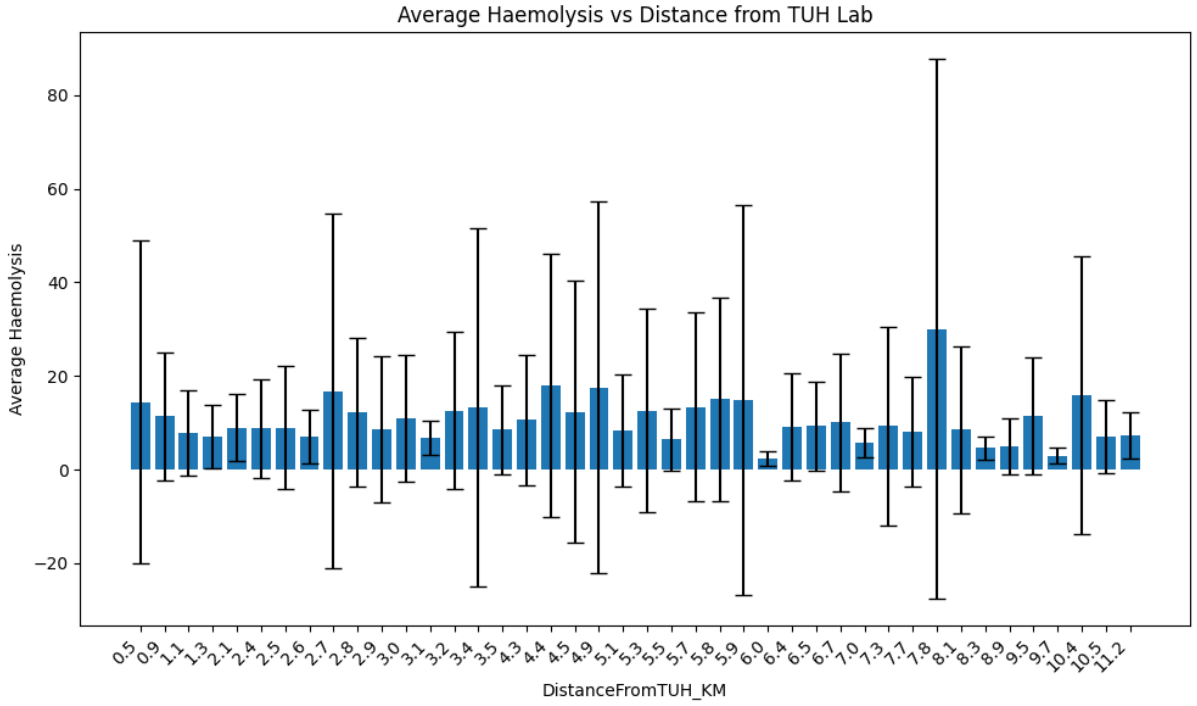

Figure 3.4.3. Average haemolysis by distance from PUH laboratory. Bars show mean HI for each recorded DistanceFromTUH_KM value. Whiskers are ±1 standard deviation.

Across the full range of distance (0.5–11.2 km), central tendency is broadly flat (most means ~5–15), indicating no visually obvious monotonic relation between distance and haemolysis. Several mid-range distances (e.g., 2.8–4.9 km and 7.8–8.2 km) show higher means and/or very wide dispersion, consistent with right-tailed outliers or small sample sizes at those bins (Figure 3.4.3). Overall, the pattern suggests that route/time-of-day factors and site-specific handling likely mediate haemolysis more than distance alone.

4. AI Modelling

The modelling strategy for this project followed an iterative and evidence-driven trajectory, designed to balance predictive performance with interpretability. Given that haemolysis arises from a complex interplay of clinical and operational factors, no single algorithm was assumed to be optimal at the outset. Instead, the process was structured as a progression: beginning with simple linear baselines to establish statistical transparency, then moving through tree-based ensembles to capture non-linear interactions, and finally adopting neural architectures once the need for modelling complex feature dependencies became evident.

4.1. Model Selection and Development

	DistanceFromTUH_KM	Potassium_mmol/L	Haemolysis	MinutesFromGPtoLab	\
0	1.3	4.2	3	0.0	
1	1.1	3.8	6	350.0	
2	6.5	3.9	4	371.0	
3	5.7	4.8	11	31.0	
4	7.8	4.3	36	77.0	

	MinutesFromLabToResults	Results
0	2336.0	0
1	1639.0	0
2	1786.0	0
3	1202.0	0
4	162.0	0

Figure 4.1.1. First 5 rows from the PUH + KNHANES dataset before merging in the courier route dataset. This shows the key variable used in models initially before the final merged dataset at the beginning of this project.

During the period where the project was pending the courier route dataset from PUH, a deep learning model was chosen as the first baseline model. The reason for this was mainly due to the fact that the variables were tabular and that results being binary (0/1), a sigmoid function could be very effective in the prediction of results. The variables used in this model can be seen in Figure 4.1.1 above. For the initial baseline, a feedforward neural network was implemented to classify haemolysis outcomes using available clinical and operational features. The model architecture consisted of three hidden layers (128, 64, and 32 neurons) with ReLU activation functions, followed by a SoftMax output layer for multi-class prediction. Input features were pre-processed through label encoding for categorical values and normalization of continuous variables as seen in figure 4.1.1. The model was trained with the Adam optimizer and categorical cross-entropy loss, using early experiments to benchmark performance. Training and validation accuracy and loss were monitored across 30 epochs, and performance was evaluated using classification reports and AUC-ROC scores. This baseline provided a reference point for model performance prior to the integration of the hospital's routes and delivery data due to the pending dataset from the hospital.

