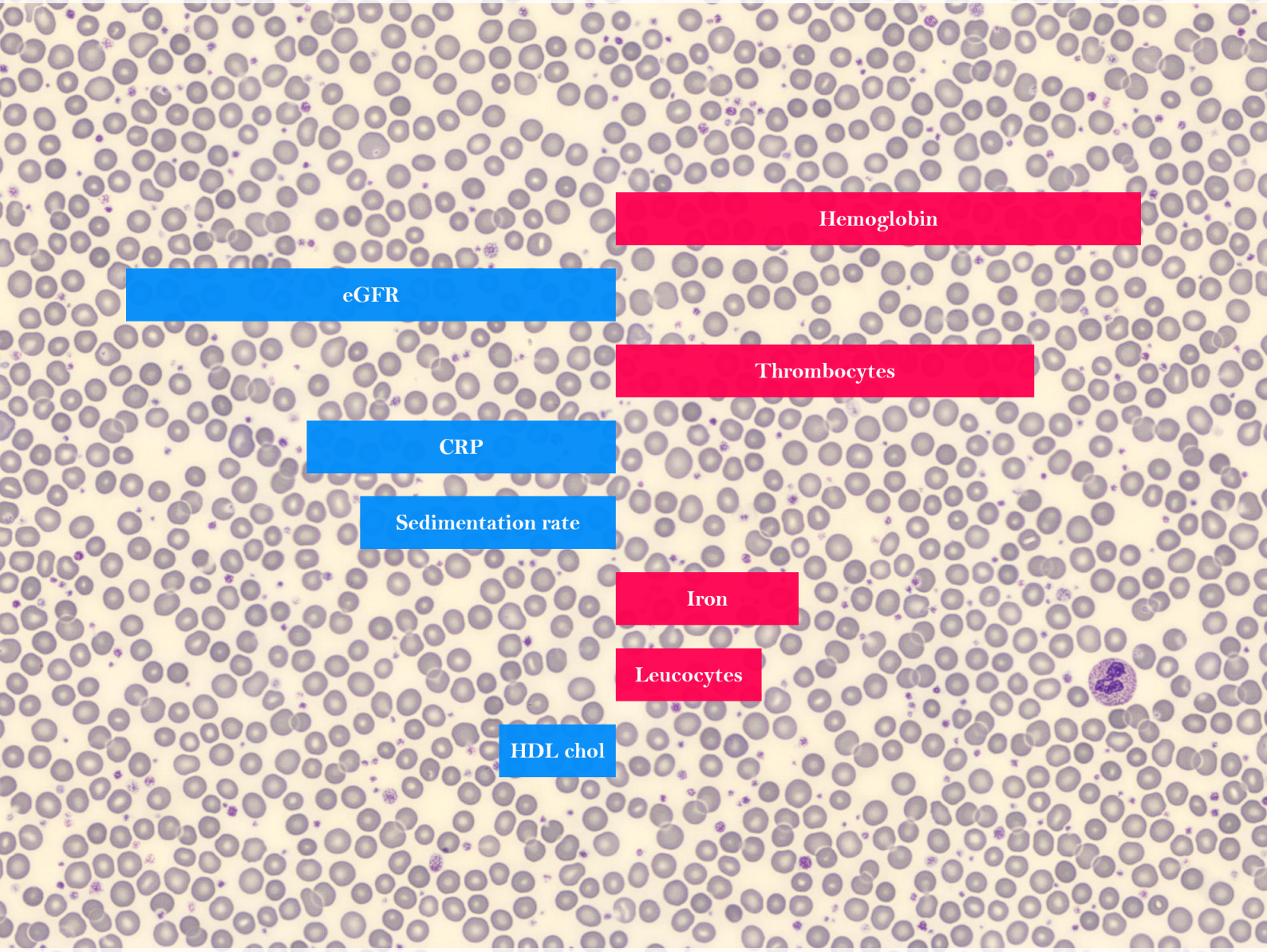
Early MPN detection in laboratory setting

Paul Nijsse



Universiteit Leiden
TU Delft
Erasmus Universiteit Rotterdam
Albert Schweitzer ziekenhuis

This page was intentionally left blank.

Early MPN detection in laboratory setting

Paulus Nijsse

Student Number: 4669746

27-09-2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)
06-03-2023 - 27-09-2023

Supervisors:

dr. Dieckens, Dennis
dr. Riedl, Jurgen
dr. Veenland, Jifke

Thesis committee members:

dr. Veenland, Jifke - Erasmus MC (chair)
dr. Dieckens, Dennis - Albert Schweitzer hospital
dr. Riedl, Jurgen - Result Laboratorium
dr. Stoel, Berend - LUMC
dr. Yavuzyigitoglu, Serdar - Erasmus MC

Universiteit Leiden

TU Delft

Erasmus Universiteit Rotterdam

Preface

This master thesis denotes the end of my time as a Technical Medicine student. Four years of studying on campus, two years of internships in Dordrecht, Rotterdam and Uganda: it has passed by like a breath. Although it was not always easy, it were good years; years of personal and academic development. Thanks to all people involved in this process!

Besides the end of my time as a student in general, this thesis also marks the end of my master thesis internship in the Albert Schweitzer hospital in Dordrecht, the Netherlands. It was good to be there, working together with all those friendly and inspiring colleagues! Many people have been involved in this project or in other side-projects I have done during this internship. Although it is not a complete list, I would like to highlight a few names below.

Thanks to...

- ... Dennis, Jurgen and Jifke for being my supervisors during this internship; thanks for critical perspective, supportive words and inspiring meetings.
- ... Peter, Ruben, Jurgen, Dennis and Sara for forming the basis for this project and continuous involvement as project team.
- ... All people involved in this project connected to the Result Laboratory, including Chantal, Ria, Karlijn and Joosje. The source of this work, all data, would not have been there without your hard work! Also your explanations of methods and results were very insightful and contributed to this project. Also Robin should be mentioned here for kindly providing technical support in making the data automatically available without the need of excessive manual labor.
- ... Colleagues from the Medical Physics and technology department; the friendly atmosphere made me to feel at home. Thanks for teasing me, eating together and all those moments when it was time for 'just a little talk'.
- ... Colleagues of the Radiology department; you were not involved in this MPN project, but you gave me the opportunity to develop my medical skills. Those 'days in white' where a welcome variation next to those days when I barely left my desk while working on this project. You taught me a lot and were very friendly to me: thank you!
- ... all those other people involved in this project or supporting me during this internship (friends, family, colleagues, Technical Medicine study board, etc..). Your names are not mentioned here explicitly, but thank you nonetheless!

With this master thesis, the circle of my study time is closed. Now it is time for the 'real world': working as a Technical Physician. New opportunities and struggles will arise, but I will go in the knowledge of being carried and with the good feeling of having a firm academical basis laid during those six years of education and internships!

Paul Nijsse
Dordrecht, September 2023

Summary

Myelo-Proliferative Neoplasms (MPNs) are a group of bone marrow diseases with potentially lethal cardio-vascular complications. Two sub-diseases of MPN are Essential Thrombocytosis (ET) and Polycythemia Vera (PV), which are recognised by an abnormal blood count of respectively thrombocytes and red blood cells.

If an MPN is treated appropriately, complications for patients are reduced, leading to a relative increase of patients life expectancy. However, MPN is often recognised long after the first clinical signs. 1/4 of MPN patients already had abnormal blood measurements for longer than 1 year in advance of their diagnosis.

Therefore, there is the call for methods for earlier recognition of MPN. Screening like methods could be useful to alert clinicians in case of a suspected case. Although genetic testing is conclusive in recognising MPN, high costs make that they are only applied in case of already clinically suspected MPN.

In this thesis, the outlines of a method are proposed for early detection of MPN patients based on blood measurements in the general hospital laboratory workflow. A two stage solution is proposed:

- Stage 1: Filter on regular blood measurements (combined with demographic data);
- Stage 2: Filter based on microscopy imaging of blood.

The primary scope of this thesis is the development of the first stage for ET and PV subtypes of MPN. A machine learning algorithm called XGBoost is utilized to develop classification algorithms for ET and PV in this stage. Patients with elevated blood platelet counts (ET marker) or elevated red blood cell indicators (PV marker) were separately included in a nested cross validation setup for training and testing of the algorithms. For ET vs control classification, mean metrics obtained during cross validation are AUC: 0.87, recall (sensitivity): 0.74 and specificity: 0.84. For PV vs control corresponding metrics are respectively 0.86, 0.66 and 0.87.

Regarding the development of methods for stage 2, a first step is set. A XGBoost model using cell counts from microscopy images as features results in mean AUC, recall and specificity scores of 0.67, 0.78 and 0.80 respectively when trained and tested using nested cross validation. Training a Convolutional Neural Network (CNN) to take microscopy images as input and return MPN vs control classification resulted in an algorithm which only predicted control cases. These results give an indication of the potential of microscopy for automated MPN recognition, calling for further development of the stage 2 filter.

With this proposed laboratory population screening method and the developed blood measurement based filtering, a next step is set toward early detection of MPN in order to prevent (lethal) MPN related complications.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
AP	Average Precision
AUC	Area Under the receiver-operating-characteristic Curve
BA	Basophil
BL	Blast
CBC	Complete Blood Count
CML	Chronic Myeloid Leukemia
CNN	Convolutional Neural Network
CRP	C-Reactive Protein
eGFR	estimated Globular Filtration Rate
EO	Eosinophil
ERB	Erythroblast
ERC	Thrombocyte aggregation
ET	Essential Thrombocytosis
GT	Giant thrombocyte
HDL	High-Density Lipoprotein
LY	Lymphocyte
MF	Myelo-Fibrosis
MMY	Metamyelocyte
MO	Monocyte
MPD	Myelo-Proliferative Disorder
MPN	Myelo-Proliferative Neoplasm(s)
MY	Myelocyte
PC	Plasma cell
Ph chromosome	Philadelphia chromosome
PMY	Promyelocyte
PR curve	Precision-Recall curve
PV	Polycythemia Vera
ROC curve	Receiver Operating Characteristic curve
SHAP	SHapley Additive exPlanations
SMU	Smudge cell
SNE	Segmented neutrophil
XGBoost	eXtreme Gradient Boosting

Evaluation metrics

Metric	Definition
True positives (TP)	The number or fraction of positive predictions by a model which are also really positive (according to ground truth labeling)
True negatives (TN)	The number or fraction of negative predictions by a model which are also really negative (according to ground truth labeling)
False positives (FP)	The number or fraction of positive predictions by a model which are in reality negative (according to ground truth labeling)
False negatives (FN)	The number or fraction of negative predictions by a model which are in reality positive (according to ground truth labeling)
Precision	(TP) / (TP + FP)
Recall (Sensitivity)	(TP) / (TP + FN)
Specificity	(TN) / (TN + FP)
Accuracy	(TP + TN) / (TP + TN + FP + FN)
F1 score	(2TP) / (2TP + FP + FN)
AUC	Area under the ROC curve; Measure for trade-off between sensitivity and specificity in model. AUC = 0.5: random model; AUC = 1.0: sensitivity and specificity are both 1.0

1

MPN – Clinical background

Myelo-Proliferative Neoplasms (MPN) are a group of bone marrow diseases, caused by a mutated hematopoietic (blood forming) stem cell. Most common MPN diseases are:[1,2,3]

1. Polycythemia Vera (PV)
2. Essential Thrombocytosis (ET)
3. (primary) Myelo-Fibrosis (MF)

MPN is sometimes also referred to as Myelo-Proliferative Disorder (MPD). This covers the same diseases, including Chronic Myeloid Leukemia (CML). Although multiple definitions of MPN are used, some including CML and others excluding CML, in this thesis CML is excluded from MPN. CML is caused by a well-defined mutation, called the Philadelphia chromosome. This is a translocation of the long arms of chromosomes 9 and 22, leading to the formation of the BCR-ABL1 fusion gene.[4] This mutation was one of the first DNA mutations found to cause cancer.[5] As a result of this knowledge, CML was one of the first cancer types to be treated through target cell therapy.[4,6,7,8] Because of its Philadelphia (Ph) chromosome mutation, CML is referred to as Ph-positive MPD; whereas PV, ET and MF are referred to as Ph-negative MPD's or MPN.

Multiple genetic mutations are associated with MPN. The three most common mutations are the JAK2, MPL and CALR mutations.[9,10] Although most of the MPN patients present themselves with at least one of these mutations, there are so called 'triple-negative' MPN cases.[9,10,11] These patients do have the clinical symptoms of MPN, although the three listed mutations are not found.

Symptoms of MPN are mostly non-specific, such as weakness, headache, loss of weight, sweating, bleeding and abdominal fullness.[1] Examination might show hepatosplenomegaly (enlarged liver and spleen) due to compensation of the liver and spleens for failing blood cell production in the bone marrow.[1] Vascular complications might occur due to MPN, such as myocardial, neurological and pulmonary infarctions.[12,13] These complications can be lethal, but with proper treatment complications are reduced and median life expectancy for PV and ET of 15-18 years is reported for patients older than 40 years old.[10] Younger patients have a life expectancy of 35-37 years.[10] For MF, median survival ranges of 4 to 5.5 years are reported, where it is also noted that earlier diagnosis does not influence life expectancy.[14]

MPN's can be recognized by elevated counts in one or more cell lines.[1] For PV, red blood cells (erythrocytes) are increased.[15,16] An increased platelets (thrombocyte) count is characteristic for ET.[15,16] In case of MF, there is an increased production of megakaryocytes (a progenitor cell of thrombocytes), accompanied by bone marrow fibrosis.[15,17] In addition to these typical increased cell counts per disease, also other less typical cell counts might be elevated in MPN.[17,18]

Incidence of MPN is reported to be 1.8 to 5.4 per 100.000.[18,19,20,21] Specified by disease type, both ET and PV have an incidence of 1.0-2.0 per 100.000.[22] For MF, the incidence rate is approximately 0.3 per 100.000.[22] Due to the higher incidence of ET and PV relative to MF, the scope in this thesis is limited to ET and PV.

2

Problem and proposed solution

It is known that the delay between first measured blood count abnormalities and actual MPN diagnosis is relatively long. For an Australian hospital, the median diagnostic delay for MPN patients is reported to be 723 days (n=142, min/max: 0/8731).[23] 25% of the patients included in that observation had potentially preventable thrombotic or hemorrhagic events during diagnostic delay.

In a retrospective analysis in our own hospital, a diagnostic delay is also seen, see figure 2.1. On a cohort of 534 MPN patients, 122 (23%) patients had abnormal blood counts more than one year prior to diagnosis (defined as repeated elevated thrombocyte counts for more than 3 months). Diagnostic delays of more than 3 and 5 years were found for 67 (13%) and 40 (7%) patients respectively. 101 (19%) patients had abnormal blood counts in the year prior to their MPN diagnosis, where 311 (58%) did not have repeated elevated thrombocyte counts prior to their diagnosis. Median diagnostic delay in our hospital was 413 days (min/max: 0/6345).

This delay could be explained by the non-specific symptoms of MPN. Additionally, MPN is a relatively rare disease, which might make that non-hematology specialists are inadequately trained to recognise

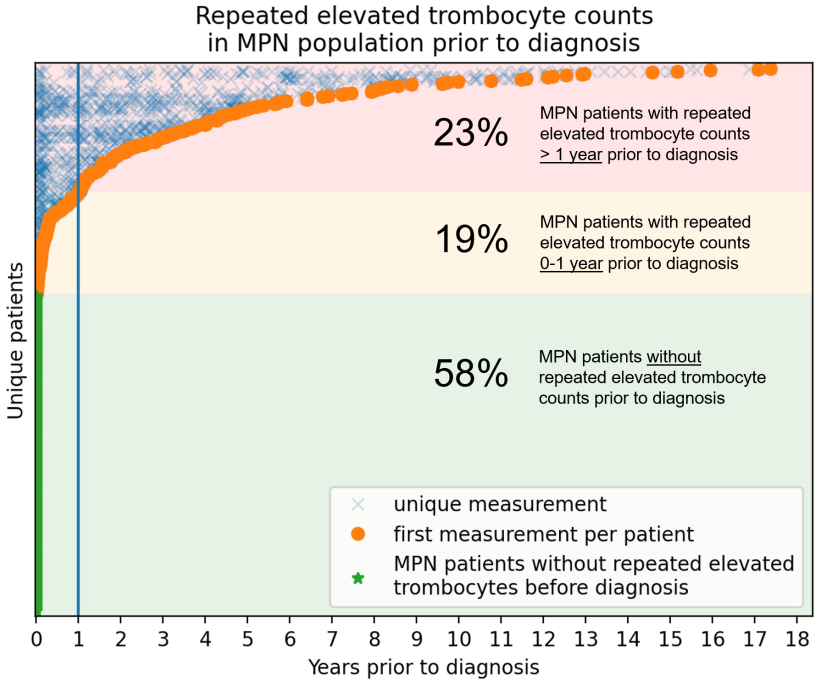


Figure 2.1: Diagnostic delay for MPN (ET/PV/MF) patients seen in the Albert Schweitzer hospital. Repeated elevated thrombocyte counts were defined as periods of repeated elevated thrombocyte counts for longer than 3 months.

potential MPN-like symptoms and laboratory results. Given that thrombotic and hemorrhagic events can be lethal, the presented diagnostic delays indicate that there are also MPN patients who have never reached their time of diagnosis due to potentially preventable death.

To prevent diagnostic delay, active screening might be performed. The gold standard for MPN diagnosis is genetic testing for MPN related mutations, such as the JAK2, MPL and CARL mutations.[24] Genetic testing, however, is expensive and it is desirable to find a low cost alternative to detect MPN patients. In this work we investigate an alternative solution using a multi-stage filter approach.

By automatically analyzing laboratory measurements this might be realized. Here we propose a multi-stage filtering methodology:

1. In the first stage, a filter should select suspected Complete Blood Counting (CBC) measurements in the routine laboratory workflow, where false positives are accepted to a certain extent.
2. The second stage consists of further microscopic analysis of the suspected cases, which can be done using the same blood sample used for CBC. This is a relatively cheap method, aiming to reduce the number of false positives in the suspected population, while keeping the true MPN patients.

If a patient's sample both passes the filters of stage 1 (CBC) and 2 (microscopy), genetic testing can be performed and/or an explicit notice to the requesting physician can be given that their patient is highly suspected for MPN and should be seen by an hematologist.

This method supports clinicians in recognising MPN, with only limited impact on normal workflow and low costs compared to screening-wise genetic testing. It aims to decrease diagnostic delay and thus reduce potentially preventable complications and deaths.

3

Stage 1: blood measurement filter

Blood measurements, such as a Complete Blood Count (CBC) are often performed in clinical practice. They give information about blood composition, metabolism and organ function. CBC is performed in a wide variety of cases, such as suspected infection, inflammatory processes, immune deficiency or anemia.[25] Also for pre-operative screening and follow-up of hematology patients CBC measurements are used.[25] The wide variety of indications for blood measurements make that this is a common diagnostic tool in clinical practice. Also, this means that more information might be present in the laboratory results than used by the clinician. Because MPN is not always recognised by clinicians, a support system might help to recognise MPN suspected laboratory results. For that reason, an algorithm is developed and tested, based on laboratory data, for the detection of MPN suspected laboratory results. Primary scope of this chapter is the development of this algorithm for detection of ET patients, as secondary secondary test the same pipeline is also applied for detection of PV patients.

3.1. Methods

3.1.1. Data collection

Laboratory results were retrospectively obtained from the routine workload of the Result Laboratory Dordrecht for ET, PV and non-MPN patients in the years 2000-2022. Patients age, gender and smoking status were coupled to the measurements. Two datasets were formed, one with ET and non-MPN patients (ET dataset), the other with PV and non-MPN patients (PV dataset). Thresholds for laboratory measurement scores in both datasets are based on the 2016 WHO classification of MPN.[26]

Data in the ET dataset was filtered based on 4 criteria:

1. Patients age above 17 and below 85;
2. Thrombocyte count in sample should be above 450;
3. Patient had no thrombocyte count below 450 in 3 months prior to sampling date;
4. Patients thrombocyte value prior to selected sample was above 450 and was measured within one year prior to sampling date of selected sample.

Data in the PV dataset was filtered based on the following criteria:

1. Patients age above 17 and below 85;
2. For male patients: hematocrit > 0.49 l/l or hemoglobin > 10.2 mmol/l;
3. For female patients: hematocrit > 0.48 l/l or hemoglobin > 9.9 mmol/l;
4. All known measurements for the patient within 3 months prior to sampling date fit the threshold for hematocrit and hemogobin;
5. Patients hematocrit or hemogobin value prior to selected sample was above the threshold.

Parameter	Base value
n_estimators	26
max_depth	6
min_child_weight	10
reg_alpha	3.1
reg_lambda	35
gamma	0.12

Table 3.1: Base parameters of XGBoost models used for initial feature selection and evaluation of the effect of hyper parameter values on AUC outcome.

Parameter	Search range
Number of estimators	2 – 200
Max depth	1 – 30
Minimum child weight	0 – 100
Alpha (L1 regularization)	0.003 – 100
Lambda (L2 regularization)	5 – 500
Gamma (min. split loss reduction)	0.01 - 316

Table 3.2: Search ranges in hyper parameter tuning.

For MPN patients in both datasets, only measurements before the date of diagnosis were included. Multiple measurements per patient were allowed to be included. Ground truth labeling was clinical diagnosis, which is based on laboratory results (including genetic testing) and further clinical information, such as physical examination and imaging. Laboratory measurements performed in less than 1% of the samples were excluded from the dataset. If possible, unknown values for measurements were imputed through last-known-data imputation. In case of no older known values, unknown values were tolerated.

3.1.2. Machine Learning model

An eXtreme Gradient Boosting (XGBoost) algorithm was trained to give an probability score to laboratory measurements (0-1: low-high ET or PV probability). XGBoost is described by its authors as 'a scalable end to-end tree boosting system'.[27] The XGBoost algorithm sequentially creates a set of decision trees (ensemble). The sum of each decision tree's outcome is used as final prediction score. Each newly created decision tree is formed in a way aiming to correctly classify those samples which were not correctly classified by the ensemble of earlier created trees. In contrast to some comparable algorithms (such as most random forest implementations), XGBoost can deal with missing values. For each node in the decision trees, branch directions are determined during training for missing values.

The model was trained and evaluated by applying outer cross validation. For each split, the following pipeline was used: data preprocessing, initial feature selection, hyper parameter tuning, definitive feature selection, model training, model performance evaluation.

3.1.3. Nested cross validation

Patients in the full dataset were split by a stratified 5 times repeated 4-fold splitting in train and test groups (outer cross validation). Inner cross validation on the train groups was performed for feature selection and hyper parameter tuning, see corresponding sections. Utilizing the selected features and hyper parameter values, training is done on the outer cross validation train groups. Evaluation of the trained models is performed on the outer cross validation test groups.

3.1.4. Under sampling

The initial train-test groups are obtained by splitting on patient level. All collected samples per patients are initially included. Random under sampling is performed to achieve an equal number of control and ET or PV samples in both the test and train sets.

ET	ET patients	ET samples	Control patients	Control samples
Train	91.5 (0.5)	519.8 (33.9)	480.4 (28.6)	519.8 (33.9)
Test	30.5 (0.5)	173.3 (33.9)	159.6 (29.2)	173.3 (33.9)

PV	PV patients	PV samples	Control patients	Control samples
Train	43.5 (0.5)	162.0 (10.1)	145.5 (8.9)	162.0 (10.1)
Test	14.5 (0.5)	54.0 (10.1)	49.2 (8.9)	54.0 (10.1)

Table 3.3: Average patient and sample numbers (sd) in outer cross validation train and test sets for ET and PV datasets.

3.1.5. Feature selection

Feature selection is performed through recursive feature elimination with 5 fold cross validation . For the initial feature selection, an XGBoost classifier with non-optimized hyper parameters is used, see table 3.1. After hyper parameter tuning using initially selected features, feature selection is again performed using the XGBoost classifier with the optimized hyper parameters.
At start of the feature selection procedure, all features are included and the model is trained. The feature with lowest feature importance is removed from the feature set, after which the model is trained again. This is repeated until only a single feature is left over. The feature set with highest accuracy is used for further training of the model.

3.1.6. Hyper parameters

Tuned hyper parameters for the XGBoost model are:

- the number of classification trees (estimators) per classifier,
- the maximal tree depth,
- minimum child weight required for the splitting of leafs,
- alpha value (L1 regularization),
- lambda value (L2 regularization) and
- gamma value (minimum split loss reduction).

For training of the models, optimal hyper parameter values were searched for using a 5-fold cross validated randomized search with folds of 50 iterations. For the search domains, see table 3.2.

3.1.7. Model performance evaluation

Outer cross validation is utilized to evaluate performance metrics. Primary evaluation metric is the Area Under the receiver-operator-Curve (AUC). Secondary metrics are Average Precision (AP), precision (positive predictive value), recall (sensitivity), accuracy, specificity and F1 score. Feature importance and hyper parameter values are analyzed for the trained classifiers.

3.1.8. Experiments

For both the ET and the PV dataset, nested cross validation is preformed using all available data in the dataset. Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were created and performance metrics obtained based on the model performance on the test sets. Feature importance

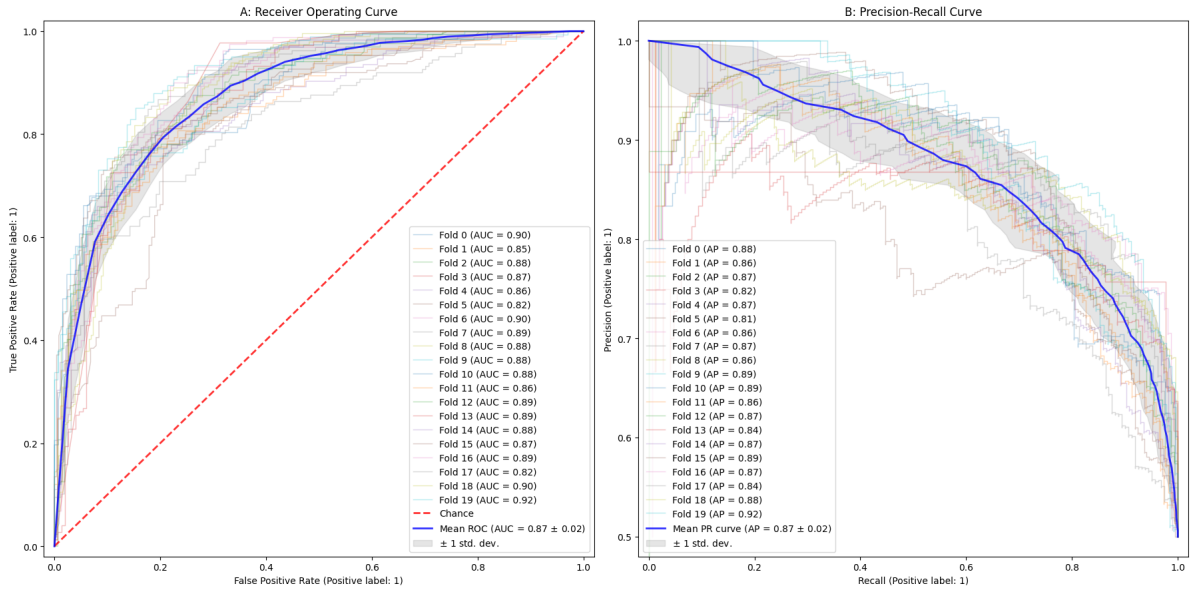


Figure 3.4: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for outer cross validated ET classifiers. A: ROC's with corresponding Area Under the Curve (AUC). B: PR-curves with corresponding Average Precision (AP).

and SHapley Additive exPlanations (SHAP) for each of the outer cross validation folds are obtained and visualized. SHAP plots visualize the impact of a feature value on the model outcome. Diagnostic delay in the test sets are analyzed, where diagnostic delay is defined as time between retrospective positive prediction by our model and actual date of diagnosis. For the ET dataset, distributions of selected hyper parameters in the trained cross validation models are obtained and visualized.

Using the ET dataset, the effect of the number of included training samples is evaluated. Three times a random split on the full dataset was performed with a test fraction of 0.3. The AUC values for the models trained for each split as function of the number of samples actually included (MPN:control ratio 1:1) is evaluated, where both AUC values for the train set and test set are visualized. Also the effect of hyper parameter values on AUC is evaluated for the ET dataset through nested cross validation. For this purpose, only single feature selection is performed and the base parameters as defined in table 3.1 are used, except for the variable feature for which the impact on AUC performance is analyzed.

For training and testing as described above, multiple samples per patient are allowed. The same methods are also applied to train and test the ET classifier in the situation when only the last known measurement per patient in the ET dataset are included. ROC curve, PR curve and evaluation metrics are obtained for this setup.

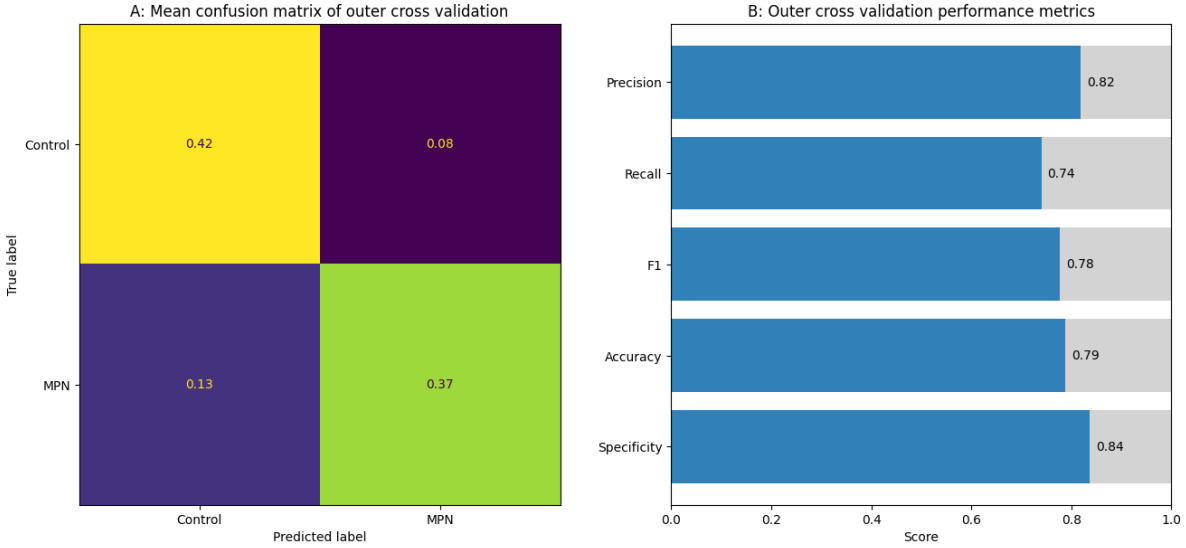


Figure 3.5: Mean confusion matrix of outer cross validation folds (A) and corresponding performance metrics (B) for ET classification.

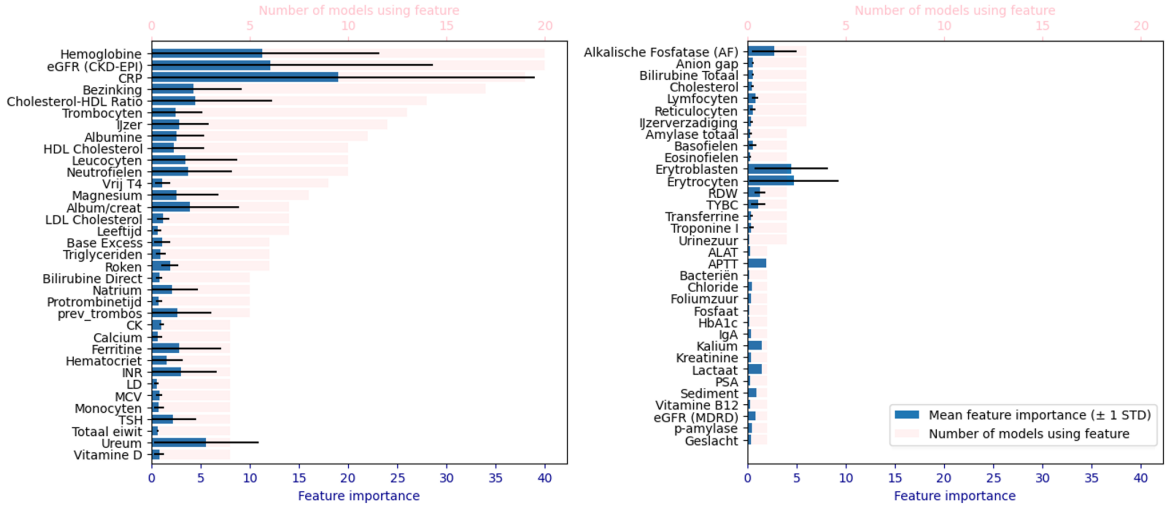


Figure 3.6: Importance of features used in outer cross validation ET classifiers and number of ET classifiers using the features. Total number of classifiers is 20. Feature importance is calculated as the average gain achieved through splits based on that feature.

