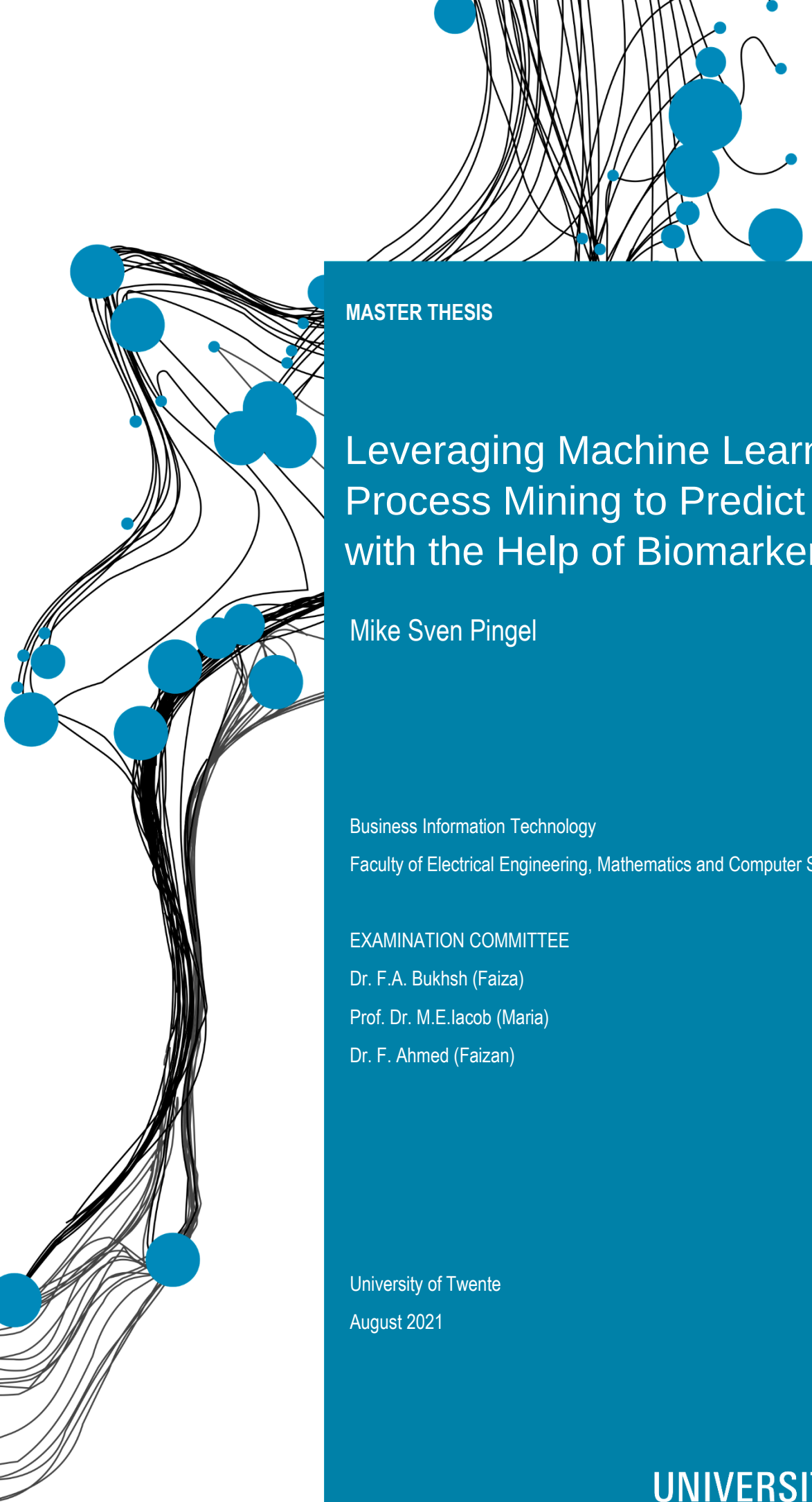

MASTER THESIS

Leveraging Machine Learning and Process Mining to Predict Anaemia with the Help of Biomarker Data

Mike Sven Pingel

Business Information Technology

Faculty of Electrical Engineering, Mathematics and Computer Science

EXAMINATION COMMITTEE

Dr. F.A. Bukhsh (Faiza)

Prof. Dr. M.E.Iacob (Maria)

Dr. F. Ahmed (Faizan)

University of Twente

August 2021

UNIVERSITY OF TWENTE.