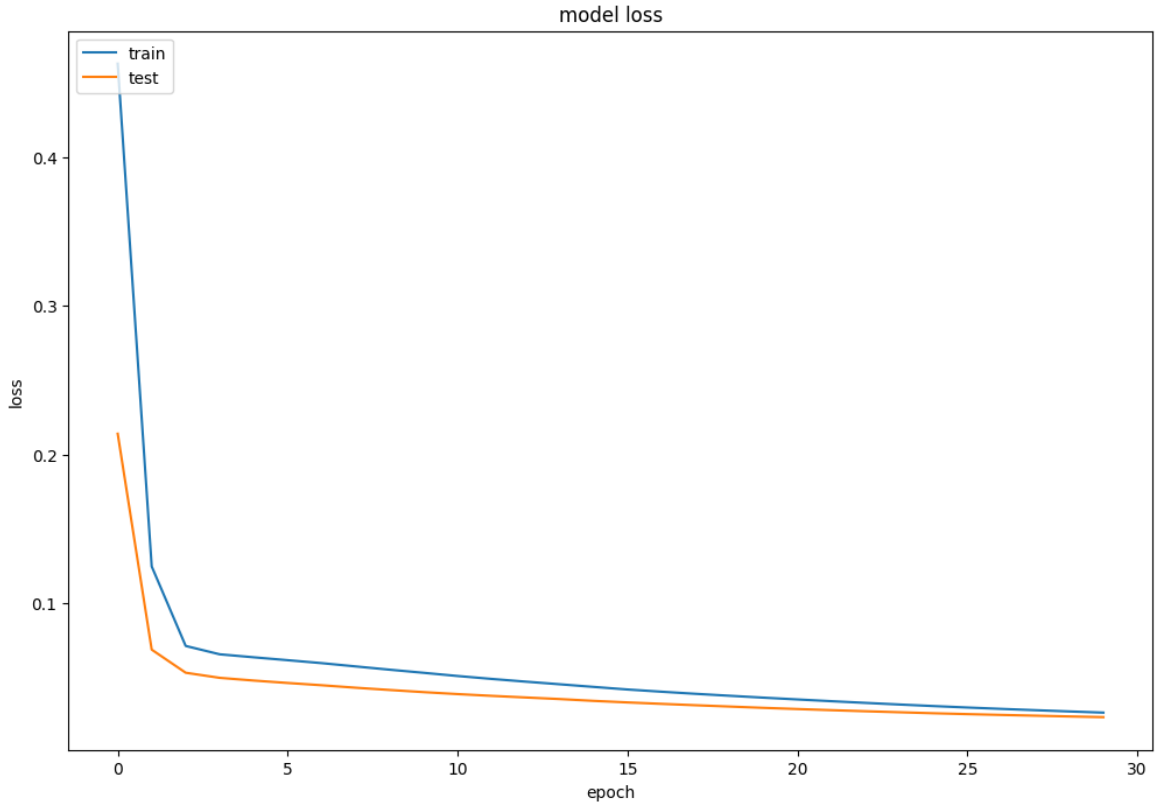

Figure 4.1.2. Train and Test Loss for the initial baseline feedforward neural network model

```
Results
0    9772
1     118
Name: count, dtype: int64
```

Figure 4.1.3. Distribution of the variable “Results” that were fed into the initial baseline neural network model

```
Classification Report:

              precision    recall  f1-score   support

         0.0       0.98      0.99      0.99      9772
         1.0       0.99      0.98      0.99      9772

    accuracy                           0.99     19544
   macro avg       0.99      0.99      0.99     19544
weighted avg       0.99      0.99      0.99     19544
```

Figure 4.1.4. Classification report for the metrics chosen for the initial baseline model

The initial multi-class neural network baseline achieved very high predictive performance on the imbalanced dataset. The training and validation loss curves (Figure 4.1.2) show rapid convergence within the first few epochs, stabilizing thereafter with no signs of overfitting. The dataset, however, was highly imbalanced, with 9,772 samples in the majority class and only 118 in the minority class (Figure 4.1.3). Despite this imbalance, the model produced a classification accuracy of 99% (Figure 4.1.4), with precision, recall, and F1-scores all reported above 0.98 for both classes. While these metrics appear exceptionally strong, they are likely highly prone to inflation due to the extreme class imbalance, meaning that the baseline model may not generalize well for rare Haemolysis outcomes.

After identifying that class imbalance and bias were going to be the main obstacle for this project, the next step was to find and explore correlation and feature importance of hemolysis in the dataset and to later experiment on class balancing techniques on these aforementioned variables and features. To establish an understanding of which features drive hemolysis, both regression and tree-based models were explored after the baseline neural network.

```
=== OLS Regression Summary ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.274
Model:                            OLS   Adj. R-squared:                  0.273
Method:                 Least Squares   F-statistic:                     931.2
Date:                Thu, 05 Jun 2025   Prob (F-statistic):               0.00
Time:                        14:20:55   Log-Likelihood:                -45550.
No. Observations:                9890   AIC:                         9.111e+04
Df Residuals:                    9885   BIC:                         9.115e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         12.2695      0.243     50.389      0.000      11.792      12.747
x1             0.0585      0.244      0.240      0.810      -0.419       0.536
x2           -14.8555      0.244    -60.949      0.000     -15.333     -14.378
x3             0.3572      0.253      1.414      0.157      -0.138       0.852
x4            -0.2660      0.253     -1.053      0.292      -0.761       0.229
==============================================================================
Omnibus:                    15959.858   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         17765176.971
Skew:                          10.383   Prob(JB):                         0.00
Kurtosis:                     209.590   Cond. No.                         1.31
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Standardized OLS Coefficients:
Potassium_mmol/L          -14.855490
MinutesFromGPtoLab          0.357198
MinutesFromLabToResults    -0.266004
DistanceFromTUH_KM          0.058512
dtype: float64

=== Lasso Nonzero Coefficients ===
Potassium_mmol/L          -14.744000
MinutesFromGPtoLab          0.203775
MinutesFromLabToResults    -0.111491
dtype: float64
```

Figure 4.1.5. *Ordinary Least Squares (OLS) regression output with standardised coefficients and Lasso feature selection. The OLS summary provides coefficient estimates, standard errors, and metric statistics, while the standardised coefficients allows for the comparison of effect sizes across key predictors. Lasso regression (bottom) identifies the subset of features with non-zero coefficients after regularisation.*

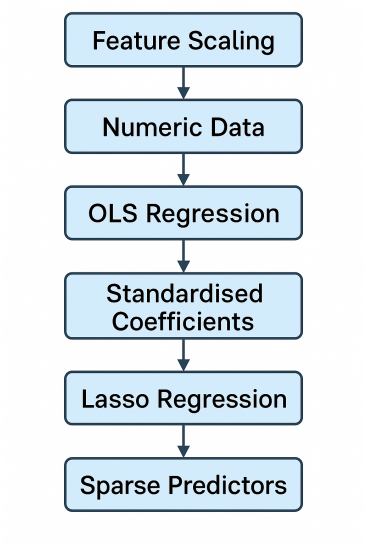

Figure 4.1.6. Flowchart of how the initial statistical and analytical framework models' function

After experimenting on the initial baseline model of the project, we selected linear regression methods to model haemolysis as a continuous outcome due to their human centered nature of the project that allows transparency, interpretability, and suitability for feature selection screening. The variable set was on purpose confined to four continuous variables considered most relevant to haemolysis risk—Potassium (mmol/L), MinutesFromLabToResults, MinutesFromGPtoLab, and DistanceFromTUH_KM, as these represent clinically meaningful and operationally factors according to the clinical professionals at PUH. Prior to modelling, missing values were imputed with the mean and all predictors were standardized using z-scores to ensure comparability and to allow regularization techniques to function accordingly. Two linear models were then developed. First, an Ordinary Least Squares (OLS) regression was fitted to show a simple and interpretable baseline where the coefficients were directly reflecting the relationship between variables and haemolysis. Second, a cross-validated Lasso regression was fitted, this is due to its nature for being able to address potential multicollinearity and reduce the chance of overfitting, thereby identifying the most stable features. This combined design ensured a balance between interpretability and robustness, with OLS offering insight into the magnitude and direction of associations, and Lasso providing a more constrained model that highlights the strongest predictors. These models (Figure 4.1.5) allow for the initial statistical and analytical framework before a more complex non-linear approaches is used in the later experiments.