3.2. Results

3.2.1. ET classification

Samples from 122 ET patients and 11962 control patients were collected. A summary of the feature values in the dataset is provided in appendix A.1. Boxplots for feature values are visualized in appendix B.1. Average number of patients and samples per group in the outer cross validation are shown in table 3.3.

The mean AUC tested through outer cross validation is 0.87 (sd: 0.02), see figure 3.4. Mean Average Precision (AP) derived from Precision-Recall curves is 0.87 (sd: 0.02), see figure 3.4. Mean confusion matrix is shown in figure 3.5; derived values for precision, recall, F1, accuracy and specificity are respectively 0.82, 0.74, 0.78, 0.79 and 0.84.

The mean feature importance in the models created in outer cross validation are visualized in figure 3.6,

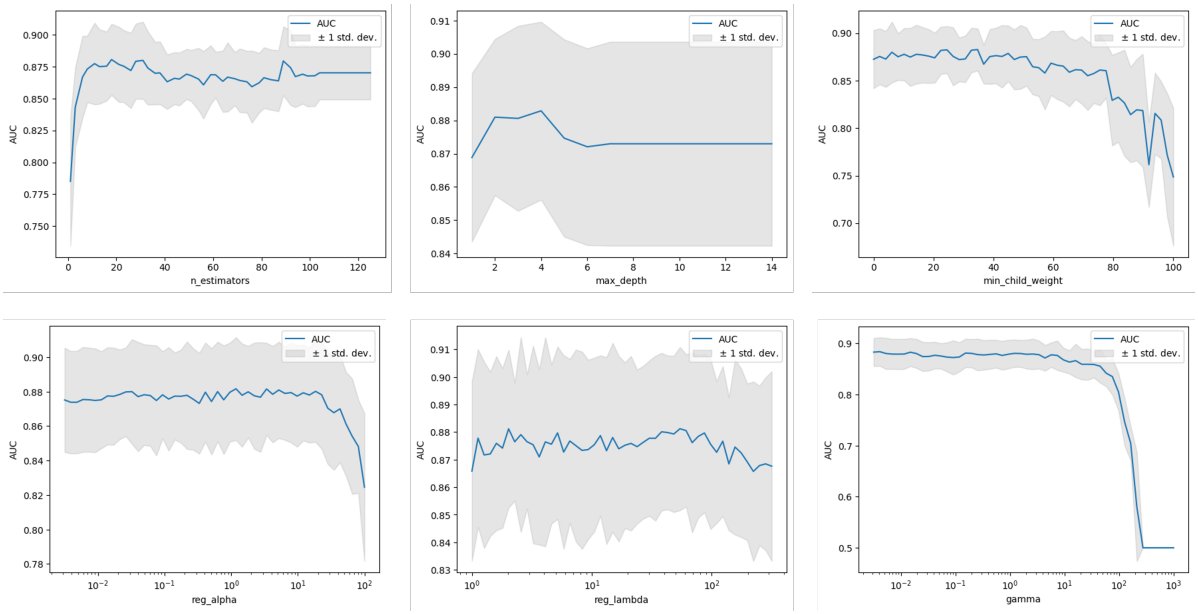


Figure 3.7: Effect of hyper parameter value on AUC with other hyper parameter values fixed on 'base parameters values' (see table 3.1) when training and testing on ET dataset.

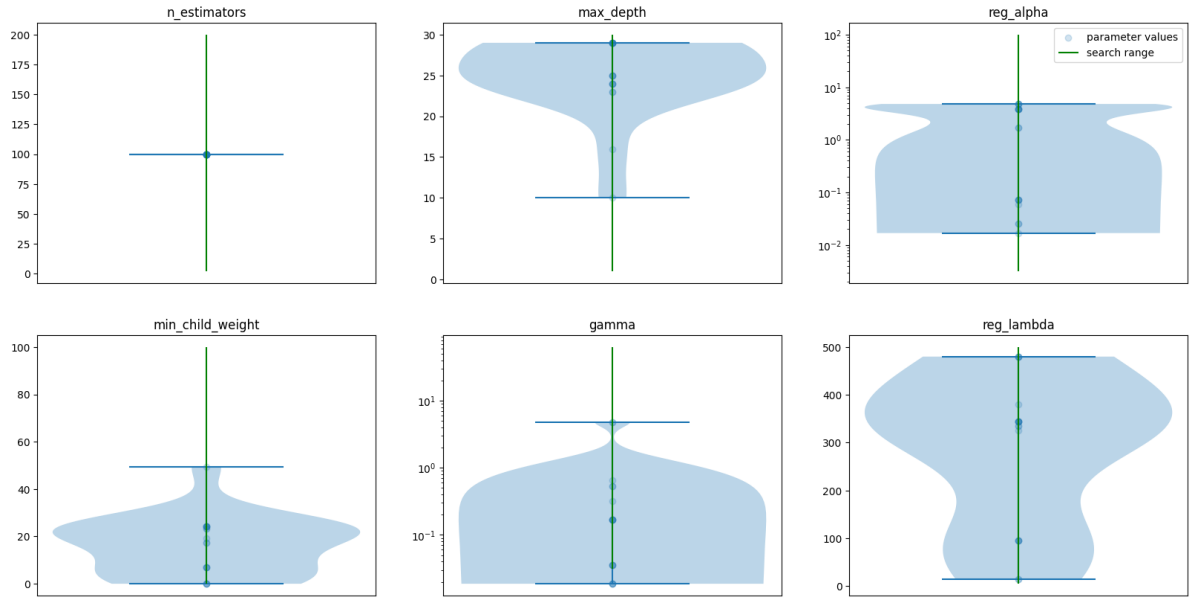


Figure 3.8: Hyper parameter distributions obtained during inner cross validation and used in the outer cross validation ET classifiers. Search range is the range of parameter values used during random grid hyper parameter search. Parameter values are the actually selected values during random grid search.

together with the number of models actually using the features. Hemoglobin and estimated Globular Filtration Rate (eGFR) were selected in all 20 models to be used. Features used in more than half of the models are C-Reactive Protein (CRP), erythrocyte sedimentation rate ('bezinking'), HDL cholesterol ratio, thrombocytes, iron ('ijzer') and albumin. SHapley Additive exPlanations (SHAP) plots for each of the outer cross validation folds are shown in appendix C.1.

Hyper parameter learning curves are shown in figure 3.7. They show the effect of a feature value on the performance of the XGBoost model. The mean AUC metric obtained by testing on the cross validation test sets is visualized, together with the standard deviation around the AUC's mean. Distribution of

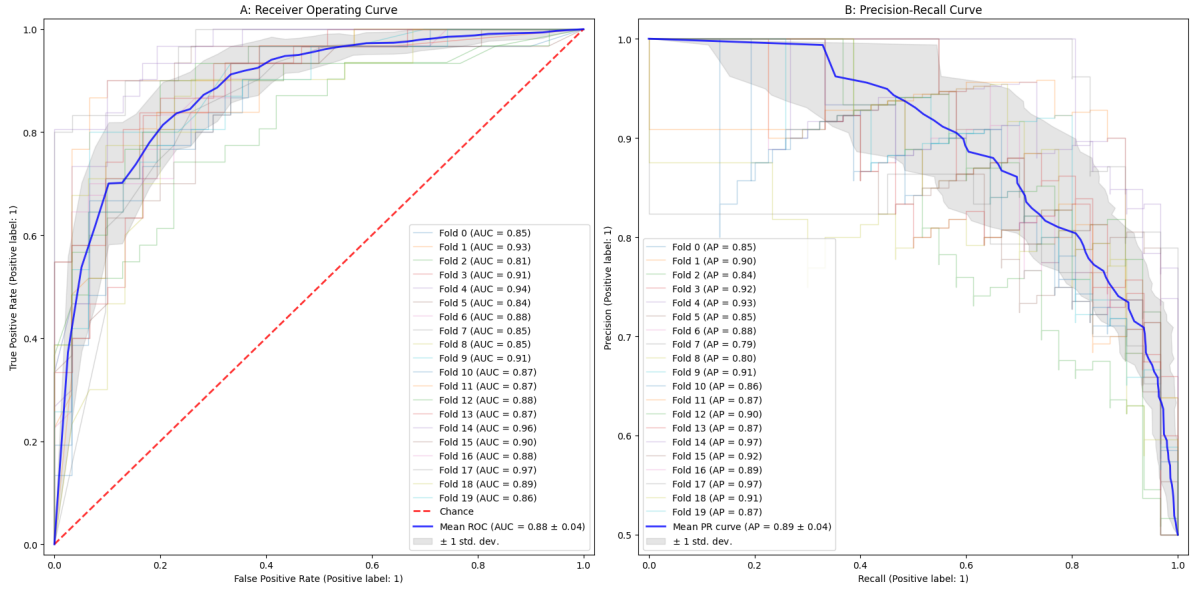


Figure 3.9: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for outer cross validated classifiers in the ET dataset when only the last known measurements per patient are used for training and testing. A: ROC's with corresponding Area Under the Curve (AUC). B: PR-curves with corresponding Average Precision (AP).

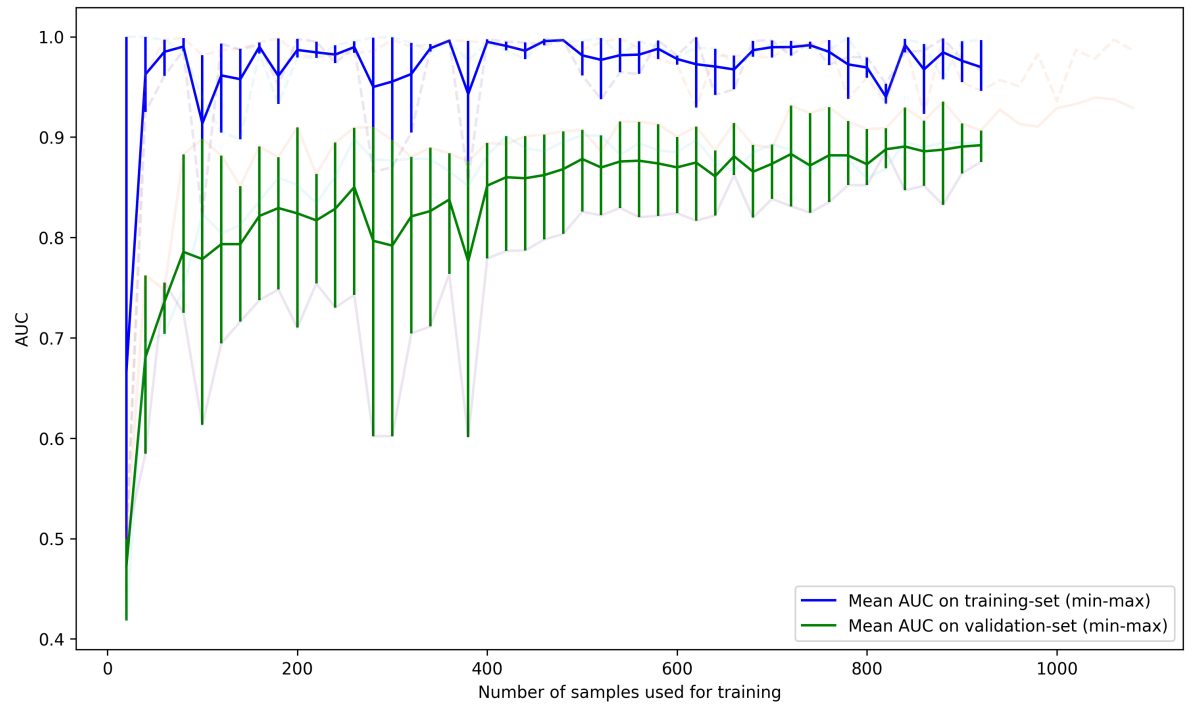


Figure 3.10: Learning curve of ET classification model for number of samples used for training. Blue line shows the mean Area Under the receiver-operating-characteristic Curve (AUC) values for the data the model was trained on, green line shows the mean AUC values on the not seen validation data. Semi-transparent lines show AUC values for each of the 3 random data splits.

hyper parameters values which are finally selected during hyper parameter tuning for each of the nested-cross validation folds are visualized in figure 3.8.

For training and testing with only the last known measurement per patient, mean number of patients in the ET and control groups is 91.5 for both groups in the train sets and 30.5 in the test sets. Mean AUC is 0.88 (sd: 0.04) and mean average precision is 0.89 (0.04). Corresponding ROC and PR curves are shown in figure 3.9. Mean precision, recall, F1 score, accuracy and specificity are respectively 0.80, 0.79, 0.80, 0.80, 0.81. Features used in more than 10 out of the 20 outer cross validation folds are Hemoglobin (20 folds), eGFR (19 folds), thrombocytes (17 folds), CRP (16 folds), previous thrombocyte count (12 folds) and erythrocyte sedimentation rate (11 folds).

The effect of the number of training samples is shown in figure 3.10, where both AUC values for the train set and test set are visualized.

3.2.2. PV classification

Samples from 58 PV patients and 2274 control patients were collected. A summary of the feature values in the dataset is provided in appendix A.2. Boxplots for feature values are visualized in appendix B.2. Average number of patients and samples per group in the outer cross validation are shown in table 3.3.

The mean AUC tested through outer cross validation is 0.86 (sd: 0.05), see figure 3.11. Mean Average Precision (AP) derived from Precision-Recall curves is 0.87 (sd: 0.05), see figure 3.11. Mean confusion matrix is shown in figure 3.12; derived values for precision, recall, F1, accuracy and specificity are respectively 0.84, 0.66, 0.74, 0.77 and 0.87.

The mean feature importance in the models created in outer cross validation are visualized in figure 3.13, together with the number of models actually using the features. estimated Globular Filtration Rate (eGFR) was selected in all 20 outer cross validation models. Features used in more than half of the models are hematocrit, hemoglobin, magnesium, total protein, ferritin, thrombocytes, phosphate, leukocytes, Red blood cell Distribution Width (RDW), Vitamin D, Urea, erythrocytes, base excess, anion gap, bilirubin, lactate and monocytes. SHapley Additive exPlanations (SHAP) plots for each of the outer cross validation folds are shown in appendix C.2.

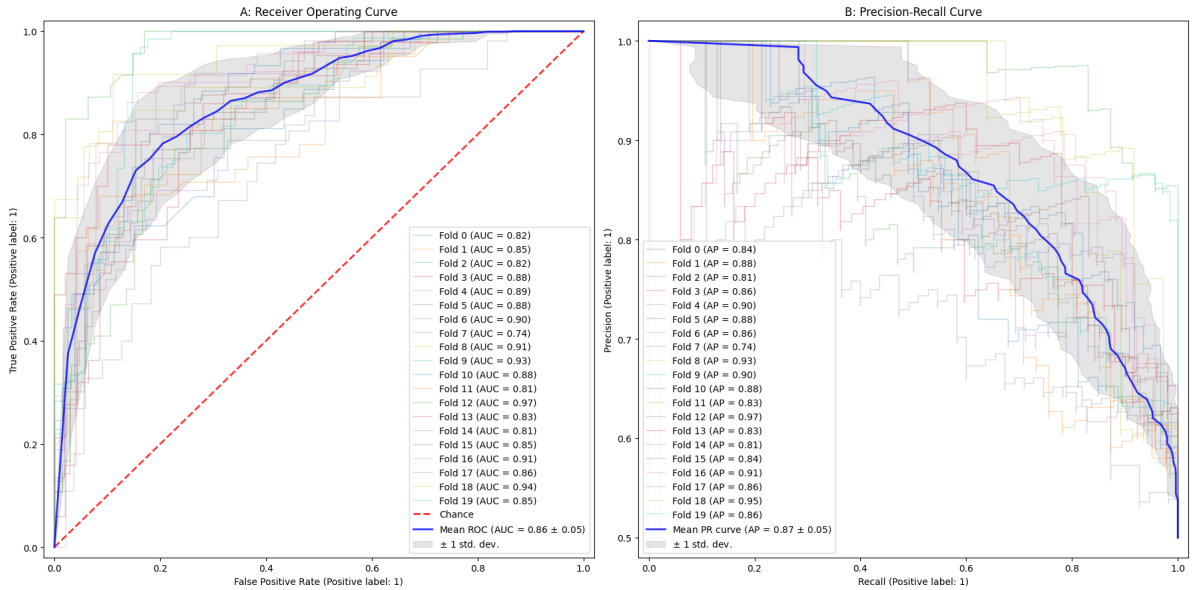


Figure 3.11: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for outer cross validated PV classifiers. A: ROC's with corresponding Area Under the Curve (AUC). B: PR-curves with corresponding Average Precision (AP).

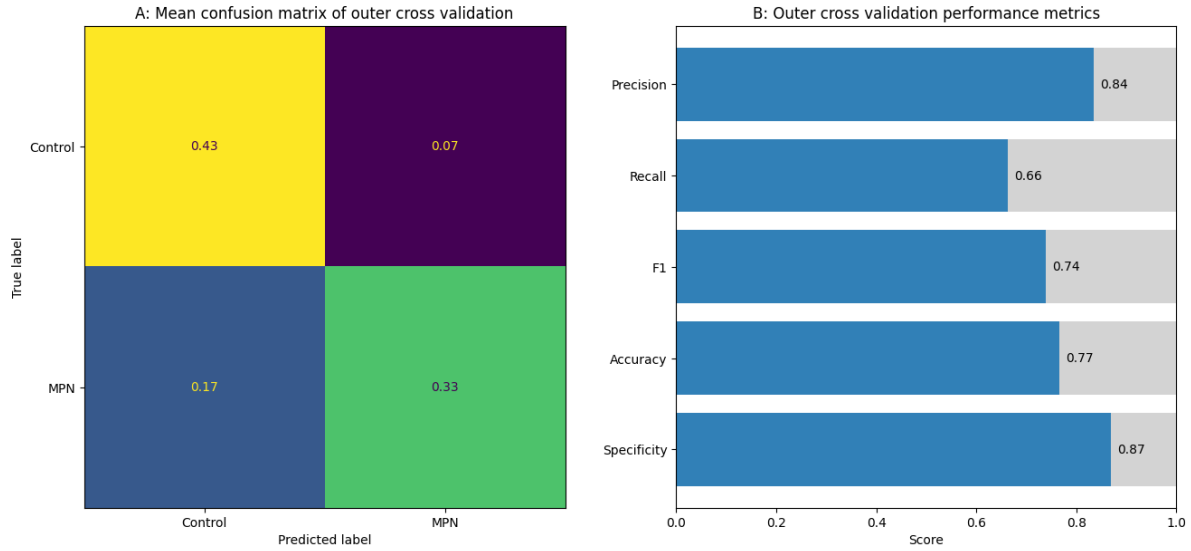


Figure 3.12: Mean confusion matrix of outer cross validation folds (A) and corresponding performance metrics (B) for PV classification.

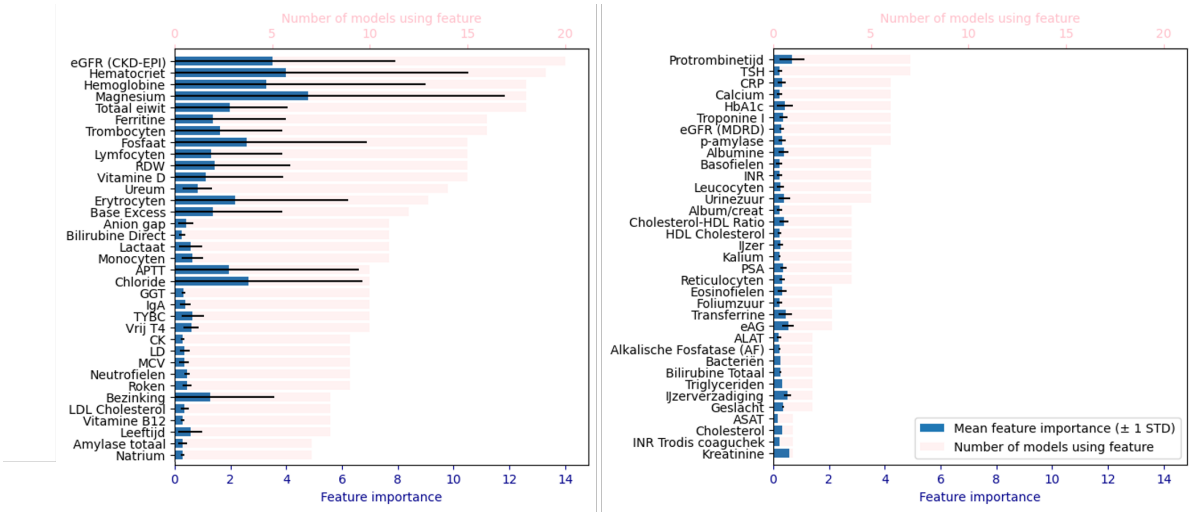


Figure 3.13: Importance of features used in outer cross validation classifiers and number of classifiers using the features for PV classification. Total number of classifiers is 20. Feature importance is calculated as the average gain achieved through splits based on that feature.

3.2.3. Retrospective diagnostic delay

For the MPN patients in the test sets, time to diagnosis and prediction value are visualized in figures 3.14 and 3.15. For the 122 ET patients included, 76 (62%) had at least one positive MPN prediction more than 1 year prior to their actual MPN diagnosis. 26 out of 58 (45%) PV patients had at least one positive retrospective model prediction more than one year prior to diagnosis.

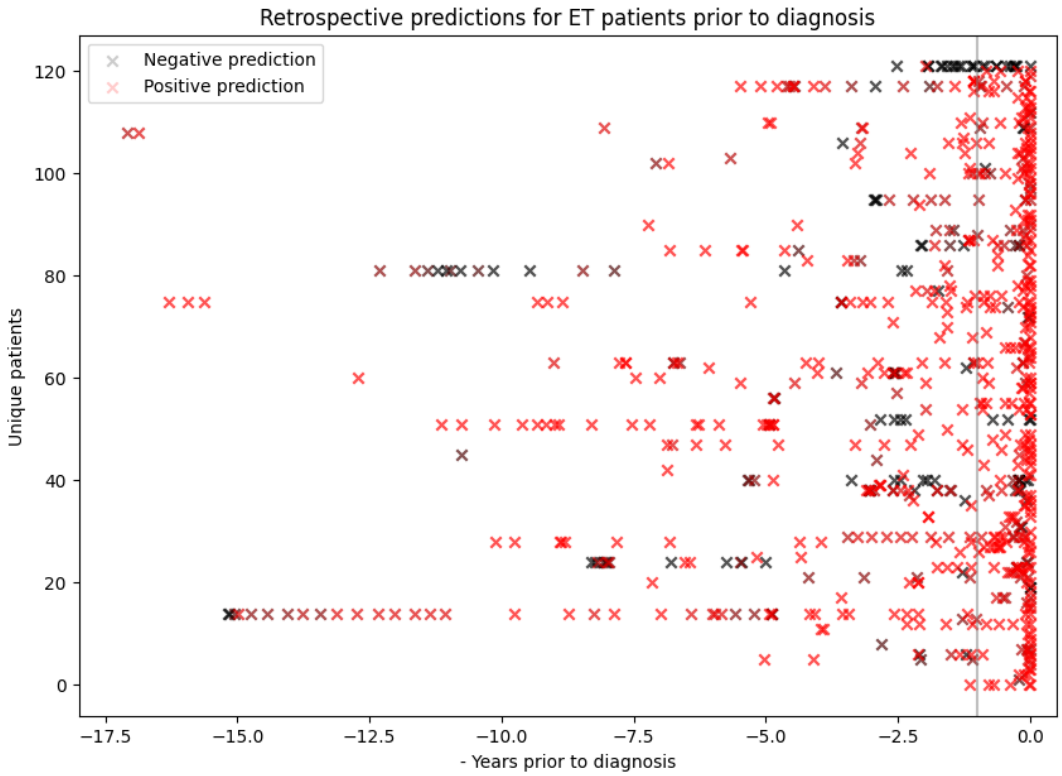


Figure 3.14: Diagnostic delay and model prediction for each unique ET sample in the test sets. Samples on the same horizontal level refer to one single patient. 76/122 patients have at least one positive ET prediction more than 1 year prior to their actual ET diagnosis.

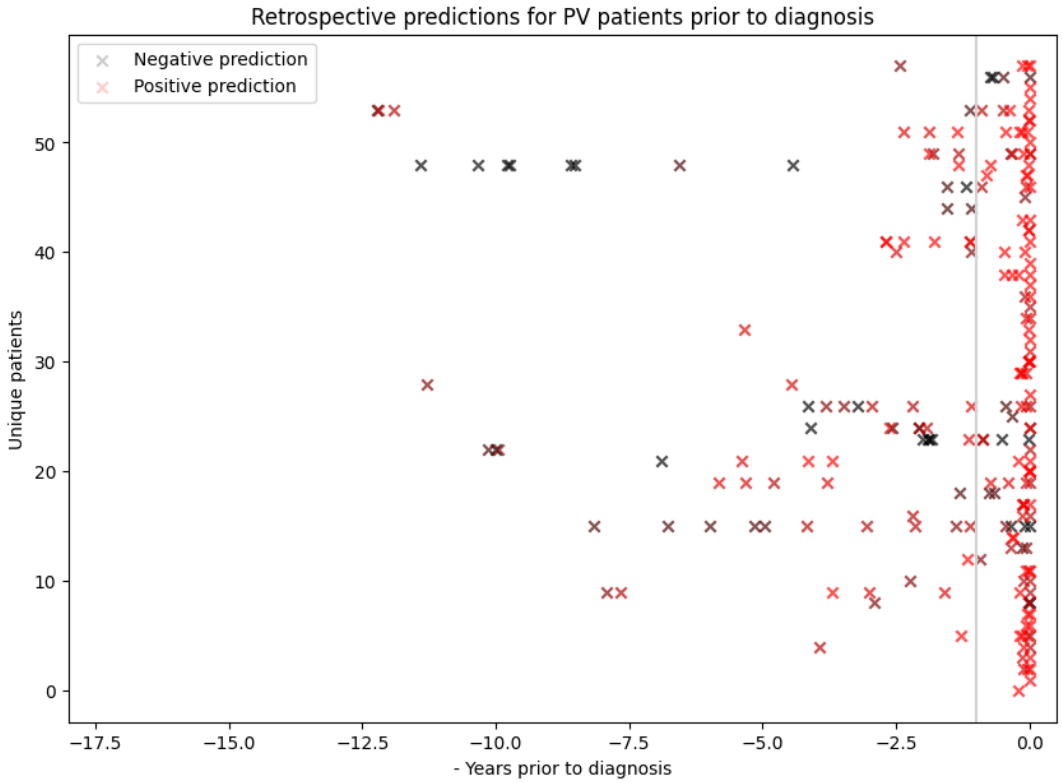


Figure 3.15: Diagnostic delay and model prediction for each unique PV sample in the test sets. Samples on the same horizontal level refer to one single patient. 26/58 patients have at least one positive PV prediction more than 1 year prior to their actual PV diagnosis.

4

Stage 2: microscopy filter

After selection of suspect cases based on general blood measurements (stage 1), a population with true MPN patients and false-positives remains. To further increase the specificity to find MPN cases, we hypothesize that further (automated) analysis of available blood samples using microscopy can boost diagnostic specificity. The costs of microscopy analysis is higher compared to CBC blood measurements, but significantly lower than the costs of mutation analysis or the intervention of human expert. Therefore, automated microscopy analysis of (already drawn) peripheral blood might be an ideal method as in-between analysis; reducing the number of false-positives with preservation of the true positive MPN patients.

This chapter shows the first steps in the development of a blood microscopy based method (stage 2) for early detection of MPN patients as follow-up to the algorithm described in the previous chapter. Multiple methods are explored, including an cell counting based XGBoost algorithm and a image based ResNet50 convolutional neural network (CNN).

4.0.1. Data collection

For training of algorithms in this stage, microscopy images were used form a historical cohort (2014-2015) of all patients clinically suspected of MPN. For samples genetically tested for MPN related mutations (JAK2, CALR, MPL) in the routine workload of the Result Laboratory Dordrecht, also peripheral blood smear microscopy images were made. Diagnosis based on molecular diagnostics and further clinical information was used for ground truth labeling of the samples (MPN / non-MPN). Microscopy of the peripheral blood samples was performed using a DI-60 digital microscope.

4.1. Cell counting based XGBoost

In clinical practice, peripheral blood microscopy is often used to count the different types of blood cells. Immature or under/over presented cell types can thus be detected. Already 20 years ago a algorithm was presented which was capable of detecting and classifying leukocytes in blood microscopy images.[28] Using cell counts from an automated image classifier, the microscopy images are simplified from roughly 12 million input values (2000x2000 pixels with 3 color channels) to 10-30 variables (the number of different cell types detected).

4.1.1. Methods

Machine Learning model

An eXtreme Gradient Boosting (XGBoost) algorithm was trained to predict if the cell count of a microscopy analysis is positive or negative for MPN.

The model was trained and evaluated by applying outer cross validation. For each split, the following pipeline was applied:

1. data preprocessing;
2. hyper parameter tuning;

3. model training;
4. model performance evaluation.

Compared to the XGBoost pipeline described in the previous chapter, feature selection is missing in this pipeline. This is due to the already limited number of leukocyte types counted in microscopy analysis.

Data preparation

Cell counts assigned by the microscopy software to the scanned blood smears are included in this study. A random selection of control patients was performed in order to include an equal number of MPN and control samples. Cell counts were normalized, by scaling the cell count so that the normalized cell counts per patient sum up to 1.

Nested cross validation

Patients in the dataset were split by a stratified 3 times repeated 3-fold splitting in train and test groups (outer cross validation). Inner cross validation on the train groups was performed for hyper parameter tuning. Using the selected hyper parameter values, training is done on the outer cross validation train groups. Evaluation of the trained models is performed on the outer cross validation test groups.

Hyper parameters

Tuned hyper parameters for the XGBoost model are:

- the number of classification trees (estimators) per classifier,
- the maximal tree depth,
- minimum child weight required for the splitting of leafs,
- alpha value (L1 regularization),
- lambda value (L2 regularization) and
- gamma value (minimum split loss reduction).

The effect of these hyper parameters on AUC-score is analyzed through nested cross validation with default hyper parameter values, except for the tested hyper parameter. For actual training of the models, optimal hyper parameter values were searched for using a 5-fold cross validated randomized search with folds of 50 iterations. For the search domains, see table 4.1.

Parameter	**Search range**
Number of estimators	2 – 30
Max depth	1 – 15
Minimum child weight	0 – 15
Alpha (L1 regularization)	0.01 – 10
Lambda (L2 regularization)	5 - 150
Gamma (min. split loss reduction)	0.01 – 10

Table 4.1: Search ranges in hyper parameter tuning of XGBoost model for microscopy white blood cell counts.

Experiments

Nested cross validation is preformed using all available data in the dataset. Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were created and performance metrics obtained based on the model performance on the test sets. Distributions of selected hyper parameters and feature importance in the trained models at outer cross validation folds are obtained and visualized.

Also the effect of hyper parameter values on AUC is evaluated through nested cross validation. For this purpose, only single feature selection is performed and the model default parameters are used, except for the variable feature for which the impact on AUC performance is analyzed.

Model performance evaluation

Outer cross validation is utilized to evaluate performance metrics. Primary evaluation metric is the Area Under the receiver-operator-Curve (AUC). Secondary metrics are Average Precision (AP), precision (positive predictive value), recall (sensitivity), accuracy, specificity and F1 score. Feature importance and used hyper parameter values are analyzed for the trained classifiers.