ACKNOWLEDGEMENTS

After half a year of thesis work, both enjoyable and challenging, my life as a student comes to an end. I am thankful for the opportunities that Utwente, my teachers, as well as my fellow students and friends have enabled me. It was a time of growth where I could develop myself and satisfy my intellectual curiosity. After completing my bachelor's in international business administration, also at Utwente, I continued my master in business information technology. Even though the change to more technical subjects proved to be challenging, I am glad about the choices I made.

While the Corona situation this and last year has not been easy, I am thankful that I could nonetheless receive a student experience, even if it was not on campus. I am grateful to all of the people that made this technically possible and that enabled the continuation of the study even during this pandemic. I would also like to thank all of my teachers who shared their knowledge to the best of their ability and helped me through projects, exam study, and also through administrative challenges. In particular, I want to thank my supervisors and all of the committee and advisory committee members that made this thesis possible. With the help from these individuals, Faiza, Faizan, Hans, and Maria, this thesis could become what it is today. Their advice and recommendations contributed a ton to the progress of this thesis, and I am thankful for the study opportunity that they have given me.

I would also like to thank all of my friends and family members that supported me during the time of my study. Especially without the help of my parents, studying at Utwente would have been a lot harder. I want to thank all of my friends and study mates I worked, people who helped me with advice, study time together, and with projects. I am glad that you all supported my ideas and helped me through the past 6 years, and I hope that the next few years will be just as fun and successful as the last few were. Finally, I want to thank my two pets who have given me comfort and companionship during the lockdown and beyond.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

With almost a third of the world population affected by anaemia, accurate diagnostic tools need to be developed that offer transparency and understandability. Diagnostic pathways are one such tool and are often used in the context of anaemia diagnostics. This thesis has the aim of assessing the anaemia standard from the Dutch Nederlands Huisartsen Genootschap (NHG). By using a combination of process mining and machine learning techniques, biomarkers are identified that have a great impact on the prediction performance of machine learning classifiers. Different pre-processing techniques are compared and their effect on the machine learning performance explained. This involves the use of various missing value imputation techniques, as well as class imbalance handling and feature selection. By doing this, biomarker sets can be defined that have the potential of complementing the already existing standard. For the assessment performance, a methodology process is developed that can also be applied to other types of diagnostic models.

1 INTRODUCTION

Anaemia is a global problem, with an estimated 32.9% of the worldwide population suffering from the disease [162]. Older estimates from the WHO (2008) estimate the anaemia prevalence in the population to be around 24.8%. While men are the least affected by anaemia, with a prevalence of around 12.7%, preschool-age children (47.4%) and pregnant women (41.8%) are affected the most [50]. Furthermore, a difference can be observed between geographical regions, with African countries most affected at around 57.1% prevalence in pregnant women and 47.5% in non-pregnant women [122]. With the significant burden that anaemia puts on healthcare systems worldwide, accurate diagnostic procedures are a necessity. To improve these procedures, machine learning techniques have been used to increase the accuracy and reliability of anaemia diagnostics.

With the advance of machine learning, application areas for its use have expanded. Ever-increasing amounts of data have opened up new technical possibilities for data analysis in a variety of domains. Machine learning can appear in many forms, such as binary classification, multiclass classification, structured estimation, regression, or novelty detection [8]. Applications in anaemia diagnostics include the prediction and classification of anaemia using various machine learning algorithms [102] [67]. This thesis expands on current work in anaemia diagnostics by making predictions using various machine learning classifiers, comparing different pre-processing techniques in relation to those classifiers, and incorporating process mining to gain further insights into current diagnostic procedures.