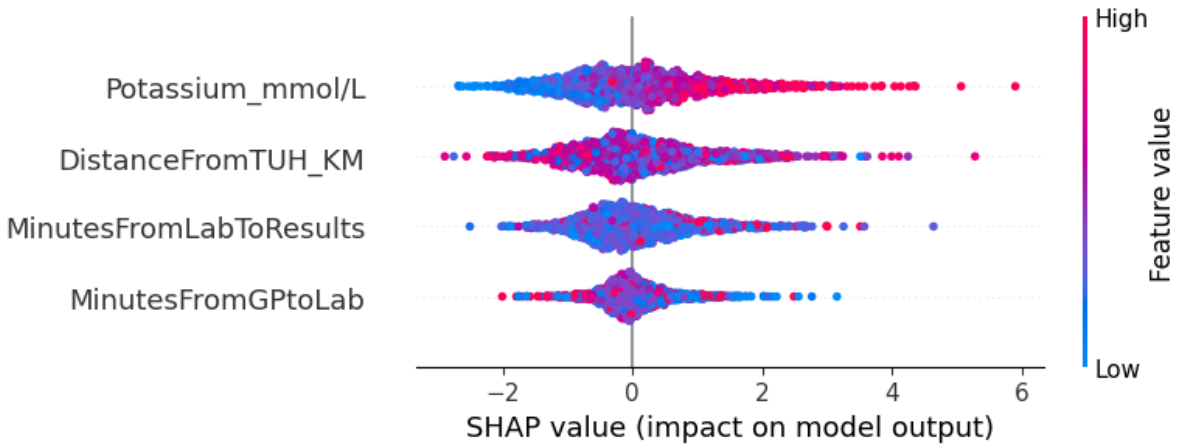

Figure 4.1.7. SHAP Analysis of Random Forest Regressor Prototype 2 with weak predictive performance

Following previous model using OLS and Lasso regression, we progressed through a series of increasingly complex modelling prototypes designed to capture both operational and clinical drivers of haemolysis. The first prototype beyond linear models employed a Random Forest Regressor. While predictive performance on unseen data was poor, SHAP analysis (Figure 4.1.7) revealed operational insights to what drives haemolysis levels.

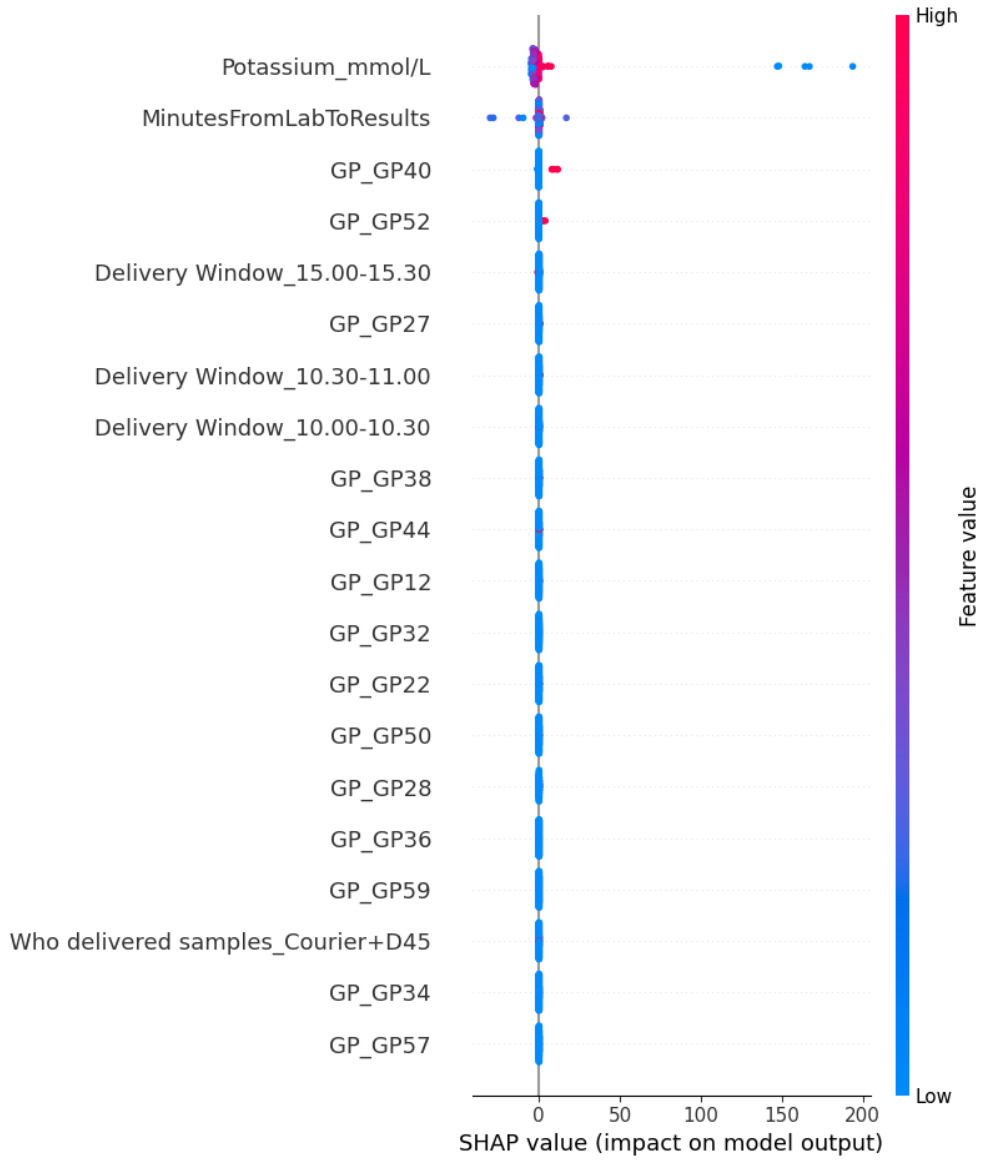

Figure 4.1.8. SHAP Analysis of Random Forest Regressor Prototype 3 with additional categorical operation variables

In the third experiment, we included categorical operational variables (e.g., GP identifier, delivery method, time window). This step showed that categorical features (Figure 4.1.8) added marginal predictive value but required careful feature selection to avoid overfitting.

The fourth prototype we explored whether converting the continuous haemolysis measurement into a binned ordinal score could improve model interpretability and reduce sensitivity to outliers. Haemolysis values greater than 90 were first removed to eliminate extreme cases, and the remaining values were divided into six bins ranging from 0 to 5, each representing a haemolysis score in 15-point intervals (e.g., 0–15, 15–30, etc.). The new variable, ScoreOfHaemolysis, aimed to simplify the prediction target and facilitate easier interpretation of severity levels.