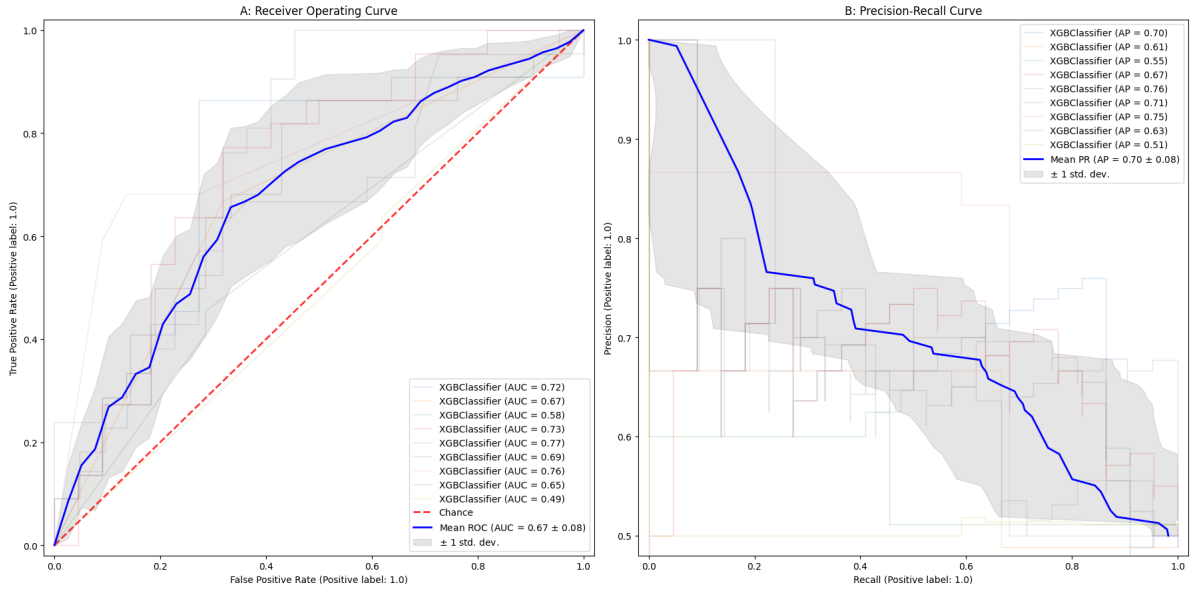


Figure 4.2: Receiver Operating Curves (ROC) and Precision-Recall (PR) curves for outer cross validated microscopy white blood cell counts based MPN classifiers. A: ROC's with corresponding Area Under the Curve (AUC). B: PR-curves with corresponding Average Precision (AP).

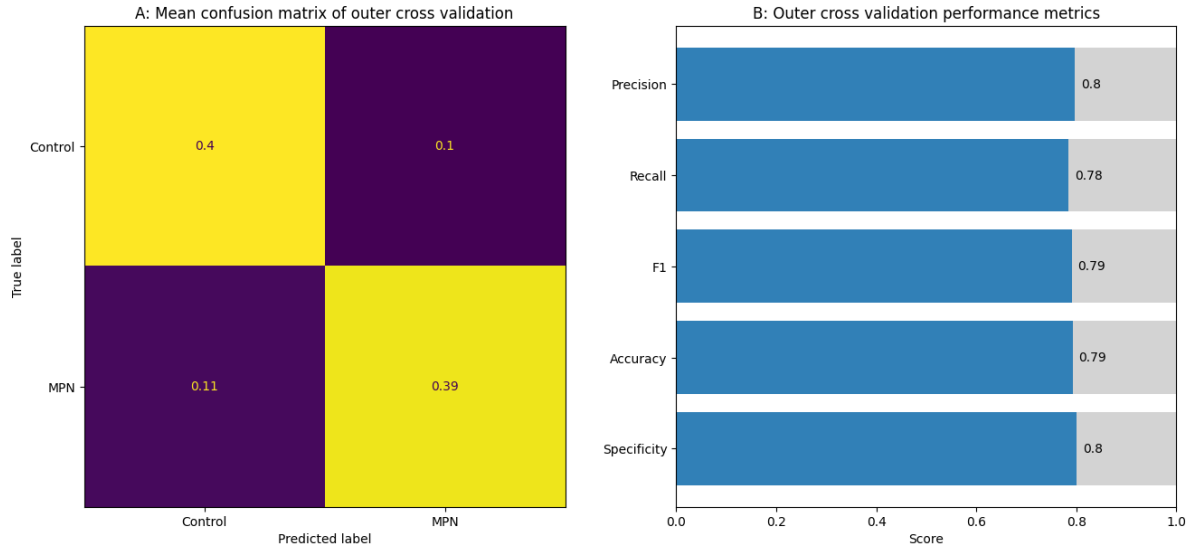


Figure 4.3: Mean confusion matrix of outer cross validation folds (A) and corresponding performance metrics (B) of microscopy white blood cell counts based MPN classifiers.

4.1.2. Results

The Receiver Operating Curve (ROC) and Precision-Recall (PR) curves are shown in figure 4.2. Mean Area Under the ROC (AUC) for the 9 outer cross validation folds is 0.67 (standard deviation: 0.08). Mean Average Precision (AP) is 0.70 (standard deviation: 0.08). Precision, recall, F1-score, accuracy and specificity based on the mean confusion matrix of the 9 outer cross validation folds are respectively 0.80, 0.78, 0.79, 0.79 and 0.80, see figure 4.3.

The effect of hyper parameter values on AUC in the test set is shown in figure 4.5. Hyper parameter values selected during hyper parameter selection in the separate cross validation folds are visualized in figure 4.4.

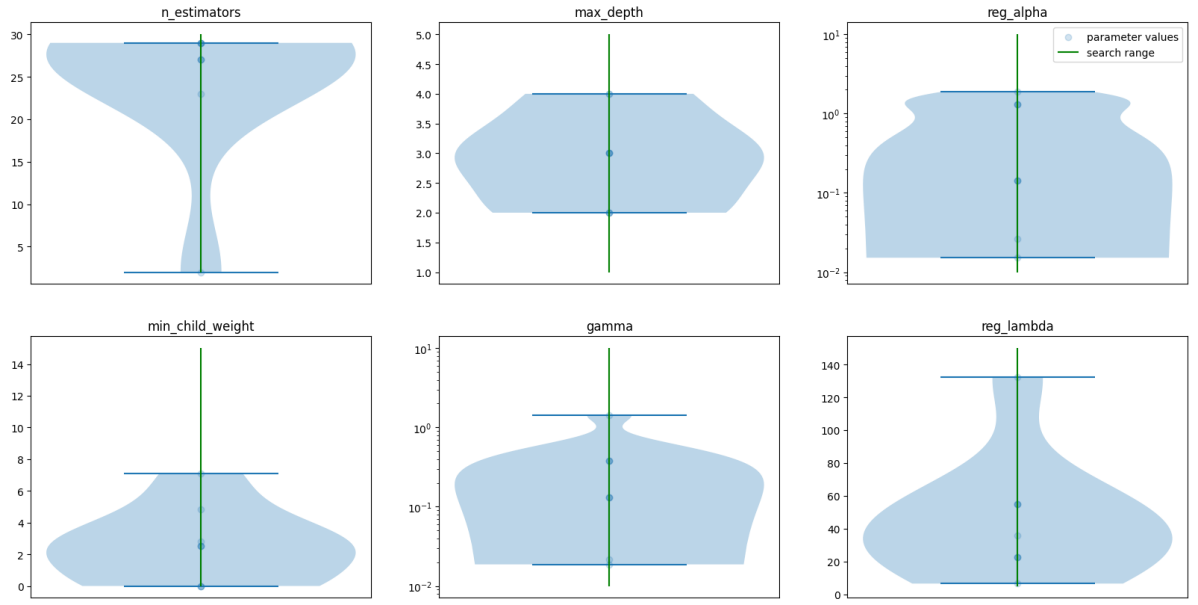


Figure 4.4: Hyper parameter distributions obtained during inner cross validation and used in the outer cross validation MPN classifiers. Search range is the range of parameter values used during random grid hyper parameter search. Parameter values are the actually selected values during random grid search.

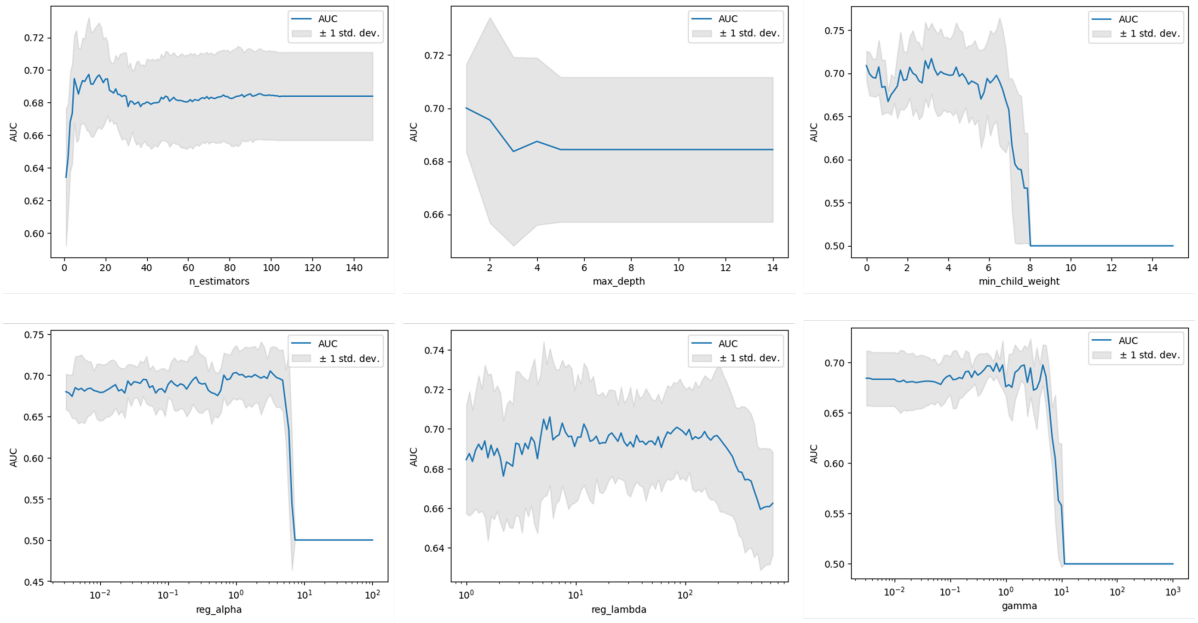


Figure 4.5: Effect of hyper parameter value on AUC with other hyper parameter values fixed to default values.

Mean feature importance of the features in the 9 cross validation folds are shown in figure 4.6. In this figure also the number of models using a cell type as feature are visualized. Although no feature selection is performed, it is possible that the classification trees in the XGBoost model do not contain splits based on a certain feature. Figure shows thus that GT (giant thrombocyte celltype) is used as feature in all cross validation models, where ERB and PMY (respectively erythroblast and promyelocyte celltypes) are used in none of the models.

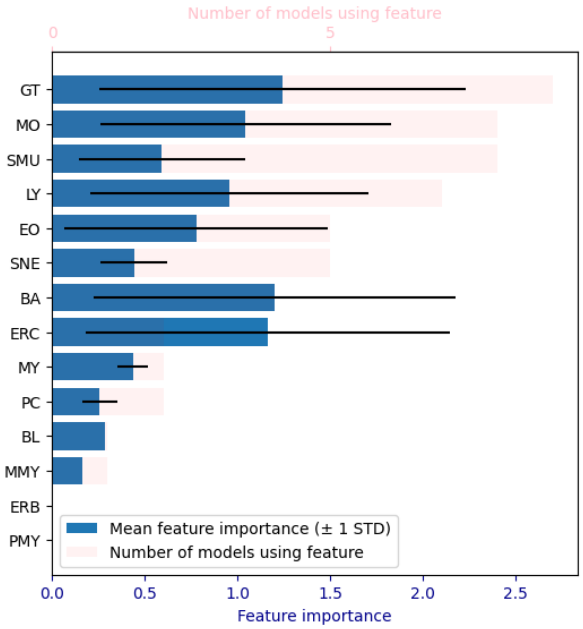


Figure 4.6: Importance of features used in outer cross validation classifiers and number of MPN classifiers using the features. Total number of classifiers is 9. Feature importance is calculated as the average gain achieved through splits based on that feature.

4.2. Image based ResNet50

To use whole microscopy images for MPN / non-MPN classification, a ResNet50 architecture is used in this section. This Convolutional Neural Network (CNN) contains shortcuts around layers.[29] The 50 layers in this architecture provide high memory and filter capacity for complex tasks, where the shortcuts make it possible to skip superfluous layers for simpler image recognition tasks.

4.2.1. Methods

Data preparation

Overview microscopy images were used with square dimensions of approximately 2800x2800 pixels ($\sim$ 600*600 µm). All images were resized to 1024x1024 pixels; after which 15 random patches with size 256x256 were extracted. Each patch was labeled according to the ground truth classification of either control or MPN. Patches were randomly assigned to the train or test group (80% train, 20% test). Random undersampling in both the train and test groups is performed to include an equal number of control and MPN patches.

ResNet50 setup

The TensorFlow Keras build-in application of ResNet50 was used to build a model architecture. Input shape of the patches is 256x256x3; the final layer is a dense layer with binary output and sigmoid activation. Training in 200 epochs is performed using the Adam optimizer with accuracy as evaluation metric. Training was performed on the train set. After training of the model, model performance is evaluated using the test set with accuracy as primary outcome.

4.2.2. Results

A total of 3390 patches are included. For a visualization of randomly selected patches, see figure4.7. After randomisation to test and train sets and undersampling, the train-set consists of 480 control and 480 MPN patches. The test-set contains 120 control and 120 MPN patches.

The accuracy on the train and validation set during training of the model is shown in figure 4.8. The confusion matrix for model predictions on the test set are shown in figure 4.9, which shows that the model tends to assign all new patches to the MPN negative control group. Corresponding accuracy is 0.5.

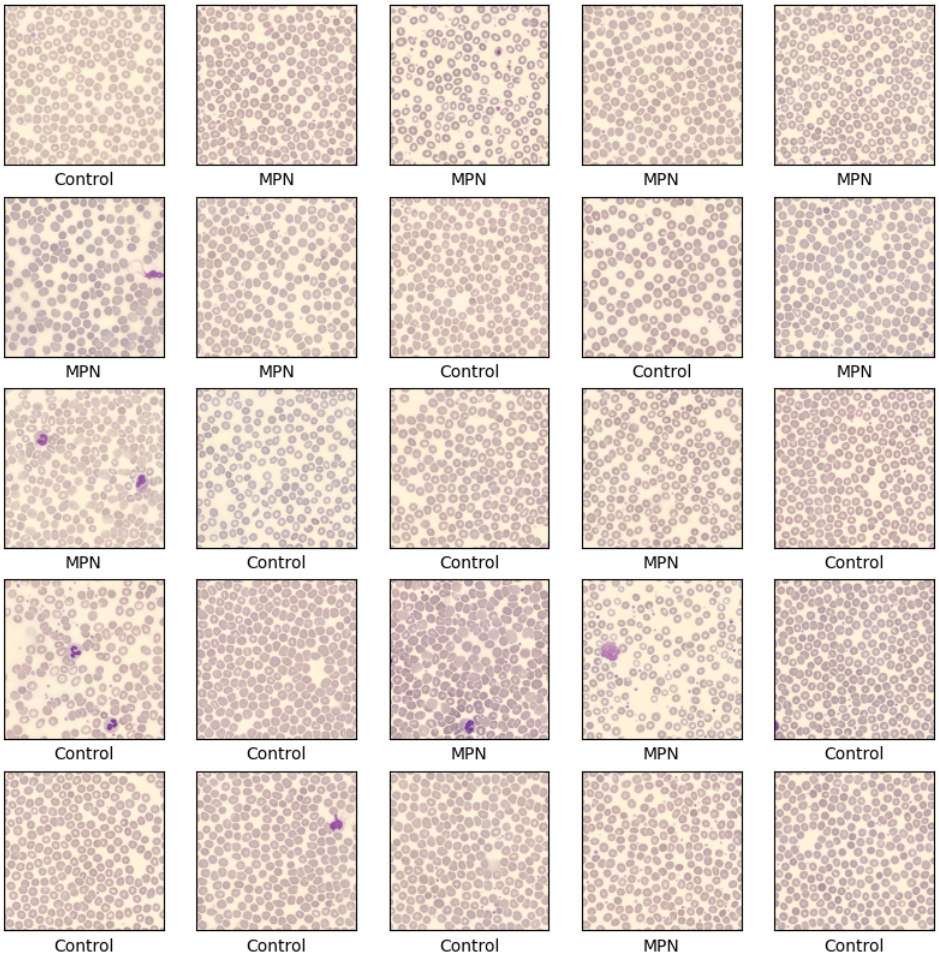


Figure 4.7: Visualisation of randomly selected patches used as input for the ResNet50 model with their ground truth labels.

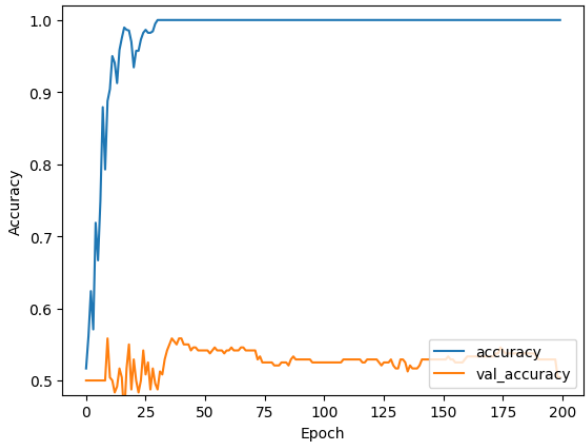


Figure 4.8: Accuracy of ResNet50 model evaluated during training.

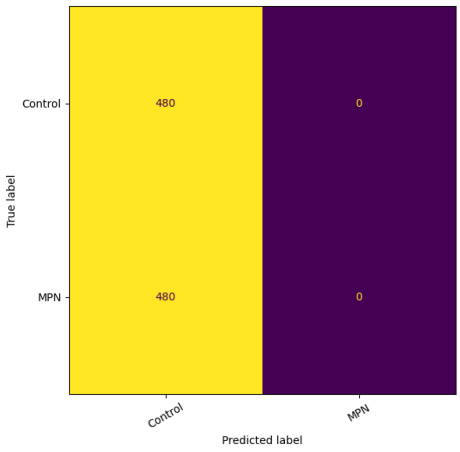


Figure 4.9: Confusion matrix for predictions on test set of ResNet50 model.

5

Discussion

Long diagnostic delays to MPN diagnosis might cause potentially preventable complications and deaths. In this thesis, methods are explored to assist clinicians in recognising MPN patients earlier. A system of two filters is proposed, where regular blood measurements are used in a first filtering method. Suspected cases identified by the first filter would subsequently be filtered by a second, microscopy based filter.

5.1. Stage 1: Regular blood measurements

5.1.1. Model performance

The given metrics for the model show that the model is working as a filter which is capable of detecting most of the MPN patients in a dataset (8 out of 10 patientsamples with a future MPN diagnosis were retrospectively detected in the test sets). See section 5.3 for an estimation regarding the meaning of the model metrics in clinical practice. For now however, it can be said that the model is a fairly good first filter to select a suspected population from the routine laboratory blood measurement workload. When looking at ET and PV patients, the models were capable to detect respectively 62% and 45% of the patients more than one year prior to diagnosis.

Depending on the demands of clinicians, the model might still be optimized. The cost and burden of additional measurements and clinical follow-up for false positive individuals could weight so much that a low false-positive rate is preferred over a perfect detection of all MPN patients. In other cases it might be that false-negative findings are seen as unacceptable, even if this means that a lot of false positive findings should be found. For both situations, the model can be fit to the clinical need in multiple ways. For the model described in this thesis, the default objective of the XGBoost python package for binary classification is used (binary logistic objective with a negative log-likelihood evaluation metric). Re-training could be performed using another combination of objective and the evaluation metric it is based on, in order to fit the clinical question. A simpler method is to tune the classification threshold; the XGBoost model returns a value in a range of 0-1 which can be interpreted as an MPN probability score. By default, a score above 0.5 is classified as MPN and a score below 0.5 as non-MPN. However, by changing this score, only the very suspected cases could be selected (high threshold) or a higher sensitivity can be achieved (lower threshold).

It was shown that mean performance of the ET model increased slightly in case of training only using the last known measurement per patient. The mean AUC increased from 0.87 to 0.88, though it should be noted that these mean values fall within each others standard deviation (0.02 and 0.04 respectively). Based on the increasing mean AUC in case of less samples used for training, it might be said that it is easier to predict if a patient has ET at the very moment a clinician also diagnosis the patient as having ET. But this does not mean that a model trained using only last measurements before diagnosis fits the objective of this study. Here, a method is required for *early* detection in order to reduce diagnostic delay and potentially preventable complications. Thus a model with a slightly lower AUC performance is seen as better tool than a model which is explicitly trained to detect MPN at the moment of diagnosis.

5.1.2. Important features

The most important features found during training of the cross validation ET models are Hemoglobin, estimated Globular Filtration Rate (eGFR), C-Reactive Protein (CRP), erythrocyte sedimentation rate, HDL cholesterol ratio, thrombocytes and iron. Thrombocytes are directly linked to ET, because they belong to the affected cell types. Hemoglobin and iron are key components of red blood cells, which are elevated in PV, but also a portion of ET patients. The models pick this up by giving a higher MPN probability for patients with elevated hemoglobin and/or iron levels.

CRP and erythrocyte sedimentation rate are negatively associated with ET. In case of an infection, thrombocytes are also elevated, but in that case CRP and erythrocyte sedimentation rate are also increased. Thus, elevated thrombocytes without increased CRP and erythrocyte sedimentation rate are predictive for ET. As could be expected, this effect is not seen for the PV dataset. Inclusion for PV is based on red blood cell counts, which are not directly associated with an infection, eliminating the need of exclusion of infection samples.

Regarding eGFR, renal failure in MPN patients is reported.[30] This is also seen in some ET and PV models where a high renal function is considered as a negative predictor for MPN. However, in most cases, the absence of eGFR measurement seems to be a negative predictor for MPN. The absence of data thus seems to contain information. These findings at least suggest that there is a group of patients who do have increased thrombocyte or red blood cell levels, but are not tested for their renal function. These patients might for example be patients with a known disease for which it is not needed to monitor their renal function.

HDL cholesterol ratio might have an metabolic connection with MPN. Cholesterol uptake of leukemic cells is reported to be abnormal.[31] Also suggestions of JAK2 activation by fat molecules are done.[32] These findings however do not explicitly explain the impact of HDL cholesterol in ET. Our models suggest a relation between decreased HDL cholesterol ratio and ET probability. A low HDL cholesterol ratio is considered to be healthy and the control patients more often had a high HDL cholesterol ratio. The most probable explanation of the predictive value of HDL cholesterol ratio and triglyceride would thus be that these measurements are used to differentiate between MPN and other (lifestyle related) diseases which present comparable blood measurements and symptoms. It should however be noted that HDL cholesterol ratio and triglycerides only have little predictive value compared to features such as CRP, hemoglobin, thrombocytes and eGFR.

5.1.3. Similar work

From our review of the literature (see appendix F we did not find reports of blood measurement based MPN vs. non-MPN classification. Kimura et al. have shown that it is possible to distinguish between MPN subtypes (PV, ET and MF) using laboratory measurements and imaging together with an XGBoost algorithm.[33] With sensitivity and specificity scores above 0.9 and AUC values of 0.97-0.99 their method was highly successful. In this thesis, a comparable method is applied to distinguish between MPN and non-MPN patients. The target group in this thesis' study is far more diverse, due to the fact that the control group comes from the general hospital population having a large variety of complaints. The AUC, sensitivity (recall) and specificity scores found for our models are 0.86-0.87, 0.66-0.74 and 0.84-0.87. These scores are lower compared to the scores in the work of Kimura et al., which can be explained by the heterogeneous control population and the usage of microscopy imaging data by Kimura et al., which was not used in this filter.

5.2. Stage 2: Microscopy based selection

To reduce the number of false positives after initial filtering on blood measurements, usage of microscopy data was proposed. The results shown in this thesis indicate that this data has the potential of being predictive for MPN.

5.2.1. Cell counting based

Using the labels provided by the machine learning algorithm of the microscope viewer software, an AUC of 0.67 (standard deviation 0.08) was found. With a sensitivity (recall) of 0.78 and a specificity of 0.80, it is proven that an XGBoost model is capable of predicting if a sample belongs to an MPN patient or not. The algorithm labeling the white blood cells does only take the cell images as input, which means that the trained XGBoost classifier is fully microscopy imaging based. This also means that the microscopy images contain information which indicates if a patient might have MPN.

5.2.2. Image based

The reported results based on the ResNet50 neural network, do however not show the capability of this model to predict MPN (accuracy of 0.50, model classifies all test samples as non-MPN). Figure 4.8 shows the ResNet50 model was of over-training (train-accuracy of 1.0), which means that the model at least has the capability of memorizing image features in order to perform perfect predictions on train data. This indicates that the model architecture is appropriate for the input data, but the lack of any discernibility shows the input data is not matching the task for this given algorithm.

5.2.3. Future prospective

The patches used for training and testing of the ResNet50 model are overview images, where most of the area is filled with background and red blood cells; little area is covered by white blood cells and platelets, see figure 4.7. It might be that these patches do not contain enough information to be usable for control/MPN classification, or not enough images are included for the model to detect predictive features (here training was performed on 480 patches per group, which is a relatively low number for neural network training tasks). Training on the overview images does not provide much information regarding white blood cells and platelets, whereas the labeled subtypes of white blood cells have shown to be predictive and platelets are also reported to possibly have changed morphology in MPN patients.[34] Thus, taking features regarding white blood cells and platelets might be used for MPN detection in future research. The model however should also be improved, for simply feeding white blood cell images instead of red blood cell patches to a ResNet50 showed the same overtraining and single output prediction class on test data as observed for the red blood cell patches, see appendix D for results of lymphocyte images as input for the ResNet50 model.

Where the application of ResNet50 in this study takes images as input and a classification as output, the study by Kimura et al. took images as input and image features as output of their neural network.[33] The features obtained by their network they then fed to an XGBoost algorithm, together with blood measurement values. This method utilizes the strength of a convolutional neural network for image feature recognition, but prevents over-training by using a simpler XGBoost model for actual classification. This also gave them the possibility to combine blood measurements and microscopy imaging data in a single prediction algorithm. For further development of a microscopy imaging based algorithm for MPN detection, a comparable method might be applied. Morphological properties of cells (such as shape, size and granulation), could be extracted by morphological operations or a neural network. A mean and standard deviation of the radiomic features for multiple cells of different cell types makes it possible to combine an unknown number of cells in a structured datatype required for an classification model. Cell counts, laboratory measurements and demographic information could be added to provide additional information in order to increase model performance. An XGBoost model has shown to be appropriate for such an task, both in this study as well as in the study by Kimura et al.[33] A separate neural network taking the radiomic features and additional information as input might also be promising for this task, due to the capability of a neural network to combine the meaning of multiple features (eg the given standard deviation of a feature is dependent on the 'real' spread of the feature value, but also on the number of samples the standard deviation is calculated on).

It should be noted that ground truth labeling of MPN patients is not a straight forward process. If a patient has one of the common MPN mutations, the diagnosis is set. However, 10-15% of the MPN patients have MPN symptoms, but do not present one of the common MPN mutations.[35] For the datasets used in this thesis the diagnosis given by hematologists was used as ground truth, even if no MPN mutation could be found. This is done because also the patients without MPN mutation need appropriate treatment to reduce the risk of cardiovascular complications. Nevertheless his also means that the MPN dataset was more heterogeneous compared to a dataset with ground truth labeling based on mutation analysis, due to the possibility that clinicians have diagnosed a non-MPN patient as having an MPN.

5.3. Clinical implementation

5.3.1. Extrapolation to real world situation

For development of the MPN detection algorithms, balanced groups with an equal number of control and MPN patients were used. In clinical practice however, only a small fraction of the total population has an MPN. A multinational registry study showed that the incidence of MPN in a hospital population is 12-15 per 100 000 hospital patients.[36] This does not directly mean that the MPN incidence rate in the laboratory is 12-15 / 100 000, but it does give an (under)estimation of the true laboratory incidence

rate. If the given incidence rate would be used and only the ET filter on regular blood measurements would be taken into consideration, approximately 16 000 per 100 000 patients would be labeled as positive. Only 9 per 16 000 patients would be true positives, the rest would be false positive. For each positive patient, around 1800 patients would be false positive. This scenario is not completely realistic due to the increased MPN probability when a patient both gets blood measurements ánd has a consecutive increase thrombocyte count. In case this would increase the probability of an sample to be MPN positive to 500 / 100 000 (5%), then the number of true positives would increase to such an extend that for each true positive patient, 43 false positives are found.

5.3.2. Next steps toward clinical implementation

Above estimations show that a filter based on blood measurements is not directly useful in clinical practice due to the large portion of false positive predictions. This supports the need for an additional filter with high sensitivity to maintain true positive cases and a moderate to good specificity to eliminate most of the false positive cases. Microscopy based filtering has shown potential for MPN selection and could thus be further explored in order to create a filtering model. Besides the development of an appropriate second filter, the first filter based on blood measurements might be optimized for exact clinical need as described in section 5.1.1.

Before routine clinical implementation of the proposed automated laboratory measurement and imaging filters, both proper validation and workflow integration are required. Validation might be performed on external data of another hospital and/or in a monitored study situation where new laboratory measurements are fed to the filter(s) and closely monitored by clinicians. Integration in workflow requires a system where real time data of blood measurements and microscopy imaging can be loaded and processed. Output of the system is ideally integrated with currently used systems. A dashboard for research purposes has been developed during this thesis project in order to give an impression of a visual representation of model output, see appendix E. During research and validation this is an appropriate tool for real time model visualisation, for clinical implementation the dashboard should be integrated in the systems already used.

6

Conclusion

In this thesis, a two stage machine learning based filter method is proposed for early detection of MPN in the laboratory workflow in order to prevent a long diagnostic delay and associated complications.

An appropriate first filter on common laboratory measurements has been developed, with a mean AUC score of 0.87 in outer cross validation testing. This filter is used to select suspected ET / PV cases from the routine laboratory workflow. A second filter based on blood smear microscopy imaging is suggested. Microscopy based white blood cell counts show to have a predictive value for MPN, but the applied neural network in this thesis project was not capable of differentiating between MPN and non-MPN microscopy images. Further research should be done, developing an microscopy based algorithm for MPN prediction, in order to have a clinically applicable MPN prediction tool.

This work shows the strength and potential of laboratory data combined with machine learning methods for early detection of MPN patients and thus the fight for a shorter diagnostic delay, less preventable complications and ultimately reduction of MPN related mortality.

References

[1] B. Meier and J. H. Burton. “Myeloproliferative disorders”. In: *Emerg Med Clin North Am* 32.3 (2014), pp. 597–612.

[2] P. J. Campbell and A. R. Green. “The myeloproliferative disorders”. In: *N Engl J Med* 355.23 (2006), pp. 2452–2466.

[3] J. Thiele et al. “The international consensus classification of myeloid neoplasms and acute Leukemias: myeloproliferative neoplasms”. In: *Am J Hematol* 98.1 (2023), pp. 166–179.

[4] E. Jabbour and H. Kantarjian. “Chronic myeloid leukemia: 2022 update on diagnosis, therapy, and monitoring”. In: *Am J Hematol* 97.9 (2022), pp. 1236–1256.

[5] P. C. NOWELL and D. A. HUNGERFORD. “Chromosome studies on normal and leukemic human leukocytes”. In: *J Natl Cancer Inst* 25 (1960), pp. 85–109.

[6] B. J. Druker et al. “Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia”. In: *N Engl J Med* 344.14 (2001), pp. 1031–1037.

[7] E. Jabbour et al. “Targeted therapy in chronic myeloid leukemia”. In: *Expert Rev Anticancer Ther* 8.1 (2008), pp. 99–110.

[8] S. G. O’Brien et al. “Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia”. In: *N Engl J Med* 348.11 (2003), pp. 994–1004.

[9] J. L. Spivak. “Myeloproliferative Neoplasms”. In: *N Engl J Med* 376.22 (2017), pp. 2168–2181.

[10] A. Tefferi and T. Barbui. “Polycythemia vera and essential thrombocythemia: 2021 update on diagnosis, risk-stratification and management”. In: *Am J Hematol* 95.12 (2020), pp. 1599–1613.

[11] J. Xie et al. “Two activating mutations of MPL in triple-negative myeloproliferative neoplasms”. In: *Cancer Med* 8.11 (2019), pp. 5254–5263.

[12] M. L. Randi et al. “Cerebral vascular accidents in young patients with essential thrombocythemia: relation with other known cardiovascular risk factors”. In: *Angiology* 49.6 (1998), pp. 477–481.

[13] E. Hachulla et al. “[What vascular events suggest a myeloproliferative disorder?]” In: *J Mal Vasc* 25.5 (2000), pp. 382–387.