As an integral part of data science, process mining is used to analyse operational processes based on event logs to gain additional insights beyond traditional data mining techniques. Event logs, consisting of activities, case IDs, and timestamps [177], are essential for this purpose. By using process discovery, conformance checking, and model enhancement, process mining can help analyse processes, detect deviations from the expected standard, and correct these deviations [181].

While machine learning is usually focused on data-oriented techniques, like simple classification, clustering, regression, or rule-learning problems, only process mining can bridge the gap between event data and end-to-end processes [178]. Even though both can create deep insight into healthcare data, current research rarely combines these two approaches. Accordingly, this thesis includes a literature review to identify how machine learning and process mining can work in conjunction (Section 3). Furthermore, research gaps are identified to determine possible research opportunities. Following this, machine learning and process mining techniques are used to determine the best biomarkers for anaemia predictions and compare these results with the current state of diagnostics.

1.1 RESEARCH PROBLEM

The main motivations for this thesis are to close the research gap that exists for the combined application of machine learning and process mining and to address the issues that come from using diagnostic standards in healthcare. In particular, the focus is placed on the Dutch Nederlands Huisartsen Genootschap (NHG)-standard on anaemia diagnostics [118]. Research has shown that diagnostic pathways, such as the one in the NHG standard, can lead to oversimplification of reality [45]. While diagnostic pathways are a helpful tool that can help in establishing transparent routine processes, their use also comes with downsides. For one, the oversimplified nature of these pathways can decrease flexibility and individuality in diagnostics, making it challenging to apply on patients with complex diseases or patients with unclear symptoms. Furthermore, while they are developed in collaboration with clinical physicians and laboratories, there is often a lack of suitable IT tool support for these pathways.

Machine learning and process mining techniques can help evaluate such standards based on their choice of biomarkers used for diagnostics. Using machine learning, various biomarkers' predictive power can be compared to increase accuracy and improve the diagnostic process. While there is an ever-increasing amount of data created in the healthcare domain [109], taking into account the full

complexity of that data can help in complementing the knowledge and experience of clinical physicians.

1.2 RESEARCH QUESTIONS

Based on the research problem, the following research questions are defined:

- **RQ1**: How can machine learning techniques/methods/approaches be used to predict anaemia?
    - **SUB 1**: What feature selection techniques should be chosen to ensure high prediction performance?
    - **SUB 2**: What is the performance difference of different machine learning techniques/methods/approaches when predicting anaemia?

- **RQ2**: How can sparse healthcare data be pre-processed by machine learning techniques/methods/approaches to achieve improved performance in machine learning models?
    - **SUB 1**: What machine learning techniques/methods/approaches are most suitable for handling the data's missing values?
    - **SUB 2**: What pre-processing techniques are most suitable for handling class imbalance within the data?

- **RQ3**: What biomarkers should be used for determining and predicting anaemia in diagnostics?
    - **SUB 1**: In what ways does the actual process differ from the process mandated by the Anaemia (M76) standard?
    - **SUB 2**: Should the Anaemia (M76) standard from the Nederlands Huisartsen Genootschap include additional biomarkers that are not part of its guidelines yet?
    - **SUB 3**: Are any biomarkers used for diagnostics in the (M76) standard redundant when considering the predictive power of machine learning techniques/methods/approaches?

Research question 1 is about predicting anaemia through the use of machine learning models. To do this, different feature selection techniques are used to compare feature sets and their effect on model performance. This comparison is handled in **sub-question 1**. **Sub-question 2** is about the performance difference between different machine learning models compared on various metrics. In total, three machine learning models are compared, KNN, Naïve Bayes, and Random Forest.