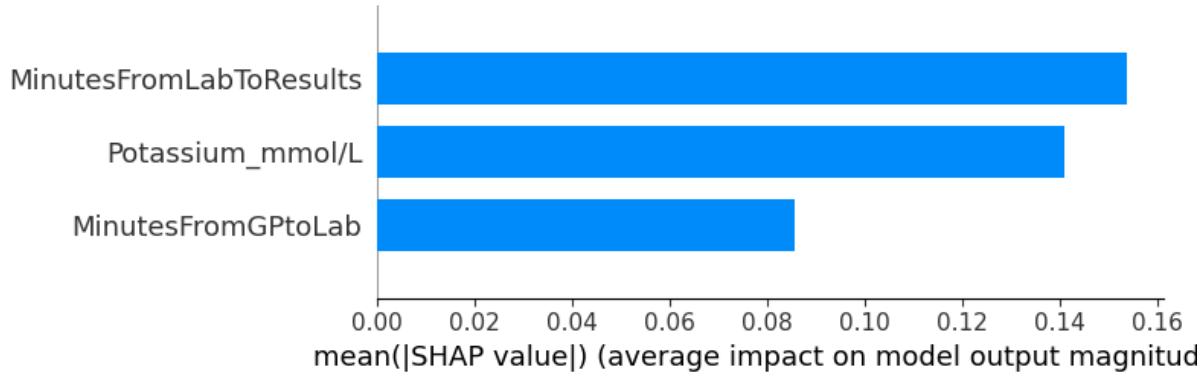

Figure 4.1.9. Individual SHAP Analysis of XGBoost Classification of Haemolysis Severity Prototype 6 for "Mild" Haemolysis levels

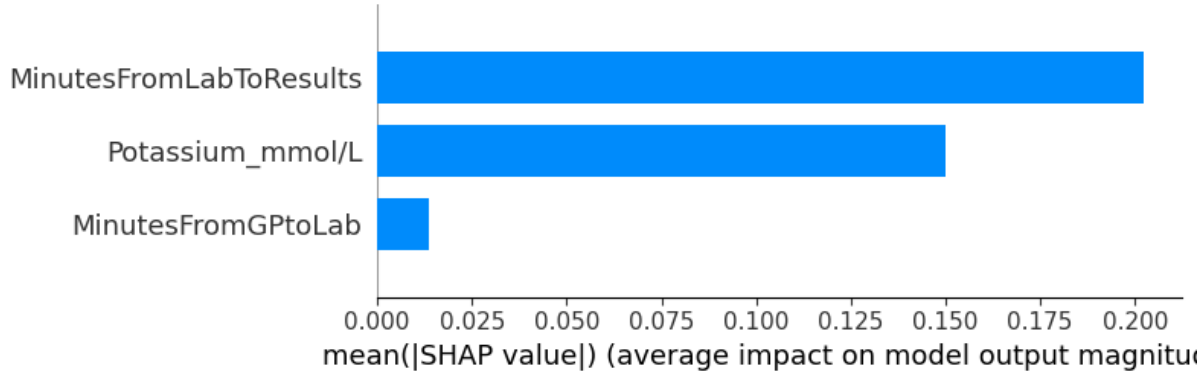

Figure 4.1.10. Individual SHAP Analysis of XGBoost Classification of Haemolysis Severity Prototype 6 for "Moderate" Haemolysis levels

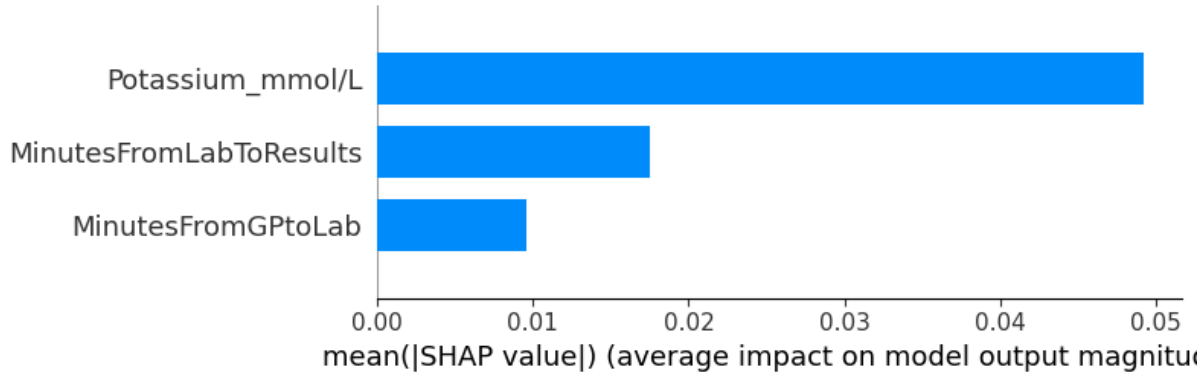

Figure 4.1.11. Individual SHAP Analysis of XGBoost Classification of Haemolysis Severity Prototype 6 for "Severe" Haemolysis levels

In later prototypes, I reframed the task as multi-class classification of haemolysis severity. An XGBoost classifier (Figure 4.1.9,10,11) mapped haemolysis into Mild, Moderate, and

Severe categories, achieving modest accuracy but offering clearer class- interpretability. To address imbalance, the final prototype (Experiment 7) applied SMOTE resampling prior to XGBoost training. This improved classification of the majority (Moderate) class but continued to struggle with minority classes (mild and severe).

4.2. Final Model Selection

For the final stage of this project, a Multi-Layer Perceptron (MLP) was chosen (Figure 4.2.2) for its ability to capture non-linear behaviours and interactions between clinical (e.g., potassium, haemolysis) and operational (e.g., delivery window, courier type, route order) features, these variables that earlier simpler regression models was not able to model effectively. The inclusion of SHAP for explainability was integrated with MLP's suitability, as SHAP provides interpretable feature contributions in high-stakes clinical settings (Ponce-Bobadilla, 2024). We have also chosen to use entries that have haemolysis levels <30 as that was where 85% of the data from the dataset were showing.

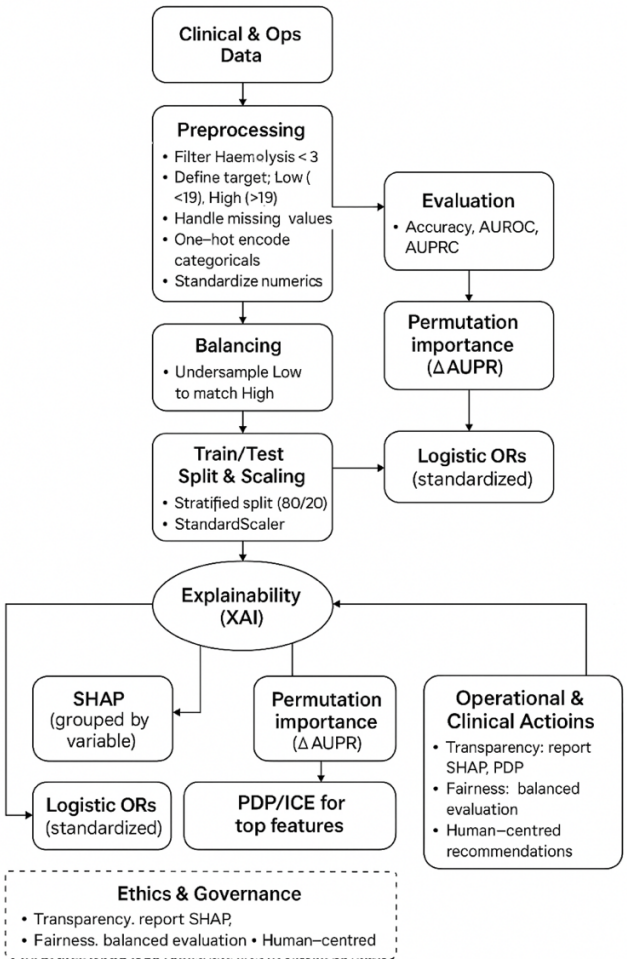

Figure 4.2.1. Methodology flowchart for the building and selection of the final MLP binary model.