[14] F. Cervantes, F. Passamonti, and G. Barosi. “Life expectancy and prognostic factors in the classic BCR/ABL-negative myeloproliferative disorders”. In: *Leukemia* 22.5 (2008), pp. 905–914.

[15] A. Tefferi and T. Barbui. “Polycythemia vera and essential thrombocythemia: 2017 update on diagnosis, risk-stratification, and management”. In: *Am J Hematol* 92.1 (2017), pp. 94–108.

[16] R. M. Shallis and N. A. Podoltsev. “Emerging agents and regimens for polycythemia vera and essential thrombocythemia”. In: *Biomark Res* 9.1 (2021), p. 40.

[17] A. Tefferi. “Primary myelofibrosis: 2023 update on diagnosis, risk-stratification, and management”. In: *Am J Hematol* 98.5 (2023), pp. 801–821.

[18] C. Forsyth et al. “Variable incidence of myeloproliferative neoplasms in Australia”. In: *Intern Med J* 51.11 (2021), pp. 1979–1980.

[19] S. A. Srour et al. “Incidence and patient survival of myeloproliferative neoplasms and myelodysplastic/myeloproliferative neoplasms in the United States, 2001-12”. In: *Br J Haematol* 174.3 (2016), pp. 382–396.

[20] M. Sant et al. “Incidence of hematologic malignancies in Europe by morphologic subtype: results of the HAEMACARE project”. In: *Blood* 116.19 (2010), pp. 3724–3734.

[21] A. Smith et al. “Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network”. In: *Br J Cancer* 105.11 (2011), pp. 1684–1692.

[22] R. M. Shallis et al. “Epidemiology of the classical myeloproliferative neoplasms: The four corners of an expansive and complex map”. In: *Blood Rev* 42 (2020), p. 100706.

[23] C. Forsyth, K. Melville, and C. Tiley. “The delayed diagnosis of myeloproliferative neoplasms is common and results in a high incidence of potentially preventable thrombotic complications”. In: *Pathology* 50.7 (2018), pp. 775–776.

[24] W. Alduaij et al. “Clinical Utility of Next-generation Sequencing in the Management of Myeloproliferative Neoplasms: A Single-Center Experience”. In: *Hemasphere* 2.3 (2018), e44.

[25] B. George-Gay and K. Parker. “Understanding the complete blood count with differential”. In: *J Perianesth Nurs* 18.2 (2003), pp. 96–114.

[26] T. Barbui et al. “The 2016 WHO classification and diagnostic criteria for myeloproliferative neoplasms: document summary and in-depth discussion”. In: *Blood Cancer J* 8.2 (2018), p. 15.

[27] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[28] B. Swolin et al. “Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks–evaluation of DiffMaster Octavia”. In: *Clin Lab Haematol* 25.3 (2003), pp. 139–147.

[29] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[30] Y. Fukuda et al. “Evidence for prevention of renal dysfunction associated with primary myelofibrosis by cytoreductive therapy”. In: *Haematologica* 104.11 (2019), e506–e509.

[31] H. Oguro. “The Roles of Cholesterol and Its Metabolites in Normal and Malignant Hematopoiesis”. In: *Front Endocrinol (Lausanne)* 10 (2019), p. 204.

[32] L. N. Griner et al. “JAK2-V617F-mediated signalling is dependent on lipid rafts and statins inhibit JAK2-V617F-dependent cell growth”. In: *Br J Haematol* 160.2 (2013), pp. 177–187.

[33] K. Kimura et al. “Automated diagnostic support system with deep learning algorithms for distinction of Philadelphia chromosome-negative myeloproliferative neoplasms using peripheral blood specimen”. In: *Sci Rep* 11.1 (2021), p. 3367.

[34] Zi Yun Ng et al. “Morphology of myeloproliferative neoplasms”. In: *International Journal of Laboratory Hematology* 45.S2 (2023), pp. 59–70. DOI: https://doi.org/10.1111/ijlh.14086. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijlh.14086. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijlh.14086.

[35] Andrew I. Schafer. “Thrombocytosis”. In: *JAMA* 314.11 (Sept. 2015), pp. 1171–1172. ISSN: 0098-7484. DOI: 10.1001/jama.2015.8515. eprint: https://jamanetwork.com/journals/jama/articlepdf/2441241/jdt150008.pdf. URL: https://doi.org/10.1001/jama.2015.8515.

[36] M. A. Yassin et al. “MERGE: A Multinational, Multicenter Observational Registry for Myeloproliferative Neoplasms in Asia, including Middle East, Turkey, and Algeria”. In: *Cancer Med* 9.13 (2020), pp. 4512–4526.

A

Characteristics of laboratory measurements dataset

A.1. ET Dataset

Table A.1: Total number of patients and samples in the laboratory measurements ET dataset (before splitting and undersampling)

	Patients	**Samples**
Control	11962	43226
ET	122	2532

Table A.2: Summary of feature values in the used laboratory measurements ET dataset. The control and ET columns contain the median and interquartile range of the specified features or the percentage of occurrences in case of binary values. The column 'n' provides the number of control / ET samples in the ET dataset for which a value of the given feature is available.

	Control	n	ET
	Median (IQR) / %	*control* / ET	*Median (IQR) / %*
ALAT	20.0 (14.0 - 31.0)	39897 / 2368	22.0 (17.0 - 30.0)
APTT	30.0 (27.0 - 35.0)	14562 / 748	28.0 (27.0 - 31.0)
ASAT	20.0 (15.0 - 29.0)	37842 / 2209	20.0 (15.0 - 26.0)
Album/creat	2.0 (1.0 - 5.0)	13235 / 943	1.0 (1.0 - 3.0)
Albumine	34.0 (28.0 - 39.0)	36838 / 2119	39.0 (36.0 - 42.0)
Alkalische Fosfatase (AF)	92.0 (71.0 - 128.0)	37536 / 2178	78.0 (63.0 - 101.0)
Amylase totaal	50.0 (35.0 - 72.0)	9129 / 638	52.0 (44.0 - 72.0)
Anion gap	8.0 (6.0 - 11.0)	11216 / 463	7.0 (5.0 - 9.0)
Bacteriën	100.0 (100.0 - 100.0)	1639 / 109	100.0 (100.0 - 100.0)
Base Excess	1.0 (-2.0 - 3.0)	17246 / 768	2.0 (1.0 - 4.0)
Basofielen	0.0 (0.0 - 0.0)	35313 / 2476	0.0 (0.0 - 0.0)
Bezinking	32.0 (13.0 - 66.0)	35269 / 2155	10.0 (5.0 - 21.0)
Bicarbonaat Arterieel	25.0 (22.0 - 28.0)	17275 / 768	26.0 (26.0 - 28.0)
Bilirubine Direct	5.0 (2.0 - 18.0)	7287 / 247	3.0 (2.0 - 30.0)
Bilirubine Totaal	8.0 (5.0 - 11.0)	36363 / 2148	8.0 (6.0 - 10.0)
CK	61.0 (36.0 - 103.0)	25085 / 1498	69.0 (52.0 - 100.0)
CRP	29.0 (7.0 - 94.0)	41309 / 2305	5.0 (4.0 - 14.0)
Calcium	2.0 (2.0 - 2.0)	35191 / 2096	2.0 (2.0 - 2.0)
Chloride	102.0 (99.0 - 106.0)	24177 / 1290	104.0 (102.0 - 107.0)

	Control *Median (IQR) / %*	n *control* / ET	ET *Median (IQR) / %*
Cholesterol	5.0 (4.0 - 6.0)	28381 / 2275	5.0 (4.0 - 6.0)
Cholesterol-HDL Ratio	4.0 (3.0 - 5.0)	24616 / 2227	4.0 (3.0 - 5.0)
Eiwit kwalitatief	0.0 (0.0 - 0.0)	2 / 0	nan (nan - nan)
Eosinofielen	0.0 (0.0 - 0.0)	35353 / 2476	0.0 (0.0 - 0.0)
Erytroblasten	1.0 (0.0 - 1.0)	11494 / 813	0.0 (0.0 - 1.0)
Erytrocyten	4.0 (4.0 - 4.0)	37109 / 2304	4.0 (4.0 - 5.0)
Ferritine	80.0 (26.0 - 238.0)	26735 / 2007	77.0 (33.0 - 135.0)
Foliumzuur	13.0 (9.0 - 23.0)	23298 / 1762	15.0 (11.0 - 23.0)
Fosfaat	1.0 (1.0 - 1.0)	29618 / 1527	1.0 (1.0 - 1.0)
GGT	33.0 (18.0 - 76.0)	38753 / 2278	22.0 (16.0 - 34.0)
Glucose	6.0 (5.0 - 8.0)	37112 / 2395	6.0 (5.0 - 7.0)
Glucose (bloedgas)	8.0 (6.0 - 9.0)	3925 / 20	10.0 (8.0 - 10.0)
Glucose 10 uur	7.0 (6.0 - 9.0)	14327 / 424	8.0 (5.0 - 9.0)
Glucose 14 uur	8.0 (6.0 - 10.0)	14145 / 395	9.0 (8.0 - 10.0)
Glucose 22 u	7.0 (6.0 - 10.0)	7615 / 153	14.0 (10.0 - 14.0)
Glucose POC	7.0 (6.0 - 10.0)	10617 / 428	8.0 (6.0 - 10.0)
Glucose nuchter	6.0 (5.0 - 7.0)	29362 / 2025	5.0 (5.0 - 6.0)
HDL Cholesterol	1.0 (1.0 - 2.0)	25493 / 2235	1.0 (1.0 - 2.0)
HbA1c	45.0 (39.0 - 55.0)	14398 / 928	42.0 (38.0 - 51.0)
Hematocriet	0.0 (0.0 - 0.0)	42521 / 2529	0.0 (0.0 - 0.0)
Hemoglobine	7.0 (6.0 - 8.0)	43198 / 2529	8.0 (8.0 - 9.0)
IJzer	7.0 (4.0 - 13.0)	22159 / 1646	14.0 (10.0 - 20.0)
INR	1.0 (1.0 - 3.0)	15678 / 569	2.0 (1.0 - 3.0)
INR Trodis coaguchek	3.0 (2.0 - 3.0)	2033 / 18	4.0 (3.0 - 4.0)
IgA	2.0 (2.0 - 4.0)	12184 / 964	2.0 (2.0 - 3.0)
Innametijd	14.0 (13.0 - 18.0)	6187 / 202	14.0 (12.0 - 21.0)
Kalium	4.0 (4.0 - 4.0)	41529 / 2417	4.0 (4.0 - 4.0)
Kalium bloedgas	4.0 (4.0 - 4.0)	4147 / 42	4.0 (3.0 - 4.0)
Kreatinine	71.0 (57.0 - 89.0)	42328 / 2516	74.0 (62.0 - 90.0)
LD	266.0 (193.0 - 368.0)	37434 / 2427	234.0 (192.0 - 328.0)
LDL Cholesterol	3.0 (2.0 - 4.0)	24536 / 2209	3.0 (2.0 - 4.0)
Lactaat	1.0 (1.0 - 2.0)	11823 / 356	1.0 (1.0 - 2.0)
Leucocyten	11.0 (9.0 - 15.0)	43105 / 2523	9.0 (7.0 - 11.0)
Leukocyten est.	500.0 (38.0 - 500.0)	23 / 26	25.0 (25.0 - 25.0)
Lymfocyten	2.0 (1.0 - 3.0)	35354 / 2476	2.0 (1.0 - 2.0)
MCV	88.0 (83.0 - 92.0)	42369 / 2489	92.0 (87.0 - 101.0)
Macrocytose	1.0 (1.0 - 1.0)	1 / 0	nan (nan - nan)
Magnesium	1.0 (1.0 - 1.0)	22014 / 883	1.0 (1.0 - 1.0)
Microalbumine in urine	10.0 (5.0 - 29.0)	16009 / 1289	8.0 (5.0 - 23.0)
Microcytose	1.0 (1.0 - 1.0)	14 / 0	nan (nan - nan)
Monocyten	1.0 (1.0 - 1.0)	35337 / 2476	1.0 (0.0 - 1.0)
Natrium	137.0 (135.0 - 139.0)	41538 / 2415	139.0 (137.0 - 140.0)
Natrium bloedgas	137.0 (134.0 - 140.0)	4177 / 42	138.0 (131.0 - 138.0)
Neutrofielen	7.0 (5.0 - 11.0)	35352 / 2476	5.0 (4.0 - 7.0)
Nitriet	1.0 (1.0 - 1.0)	8 / 0	nan (nan - nan)
Prostaat Specifiek Antigeen	2.0 (1.0 - 4.0)	5811 / 337	1.0 (0.0 - 3.0)
Protrombinetijd	11.0 (11.0 - 12.0)	12894 / 781	11.0 (10.0 - 11.0)

	Control *Median (IQR) / %*	n *control* / ET	ET *Median (IQR) / %*
RDW	15.0 (14.0 - 17.0)	42120 / 2484	14.0 (13.0 - 17.0)
Reticulocyten	1.0 (1.0 - 2.0)	16722 / 1203	1.0 (1.0 - 2.0)
Sediment	0.0 (0.0 - 0.0)	48 / 14	0.0 (0.0 - 0.0)
TSH	2.0 (1.0 - 2.0)	31101 / 2160	2.0 (1.0 - 3.0)
TYBC	60.0 (48.0 - 73.0)	19791 / 1443	64.0 (57.0 - 72.0)
Temperatuur	37.0 (37.0 - 37.0)	17082 / 768	37.0 (37.0 - 37.0)
Totaal eiwit	38.0 (1.0 - 71.0)	18786 / 1276	60.0 (1.0 - 72.0)
Transferrine	2.0 (2.0 - 3.0)	20754 / 1594	3.0 (2.0 - 3.0)
Triglyceriden	2.0 (1.0 - 2.0)	26437 / 2244	1.0 (1.0 - 2.0)
Trombocyten	523.0 (479.0 - 606.0)	43226 / 2532	568.0 (506.0 - 679.0)
Troponine I	0.0 (0.0 - 0.0)	7422 / 507	0.0 (0.0 - 0.0)
Ureum	6.0 (4.0 - 8.0)	37557 / 2109	6.0 (4.0 - 7.0)
Urinezuur	0.0 (0.0 - 0.0)	12237 / 1323	0.0 (0.0 - 0.0)
IJzerverzadiging	13.0 (8.0 - 21.0)	18973 / 1425	22.0 (15.0 - 29.0)
Vitamine B12	294.0 (211.0 - 446.0)	26317 / 1894	295.0 (219.0 - 427.0)
Vitamine D	51.0 (31.0 - 69.0)	17009 / 1125	51.0 (33.0 - 68.0)
Vrij T4	15.0 (13.0 - 17.0)	17971 / 1501	15.0 (13.0 - 16.0)
eAG	7.0 (6.0 - 9.0)	14330 / 928	7.0 (6.0 - 8.0)
eGFR (CKD-EPI)	78.0 (60.0 - 90.0)	19556 / 2335	73.0 (56.0 - 87.0)
eGFR (MDRD)	60.0 (60.0 - 60.0)	31429 / 2402	60.0 (60.0 - 60.0)
p-amylase	17.0 (12.0 - 26.0)	15224 / 876	20.0 (15.0 - 24.0)
pCO2	6.0 (5.0 - 7.0)	17498 / 800	6.0 (5.0 - 7.0)
pH	7.0 (6.0 - 7.0)	27319 / 1698	6.0 (6.0 - 7.0)
pO2	10.0 (8.0 - 15.0)	13004 / 499	9.0 (7.0 - 11.0)
sO2	1.0 (1.0 - 1.0)	13068 / 488	1.0 (1.0 - 1.0)
Diagnose	0.0 (0.0 - 0.0)	43226 / 2532	1.0 (1.0 - 1.0)
Leeftijd	65.0 (52.0 - 75.0)	43226 / 2532	64.0 (52.0 - 73.0)
prev trombos	524.0 (480.0 - 607.0)	43226 / 2532	567.0 (505.0 - 682.0)
Male	35.0%	15055 / 677	27.0%
Female	65.0%	28171 / 1855	73.0%
Smoking	10.0%	4169 / 191	8.0%

A.2. PV Dataset

Table A.3: Total number of patients and samples in the laboratory measurements PV dataset (before splitting and undersampling)

	Patients	**Samples**
Control	2274	8932
PV	58	180

Table A.4: Summary of feature values in the used laboratory measurements PV dataset. The control and PV columns contain the median and interquartile range of the specified features or the percentage of occurrences in case of binary values. The column 'n' provides the number of control / PV samples in the PV dataset for which a value of the given feature is available.

	Control	n	PV
	Median (IQR) / %	*control* / PV	*Median (IQR) / %*
ALAT	27.0 (20.0 - 39.0)	8709 / 173	28.0 (23.0 - 35.0)
APTT	30.0 (27.0 - 34.0)	3913 / 29	28.0 (28.0 - 31.0)
ASAT	24.0 (18.0 - 31.0)	8416 / 157	22.0 (16.0 - 27.0)
Afnametijd	10.0 (8.0 - 15.0)	2193 / 11	14.0 (14.0 - 14.0)
Album/creat	2.0 (1.0 - 7.0)	3924 / 69	2.0 (0.0 - 2.0)
Albumine	39.0 (35.0 - 42.0)	8017 / 143	40.0 (38.0 - 43.0)
Alkalische Fosfatase (AF)	82.0 (67.0 - 104.0)	8325 / 149	75.0 (58.0 - 95.0)
Amylase totaal	51.0 (37.0 - 76.0)	2489 / 33	63.0 (53.0 - 100.0)
Anion gap	8.0 (6.0 - 10.0)	3410 / 7	3.0 (3.0 - 7.0)
Bacteriën	100.0 (100.0 - 100.0)	296 / 0	nan (nan - nan)
Base Excess	1.0 (-1.0 - 3.0)	4959 / 36	3.0 (1.0 - 3.0)
Basofielen	0.0 (0.0 - 0.0)	7785 / 177	0.0 (0.0 - 0.0)
Bezinking	7.0 (3.0 - 16.0)	8119 / 159	5.0 (2.0 - 6.0)
Bicarbonaat Arterieel	25.0 (23.0 - 28.0)	4966 / 36	28.0 (26.0 - 30.0)
Bilirubine Direct	7.0 (2.0 - 22.0)	2322 / 5	1.0 (1.0 - 1.0)
Bilirubine Totaal	10.0 (8.0 - 14.0)	8163 / 151	8.0 (8.0 - 11.0)
CK	82.0 (51.0 - 137.0)	6448 / 119	66.0 (45.0 - 108.0)
CRP	5.0 (5.0 - 19.0)	8608 / 169	5.0 (3.0 - 5.0)
Calcium	2.0 (2.0 - 2.0)	7737 / 145	2.0 (2.0 - 2.0)
Chloride	103.0 (100.0 - 106.0)	6044 / 72	104.0 (102.0 - 105.0)
Cholesterol	5.0 (4.0 - 6.0)	7417 / 175	5.0 (4.0 - 6.0)
Cholesterol-HDL Ratio	4.0 (3.0 - 5.0)	6816 / 175	4.0 (4.0 - 5.0)
Doseeradvies antibiotica	21.0 (13.0 - 27.0)	22 / 0	nan (nan - nan)
Eosinofielen	0.0 (0.0 - 0.0)	7793 / 177	0.0 (0.0 - 0.0)
Erytroblasten	0.0 (0.0 - 1.0)	2267 / 43	0.0 (0.0 - 1.0)
Erytrocyten	5.0 (4.0 - 6.0)	7422 / 154	6.0 (5.0 - 6.0)
Ferritine	107.0 (42.0 - 228.0)	4955 / 152	80.0 (22.0 - 144.0)
Foliumzuur	14.0 (10.0 - 22.0)	4371 / 152	15.0 (10.0 - 23.0)
Fosfaat	1.0 (1.0 - 1.0)	6791 / 79	1.0 (1.0 - 1.0)
GGT	35.0 (22.0 - 62.0)	8502 / 167	26.0 (20.0 - 32.0)
Glucose	6.0 (5.0 - 8.0)	8297 / 169	5.0 (5.0 - 6.0)
Glucose (bloedgas)	8.0 (6.0 - 10.0)	1299 / 0	nan (nan - nan)
Glucose 10 uur	7.0 (6.0 - 9.0)	3015 / 1	9.0 (9.0 - 9.0)
Glucose 14 uur	8.0 (6.0 - 10.0)	3086 / 1	12.0 (12.0 - 12.0)
Glucose 22 u	7.0 (6.0 - 9.0)	1707 / 1	12.0 (12.0 - 12.0)
Glucose POC	7.0 (6.0 - 10.0)	2866 / 26	6.0 (5.0 - 6.0)
Glucose nuchter	6.0 (5.0 - 7.0)	6892 / 163	5.0 (4.0 - 6.0)
HDL Cholesterol	1.0 (1.0 - 1.0)	6973 / 175	1.0 (1.0 - 1.0)
HbA1c	45.0 (39.0 - 52.0)	4081 / 71	39.0 (37.0 - 41.0)
Hematocriet	0.0 (0.0 - 1.0)	8909 / 180	1.0 (0.0 - 1.0)
Hemoglobine	10.0 (10.0 - 11.0)	8932 / 180	10.0 (10.0 - 11.0)
IJzer	12.0 (6.0 - 18.0)	3670 / 119	15.0 (8.0 - 20.0)
INR	1.0 (1.0 - 3.0)	3608 / 29	3.0 (2.0 - 3.0)
INR Trodis coaguchek	3.0 (2.0 - 3.0)	478 / 7	4.0 (3.0 - 4.0)

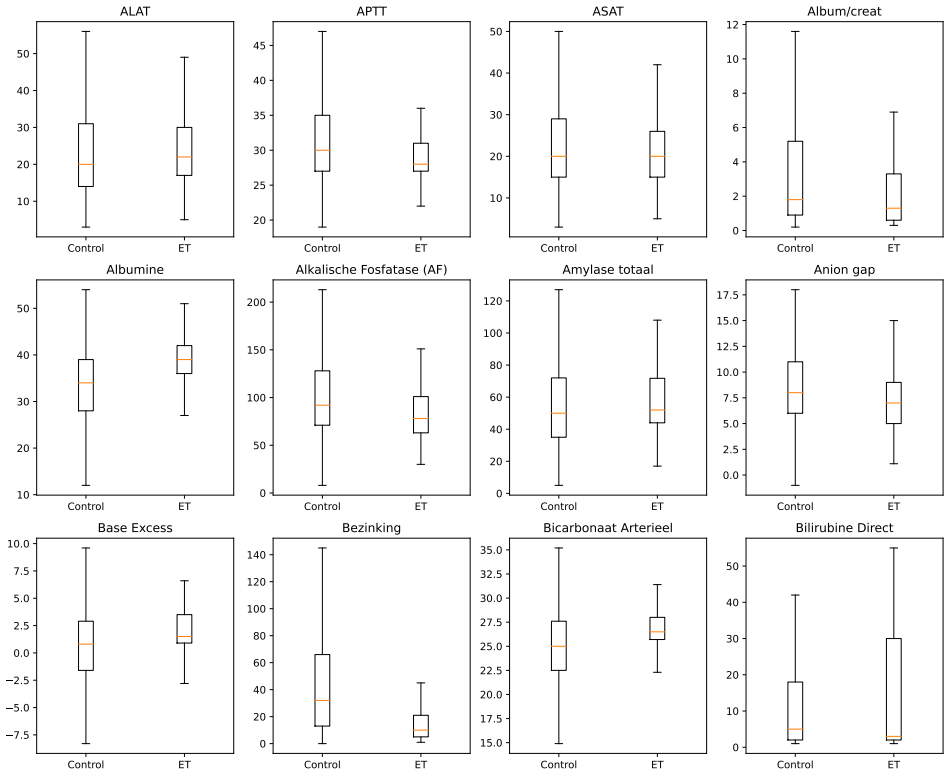
	Control	n	PV
	Median (IQR) / %	*control* / PV	*Median (IQR) / %*
IgA	2.0 (2.0 - 3.0)	2672 / 47	2.0 (2.0 - 2.0)
Innametijd	14.0 (13.0 - 19.0)	1685 / 22	22.0 (14.0 - 22.0)
Kalium	4.0 (4.0 - 4.0)	8774 / 167	4.0 (4.0 - 4.0)
Kalium bloedgas	4.0 (4.0 - 4.0)	1300 / 0	nan (nan - nan)
Kreatinine	81.0 (66.0 - 98.0)	8888 / 180	81.0 (72.0 - 93.0)
LD	287.0 (208.0 - 368.0)	8294 / 177	234.0 (204.0 - 321.0)
LDL Cholesterol	3.0 (2.0 - 4.0)	6700 / 173	3.0 (2.0 - 3.0)
Lactaat	1.0 (1.0 - 2.0)	3638 / 14	2.0 (1.0 - 2.0)
Leucocyten	9.0 (7.0 - 11.0)	8892 / 180	9.0 (8.0 - 11.0)
Leukocyten est.	20.0 (20.0 - 20.0)	6 / 0	nan (nan - nan)
Lymfocyten	2.0 (2.0 - 3.0)	7793 / 177	2.0 (1.0 - 2.0)
MCV	91.0 (88.0 - 95.0)	8830 / 179	89.0 (84.0 - 97.0)
Macrocytose	nan (nan - nan)	0 / 0	nan (nan - nan)
Magnesium	1.0 (1.0 - 1.0)	5163 / 39	1.0 (1.0 - 1.0)
Microalbumine in urine	13.0 (6.0 - 46.0)	4574 / 74	14.0 (10.0 - 30.0)
Microcytose	1.0 (1.0 - 1.0)	10 / 0	nan (nan - nan)
Monocyten	1.0 (1.0 - 1.0)	7793 / 177	1.0 (1.0 - 1.0)
Natrium	139.0 (137.0 - 140.0)	8773 / 167	139.0 (138.0 - 140.0)
Natrium bloedgas	138.0 (135.0 - 140.0)	1307 / 0	nan (nan - nan)
Neutrofielen	6.0 (4.0 - 8.0)	7793 / 177	6.0 (5.0 - 9.0)
PSA	1.0 (1.0 - 3.0)	3002 / 89	1.0 (1.0 - 2.0)
Protrombinetijd	11.0 (11.0 - 12.0)	3673 / 25	12.0 (12.0 - 12.0)
RDW	14.0 (13.0 - 15.0)	8773 / 175	15.0 (14.0 - 18.0)
Reticulocyten	1.0 (1.0 - 2.0)	2448 / 83	2.0 (1.0 - 2.0)
Sediment	0.0 (0.0 - 0.0)	3 / 0	nan (nan - nan)
TSH	2.0 (1.0 - 2.0)	7454 / 159	2.0 (1.0 - 2.0)
TYBC	64.0 (55.0 - 74.0)	2983 / 114	63.0 (58.0 - 69.0)
Temperatuur	37.0 (37.0 - 37.0)	4936 / 36	37.0 (37.0 - 37.0)
Totaal eiwit	24.0 (1.0 - 72.0)	4791 / 62	68.0 (1.0 - 74.0)
Transferrine	2.0 (2.0 - 3.0)	3274 / 119	3.0 (2.0 - 3.0)
Triglyceriden	2.0 (1.0 - 2.0)	7024 / 173	2.0 (1.0 - 2.0)
Trombocyten	269.0 (221.0 - 332.0)	8861 / 180	544.0 (440.0 - 702.0)
Troponine I	0.0 (0.0 - 0.0)	2339 / 54	0.0 (0.0 - 0.0)
Ureum	6.0 (4.0 - 8.0)	8191 / 123	5.0 (5.0 - 6.0)
Urinezuur	0.0 (0.0 - 0.0)	3573 / 112	0.0 (0.0 - 0.0)
IJzerverzadiging	19.0 (10.0 - 28.0)	2866 / 110	22.0 (10.0 - 29.0)
Vitamine B12	302.0 (219.0 - 446.0)	5386 / 154	300.0 (252.0 - 391.0)
Vitamine D	48.0 (32.0 - 67.0)	3823 / 47	42.0 (30.0 - 54.0)
Vrij T4	15.0 (13.0 - 17.0)	4273 / 88	16.0 (13.0 - 18.0)
eAG	7.0 (6.0 - 9.0)	4072 / 71	6.0 (6.0 - 7.0)
eGFR (CKD-EPI)	73.0 (57.0 - 88.0)	5432 / 178	77.0 (62.0 - 85.0)
eGFR (MDRD)	60.0 (60.0 - 60.0)	7928 / 164	60.0 (60.0 - 60.0)
p-amylase	17.0 (12.0 - 27.0)	4213 / 43	15.0 (14.0 - 16.0)
pCO2	6.0 (5.0 - 7.0)	5001 / 36	6.0 (5.0 - 54.0)
pH	6.0 (6.0 - 7.0)	6905 / 98	6.0 (6.0 - 7.0)
pO2	10.0 (8.0 - 13.0)	4104 / 30	10.0 (10.0 - 10.0)
sO2	1.0 (1.0 - 1.0)	4150 / 30	1.0 (1.0 - 1.0)

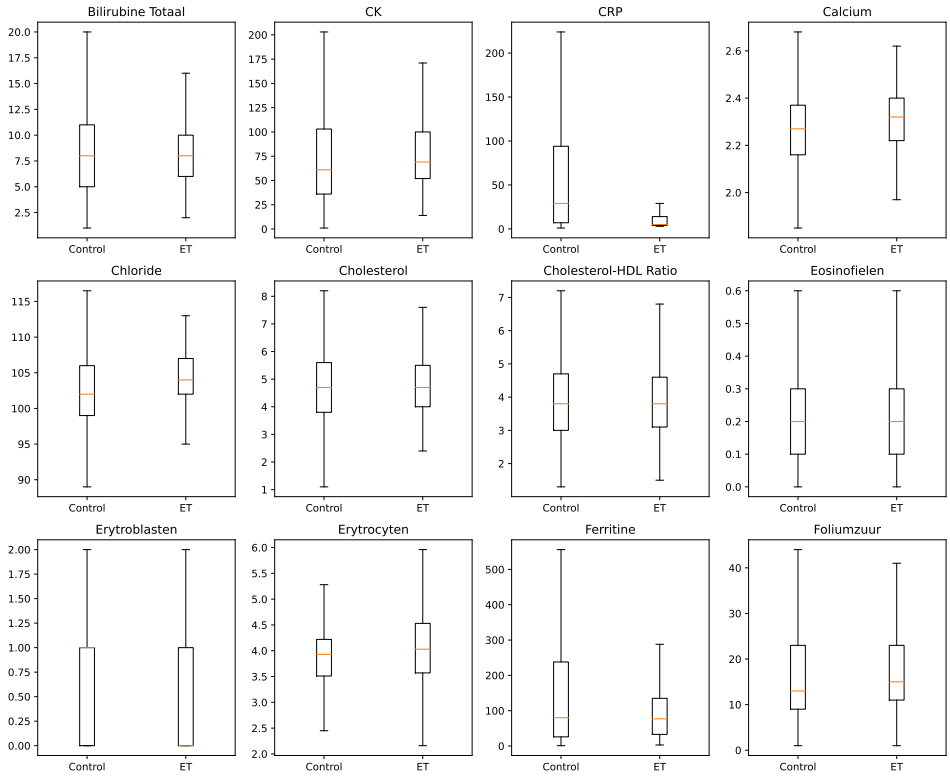
	Control *Median (IQR) / %*	n *control* / PV	PV *Median (IQR) / %*
Diagnose	0.0 (0.0 - 0.0)	8932 / 180	1.0 (1.0 - 1.0)
Leeftijd	64.0 (54.0 - 72.0)	8932 / 180	63.0 (53.0 - 72.0)
Male	72.0%	6410 / 131	73.0%
Female	28.0%	2522 / 49	27.0%
Smoking	15.0%	1369 / 32	18.0%