Research question 2 focuses on the dataset used and the pre-processing steps needed to prepare it for use in machine learning models. Due to the sparse nature of many datasets in healthcare, including this project's data, methods need to be chosen to address this issue. Research has shown that sparse data can lead to machine learning performance loss, resulting in less accurate predictions [113]. Since the dataset used in this thesis contains many missing values, different techniques for imputation are used and compared, which is the focus of **sub-question 1**. **Sub-question 2**, on the other hand, deals with the class imbalance issues present in the data. Just as with the first sub-question, different techniques are being compared to answer this question. The need for such pre-processing techniques becomes apparent when looking at recent studies [133], highlighting the performance boost such approaches produce.

As the last research question, **research question 3** focuses on evaluating the results collected in research question 2. First, a comparison is made between the actual process and the ideal biomarkers found from answering research question one (**Sub-question 1**). Secondly, an assessment is made on whether different biomarkers should be used to diagnose anaemia (**Sub-question 2**). Following this, an evaluation is done on the possible redundancy of different biomarkers present in the diagnostic standard (**Sub-question 3**). Both steps are executed with the help of an expert that assists by giving medically sound advice.

1.3 RESEARCH APPROACH & THESIS STRUCTURE

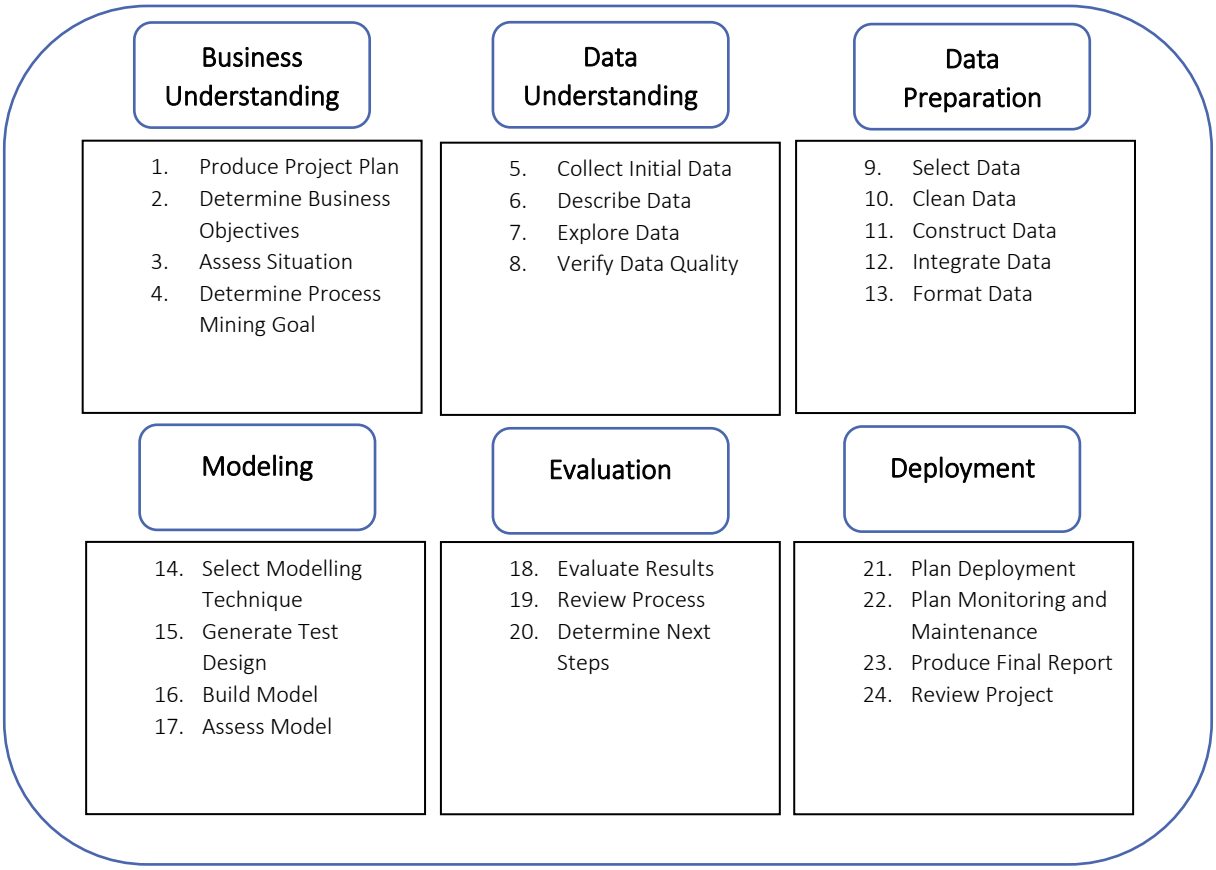

FIGURE 1: CRISP-DM TASKS FOR THIS RESEARCH THESIS

The CRISP-DM (Cross Industry Standard Process for Data Mining) research approach is used to answer the research questions in section 1.2. CRISP-DM is a generic model that defines the necessary steps for the development of a specialized process model. The steps defined in the model are business understanding, data understanding, data preparation, modelling, evaluation, and deployment. These steps are executed by using an anaemia dataset containing various biomarkers for possibly anaemic patients. How the research questions are applied to each step in this approach can be read in section 4.1. While the previously mentioned steps define the generic steps that should be taken during the research project, specific tasks must be defined that need to be executed for each step. An overview of all of the tasks relevant for each process step can be seen in Figure 1.