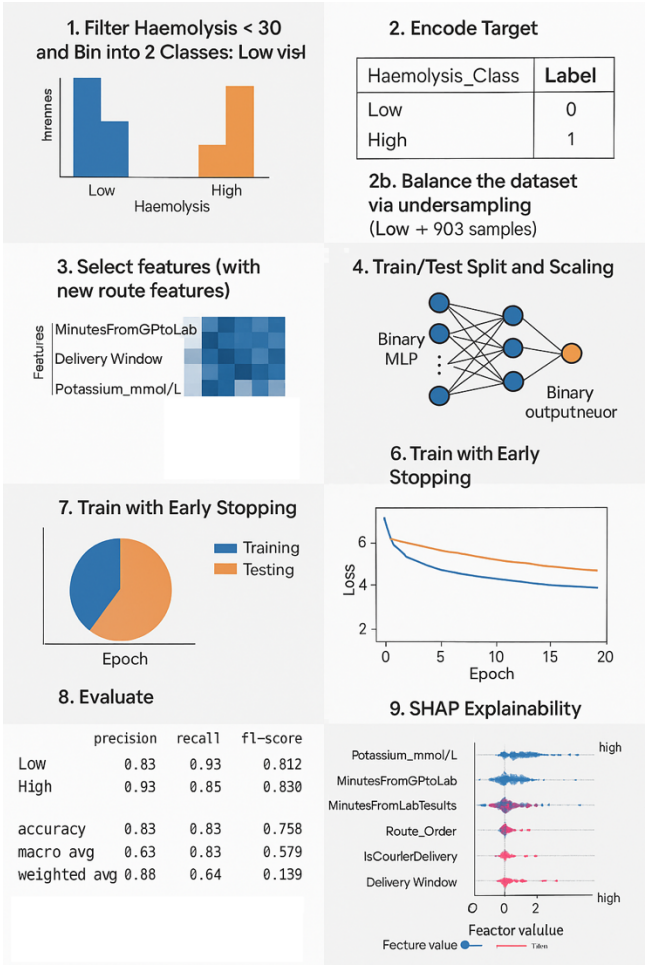

Figure 4.2.2. Summary of the final MLP model

4.2.1 Hyperparameter Tuning In Final Model

The MLP architecture has two hidden layers (128 and 64 units) with ReLU activations—was refined through experimentation. Dropout (0.4, 0.2) and batch normalization inserted between layers helped avoid overfitting, which tends to happen in complex networks. Early stopping (patience=5) and the Adam optimizer with binary cross-entropy loss further regularised training and preserved generalizability.

4.2.2 Core Technologies

Core tools that were used relied on already well-established libraries. TensorFlow and Keras for deep learning architectures and training, scikit-learn for preprocessing and evaluation, Pandas and NumPy for data moulding and manipulation, and SHAP for finding feature importance (Feretzakis, 2024).

4.2.3 Model Evaluation

Accuracy was included because it provides a straightforward measure of overall performance.

Precision, recall, and F1-scores were reported for each class to ensure balanced evaluation. Confusion matrix was used as a complementary tool, offering a transparent view of misclassifications. This allowed inspection of whether the model systematically favoured one class over another, which is key for fairness and trust in clinical AI (Powers, 2011).
SHAP analysis was included not as a performance metric but as an interpretability framework.

5. Results and Discussion

The results of this project are presented as a structured progression of modelling experiments, reflecting the iterative approach taken to address the complexity of haemolysis prediction. Rather than relying on a single model, successive prototypes were developed and evaluated to test different assumptions about the data, ranging from linear regression baselines to tree-based ensembles and deep learning classifiers. Each prototype was designed to answer specific methodological questions: Can haemolysis be modelled as a continuous outcome? Do categorical operational variables add predictive power? Does reframing haemolysis into discrete classes improve performance? And finally, can a balanced and explainable neural model provide clinically actionable outputs?
Performance was reported using appropriate metrics for regression and classification tasks, including R^2, mean absolute error (MAE), accuracy, precision, recall, and F1-scores. In addition, SHAP explainability was integrated to interpret feature contributions across models. Tables 1 and 2 summarise the evolution of models and their performance, while Figure 5.1.1 highlights the SHAP analysis for the final MLP. Together, these results provide a clear picture of both the challenges—such as class imbalance and overfitting—and the actionable insights that emerged from the modelling process.

5.1. Results

Prototype	Model Type	Target Variable	Features Used	Performance
1	Random Forest Regressor	Haemolysis (continuous)	Delivery time, distance, Potassium	R^2 = -0.071 MSE = 34.309 MAE = 4.452
3	OLS & Random Forest Regressor	Haemolysis (continuous)	Categorical + numeric (Potassium, time, distance)	OLS R^2 = 0.299 RF R^2 = 0.516 → 0.654 (reduced), MAE = 7.45

4	Random Forest Regressor	ScoreOfHaemolysis (0–5 bins)	Delivery time, distance, Potassium	R^2 = -0.086 MSE = 0.485 MAE = 0.393
5	OLS & Random Forest Regressor	Haemolysis (continuous)	Delivery time, distance, Potassium	OLS R^2 = 0.034 RF train R^2 = 0.849 (overfit)
6	XGBoost Classifier (weighted)	Haemolysis class (Mild, Moderate, Severe)	Delivery time, distance, Potassium	Accuracy = 0.342 Weighted F1 = 0.351
7	XGBoost Classifier + SMOTE	Haemolysis class (Mild, Moderate, Severe)	Delivery time, distance, Potassium	Accuracy = 0.509 Weighted F1 = 0.426
8	MLP (Neural Network) + SMOTE	Haemolysis class (Mild, Moderate, Severe)	Delivery time, Potassium, categorical delivery metadata	Accuracy = 0.368 Precision (Mild) = 0.277 Precision (Moderate) = 0.577 Precision (Severe) = 0.297
9	Binary MLP Classifier	Haemolysis risk (Low vs High)	Full numeric + categorical delivery metadata	Accuracy = 0.691 Precision (High) = 0.130 Precision (low) = 0.798
10	Binary MLP Classifier (Final Model)	Haemolysis risk (Low vs High) with undersampling	Full numeric + categorical delivery metadata	Accuracy = 0.602 Precision (High) = 0.609 Precision (Low) = 0.596

Table 1. Overview of previous experimental prototypes for haemolysis prediction, showing model types, target variables, features and performance.