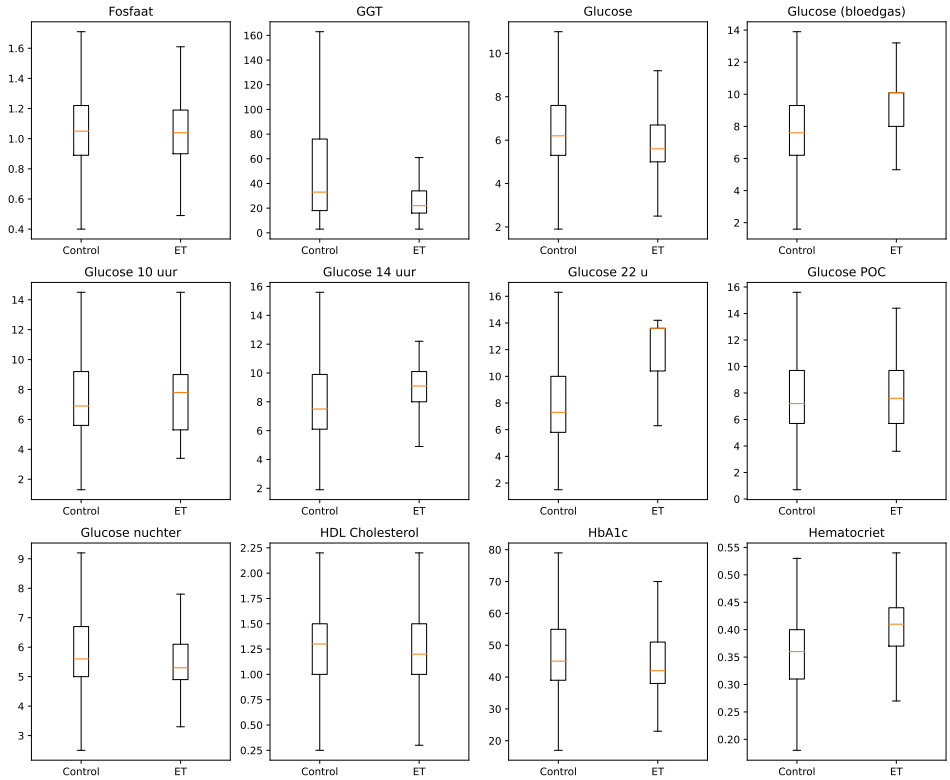
B

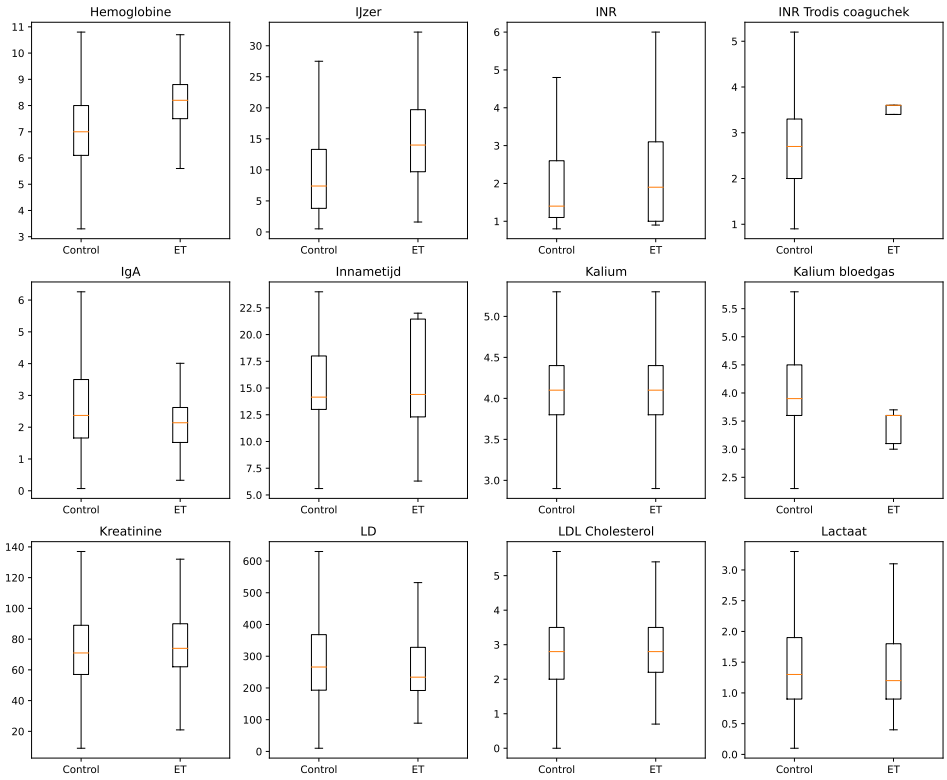
Boxplot of blood measurement dataset values

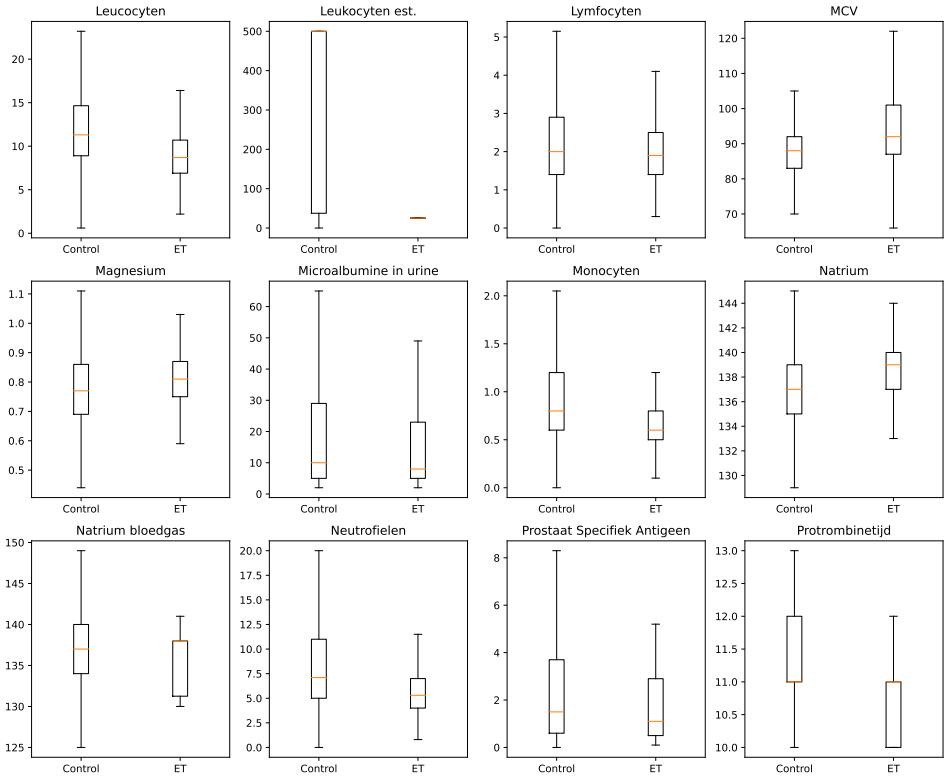
B.1. Boxplots for ET dataset

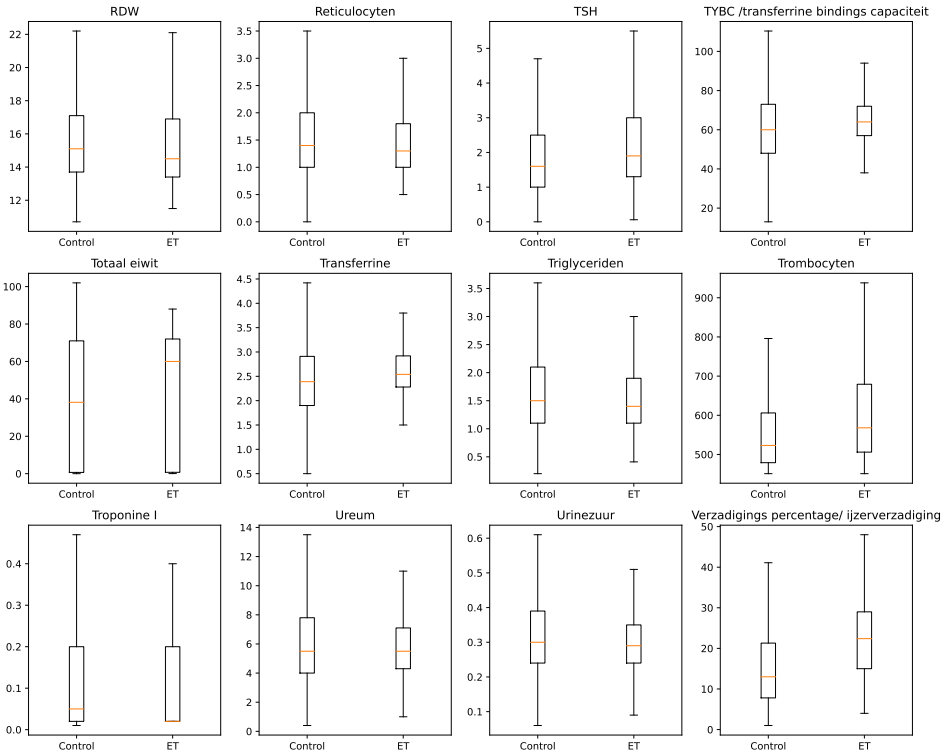


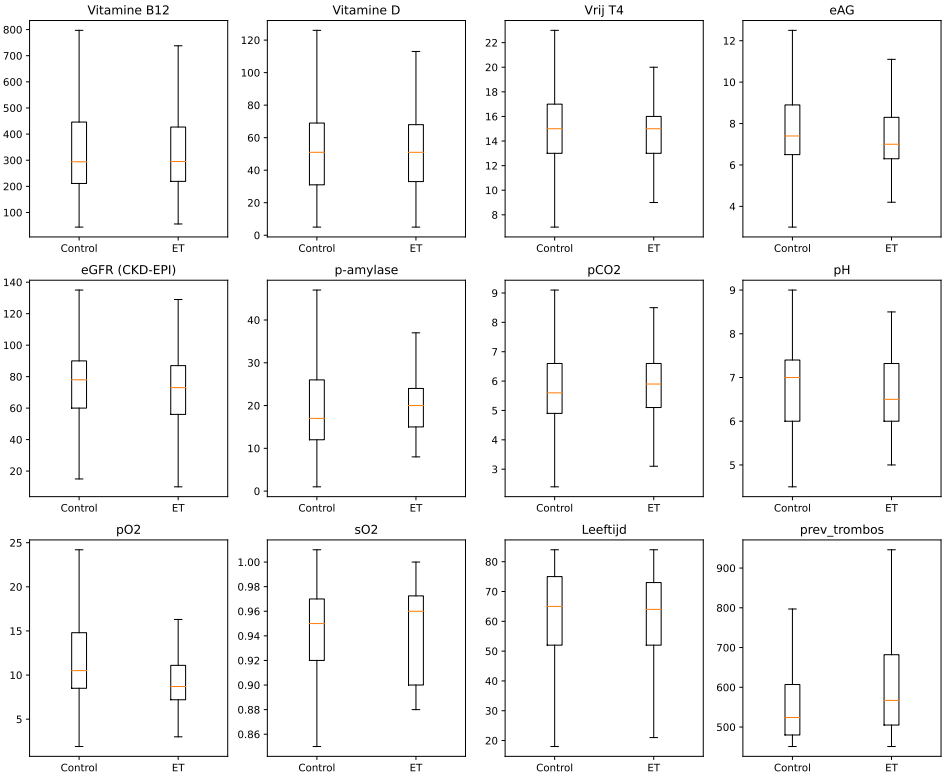

Bilirubine Totaal
CK
CRP
Calcium
Chloride
Cholesterol
Cholesterol-HDL Ratio
Eosinofielen
Erytroblasten
Erytrocyten
Ferritine
Foliumzuur
Control
ET


Fosfaat
GGT
Glucose
Glucose (bloedgas)
Glucose 10 uur
Glucose 14 uur
Glucose 22 u
Glucose POC
Glucose nuchter
HDL Cholesterol
HbA1c
Hematocriet
Control
ET

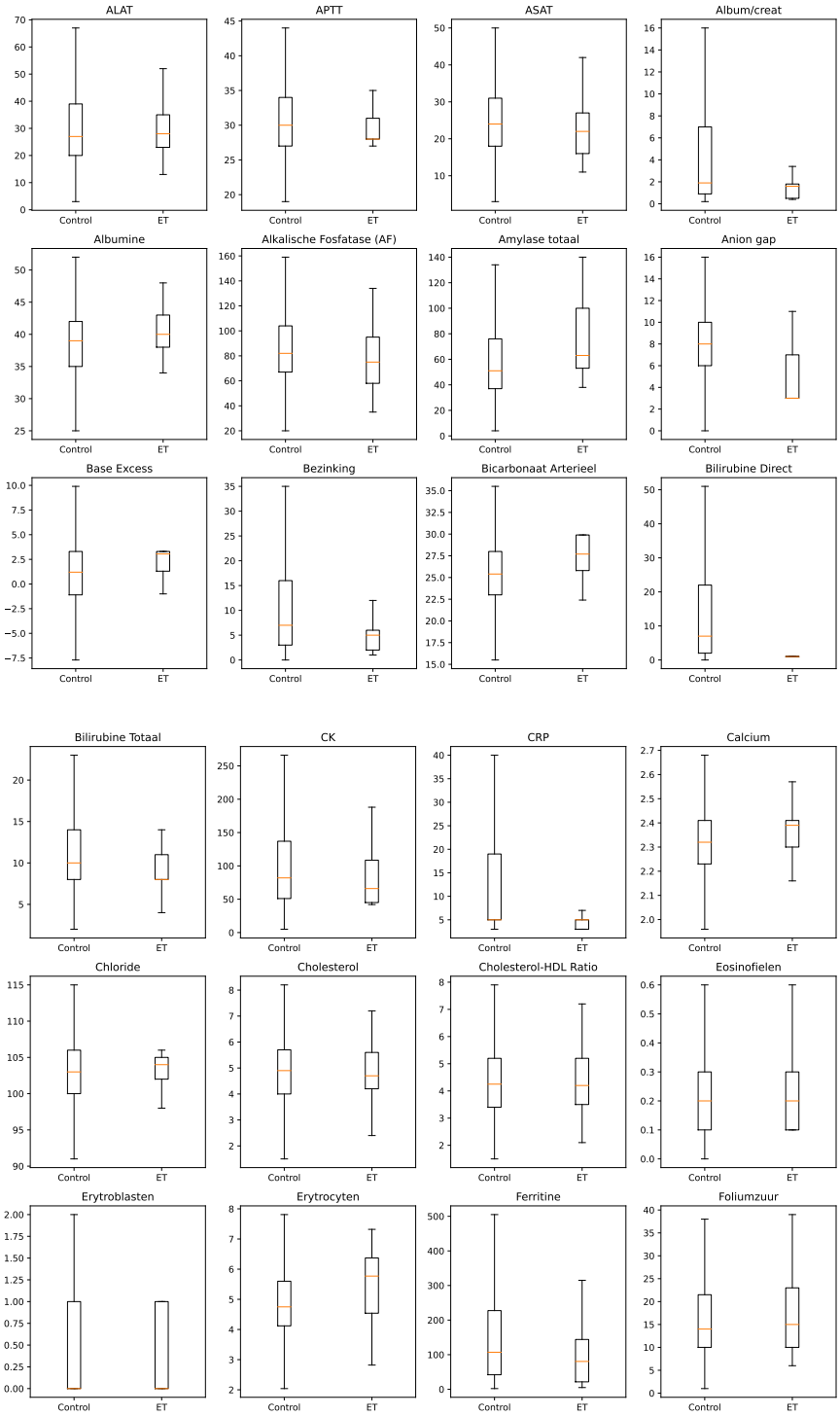

Hemoglobine
IJzer
INR
INR Trodis coaguchek
IgA
Innametijd
Kalium
Kalium bloedgas
Kreatinine
LD
LDL Cholesterol
Lactaat
Control
ET

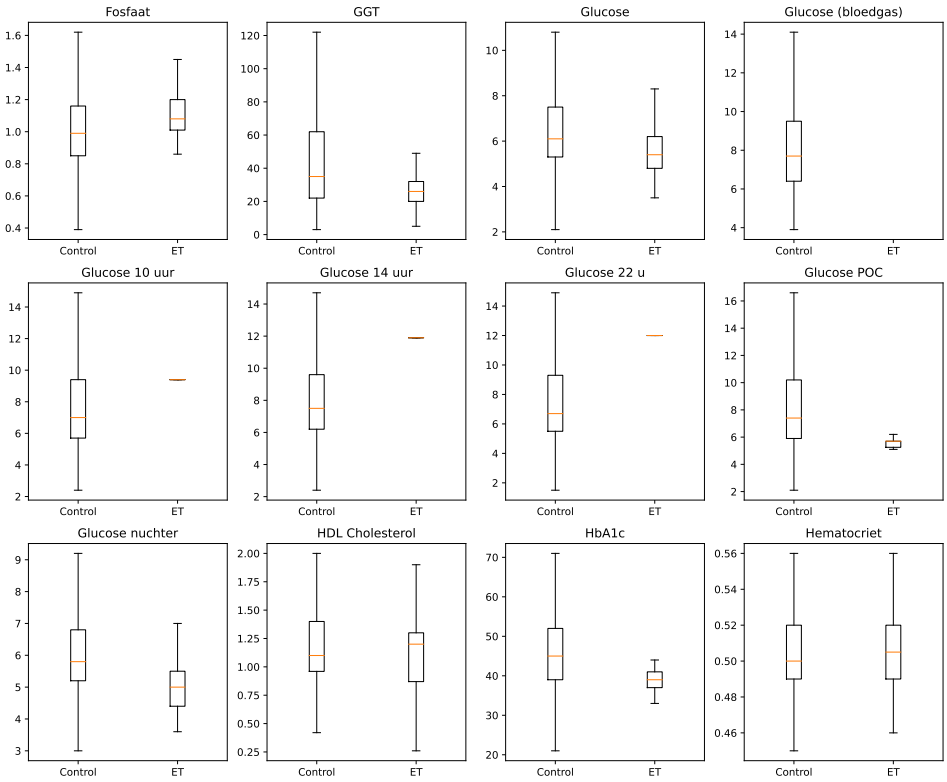

Leucocyten
Leukocyten est.
Lymfocyten
MCV
Magnesium
Microalbumine in urine
Monocyten
Natrium
Natrium bloedgas
Neutrofielen
Prostaat Specifiek Antigeen
Protrombinetijd
Control
ET

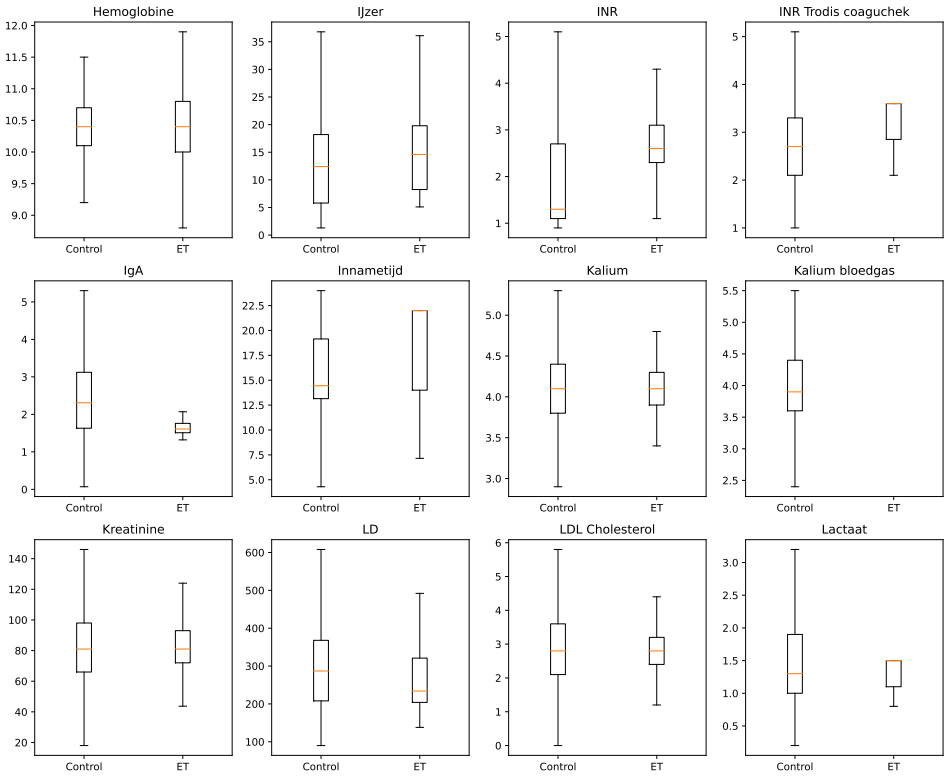

RDW
Reticulocyten
TSH
TYBC /transferrine bindings capaciteit
Totaal eiwit
Transferrine
Triglyceriden
Trombocyten
Troponine I
Ureum
Urinezuur
Verzadigings percentage/ ijzerverzadiging
Control
ET

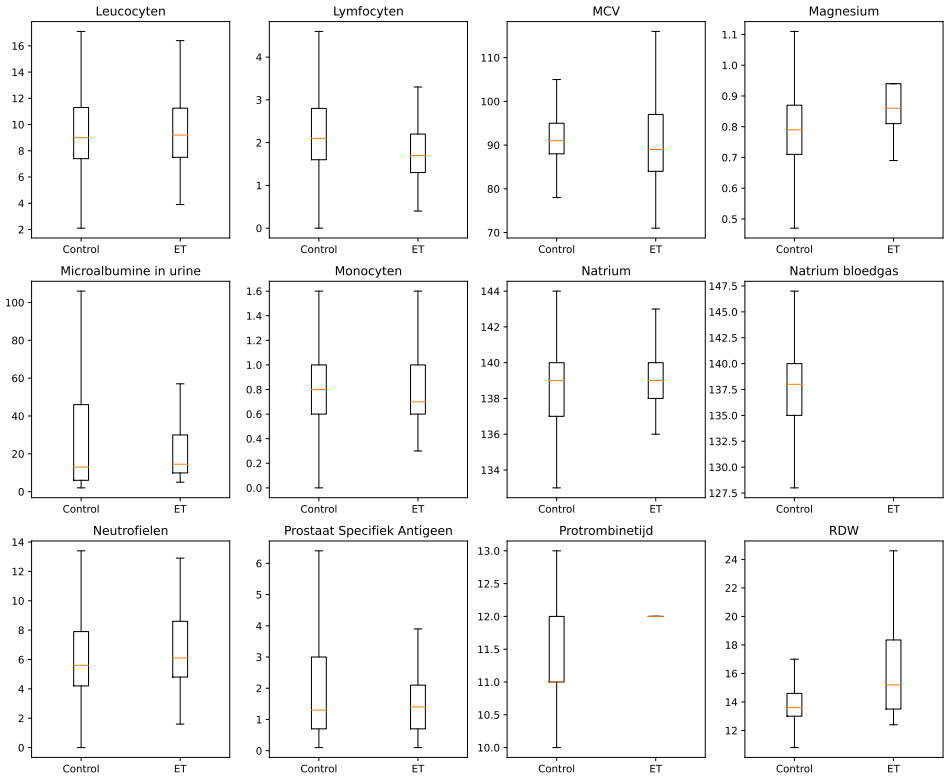

Vitamine B12
Vitamine D
Vrij T4
eAG
eGFR (CKD-EPI)
p-amylase
pCO2
pH
pO2
sO2
Leeftijd
prev_trombos
Control
ET

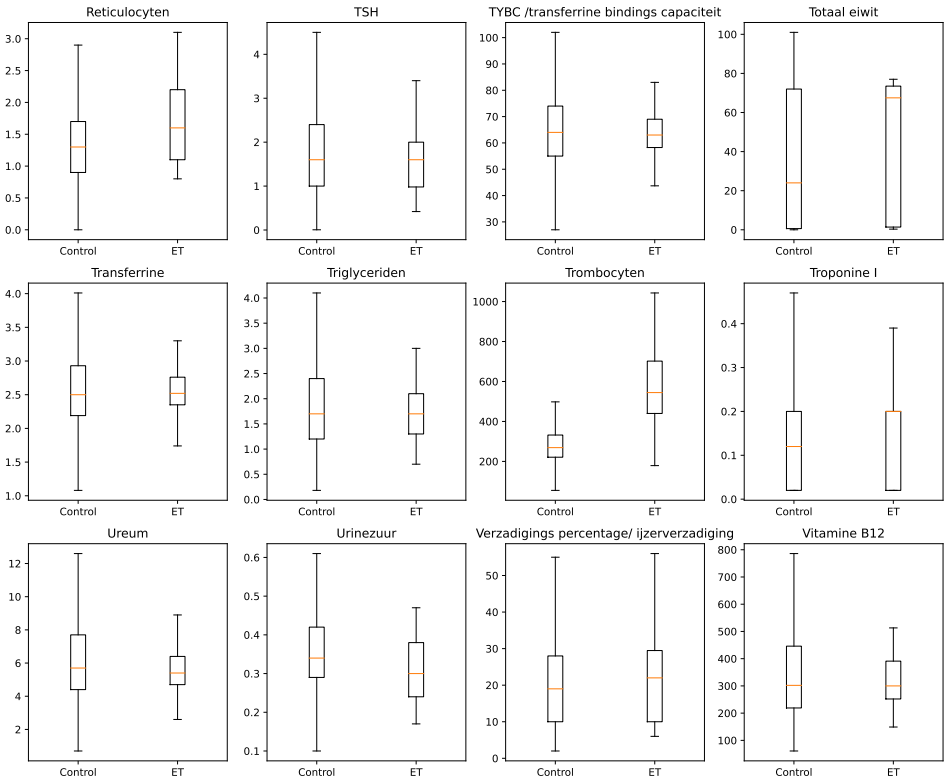
B.2. Boxplots for PV dataset

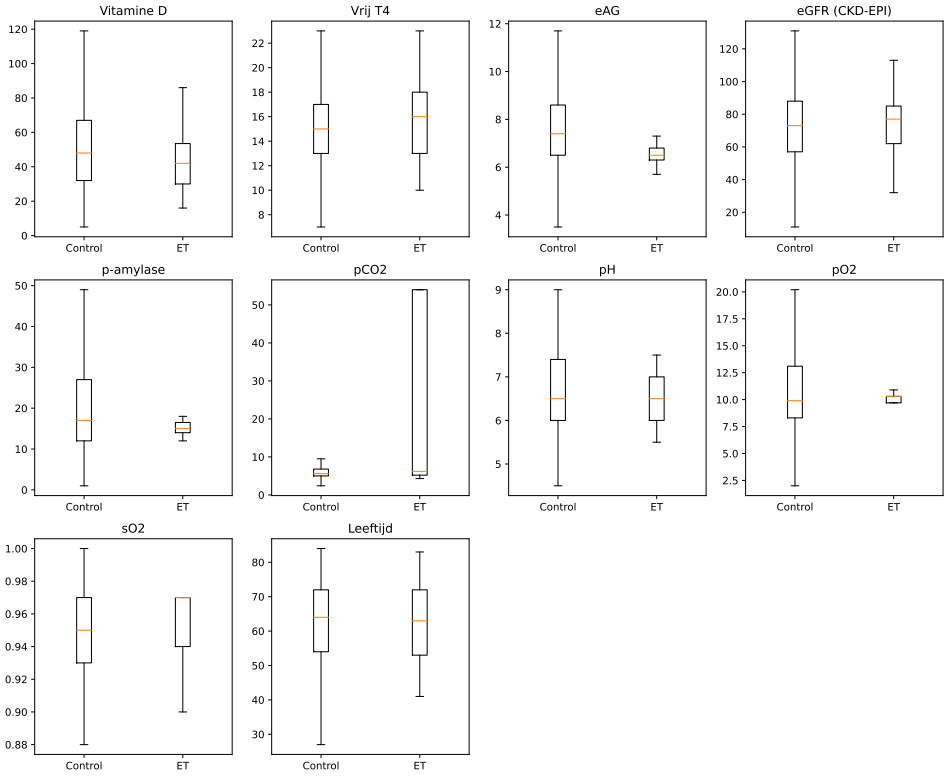



Fosfaat
GGT
Glucose
Glucose (bloedgas)
Glucose 10 uur
Glucose 14 uur
Glucose 22 u
Glucose POC
Glucose nuchter
HDL Cholesterol
HbA1c
Hematocriet
Control
ET


Hemoglobine
IJzer
INR
INR Trodis coaguchek
IgA
Innametijd
Kalium
Kalium bloedgas
Kreatinine
LD
LDL Cholesterol
Lactaat
Control
ET


Leucocyten
Lymfocyten
MCV
Magnesium
Microalbumine in urine
Monocyten
Natrium
Natrium bloedgas
Neutrofielen
Prostaat Specifiek Antigeen
Protrombinetijd
RDW
Control
ET


Reticulocyten
TSH
TYBC /transferrine bindings capaciteit
Totaal eiwit
Transferrine
Triglyceriden
Trombocyten
Troponine I
Ureum
Urinezuur
Verzadigings percentage/ ijzerverzadiging
Vitamine B12
Control
ET


Vitamine D
Vrij T4
eAG
eGFR (CKD-EPI)
p-amylase
pCO2
pH
pO2
sO2
Leeftijd
Control
ET

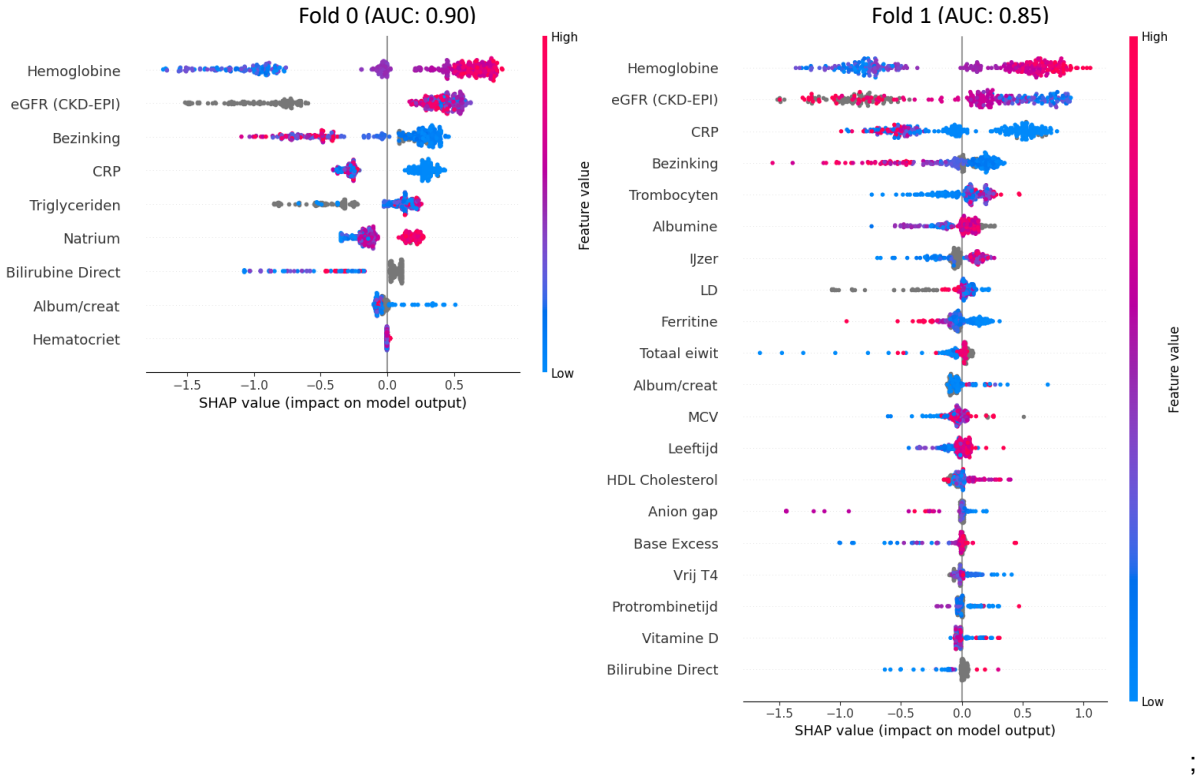
C

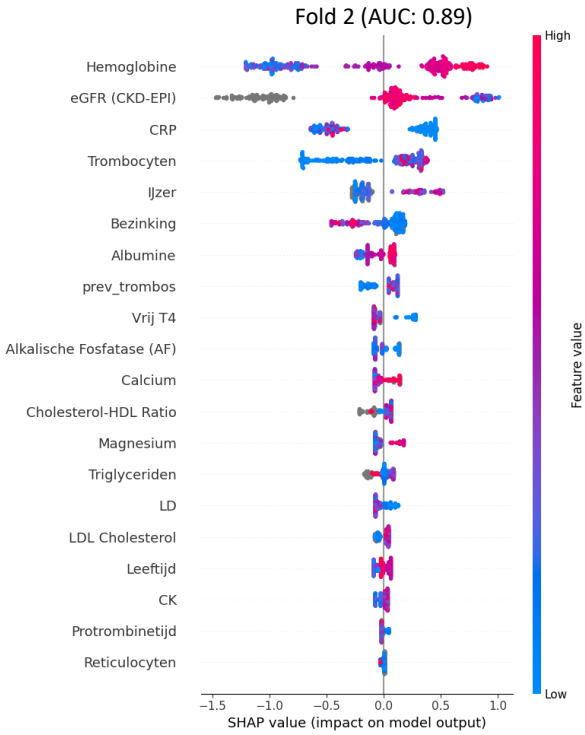
SHAP plots per outer crossvalidation fold in blood measurements filter

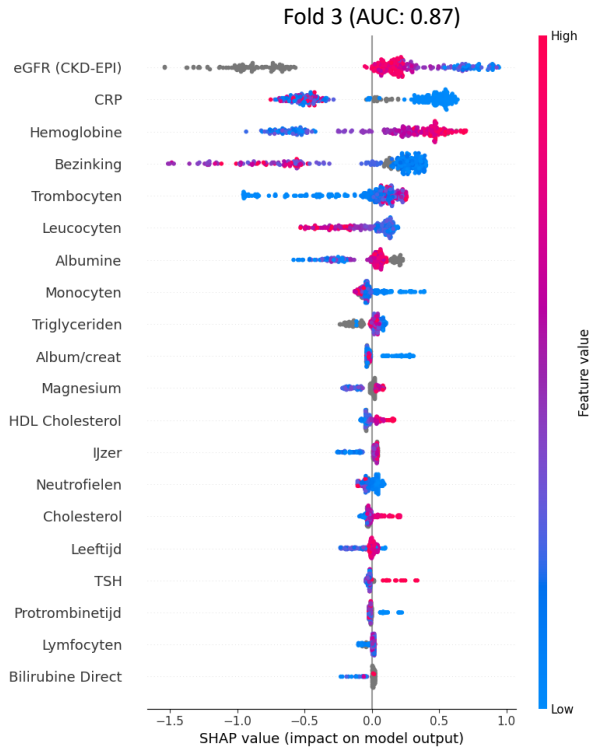
SHapley Additive exPlanations (SHAP) plots for each of the outer cross validation folds of the blood measurement filter. Gray points indicate unknown measurement values (measurment not performed), blue-red points indicate low-high feature value. Positive SHAP values indicate positive predictive value of feature value for prediction, negative SHAP values indicate negative ET or PV predictive value. A larger deviation from 0 means a larger inpact on model outcome.

AUC scores for the test set are shown together with the SHAP plot for each fold.

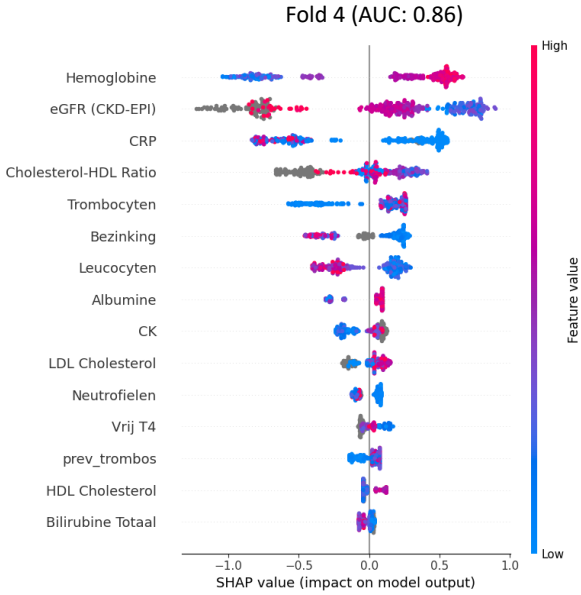
C.1. ET prediction models

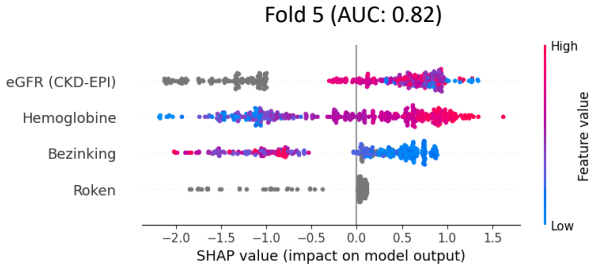



Fold 2 (AUC: 0.89)
Hemoglobine
eGFR (CKD-EPI)
CRP
Trombocyten
IJzer
Bezinking
Albumine
prev_trombos
Vrij T4
Alkalische Fosfatase (AF)
Calcium
Cholesterol-HDL Ratio
Magnesium
Triglyceriden
LD
LDL Cholesterol
Leeftijd
CK
Protrombinetijd
Reticulocyten
−1.5 −1.0 −0.5 0.0 0.5 1.0
SHAP value (impact on model output)
High
Feature value
Low

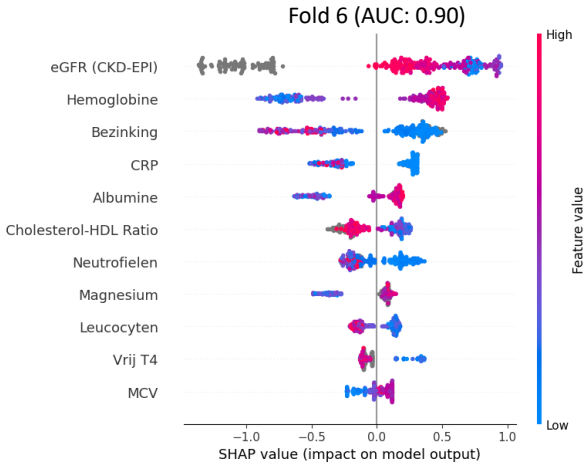

Fold 3 (AUC: 0.87)
eGFR (CKD-EPI)
CRP
Hemoglobine
Bezinking
Trombocyten
Leucocyten
Albumine
Monocyten
Triglyceriden
Album/creat
Magnesium
HDL Cholesterol
IJzer
Neutrofielen
Cholesterol
Leeftijd
TSH
Protrombinetijd
Lymfocyten
Bilirubine Direct
−1.5 −1.0 −0.5 0.0 0.5 1.0
SHAP value (impact on model output)
High
Feature value
Low

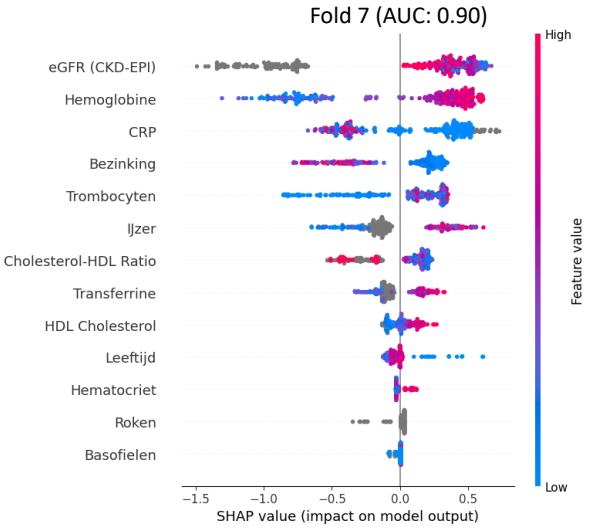
;


Fold 4 (AUC: 0.86)
Hemoglobine
eGFR (CKD-EPI)
CRP
Cholesterol-HDL Ratio
Trombocyten
Bezinking
Leucocyten
Albumine
CK
LDL Cholesterol
Neutrofielen
Vrij T4
prev_trombos
HDL Cholesterol
Bilirubine Totaal
−1.0 −0.5 0.0 0.5 1.0
SHAP value (impact on model output)
High
Feature value
Low

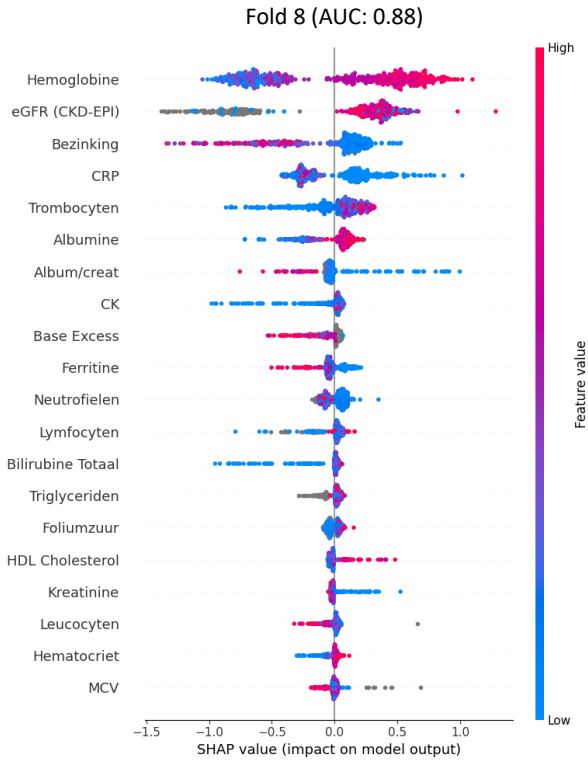

Fold 5 (AUC: 0.82)
eGFR (CKD-EPI)
Hemoglobine
Bezinking
Roken
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
SHAP value (impact on model output)
High
Feature value
Low

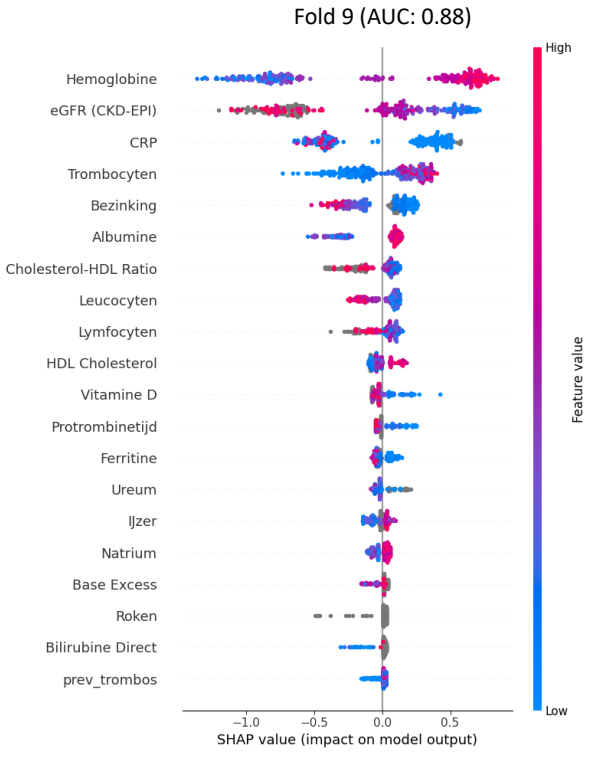
;


Fold 6 (AUC: 0.90)
eGFR (CKD-EPI)
Hemoglobine
Bezinking
CRP
Albumine
Cholesterol-HDL Ratio
Neutrofielen
Magnesium
Leucocyten
Vrij T4
MCV
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low


Fold 7 (AUC: 0.90)
eGFR (CKD-EPI)
Hemoglobine
CRP
Bezinking
Trombocyten
IJzer
Cholesterol-HDL Ratio
Transferrine
HDL Cholesterol
Leeftijd
Hematocriet
Roken
Basofielen
−1.5
−1.0
−0.5
0.0
0.5
SHAP value (impact on model output)
High
Feature value
Low

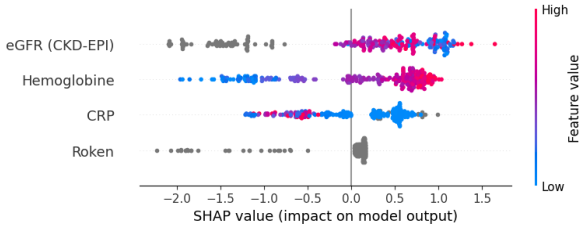
;


Fold 8 (AUC: 0.88)
Hemoglobine
eGFR (CKD-EPI)
Bezinking
CRP
Trombocyten
Albumine
Album/creat
CK
Base Excess
Ferritine
Neutrofielen
Lymfocyten
Bilirubine Totaal
Triglyceriden
Foliumzuur
HDL Cholesterol
Kreatinine
Leucocyten
Hematocriet
MCV
−1.5
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low

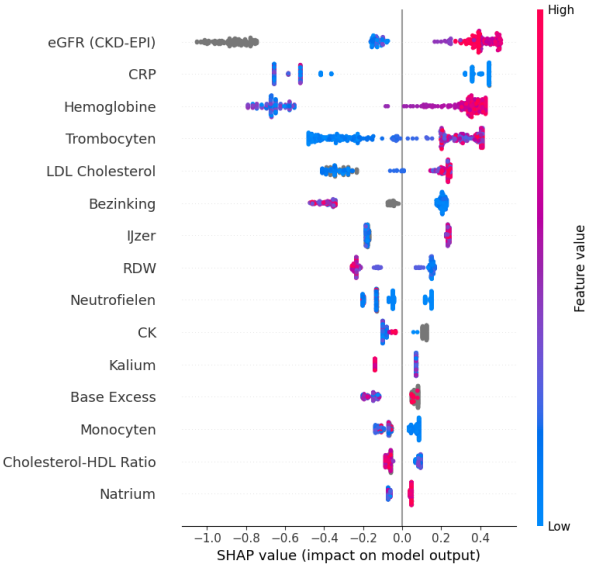

Fold 9 (AUC: 0.88)
Hemoglobine
eGFR (CKD-EPI)
CRP
Trombocyten
Bezinking
Albumine
Cholesterol-HDL Ratio
Leucocyten
Lymfocyten
HDL Cholesterol
Vitamine D
Protrombinetijd
Ferritine
Ureum
IJzer
Natrium
Base Excess
Roken
Bilirubine Direct
prev_trombos
−1.0
−0.5
0.0
0.5
SHAP value (impact on model output)
High
Feature value
Low

;

Fold 10 (AUC: 0.88)

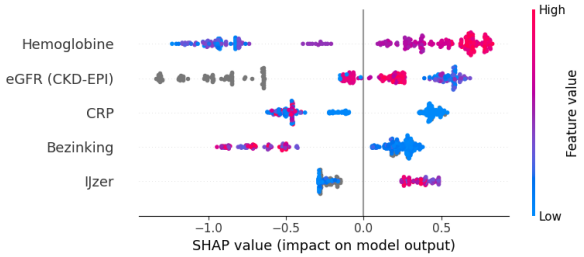


Fold 11 (AUC: 0.86)

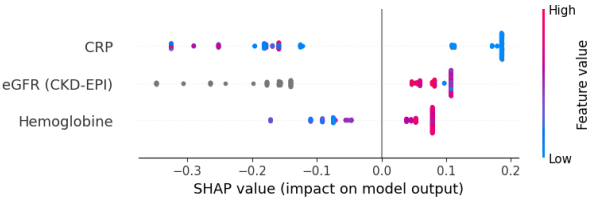


;

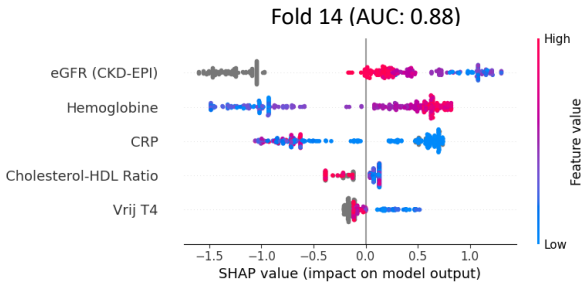
Fold 12 (AUC: 0.89)

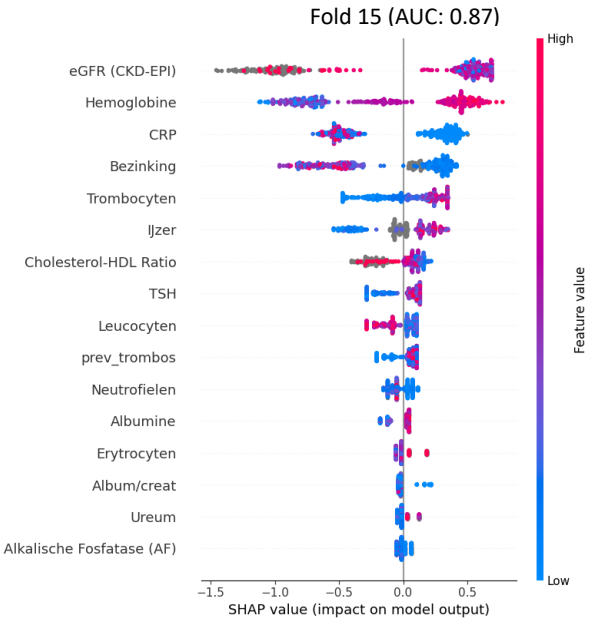


Fold 13 (AUC: 0.89)

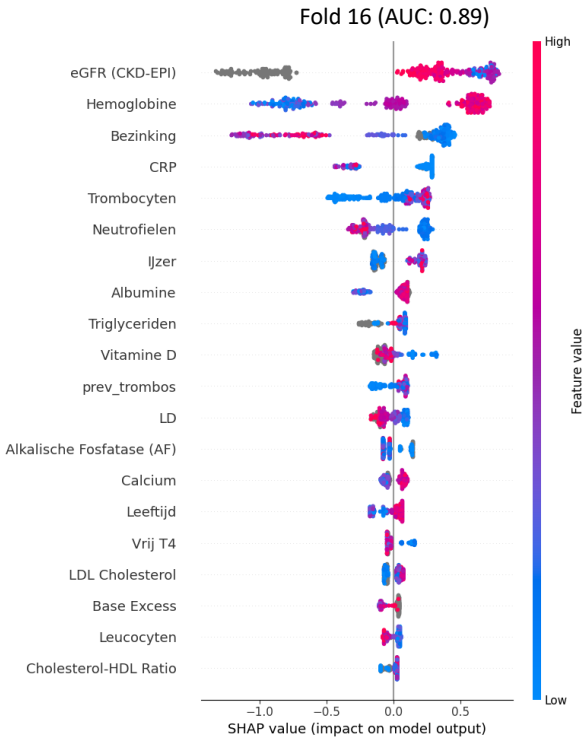


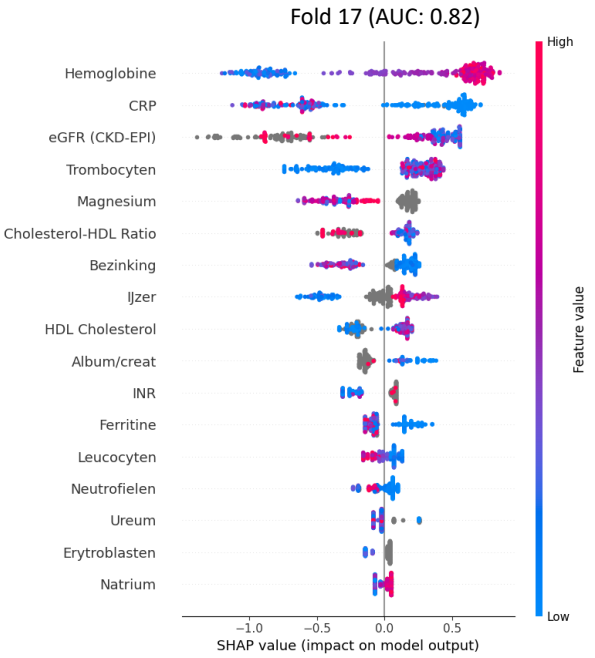
;


Fold 14 (AUC: 0.88)
eGFR (CKD-EPI)
Hemoglobine
CRP
Cholesterol-HDL Ratio
Vrij T4
−1.5 −1.0 −0.5 0.0 0.5 1.0
SHAP value (impact on model output)
High
Feature value
Low

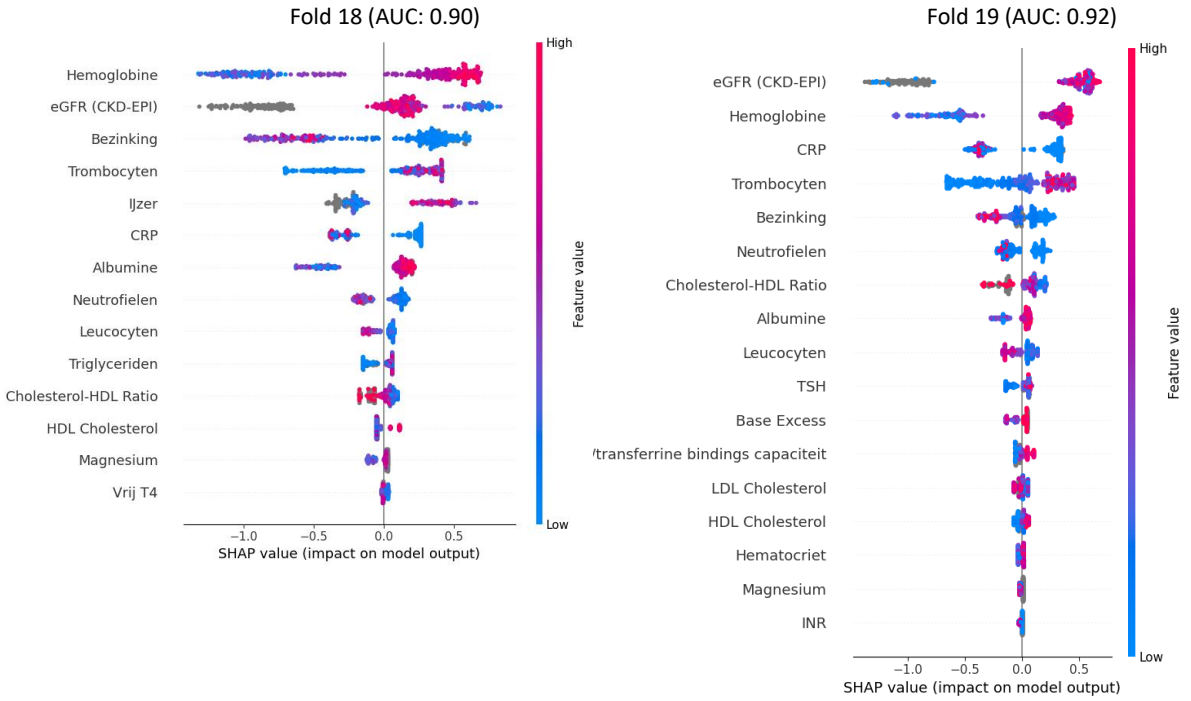

Fold 15 (AUC: 0.87)
eGFR (CKD-EPI)
Hemoglobine
CRP
Bezinking
Trombocyten
IJzer
Cholesterol-HDL Ratio
TSH
Leucocyten
prev_trombos
Neutrofielen
Albumine
Erytrocyten
Album/creat
Ureum
Alkalische Fosfatase (AF)
−1.5 −1.0 −0.5 0.0 0.5
SHAP value (impact on model output)
High
Feature value
Low

;


Fold 16 (AUC: 0.89)
eGFR (CKD-EPI)
Hemoglobine
Bezinking
CRP
Trombocyten
Neutrofielen
IJzer
Albumine
Triglyceriden
Vitamine D
prev_trombos
LD
Alkalische Fosfatase (AF)
Calcium
Leeftijd
Vrij T4
LDL Cholesterol
Base Excess
Leucocyten
Cholesterol-HDL Ratio
−1.0 −0.5 0.0 0.5
SHAP value (impact on model output)
High
Feature value
Low

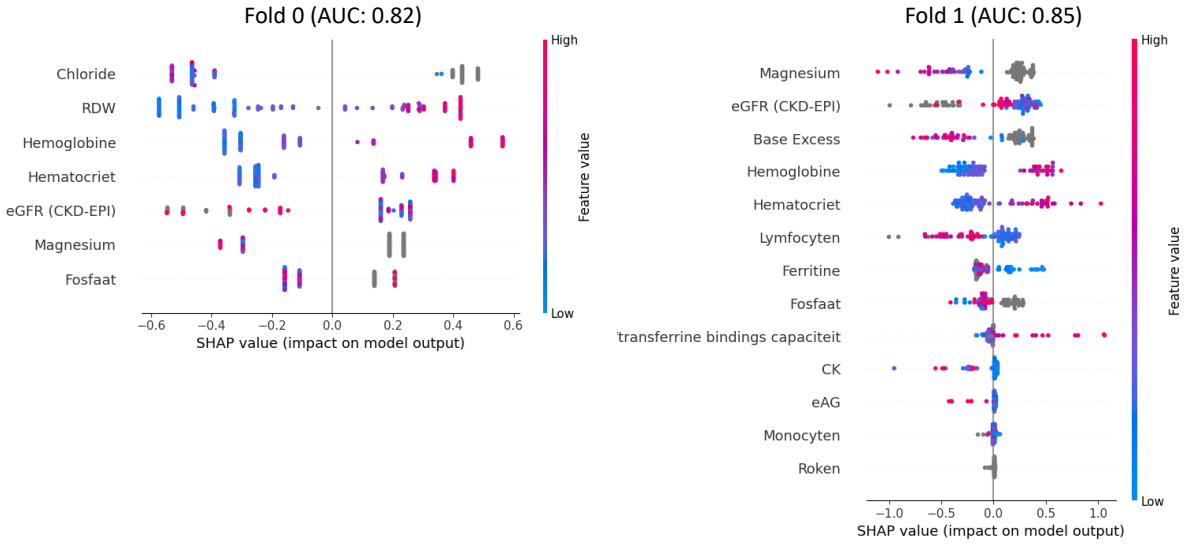

Fold 17 (AUC: 0.82)
Hemoglobine
CRP
eGFR (CKD-EPI)
Trombocyten
Magnesium
Cholesterol-HDL Ratio
Bezinking
IJzer
HDL Cholesterol
Album/creat
INR
Ferritine
Leucocyten
Neutrofielen
Ureum
Erytroblasten
Natrium
−1.0 −0.5 0.0 0.5
SHAP value (impact on model output)
High
Feature value
Low

;



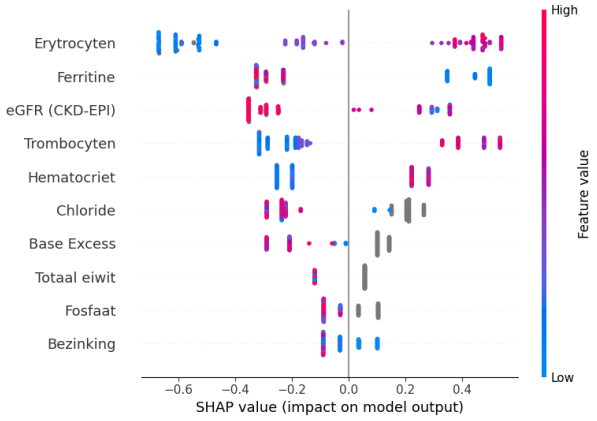
;

C.2. PV prediction models

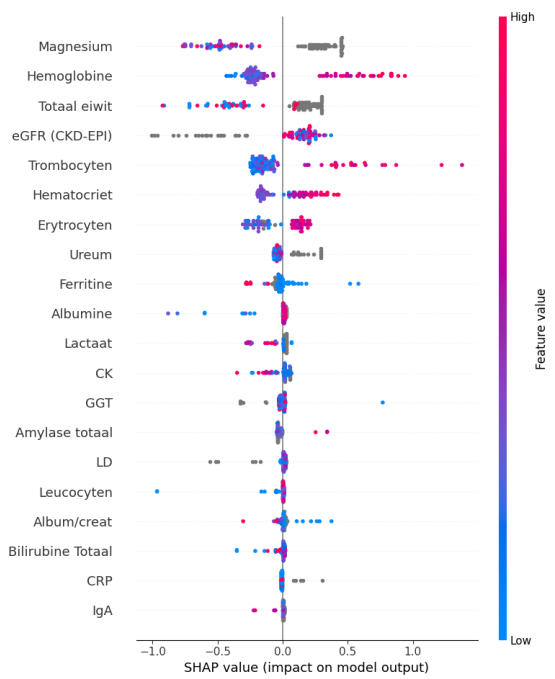


;

Fold 2 (AUC: 0.82)

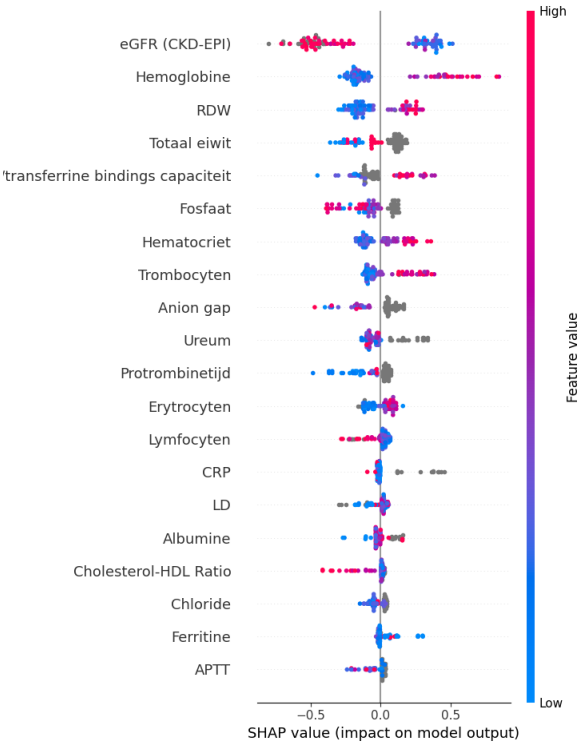


Fold 3 (AUC: 0.88)

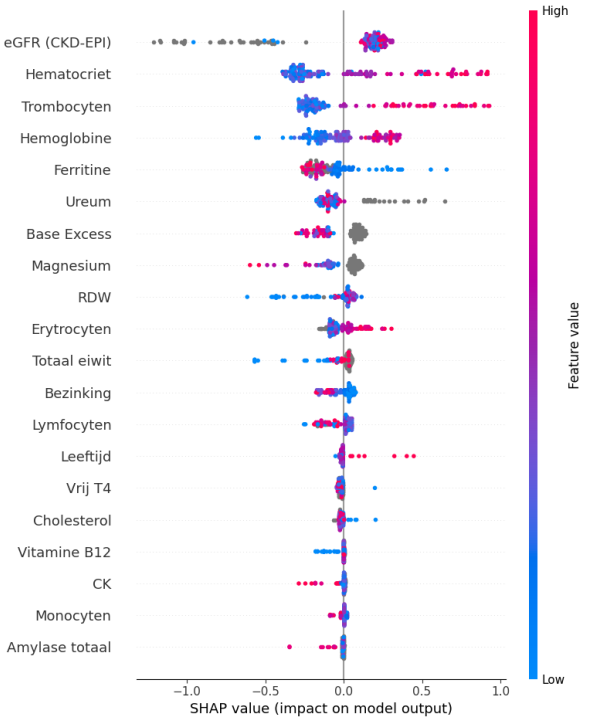


;

Fold 4 (AUC: 0.89)

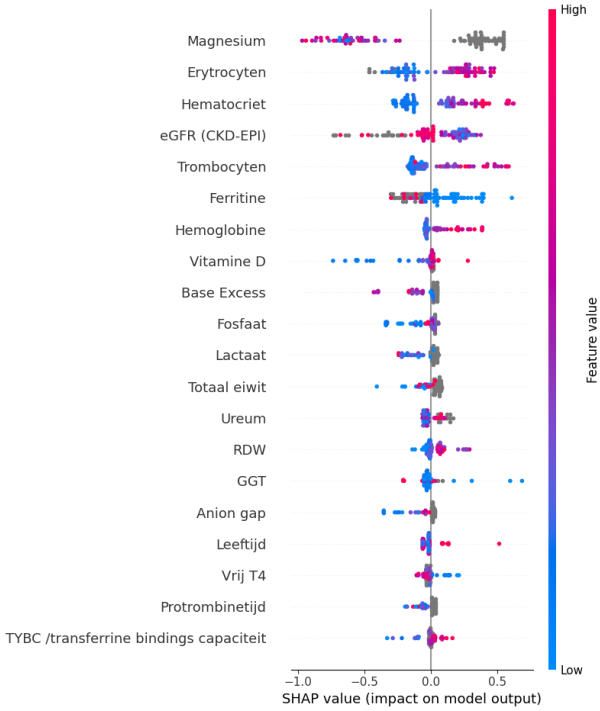


Fold 5 (AUC: 0.88)