Class	Precision	Recall	F1-score	Support
High	0.609	0.569	0.589	181
Low	0.596	0.635	0.615	181
Accuracy			0.602	362
Macro avg	0.603	0.602	0.602	362
Weighted avg	0.603	0.602	0.602	362

	Predicted High	Predicted Low
Actual High	103	78
Actual Low	66	115

Table 2. Overview of experiments for haemolysis prediction, showing model types, target variables, features and performance.

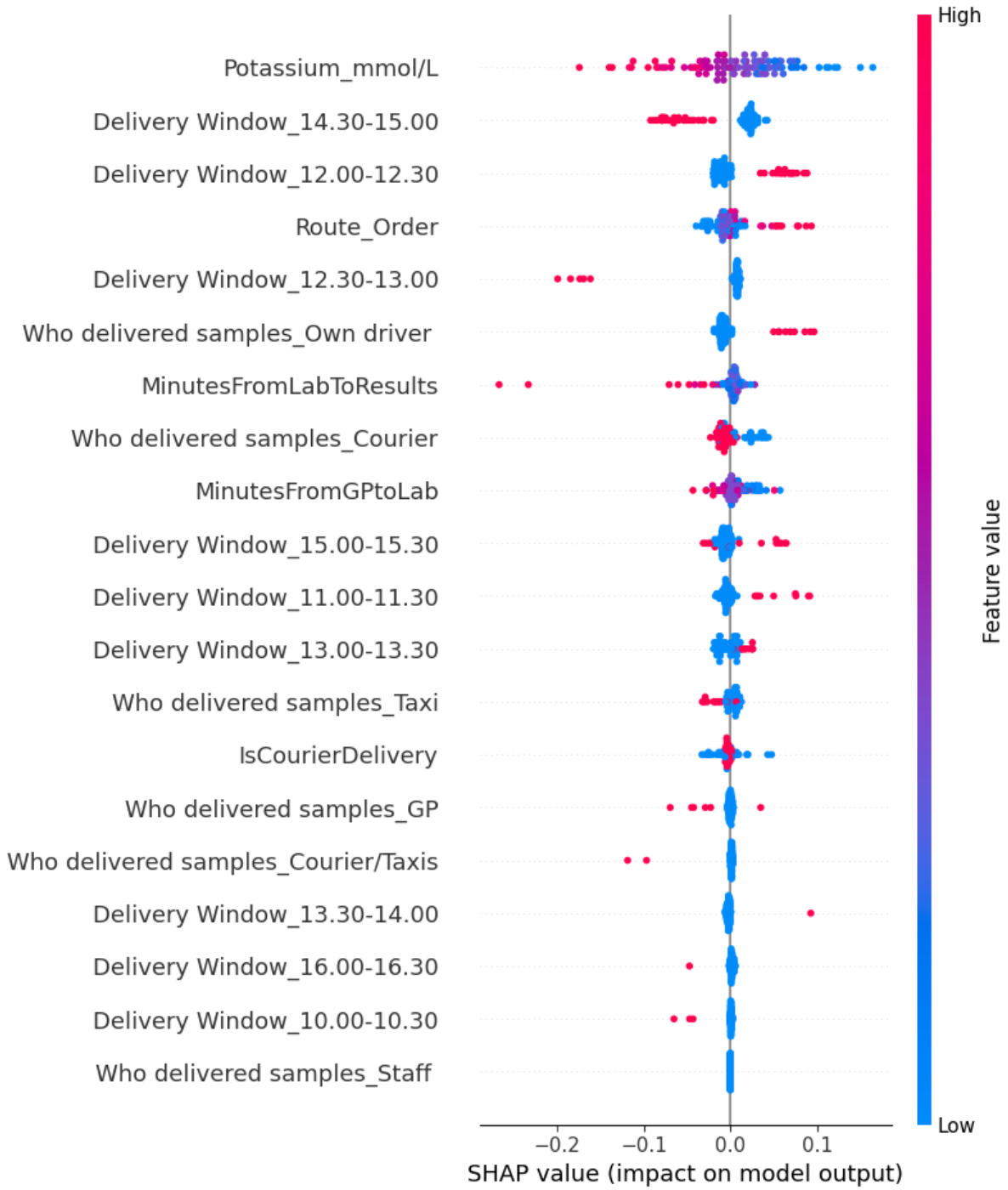

Figure 5.1.1. SHAP Analysis of Final MLP with all categorical operation variables

Results for the initial first baseline deep learning model (Figure 4.1.4) had metrics that showed evidence of inflation (Precision, recall, F1, and accuracy of > 98%) due to class imbalance. Following that, we have decided to move on to modelling a series of experiments to evaluate the factors that contribute to haemolysis and to therefore determine which predictive approaches are suitable (Table 1).

Initial regression models showed performance that were poor, prototype 1 showed a negative explanatory power ($R^2 = –0.071$), suggesting that haemolysis could not be reliably modelled as a continuous outcome. When the introduction of additional categorical features in prototype 3, it modestly improved its linear regression performance ($R^2 = 0.299$), while Random Forest models evidently showed higher explanatory power ($R^2 = 0.654$ after feature reduction), which showed that a non-linear approaches could potentially benefit the project further. Prototype 4 restructured the continuous haemolysis levels as an ordinal score, however, predictive performance remained very low due to the severe class imbalance (Figure 4.1.3).

Prototypes following the previous ones shifted to classification. Weighted XGBoost models (Prototype 6) achieved slight improvement in performance (accuracy = 0.342, weighted F1 = 0.351), however when we balance it with SMOTE (Experiment 7) improved results (accuracy = 0.509, weighted F1 = 0.426). Neural networks (Experiment 8) showed unique and complete new insights into class-specific performance (Figure 4.1.9,10,11), with better recall for moderate haemolysis cases. Binary classification approaches (Prototype 9 and 10) simplified the problem to predicting only Low vs High haemolysis that were < 30. The final undersampled MLP (Prototype 10) achieved balanced accuracy (0.602) and near-symmetric precision across classes (Precision for High = 0.609 and Precision for Low = 0.596), making it the most clinically interpretable and operationally feasible model.

5.2. Statistics

Statistical analyses were calculated and performed by using Python programming language on Google Colab. For a comparison of continuous variables based on 2 classes Low and High, we used the t test, Mann Whitney U test and Welch's t-test. For categorical variables, we used the Chi-squared test, and a P value of <.05 was considered to be statistically significant. This was mainly for statistical hypotheses testing on whether individual features differed between the low and high haemolysis groups.