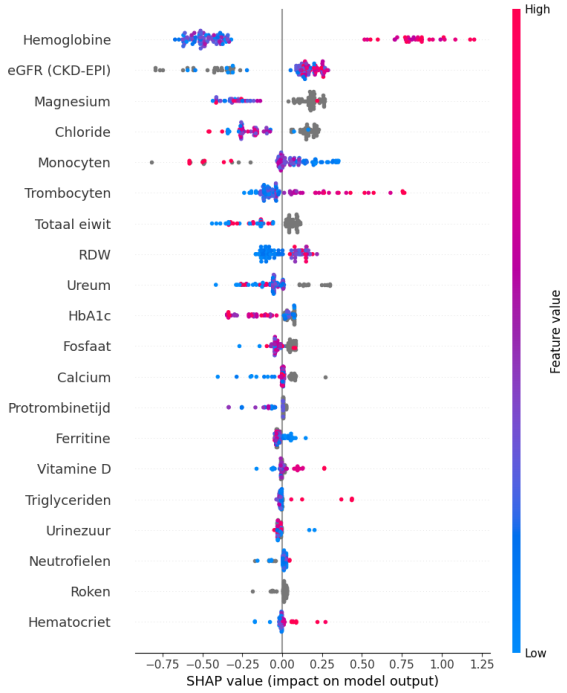


;

Fold 6 (AUC: 0.90)

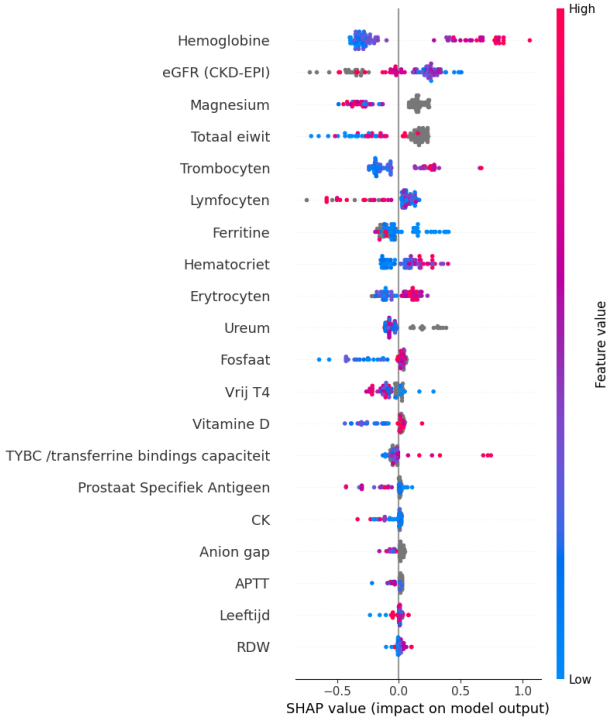


Fold 7 (AUC: 0.74)

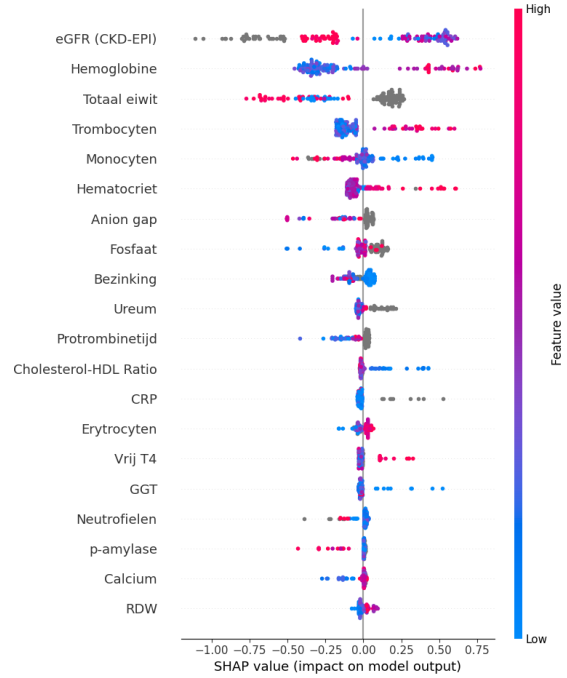


;

Fold 8 (AUC: 0.91)

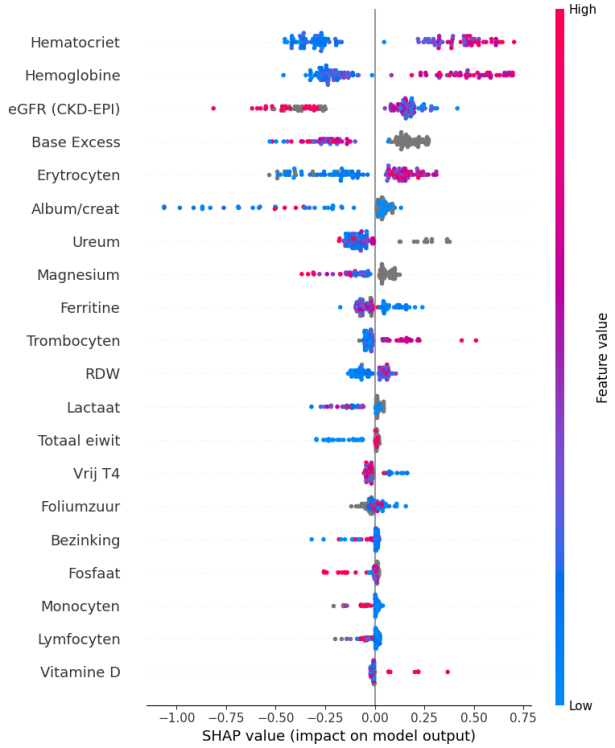


Fold 9 (AUC: 0.93)

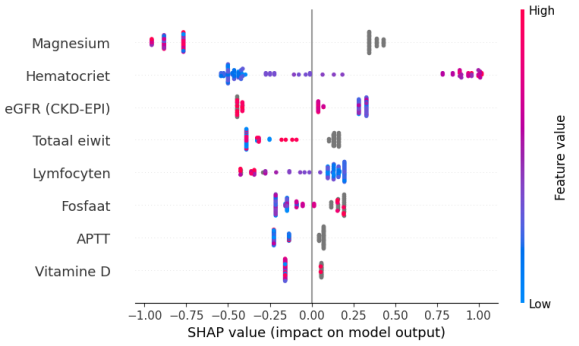


;

Fold 10 (AUC: 0.88)

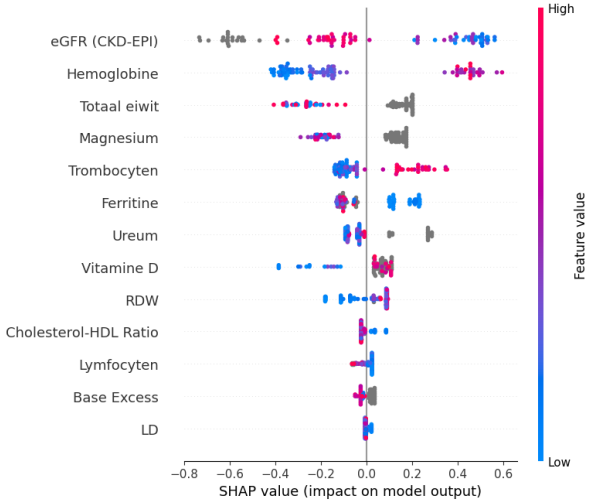


Fold 11 (AUC: 0.81)

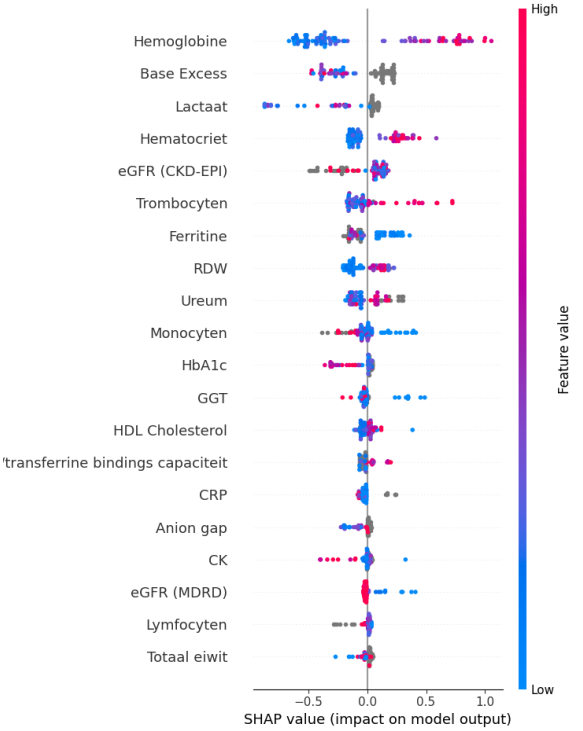


;

Fold 12 (AUC: 0.97)

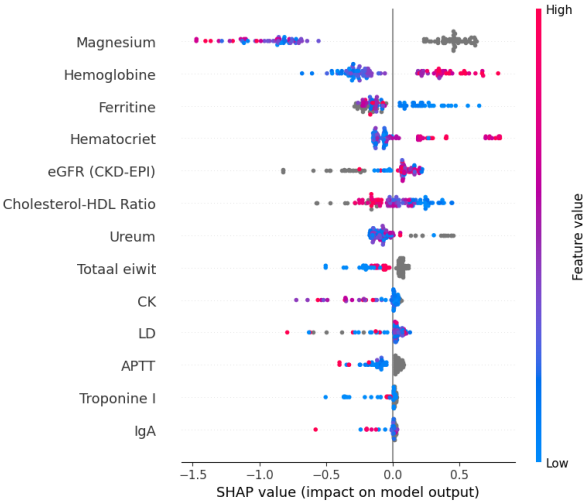


Fold 13 (AUC: 0.83)

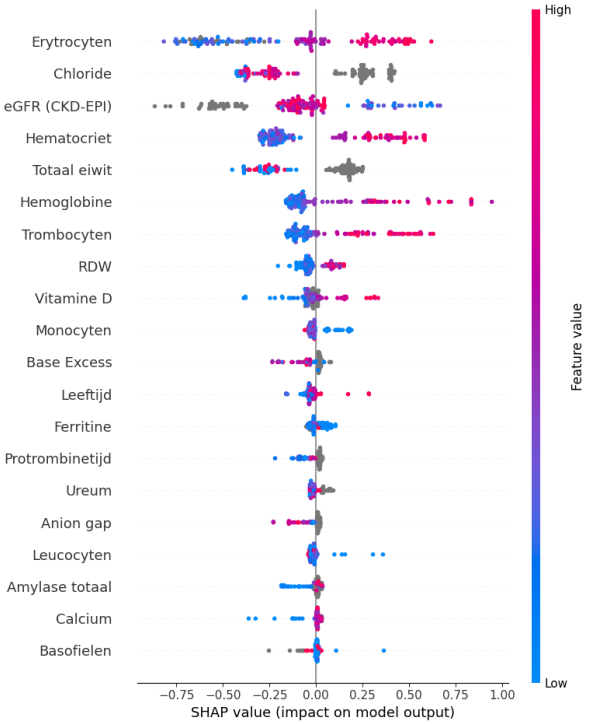


;

Fold 14 (AUC: 0.81)

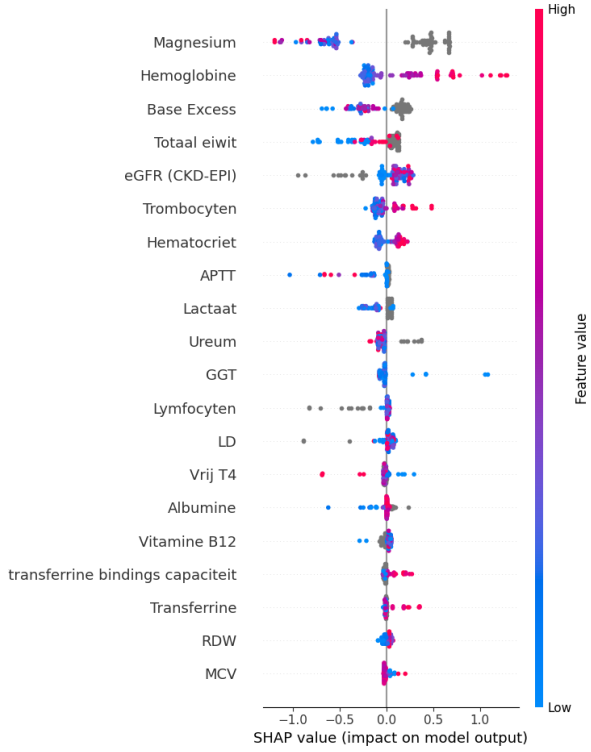


Fold 15 (AUC: 0.85)

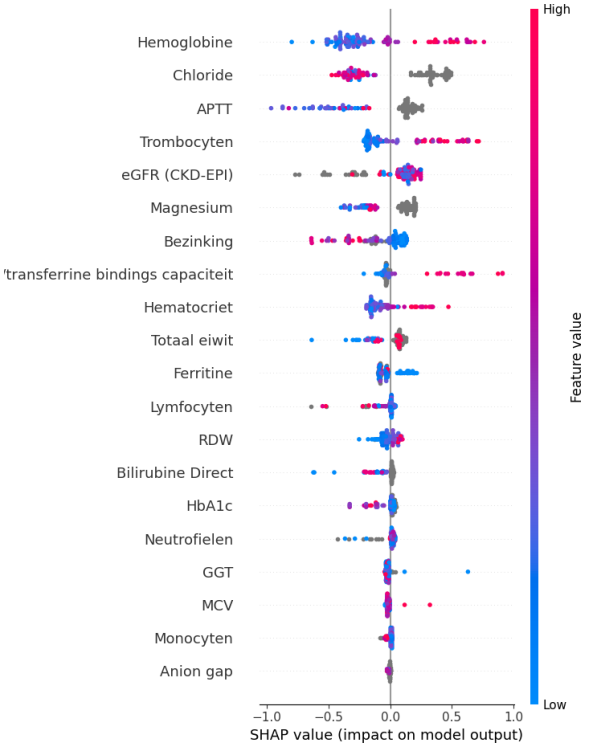


;

Fold 16 (AUC: 0.91)

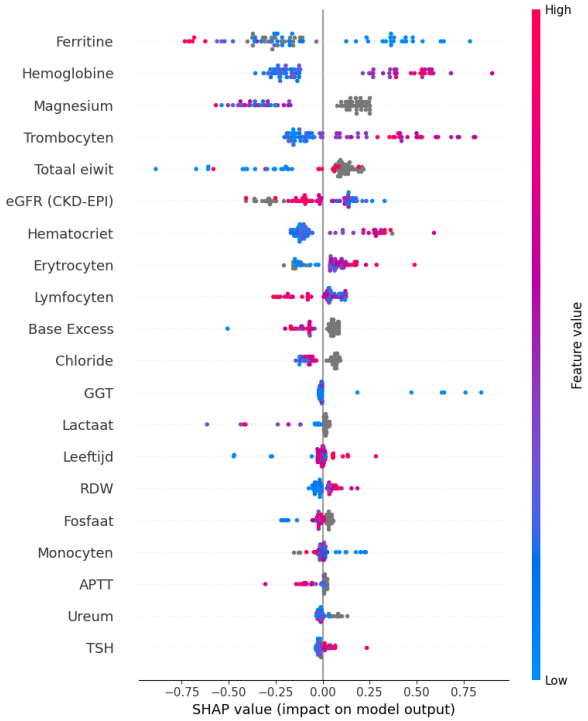


Fold 17 (AUC: 0.86)

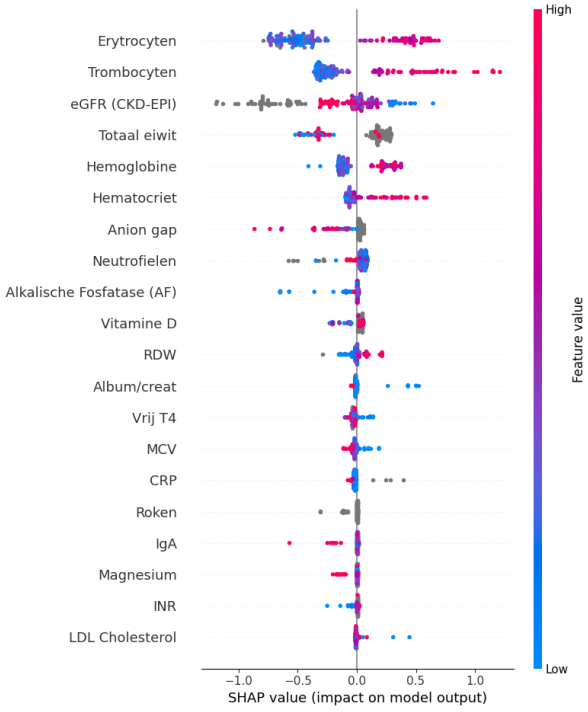


;

Fold 18 (AUC: 0.94)



Fold 19 (AUC: 0.85)



;

D

ResNet50 with lymphocyte images

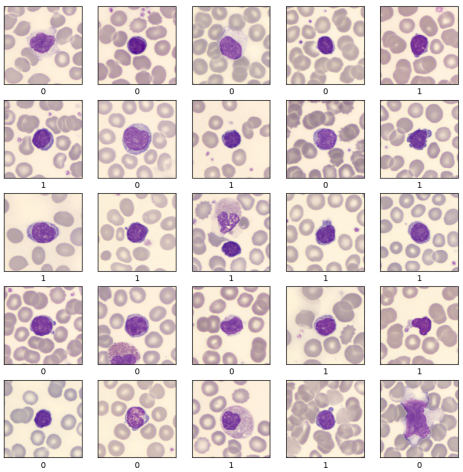


Figure D.1: Visualisation of randomly selected lymphocyte images used as input for the ResNet50 model with their ground truth labels (0: control, 1: MPN).

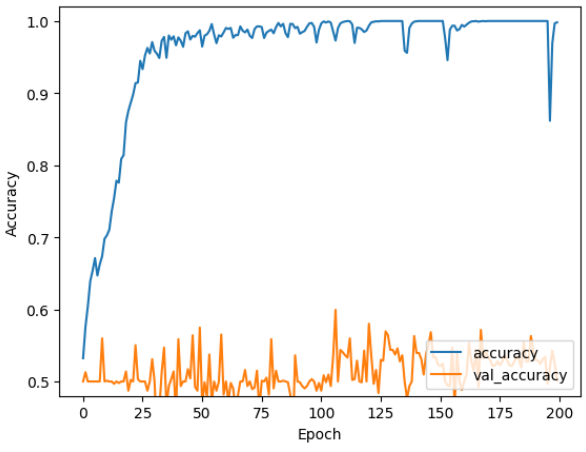


Figure D.2: Accuracy of ResNet50 model evaluated during training with lymphocyte images.

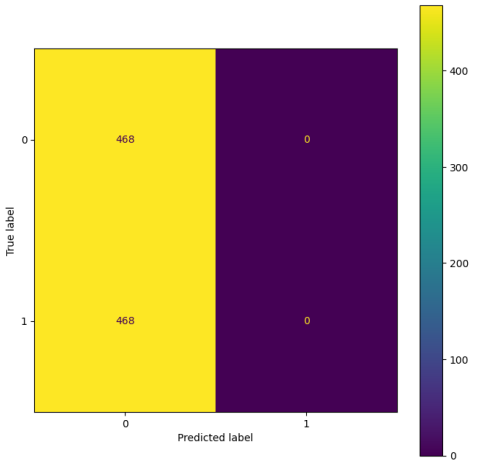


Figure D.3: Confusion matrix for predictions on test set of ResNet50 model with lymphocyte images.

E

MPN Dashboard

This dashboard is developed as an example of visual output of a blood measurement based filter. The used filter is an earlier version of the blood measurement filter as described in chapter 3 of this thesis.

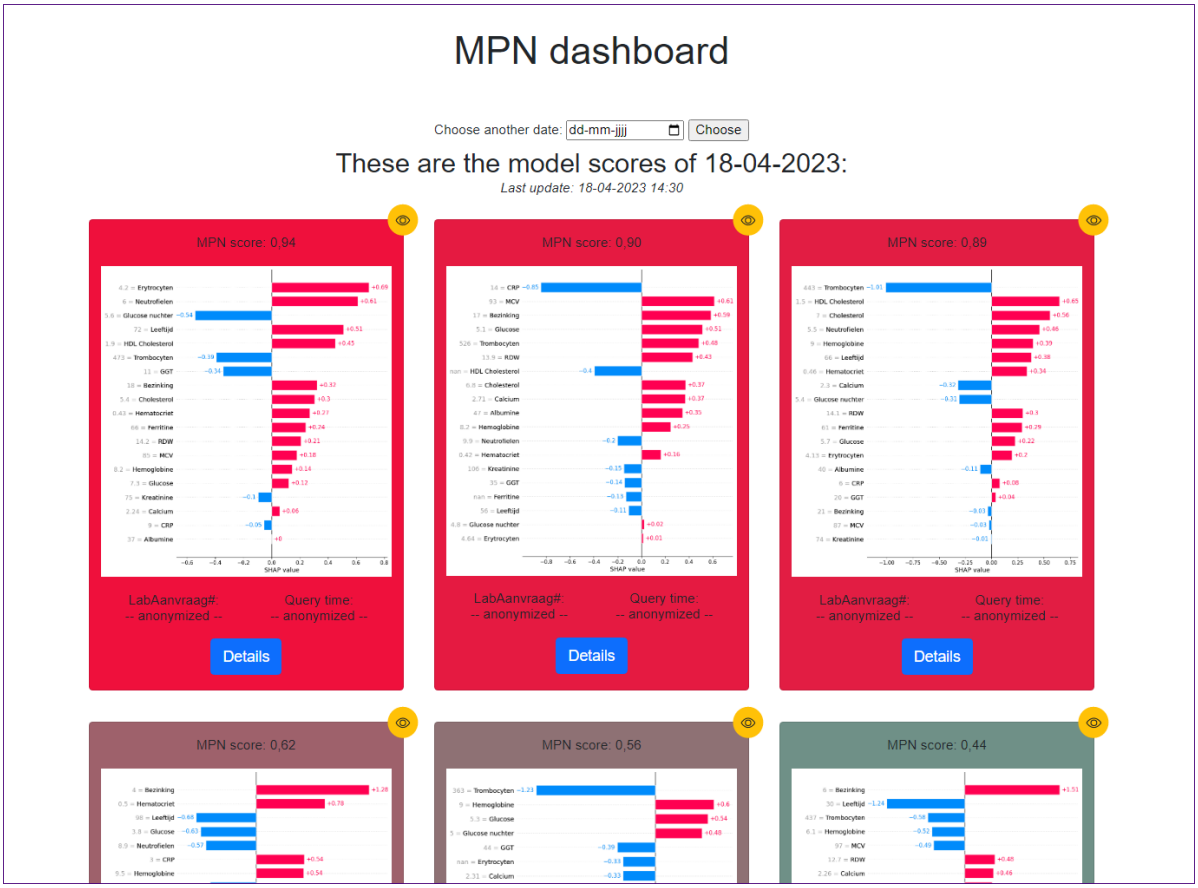


Figure E.1: MPN dashboard

F

Literature review: Machine Learning in Diagnosis and Prognosis of Myeloproliferative Disorders

A literature review was conducted prior to this master thesis project to identify the usage of machine learning in diagnosis and prognosis of Myeloproliferative Disorders (MPD; a combination term for MPN and Chronic Myeloid Leukemia (CML)).

See next pages for the review text.

Note that this review is already assessed and not part of this master thesis project.

Machine Learning in Diagnosis and Prognosis of Myeloproliferative Disorders

Paul Nijsse, intern Technical Medicine,
Medical Physics department Albert Schweitzer hospital, the Netherlands
p.nijsse@asz.nl

Abstract

Introduction: Myeloproliferative Disorders (MPD's) are a cluster of disorders related to increased proliferation of mature blood cells. The most frequent MPD's are Chronic Myeloid Leukemia (CML), Polycythemia Vera (PV), Essential Thrombocythemia (ET) and Primary Myelo-Fibrosis (MF). Diagnosis is based on blood counts, microscopy and genetics. Machine Learning (ML) might help in dealing with the increasing data obtained during the diagnostic process.
Methods: A search in the PubMed database was performed to systematically select articles relating to machine learning in diagnosis or prognosis for MPD's.
Results: 20 articles were included in this review, of which 14 describe ML in diagnostics and 6 in prognosis for MPD patients.
Conclusion: Several machine learning methods have been developed for diagnosis and prognosis in MPD patients, mainly relying on deep learning or tree based algorithms.

Keywords

Myeloproliferative disorders; Machine learning; Diagnosis; Prognosis

Introduction

Myelo-Proliferative Disorders (MPD's) are a cluster of disorders related to increased proliferation of mature blood cells[(1)]. This is caused by a hematopoietic stem cell located in the bone marrow (myelum) with increased proliferation rate, mainly because of an acquired mutation[(2)]. The World Health Organisation (WHO) defines the following disorders in the cluster of Myeloproliferative neoplasms[(3)]:

- Chronic Myeloid Leukemia (CML)
- Chronic neutrophilic leukemia
- Chronic eosinophilic leukemia
- Polycythemia Vera (PV)
- Essential Thrombocythemia (ET)
- Primary Myelo-Fibrosis (MF)

The most common diseases form this list are CML, PV, ET and MF[(1)]. In 1960, Nowell and Hungerford discovered a chromosomal aberration in CML patients, which was not present in patients with other types of leukemia[(4)]. This chromosome was named after the city where it was found: Philadelphia[(5)]. This Philadelphia (Ph) chromosome was one of the first proofs of genetic causes in cancer. The translocation between the long arms of chromosomes 9 and 22 leads to a fusion gene of BCR and ABL (BCR-ABL1), which encodes for a unregulated tyrosine kinase causing cancerous cell growth[(6)]. Knowing the molecular mechanism, targeted therapy through tyrosine kinase inhibitors is possible. This leads for most CML patients to a normal life expectancy[(7)].

Polycythemia Vera, Essential Thrombocythemia and Primary Myelofibrosis belong to the Ph-negative MPD's. To distinguish them from CML, these diseases are grouped together in this and other papers as Myelo-Proliferative Neoplasms (MPN's)[(1, 8)]. PV is characterized by an increased production of

erythrocytes (red blood cells). In ET there is a overproduction of thrombocytes (blood platelets). For MF the bone marrow becomes fibrotic, due to increased megakaryocyte proliferation. All three MPN types have in common that in a majority of the cases a JAK2 mutation is found[2, 9, 10]. Next to JAK2, also CALR and MPL mutations are found in MPN patients[2]. A small number of patients has the phenotype of an MPN, without a JAK2, CALR or MPL mutation [2, 11]. These are referred to as triple-negative MPN's.

Symptoms of MPD's are generally non-specific, such as headache, weakness, sweating, bleeding, weight loss and abdominal fullness[12]. Abdominal fullness is caused by an increased liver and/or spleen volume, due to accumulation of blood cells in these organs. Additionally, vascular problems might occur such as erythromelalgia, (transient) ischemic attacks, myocardial infarctions, pulmonary emboli and arterial thrombosis[12-14].

Diagnosis of MPD's is based on blood counting, microscopy and genetics. These processes have been automated more and more over the last decades. With the advent of flowcytometry based Complete Blood Counting (CBC) machines in the second half of the 20th century, more blood counting is performed[15, 16]. This has assumably lead to earlier diagnosis of MPD[17-20]. Microscopy of peripheral blood smears and bone marrow is also performed for blood cell counting and is additionally used for morphological characterization of blood cells and bone marrow[21-23]. Ongoing robotization of microscopy and automation of image analysis has shown to be useful in hematology[24-26]. Genetics has probably known the fastest growth in the last decades. Where Nowell and Hungerford applied manual karyotyping when they found the Philadelphia chromosome in 1960, 60 years later new techniques have been developed, with the possibility to analyze whole genomes through Next-Generation Sequencing (NGS)[4, 27, 28].

With the advent of these novel diagnostic techniques, the amount of available diagnostic data has increased. Machine Learning (ML) has been proposed as a method to automate (parts of) the diagnostic decision-making process[26, 29, 30]. Machine learning is the field of algorithms which are trained to find structures in data[26]. Based on the trained 'experience', artificially intelligent outputs are given[31] . This makes ML one of the methods used in Artificial Intelligence (AI), which is the field of science aiming to automate intelligent processes[31].

Different approaches are applied in ML such as Random Forest (grouped decision trees), Support Vector Machine, Bayesian Networks, and Nearest-Neighborhood classifiers[32]. A specific, upcoming field within ML is the application of Deep Learning (DL)[26]. This makes use of neural networks, inspired by information processing of neurons in a living brain. Generally spoken, conventional ML algorithms are (to some extend) explainable, whereas DL is often seen as a 'black box' algorithm[26]. To help users to understand what the algorithm does, explainable AI is introduced[33]. This is a set of methods which help users to either understand how the model works or which features where important for the model to come to the given output. Examples of this are SHapley Additive exPlanation (SHAP) values (a measure for features to show in which extend they contributed to the models outcome) and Class Activation Mapping (CAM; highlighting areas in an image which were important for the model to classify an image as it did)[34, 35].

The common workflow for development of ML algorithms is based on three datasets: train set, test set and validation set. The train and test set are used for the development of the model. Training data is used as input data for the algorithm to learn what it should do, the test data is used to measure the performance of the algorithm. This process can be repeated multiple times for different algorithms and parameters. When a final algorithm is chosen, it is validated on a non-seen dataset. Preferably this dataset is provided by an external party (external validation) to show the generalizability of the

algorithm. In practice however, often only internal validation is performed or even only the performance is reported for the test set.

This systematic review aims to explore the use of machine learning in diagnosis and prognosis for myeloproliferative disorders.

Methods

A search in the PubMed database was performed on the 18th of January 2023. Articles mentioning terms related to both machine learning and myeloproliferative disorders in their title or abstract were queried (see appendix for the full query). These articles were first screened on title and abstract, possible eligible articles were subsequently screened on full text. Inclusion criteria were that the article should be on one or more myeloproliferative disorders and it should describe the use of machine learning methods in the diagnostic or prognostic process of MPD's. Reviews, articles without description of used ML method and articles describing techniques which are not specifically developed for the use in MPD were excluded. Only full text available articles were included.

Results

The search query in the PubMed database resulted in 213 results. A flowchart of the article selection is shown in figure 1. 46 articles were selected after title and abstract screening. 20 articles remained after full text screening[8, 36-54]. An overview of the ML methods and their performance described in these articles is shown in table 1.

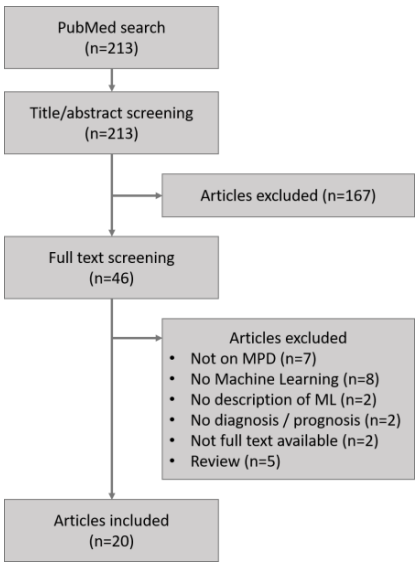


Figure 1: Flowchart of article selection

Microscopy diagnostics

Nine included articles describe Machine Learning algorithms using microscopy imaging as input[36-43, 47]. One article did not solely use microscopy, but also added CBC data and is thus covered at the 'Multi-input diagnostics' section[47]. In general, the papers regarding microscopy can be classified in two ways. For the first classification, we see that some articles use peripheral blood smears[38, 42, 43], where the others use bone marrow biopsies as imaging samples[36, 37, 39-41]. Another classification of the articles is the way they apply ML to the imaging data. In most of the cases, segmentation of individual cells (eg leukocytes, megakaryocytes and red blood cells) is performed, followed by cell-based analysis[37-39, 41, 42]. This is comparable to manual microscopy analysis and is thus integrable in the manual workflow. Others use the whole microscopy image as input for their classification algorithms[36, 40, 43].

Swolin et al. used blood smear microscopy, performing an initial segmentation and classification of white blood cells (leukocytes), followed by a classification of normal versus abnormal leukocyte counts[42]. Segmentation of cells is done using conventional techniques such as applying a threshold and watershed. Calculated image features of the cells were fed to an artificial neural network, which assigned a type to the leukocyte. Counts of cell types per sample were used to distinguish between

Table 1: Overview of machine learning methods and performance in included articles

First Author	Input Data	Output	ML Algorithm	Train/Test/Validation Method	Performance
Swolin (2003)	Blood smear microscopy imaging	Leukocyte counts, normal/abnormal classification	Artificial Neural Network	Validation on 322 blood samples	Sensitivity: 97.9% Specificity: 82.2%
Egelé (2015)	Blood smear microscopy imaging	Teardrop cell detection	Artificial Neural Network	Validation on teardrop blood samples (n=46) and normal blood samples (n=10)	Sensitivity: 100% Specificity: 45%
Ballarò (2008)	Bone marrow biopsy microscopy imaging	Megakaryocyte classification (normal/ET/MF)	Nearest neighborhood clustering	Leave-one-out validation	Sensitivity: 88.2 - 90.2%
Sirinukunwattana (2020)	Bone marrow biopsy microscopy imaging	Megakaryocyte segmentation Reactive versus ET / PV / MF classification	Deep Neural Network (Single Shot Multibox Detector); U-Net; Autoencoder Neural Network; Random forest classifier	5-fold cross validation; Population: reactive (n=43), ET/PV/MF (n=88)	AUC: 0.96-0.98
Zhang (2022)	Bone marrow biopsy microscopy imaging	Megakaryocyte segmentation Healthy versus CML classification	conditional Generative Adversarial Network (cGAN); Linear Support Vector Machine classifier	3-fold cross validation; Population: CML (n=58), control (n=31)	AUC: 0.84
Ryou (2023)	Bone marrow biopsy microscopy imaging	Fibrosis indexing Reactive versus ET / PV / MF classification	Learning To Rank algorithm (RankNet) with Convolutional Neural Network; Random Forest Classifier	Train (n=39), Test (n=18)	Reactive versus MPN: AUC: 0.62
Huang (2020)	Bone marrow biopsy microscopy imaging	Direct classification of Healthy / AML / ALL / CML	Convolutional Neural Network (DenseNet-121)	3:1 Train/test split (n=380)	Accuracy: 95.3%
Bibi (2020)	Blood smear microscopy imaging	Direct classification of Healthy / AML / ALL / CML / CLL	Convolutional Neural Networks (ResNet-34 & DenseNet-121)	Unknown	Accuracy: 99.6 - 99.9%
Hauser (2021)	Health record data	Classification of Healthy vs CML at different time points to diagnosis (range 5 to 0 years before diagnosis)	Extreme Gradient Boosting (XGBoost); Least Absolute Shrinkage and Selection Operator (LASSO)	N=1623, of with 6.2% CML positive; 80-20% training-validation randomization	XGBoost - AUC: 0.55-0.95 LASSO - AUC: 0.52-0.96
Radakovich (2021)	NGS, Health record data	Classification of MDS / 'MDS-MPN/CMML' / MPN / ICUS / CCUS	Extreme Gradient Boosting (XGBoost)	Multi-center train-test set (n=1190) with 5-fold cross validation 80-20% train-test splits Another centers data used for validation (n=1509)	Validation AUC: 0.92-0.94
Kimura (2021)	CBC, Blood smear microscopy	Classification of PV / ET / MF	Convolutional Neural Network; Extreme Gradient Boosting (XGBoost)	Training group: PV (n=23), ET (n=101), MF (n=36) Test group: PV (n=9), ET (n=53), MF (n=12)	AUC: 0.97-0.99

Table 1 (continued): Overview of machine learning methods and performance in included articles

First Author	Input data	Output	ML algorithm	train/test/validation method	Performance
Ni (2013)	Neutrophil Immunophenotyping data	Classification CML versus Normal	Support Vector Machine	Training group: CML (n=9) and healthy (n=9) Test group: CML (n=24) and non-CML (n=43)	AUC: 0.97 Sensitivity: 95.8% Specificity: 95.3%
Liu (2019)	Single Cell Mass Spectroscopy of cultured CML cells	Classification CML cells adhering to fibronectin yes/no (resp. phenotype I or II)	Random Forest (RF); Penalized Logistic Regression (PLR); Artificial Neural Network (ANN)	Samples: Phenotype I (100 cells), Phenotype II (108 cells); 80-20% training-validation randomization	RF - AUC: 0.95 PLR - AUC: 0.99 ANN - AUC: 1.00
Faisal (2019)	Bone marrow biopsy sequencing data	Classification aCML vs CMML	Logistic Regression	Leave-one-out cross-validation; Population: aCML (n=26), CMML (n=59)	Correct classification: 73% for aCML, 92% for CMML
Banjar (2017)	Health record data	Classification responders vs non-responders to imatinib treatment	Regression tree	Internal train/test/validation: responders (n=102), non-responders (n=71); External validation: responders (n=78), non-responders (n=31)	Sensitivity: 55% Specificity: 35% PPV: 68% NPV: 24%
Hoffmann (2021)	Artificially generated data	Relapse prediction after treatment for CML or AML	Mechanistic models; Generalized Linear Model; Neural Network (bidirectional Long-Short-Term-Memory)	10-fold cross validation	Accuracy 'up to 70%'
Yen (2022)	Microfluidic quantitative miRNA PCR and Colony-forming cell assay	Classification responders vs non-responders to Nilotinib treatement in CML patients	Random Forest; Naïve-Bayes	10fold crossvalidation train/test cohort (n=58)	Random Forest - AUC: 0.72 Naïve-Bayes - AUC: 0.74
Sasaki (2021)	Health record data	Hazard ratio for overall survival	Extreme Gradient Boosting (XGBoost)	3fold crossvalidation train/test cohort (n=524); Valdiation (n=126)	AUC: 0.82
Shanbehzadeh (2022)	Health record data	Prediction of 5-year survival chance for CML patients	Extreme Gradient Boosting (XGBoost); k-nearest neighborhood (KNN); pattern recognition network (PRN); probabilistic neural network (PNN); multilayer perceptron (MLP); support vector machines (SVM); J-48	10fold crossvaliation on dataset with 5-year survivers (n=740) and non-survivers (n=97)	XGBoost - AUC: 0.76 KNN - AUC: 0.69 PRN - AUC: 0.69 PNN - AUC: 0.70 MLP - AUC: 0.76 SVM - AUC: 0.83-0.86 J-48 - AUC: 0.83
Mosquera-Orgueira (2023)	Health record data	Overall and leukemia-free survival in myelofibrosis patients	Random Forest	Train/test 500 cycles of 75%-25% crossvalidation (n=1109); Validation (n=277)	Overall survival - c-index: 0.74 Leukemia free survival - c-index: 0.70

normal and abnormal leukocyte presentation, based on a set of reference values. Another article describes the same method of cell segmentation and classification, applied for red blood cells (erythrocytes)[38]. Here Egelé et al. mention the number of image features extracted from the red blood cell segmentations and fed to the artificial neural network to be 80. Examples of these features are cell size and circularity. Sensitivity for normal/abnormal classification of both the leukocyte and the erythrocyte algorithms was high, respectively 97.9% and 100%. Specificity was lower, especially for the red blood cell abnormalities with a value of 45%; specificity for the leukocyte classification was 82.2%.

Three articles described segmentation of megakaryocytes in bone marrow microscopy for diagnostic classification[37, 39, 41]. Ballarò et al. used morphological operations and wavelet transforms to segment the megakaryocytes[39]. In contrast to the conventional techniques used by Bollarò et al., Zhang et al. and Sirinukunwattana et al. used deep learning methods for megakaryocyte segmentation. Both applied a U-Net architecture. Sirinukunwattana et al. first applied a detection algorithm called Single Shot Multibox Detector which predefined which areas contained megakaryocytes; these areas were used as input for the U-Net to segment the megakaryocytes[37]. Zhang et al. fed the whole microscopy images to the U-Net for segmentation[41].

In all of the three studies on megakaryocytes, features from the segmentations were extracted and used to give a diagnostic classification. Ballarò et al. used a 3-nearest neighbor algorithm to classify the megakaryocyte as being normal, ET or MF and achieved a sensitivity of 88.2% for ET and 90.2% for MF. Sirinukunwattana et al. implemented a combination of an autoencoder neural network for feature extraction, a Principal Component Analysis (PCA) for reduction of feature dimensionality and a random forest classifier for reactive/ET/PV/MF classification. Their pipeline resulted in a classification algorithm with an Area Under the receiver operating characteristic Curve (AUC) ranging from 0.96 to 0.98. Zhang et al. achieved an AUC of 0.84 for classification in classes healthy versus CML, applying a linear support vector machine classifier on size, density and cell counts of segmented megakaryocytes.

In three selected articles, no segmentation of individual cells was performed, but the whole microscopy image was used for diagnostic classification[36, 40, 43]. An algorithm to grade the degree of fibrosis in bone marrow was developed by Ryou and colleagues[40]. The algorithm pipeline was as follows: microscopy images were split into small tiles, a U-Net was applied to detect and exclude bony structures, followed by a learning to rank method making use of a Convolutional Neural Network (CNN). This resulted in a ranking of microscopy tiles based on their degree of fibrosis. Statistics of the fibrosis throughout the sample was used to make a reactive/ET/PV/MF classification, using a random forest classifier. Although this method intuitively has the most potential for detection of MF, no statistics are given for reactive versus MF classification. Distinguishing Reactive from MPN samples (subgroups ET, PV and MF combined) resulted in a AUC of 0.62. Ryou et al. also combined their work with the earlier discussed work of Sirinukunwattana et al. by combining megakaryocyte and fibrosis features for diagnostic classification. This addition gave a boost to the AUC scores for classification, but there was only little difference between solely using megakaryocyte properties and the combination of fibrosis and megakaryocytes (AUC scores of 0.97 and 0.96 respectively for healthy vs. MPN). Also in this combined case, no analysis was performed on healthy versus MF classification.

Direct classification between healthy and leukemia subtypes (including CML) on microscopy imaging was done by the groups of Huang and Bibi[36, 43]. In both projects, a CNN was applied called DenseNet-121, where Bibi et al. also tested another CNN: ResNet-34. Huang et al. used microscopy imaging of bone marrow biopsies and achieved a accuracy of 95.3%[36]. An accuracy of 99.6-99.9% is reported by Bibi et al., who used blood smear microscopy images[43].

Molecular diagnostics

Molecular diagnostics was mentioned in 4 included articles[8, 44-46], one article also used demographics and blood counts and is thus covered in section 'Multi-input diagnostics'[8]. Ni et al. applied a Support Vector Machine method to classify the results of Single Cell Mass Spectroscopy on neutrophils as CML or normal neutrophils[45]. Training on data of 9 CML patients and 9 healthy controls, they achieved a sensitivity of 95.8% and a specificity of 95.3% in a test population with 24 CML and 43 normal cases.

Liu and colleagues differentiated between two CML phenotypes, using Single Cell Mass Spectroscopy[46]. As ground-truth, these phenotypes were defined by the capability of cells to adhere to fibronectin (a glycoprotein in extracellular matrix). Those cells adhering to fibronectin belonged to phenotype I and those not adhering belonged to phenotype II. For classification, they compared three ML algorithms, namely Random Forest, Penalized Logistic Regression and Artificial Neural Network. AUC scores of 0.95, 0.99 and 1.0 are reported for the respective algorithms.

Another subclassification of CML was performed by Faisal et al., they used Next Generation Sequencing data to distinguish atypical CML (aCML) from Chronic Myelo-Monocytic Leukemia (CMML)[44]. Genetic analysis op patients was fed to a Logistic Regression algorithm which classified the patients as either aCML or CMML. Using Leave-one-out cross-validation on a population with 26 aCML and 59 CMML cases, they achieved a correct classification of 73% and 92% in the aCML and CMML cases, respectively.

Multi-input diagnostics

Three studies used multiple modality data as input for diagnostic classification algorithms[8, 47, 48]. They all made use of extreme gradient boosting (XGBoost), which is a classification algorithm based on the Random Forest technique.

Radakovich et al. used clinical data and mutation analysis to differentiate between myelodysplastic syndromes (MDS), MPN, CML, Idiopathic Cytopenia of Undetermined Significance (ICUS) and Clonal Cytopenia of Unknown Significance (CCUS)[8]. Data of two centers was used for training and testing the XGBoost algorithm (n=1190), validation was performed on data from a third center (n=1509). All included patients were known to have one of the mentioned diseases. AUC for test and validation sets were 0.93-0.97 and 0.92-0.94 respectively. (Note that the validation results are very probably confused in the article by calling them accidentally training results. The corresponding author is asked for clarification, but without response.) A method for explainable AI is applied, providing visualization to the user regarding the impact of input variables on individual classification results.

Kimura et al. developed a CNN to segment and classify blood cells in blood smear microscopy[47]. The output of this CNN was combined with Complete Blood Count (CBC) variables and fed to a XGBoost machine. Doing this, they were able to subclassify MPN patients to PV, ET or MF with a AUC of 0.97-0.99 in a test group of 9 PV, 53 ET and 12 MF patients.

A healthy versus CML classification based on health record data was proposed by Hauser et al., where they retrospectively analyzed their results on data grouped by time to actual diagnosis ranging from 5 to 0 years before CML diagnosis[48]. Health record data was defined as 'laboratory results, patient demographics, and clinical encounter information'. Besides the use of XGBoost, also a logistic regression approach named LASSO was used. AUC for XGBoost at different times before diagnosis were 0.55-0.95, for LASSO the AUC values were 0.52-0.96. Best AUC's were found for datapoints taken at time of diagnosis, with generally decreasing AUC values when time to diagnosis became more.

Prognostic modeling

Six included articles reported the use of Machine Learning in prognostic modeling for patients with MPD diagnosis[49-54]. These articles could be divided into two groups: articles on prediction of survival and articles on prediction of response to therapy.

Three articles reported models to predict treatment response in MPD patients[49-51]. Yen and colleagues developed classification algorithms to predict response to Nilotinib treatment in CML patients, based on micro RNA expression[50]. Both a Random Forest and a Naïve-Bayes classifier were applied. In a cohort of 58 patients, using 10 fold cross-validation, AUC scores of 0.72 and 0.74 were achieved for the Random Forest and the Naïve-Bayes classifier respectively.

Response to imatinib treatment in CML patients was predicted by Benjar et al., using a regression tree[51]. They made use of demographic, clinical and laboratory data and imputed missing data through linear interpolation. Testing on an external dataset (78 responders and 31 non-responders), they found a sensitivity of 55% and a specificity of 35%. Compared with conventional scores (Sokal, Hasford and EUTOS scores), the specificity of the new model is higher (35% versus 6-19%), but a lower sensitivity (55% versus 83-92%).

Hoffmann et al. predicted relapse after treatment in AML and CML patients, using artificially generated data[49]. Applying mechanistic models (making use of biological knowledge), generalized linear models and neural networks, they report accuracies 'up to 70%', specific accuracies per model type are not given.

Three other articles described models to predict survival of MPD patients, making use of health record data[52-54]. For myelofibrosis patients, Mosquera-Orgueira et al. applied a Random Forest for prediction of overall and leukemia free survival, using clinical and laboratory data[54]. On a validation cohort of 277 patients, c-indices of 0.74 and 0.70 were found for overall survival and leukemia free survival. Overall survival is defined as the time from myelofibrosis diagnosis to death, and Leukemia free survival was defined as the time from myelofibrosis diagnosis to either leukemia diagnosis, death or last contact. Kaplan-Meier curves in the article however show a higher leukemia free survival probability compared to overall survival, which is contradicting the definition of both survivals (leukemia free survival has added requirements for survival, making the probability of survival at least equal or lower to overall survival). It might be concluded that leukemia free survival is thus in fact defined as the leukemia freeness, given that the patient is still alive. This probably inadequately described definition creates unclearness regarding methodology and meaning of results.

Survival of CML patients was predicted by Shanbehzadeh et al., based on clinical history, clinical measurements and lab results[52]. They simplified survival prediction to survival after 5 year as a binary outcome (surviving or not surviving). To select their features, they used a minimal-redundancy-maximal-relevance approach. This does not select the best features with highest individual predictiveness, but iteratively adds the feature with best added predictiveness to the feature selection until the desired amount of features is reached[55, 56]. Multiple networks were applied, namely XGBoost, k-nearest neighborhood (KNN), pattern recognition network (PRN), probabilistic neural network (PNN), multilayer perceptron (MLP), support vector machine (SVM), and J-48. Mean evaluation metrics after 10-fold cross validation were highest for SVM and J-48 with AUC values of 0.83-0.86.

Sasaki et al. predicted overall survival for several treatment options in CML patients, to recommend the best treatment option (defined as treatment with highest hazard ratios for overall survival)[53]. Survival prediction was performed using XGBoost, based on health record data. Performance on their validation cohort (n=126) resulted in an AUC value of 0.82. In order to make individual predictions

more interpretable for clinicians, visual plots showing the effect of input variables on prediction were created, using SHAP values[57].

Discussion

In this review, an overview of literature regarding machine learning for diagnosis and prognosis of MPD is given. Both for diagnosis as well as for prognosis, literature stating adequate performance using machine learning is found.

Due to the variety of use cases, methods and evaluation measures, comparison of model performance is irrelevant. However, assuming that the best models are published, analysis of used models for different use cases is relevant. It gives an indication of the best models for given situations. In 8 out of 9 articles using microscopy images as input for diagnostic classification, deep learning is used[36-38, 40-43, 47]. The main purpose of these DL algorithms is image segmentation. In some cases also classification is performed using the DL algorithm. In only 3 of the remaining 11 papers DL was also applied[46, 49, 52]. These three papers also provide performance for non-DL algorithms. Only Liu et al. showed a superior performance of deep learning compared to other ML methods[46]. The most commonly applied group of ML methods is that of the tree based algorithms (e.g. random forest, XGBoost), which are used for classification tasks in 11 of the 20 included papers[8, 37, 40, 46-48, 50-54]. In a comparison between different methods for survival prediction, Shanbehzadeh et al. show that Support Vector Machines (SVM) can also be of good value[52]. However, SVM is only applied in two other articles[41, 45]. The reason for low adoption of SVM methods might lie in the fact that SVM's can only deal with numerical data, whereas tree based models can also deal with categorical data.

Regarding the used algorithms, it might thus be concluded that DL methods are currently the best models to deal with images in classification tasks. Tree based models, with XGBoost as the prince, can deal best with classification tasks based on non-image data. Given the low number of cases included in the described studies (ten to hundreds of cases) and the need for 'big data' in DL, it does not surprise that DL is mainly adopted for image segmentation and not so commonly for classification tasks in the described articles.

In the included articles, there is a wide range of methodological quality. External validation was only applied in two studies[8, 51]. Most articles only report their cross-validation results, without validation on completely unseen data[37, 39, 41, 44, 49, 50, 52, 53]. This makes that most of the described performances should be considered as overestimations of the actual clinical performance.

Also patient populations were chosen in different ways. Two articles on diagnostic modelling only include proven patients, without a non-MPD control group[44, 47]. Most articles on diagnostic modeling for CML do involve a healthy control group. For MPN, only control groups are included when bone marrow microscopy was used for diagnosis[37, 39, 40]. A bone marrow biopsy, however, is only performed when there is a high a priori chance to find an MPD[12]. This makes that these diagnostic models can be used in the later diagnostic process of MPD, but not for screening-like applications. Models using health record data or CBC results for automatically diagnosing MPN versus healthy have not been found in this search.

Explainable AI was integrated in the workflow of Kimura et al., Radakovich et al. and Sasaki et al.; they all used SHAP values[8, 47, 53]. These values help to understand which variables influenced the decision made by the model. Adding explainable AI methods to the standard ML workflow would be of great added value, especially to let clinicians adopt ML methods in a responsible way in their clinical practice.

However, the wide variety of methods for the development and testing of ML applications described in the included papers show that there is no generally adopted workflow for ML development.

Concludingly, several Machine Learning methods have been developed for diagnosis and prognosis in MPD patients, mainly relying on DL or tree based algorithms.

Lessons Learned for Project

- No laboratory measurements based healthy vs. MPN algorithm is presented in current literature.
- Based on the results of models performing MPN subclassification and models classifying CML vs MPN, it is reasonable that a healthy vs. MPN algorithm based on laboratory measurements if feasible.
- The XGBoost algorithm currently seems to be the best method for laboratory measurements based classification.
- Deep Learning methods (e.g. CNN's) currently seem to be the best method for microscopy imaging based classification.
- Application of explainable AI, such as SHAP values might be of additive value for proper model interpretation for developers and users.

References

1. Campbell PJ, Green AR. The myeloproliferative disorders. N Engl J Med. 2006;355(23):2452-66.
2. Spivak JL. Myeloproliferative Neoplasms. N Engl J Med. 2017;376(22):2168-81.
3. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, et al. WHO classification of tumours of haematopoietic and lymphoid tissues: International agency for research on cancer Lyon; 2008.
4. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst. 1960;25:85-109.
5. Haider MZ, Anwer F. Genetics, Philadelphia Chromosome. StatPearls. Treasure Island (FL): StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.; 2022.
6. Shtivelman E, Lifshitz B, Gale RP, Canaani E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature. 1985;315(6020):550-4.
7. Hehlmann R. Chronic Myeloid Leukemia in 2020. Hemasphere. 2020;4(5):e468.
8. Radakovich N, Meggendorfer M, Malcovati L, Hilton CB, Sekeres MA, Shreve J, et al. A geno-clinical decision model for the diagnosis of myelodysplastic syndromes. Blood Adv. 2021;5(21):4361-9.
9. de Lacerda JF, Oliveira SN, Ferro JM. Chronic myeloproliferative diseases. Handb Clin Neurol. 2014;120:1073-81.
10. James C, Ugo V, Le Couédic JP, Staerk J, Delhommeau F, Lacout C, et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature. 2005;434(7037):1144-8.
11. Xie J, Chen X, Gao F, Hou R, Tian T, Zhang Y, et al. Two activating mutations of MPL in triple-negative myeloproliferative neoplasms. Cancer Med. 2019;8(11):5254-63.
12. Meier B, Burton JH. Myeloproliferative Disorders. Hematol Oncol Clin North Am. 2017;31(6):1029-44.
13. Randi ML, Fabris F, Cella G, Rossi C, Girolami A. Cerebral vascular accidents in young patients with essential thrombocythemia: relation with other known cardiovascular risk factors. Angiology. 1998;49(6):477-81.
14. Hachulla E, Rose C, Trillot N, Caulier-Leleu MT, Pasturel-Michon U. [What vascular events suggest a myeloproliferative disorder?]. J Mal Vasc. 2000;25(5):382-7.
15. Don M. The Coulter Principle: Foundation of an Industry. JALA: Journal of the Association for Laboratory Automation. 2003;8(6):72-81.
16. Burrows RF. Platelet disorders in pregnancy. Curr Opin Obstet Gynecol. 2001;13(2):115-9.
17. Johansson P, Kutti J, Andréasson B, Safai-Kutti S, Vilén L, Wedel H, et al. Trends in the incidence of chronic Philadelphia chromosome negative (Ph-) myeloproliferative disorders in the city of Göteborg, Sweden, during 1983-99. J Intern Med. 2004;256(2):161-5.
18. Krause JR, Costello RT, Krause J, Penchansky L. Use of the Technicon H-1 in the characterization of leukemias. Arch Pathol Lab Med. 1988;112(9):889-94.
19. Watanabe K, Takeuchi K, Kawai Y, Ikeda Y, Kubota F, Nakamoto H. Automated measurement of reticulated platelets in estimating thrombopoiesis. Eur J Haematol. 1995;54(3):163-71.
20. Watanabe K, Kawai Y, Takeuchi K. [Reticulated platelets--automated measurement and clinical utility]. Rinsho Ketsueki. 1995;36(4):267-72.
21. Barbui T, Thiele J, Gisslinger H, Kvasnicka HM, Vannucchi AM, Guglielmelli P, et al. The 2016 WHO classification and diagnostic criteria for myeloproliferative neoplasms: document summary and in-depth discussion. Blood Cancer J. 2018;8(2):15.
22. Sangiorgio VFI, Orazi A, Arber DA. Myelodysplastic/myeloproliferative neoplasms: are morphology and immunophenotyping still relevant? Best Pract Res Clin Haematol. 2020;33(2):101139.
23. Tefferi A, Pardanani A. Myeloproliferative Neoplasms: A Contemporary Review. JAMA Oncol. 2015;1(1):97-105.

24. Kratz A, Lee SH, Zini G, Riedl JA, Hur M, Machin S. Digital morphology analyzers in hematology: ICSH review and recommendations. Int J Lab Hematol. 2019;41(4):437-47.
25. Acharya V, Kumar P. Identification and red blood cell automated counting from blood smear images using computer-aided system. Med Biol Eng Comput. 2018;56(3):483-9.
26. Obstfeld AE. Hematology and Machine Learning. J Appl Lab Med. 2023;8(1):129-44.
27. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.
28. Fournier PE, Drancourt M, Colson P, Rolain JM, La Scola B, Raoult D. Modern clinical microbiology: new challenges and solutions. Nat Rev Microbiol. 2013;11(8):574-85.
29. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019;20(5):e262-e73.
30. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.
31. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28(2):73-81.
32. Santos K, Dias JP, Amado C. A literature review of machine learning algorithms for crash injury severity prediction. J Safety Res. 2022;80:254-69.
33. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal. 2022;79:102470.
34. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
35. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, editors. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 27-30 June 2016.
36. Huang F, Guang P, Li F, Liu X, Zhang W, Huang W. AML, ALL, and CML classification and diagnosis based on bone marrow cell morphology combined with convolutional neural network: A STARD compliant diagnosis research. Medicine (Baltimore). 2020;99(45):e23154.
37. Sirinukunwattana K, Aberdeen A, Theissen H, Sousos N, Psaila B, Mead AJ, et al. Artificial intelligence-based morphological fingerprinting of megakaryocytes: a new tool for assessing disease in MPN patients. Blood Adv. 2020;4(14):3284-94.
38. Egelé A, van Gelder W, Riedl J. Automated detection and classification of teardrop cells by a novel RBC module using digital imaging/microscopy. Int J Lab Hematol. 2015;37(6):e153-6.
39. Ballarò B, Florena AM, Franco V, Tegolo D, Tripodo C, Valenti C. An automated image analysis methodology for classifying megakaryocytes in chronic myeloproliferative disorders. Med Image Anal. 2008;12(6):703-12.
40. Ryou H, Sirinukunwattana K, Aberdeen A, Grindstaff G, Stolz BJ, Byrne H, et al. Continuous Indexing of Fibrosis (CIF): improving the assessment and classification of MPN patients. Leukemia. 2023;37(2):348-58.
41. Zhang Z, Huang X, Yan Q, Lin Y, Liu E, Mi Y, et al. The Diagnosis of Chronic Myeloid Leukemia with Deep Adversarial Learning. Am J Pathol. 2022;192(7):1083-91.
42. Swolin B, Simonsson P, Backman S, Löfqvist I, Bredin I, Johnsson M. Differential counting of blood leukocytes using automated microscopy and a decision support system based on artificial neural networks--evaluation of DiffMaster Octavia. Clin Lab Haematol. 2003;25(3):139-47.
43. Bibi N, Sikandar M, Ud Din I, Almogren A, Ali S. IoMT-Based Automated Detection and Classification of Leukemia Using Deep Learning. J Healthc Eng. 2020;2020:6648574.
44. Faisal M, Stark H, Büsche G, Schlue J, Teiken K, Kreipe HH, et al. Comprehensive mutation profiling and mRNA expression analysis in atypical chronic myeloid leukemia in comparison with chronic myelomonocytic leukemia. Cancer Med. 2019;8(2):742-50.
45. Ni W, Tong X, Qian W, Jin J, Zhao H. Discrimination of malignant neutrophils of chronic myelogenous leukemia from normal neutrophils by support vector machine. Comput Biol Med. 2013;43(9):1192-5.
46. Liu R, Zhang G, Yang Z. Towards rapid prediction of drug-resistant cancer cell phenotypes: single cell mass spectrometry combined with machine learning. Chem Commun (Camb). 2019;55(5):616-9.

47. Kimura K, Ai T, Horiuchi Y, Matsuzaki A, Nishibe K, Marutani S, et al. Automated diagnostic support system with deep learning algorithms for distinction of Philadelphia chromosome-negative myeloproliferative neoplasms using peripheral blood specimen. Sci Rep. 2021;11(1):3367.
48. Hauser RG, Esserman D, Beste LA, Ong SY, Colomb DG, Bhargava A, et al. A Machine Learning Model to Successfully Predict Future Diagnosis of Chronic Myelogenous Leukemia With Retrospective Electronic Health Records Data. Am J Clin Pathol. 2021;156(6):1142-8.
49. Hoffmann H, Baldow C, Zerjatke T, Gottschalk A, Wagner S, Karg E, et al. How to predict relapse in leukemia using time series data: A comparative in silico study. PLoS One. 2021;16(11):e0256585.
50. Yen R, Grasedieck S, Wu A, Lin H, Su J, Rothe K, et al. Identification of key microRNAs as predictive biomarkers of Nilotinib response in chronic myeloid leukemia: a sub-analysis of the ENESTxtnd clinical trial. Leukemia. 2022;36(10):2443-52.
51. Banjar H, Ranasinghe D, Brown F, Adelson D, Kroger T, Leclercq T, et al. Modelling Predictors of Molecular Response to Frontline Imatinib for Patients with Chronic Myeloid Leukaemia. PLoS One. 2017;12(1):e0168947.
52. Shanbehzadeh M, Afrash MR, Mirani N, Kazemi-Arpanahi H. Comparing machine learning algorithms to predict 5-year survival in patients with chronic myeloid leukemia. BMC Med Inform Decis Mak. 2022;22(1):236.
53. Sasaki K, Jabbour EJ, Ravandi F, Konopleva M, Borthakur G, Wierda WG, et al. The LEukemia Artificial Intelligence Program (LEAP) in chronic myeloid leukemia in chronic phase: A model to improve patient outcomes. Am J Hematol. 2021;96(2):241-50.
54. Mosquera-Orgueira A, Pérez-Encinas M, Hernández-Sánchez A, González-Martínez T, Arellano-Rodrigo E, Martínez-Elicegui J, et al. Machine Learning Improves Risk Stratification in Myelofibrosis: An Analysis of the Spanish Registry of Myelofibrosis. Hemasphere. 2023;7(1):e818.
55. Huang X, Chen X, Chen X, Wang W. Screening of Serum miRNAs as Diagnostic Biomarkers for Lung Cancer Using the Minimal-Redundancy-Maximal-Relevance Algorithm and Random Forest Classifier Based on a Public Database. Public Health Genomics. 2022:1-9.
56. Jiao YS, Du PF. Prediction of Golgi-resident protein types using general form of Chou's pseudo-amino acid compositions: Approaches with minimal redundancy maximal relevance feature selection. J Theor Biol. 2016;402:38-44.
57. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:180203888. 2018.

Appendix - PubMed search query

("artificial intelligence"[MeSH Terms] OR ("artificial"[All Fields] AND "intelligence"[All Fields]) OR "artificial intelligence"[All Fields] OR ("machine learning"[MeSH Terms] OR ("machine"[All Fields] AND "learning"[All Fields]) OR "machine learning"[All Fields]) OR ("automate"[All Fields] OR "automated"[All Fields] OR "automates"[All Fields] OR "automating"[All Fields] OR "automation"[MeSH Terms] OR "automation"[All Fields] OR "automations"[All Fields] OR "automation s"[All Fields]))
AND
("myeloproliferative disorders"[MeSH Terms] OR ("myeloproliferative"[All Fields] AND "disorders"[All Fields]) OR "myeloproliferative disorders"[All Fields] OR ("polycythaemia vera"[All Fields] OR "polycythemia vera"[MeSH Terms] OR ("polycythemia"[All Fields] AND "vera"[All Fields]) OR "polycythemia vera"[All Fields]) OR ("primary myelofibrosis"[MeSH Terms] OR ("primary"[All Fields] AND "myelofibrosis"[All Fields]) OR "primary myelofibrosis"[All Fields]) OR ("thrombocythemia, essential"[MeSH Terms] OR ("thrombocythemia"[All Fields] AND "essential"[All Fields]) OR "essential thrombocythemia"[All Fields] OR ("thrombocythemia"[All Fields] AND "essential"[All Fields]) OR "thrombocythemia essential"[All Fields]) OR ("chronic myelogenous leukaemia"[All Fields] OR "leukemia, myelogenous, chronic, bcr abl positive"[MeSH Terms] OR ("leukemia"[All Fields] AND "myelogenous"[All Fields] AND "chronic"[All Fields] AND "bcr abl"[All Fields] AND "positive"[All Fields]) OR "bcr-abl positive chronic myelogenous leukemia"[All Fields] OR ("chronic"[All Fields] AND "myelogenous"[All Fields] AND "leukemia"[All Fields]) OR "chronic myelogenous leukemia"[All Fields]))