Feature	Test	Statistic	p-value
Potassium (mmol/L)	Welch's *t*-test	$t = –6.576$	< 0.001
Potassium (mmol/L)	Mann–Whitney U	$U = 333{,}936.5$	< 0.001
MinutesFromLabToResults	Welch's *t*-test	$t = –1.518$	0.129
MinutesFromLabToResults	Mann–Whitney U	$U = 396{,}328.5$	0.305

Who Delivered Samples	Chi-squared test	$\chi^2 = 22.884$	0.0035
Delivery Window	Chi-squared test	$\chi^2 = 59.266$	< 0.001

Table 3. Statistical Tests Comparing Features Between Low and High Haemolysis Groups

For potassium levels, both Welch's t-test ($t = -6.576$, $p < 0.001$) and the Mann–Whitney U test ($U = 333{,}936.5$, $p < 0.001$) indicated a significant difference between groups. For laboratory turnaround time (MinutesFromLabToResults), neither Welch's t-test ($t = -1.518$, $p = 0.129$) nor the Mann–Whitney U test ($U = 396{,}328.5$, $p = 0.305$) found any sort of evidence of a significant difference.

For categorical predictors, the Chi-squared test of independence showed a significant association between delivery method and haemolysis class ($\chi^2 = 22.884$, $p = 0.0035$). Similarly, delivery window was significantly associated with haemolysis ($\chi^2 = 59.266$, $p < 0.001$).

5.3. Discussion

The discussion section brings together the insights gained from the modelling experiments and statistical analyses, linking them back to the research questions and objectives. While the results provided quantitative evidence of model performance and feature importance, the discussion seeks to interpret these findings in a clinical and operational context, highlighting their implications for laboratory practice and patient care. By comparing the behaviour of different modelling approaches, linear, ensemble-based, and neural networks, this section evaluates how haemolysis risk can best be predicted and, more importantly, explained in a way that supports decision-making. The discussion also integrates classical statistical validation with machine learning interpretation techniques such as SHAP to ensure the conclusions are robust and trustworthy.

5.3.1. Recap of Research Question

The primary research question that acted as a guide for this project was: *Which clinical and operational factors most influence haemolysis in blood samples, and how can machine learning models be leveraged to predict haemolysis risk in a way that is both accurate and explainable for clinical use?*

To solve and address this, the study explored an array of modeling solutions, starting with regression-based approaches and slowly shifting into classification-based models and frameworks. With that said, the goal for this study was not only to build a highly predictive model but also to identify which variables were highly influential on haemolysis across both classical statistical testing and machine learning interpretation techniques like SHAP. The

secondary research aim was to also examine how operational logistics, such as courier delivery type and timing interact with variable / biological markers to affect haemolysis.

5.3.2. Summary of Findings

We had a linear approach in the methodology for experimenting with the prototype from the start of this project. Therefore, we can see clear progression in model performance and interpretability across all the prototypes.

Regression approaches (Prototype 1, 3, 4, and 5) proved to be highly inadequate for predicting haemolysis as a continuous output, with most R^2 values close to zero or negative (Table 1). Even when categorical features were added, explanatory power stayed below 30%, indicating that haemolysis levels cannot easily be modelled as a linear function when there are only a small set of variables with bias present in the class.

Classification approaches (Prototypes 6–10) proved to be much more promising in terms of accuracy and overall performance, particularly when haemolysis was binned as a binary problem (Low vs High) and not into three classes (Mild, Moderate, Severe). The final binary MLP classifier with undersampling achieved balanced precision (0.609 for High, 0.596 for Low) and an overall accuracy of 0.602, outperforming earlier regression and MLP models in terms of class balance.

SHAP explainability consistently highlighted (Figure 5.1.1) potassium concentration, delivery window, and courier-related variables as key predictors of haemolysis, which coincides with the initial prediction and hypothesis created by the clinical professionals at PUH.
Statistical validation using hypothesis tests were implemented to strengthened the final findings. Potassium levels were significantly higher in high-haemolysis cases ($p < 0.001$), while both delivery method ($\chi^2 = 22.884$, $p = 0.0035$) and delivery window ($\chi^2 = 59.266$, $p< 0.001$) were significantly influential on haemolysis class. Surprisingly, turnaround time from lab arrival to results was not statistically significant, despite having some influence in ML models.

5.3.3. Interpretation of Findings

Both machine learning models and statistical tests performed in this project has repeatedly confirmed that potassium is the most reliable predictor of haemolysis. Clinically and medically, this aligns with the understanding that haemolysis releases intracellular potassium into plasma, raising measured concentrations (Confirmed by clinical professionals at PUH). The strong statistical significance across multiple tests supports potassium's role not only as a = marker but also as an indirect cause for sample degradation. This also acts as a safety check when modelling regression or machine learning models, as if we see potassium not influencing haemolysis in the way that we know it should, it is a very clear indicator that there is something really wrong with how the model was built and created.

Delivery method and delivery window were highly associated with haemolysis, showing clear evidence that transport conditions play a crucial role in haemolysis levels in blood. The SHAP results showed that certain time slots (e.g., midday deliveries) consistently heightened haemolysis levels in the samples, this could potentially due to traffic congestion, temperature variations, or processing bottlenecks. Differences between courier, taxi, and internal driver deliveries from PUH may also show variations in handling protocols, training, or sample external exposure times.

Although longer MinutesFromLabToResults showed importance in some Random Forest and SHAP analyses, traditional hypothesis testing (Table 3) did not show any evidence of significant group-level differences in Low and High classes. This suggests that turnaround time may not independently determine haemolysis risk but instead interacts with other factors. It proves that the value of machine learning in capturing non-linear relationships that are not easily identified with classical tests.

5.3.4. Addressing Research Questions and Objectives

The study have been consistently able to address the primary research question by pin-pointing potassium concentration, delivery method, and delivery window as the strongest predictors of haemolysis. These findings were robust across different modelling prototypes and validated through a specific methodology and classical statistics.
From an operational point of view, the research demonstrates that logistical factors are not just secondary influences but significant contributors to haemolysis. These key new insights allows the hospital to come up with actionable targets, such as reviewing courier and GP handling practices, optimising delivery scheduling, and monitoring environmental conditions during peak (High influence) delivery windows.
The secondary objective was to build a highly predictive model, which was achieved with the final binary MLP. While performance was modest (accuracy = 0.602), it showed evidence of balanced class-level precision and interpretability via SHAP. Most importantly, the model's main intention is not really to replace laboratory analysis but rather to support workflow optimisation by pin-pointing high-risk samples or delivery situations / scenarios before they happen.

One of the most unexpected findings was that regression models had performed poorly across all experiments. At the start of the project it was anticipated that haemolysis as a continuous variable could be predicted with reasonable accuracy using potassium and operational timings. However, the low R^2 values demonstrated that the complexity of the process and the likelihood that haemolysis is influenced by multiple unmeasured factors (e.g., delivery windows, handling stress, or logistical delays).

5.3.5. Human-Centered Aspects Of The Project

The human-centred qualities of this project ensures that predictive models break free from the traditional “black-box” AI and become much more interpretable AI tools that laboratory staff and clinicians can not only understand but actually trust. By integrating explainable AI techniques and outputs working alongside rigorous statistical validation, identifying influential factors such as potassium concentration, delivery method, and delivery timing are additionally grounded in both data-driven insights and classical AI inference. This makes sure that the project is aligning with emerging AI best practices in human-centred clinical AI design (Gambetti, 2025).

With the incorporation SHAP values in the final model it has enabled transparent, case-level validations, evidently showing how high potassium levels or different delivery windows affect individual predictions toward higher haemolysis levels. This transparent quality in the AI model supports interpretability and fosters clinician confidence, consistent with findings that XAI techniques like SHAP enhance diagnostic safety and decision accountability in a clinical settings (Mienye, 2024).

Additionally, focusing and emphasizing the systemic factors such as delivery windows and courier types, allows the redirection and focus away from individual blame and towards structural improvements. This is in line with the philosophy of human-centred AI, which promotes system-level fairness and empowers stakeholders to act through insights rather than ambiguous and vague algorithmic judgments. Ethical frameworks promotes that AI systems should support equitable care and contextual understanding (Hoche, 2025).

From a clinical and medical perspective, while the final model performance can be considered to be moderate, it retains its operational utility. It does this by flagging any high-risk delivery windows at PUH and the system enables pre-emptive measures of processing and intervention in any of the outpatient and inpatient blood samples. This aligns with principles for explainable clinical decision support systems, where transparency and operational feasibility are not only predictive accuracy but are crucial and vital for long-term clinical machine learning integration.

5.3.6. Limitations

The main obstacles and limitation that was faces in this project was data imbalance, with 118 out of 9922 samples being 1 (Unusable samples) and the rest 0 (Usable samples). The haemolysis dataset was skewed towards low values, making classification in the model very difficult. While undersampling techniques and SMOTE were incorporated in the model as class balancing techniques, class imbalance still plays a big role in the constraining of accuracy models, especially in distinguishing severe haemolysis samples. There were only 5 cases of severe haemolysis (Haemolysis levels > 500) in the entire dataset, which meant that

the model could not feasibly predict these entries and therefore unable to find variables that affect them.

Additionally, the accuracy of the original dataset had also played a significant role as the limiting factor of this project. Clinical professionals at PUH have confirmed that sample taken outside the institutions, mainly from GP's, do not have accurate labels attached to them. In other words, they cannot confirm the accuracy of the time the samples were taken, time in transit, which batch they are from and when exactly did they arrive at the hospital collection point. This was proven true as there were some samples in the original dataset that took <=0 minutes to reach the hospital from the GP's, which is not physically feasible, and a median needed to be calculated to replace those blatant incorrect values.

6. Conclusion and Future work

This thesis set out to answer a practical and clinically relevant question for PUH: can routinely recorded clinical and logistics metadata predict haemolysis risk in a way that clinicians can understand and act upon? Across a structured sequence of modelling prototypes, the study demonstrated that haemolysis cannot be reliably predicted as a continuous outcome with the available variables, but when reframed as a binary classification problem, distinguishing low from high haemolysis under a clinically pragmatic cap (Haemolysis < 30), the signal becomes both meaningful and actionable. The final binary multi-layer perceptron (MLP), trained on engineered process features such as MinutesFromGPtoLab and MinutesFromLabToResults, operational variables including Delivery Window, Route_Order, and Who delivered samples, and the mechanistic biochemical predictor Potassium (mmol/L), achieved balanced performance with an accuracy of approximately 0.60 and near-symmetric precision across both haemolysis classes. More importantly, this model was interpretable through SHAP analysis, which consistently highlighted potassium, delivery method, and delivery window as key predictors. Statistical validation supported these findings, confirming potassium as significantly different across groups and identifying delivery method and timing as strongly associated with haemolysis. Together, these results reframe haemolysis not only as a biological artefact but also as a systemic issue influenced by transport and workflow conditions, thereby providing new operational levers for quality improvement.

Several design principles underpinned the project. Emphasis was placed on explainability over marginal accuracy gains to ensure that predictions could be transparently broken down into feature-level contributions interpretable by laboratory staff. Modelling followed an iterative trajectory from linear baselines to tree-based ensembles and finally neural networks, allowing each step to inform the next based on empirical performance and interpretability rather than algorithmic novelty. Data quality challenges, such as implausible timestamps, were addressed by transforming them into robust interval features and imputing errors with GP-level medians, while class imbalance was mitigated using undersampling and class

weighting. The end result was a pipeline capable of surfacing actionable, operation-linked insights: high-risk delivery windows could be prioritised for processing, courier practices could be reviewed, and samples flagged by potassium levels could be triaged to reduce the likelihood of degradation.

Future work should focus on prospective and multi-site validation to ensure robustness across hospitals and courier networks, alongside temporal validation to capture seasonality and workflow drift. Richer data capture, such as route instrumentation with temperature and vibration sensors, packaging metadata, and queue telemetry could significantly enhance predictive power. Model refinement should include calibrated probabilities, decision-curve analysis, and cost-sensitive objectives to align outputs with clinical utility, while uncertainty quantification would allow low-confidence cases to be flagged for manual review. Transfer learning and continual learning approaches could be adopted to adapt models to new routes or seasonal changes, supported by automated drift monitoring

In conclusion, this project has shown that hidden logistics can be transformed into measurable and explainable haemolysis risk, producing a model that is not only technically sound but also human-centred and operationally relevant. By combining statistical validation with machine learning explainability, the study provides PUH with evidence-based insights and a transparent predictive tool. With further validation, richer data, and integration into routine workflows, the approach has the potential to evolve into a cornerstone of pre-analytical quality assurance, helping laboratories reduce haemolysis rates and deliver more reliable results to support patient care.

Bibliography

Lee T, K. J. (2019). *Deep neural network for estimating low density lipoprotein cholesterol.*

Council, E. P. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation)*. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng?.com

Ponce-Bobadilla, A. V. (2024). Practical guide to SHAP analysis. *Bioinformatics Tutorial*.

Feretzakis, G. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Using SHAP to maintain interpretability while preserving predictive accuracy*.

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*.

Gambetti, A. H. (2025). A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems.

Mienye, I. D. (2024). Explainable artificial intelligence in healthcare: potential, transparency, reliability, ethical compliance. *Computer Methods and Programs in Biomedicine*.

Hoche, M. M. (2025). What makes clinical machine learning fair? A practical ethics framework. *PLOS Digital Health*.

Rabbani, N. K. (2022). Applications of machine learning in routine laboratory medicine: Current state and future directions. *Clinical Biochemistry*.

Hwang, Y. L. (2021). A deep neural network for estimating low-density lipoprotein cholesterol from electronic health records: Real-time clinical application. *JMIR Medical Informatics*.

Azarkhish, I. R. (2011). Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. *Journal of Medical Systems*.

McDermott, M. D. (2020). Using machine learning to develop smart reflex testing protocols. *NPJ Digital Medicine*.

Azarkhish, I. R. (2011). Artificial intelligence models for predicting iron deficiency anemia and iron serum level based on accessible laboratory data. *Journal of Medical Systems*, 2057–2061.

Lundberg, S. M. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.