This thesis starts with the creation of a project plan, defining the process used for the remainder of this thesis. This is covered in section 1.4. Section 2 contains essential background information on anaemia and different biomarkers commonly used to diagnose the condition. By understanding the medical background of this thesis, business objectives can be defined, and the situation can be assessed, which contributes to an improved business understanding. As mentioned before, section 3 contains a literature review that includes a study on the combined use of process mining and machine learning. While this section does not relate to any step in the CRISP-DM standard, it is nonetheless crucial for a complete understanding of the tools used in the subsequent thesis sections. Using this knowledge, section 4 contains the methodology section, explaining the processes and techniques used to answer the defined research questions.

Furthermore, section 4 contains process mining goals, describing the objectives for the process mining techniques used. Because section 4 also gives an overview of the data used, it contributes to both the business understanding and the data understanding part of the CRISP-DM model. The findings of the research are presented in three different sections, one covering each research question. Section 5 contains the information relating to research question one, while section 6 and section 7 handle research questions two and three accordingly. These three sections each include the tasks related to the

data preparation, modelling, and evaluation steps within the CRISP-DM model. The discussion section (Section 8) examines the impact that these findings have on the current state of anaemia diagnostics and gives recommendations on possible changes in the current diagnostic standard. By doing this, tasks related to the deployment step in the CRISP-DM model are answered. Finally, the conclusion section (Section 9) finalizes this thesis with a summary of these findings and some concluding remarks, including future research opportunities and limitations. Acronyms, as well as the appendix and the references, can be found at the end of this thesis.

1.4 PROJECT PLAN

As shown in Figure 2, which represents the project plan for this thesis, five distinct steps are of great importance. The first three steps involve the machine learning pipeline defined in section 4.2, covering the first two CRISP-DM cycles. These first steps aim to define alternative biomarker sets to the biomarkers defined in the diagnostic standard. This is done by comparing the prediction performance when the alternative biomarkers are used to the performance achieved when the standard biomarkers are used. From there, conclusions can be made on possible changes in the diagnostic standard by including new biomarkers or exchanging some already existing ones. The fourth step in Figure 2 covers the process mining, which checks, as described before, the conformance to the diagnostic standard in general. From there, biomarker use of the diagnostic standard biomarkers can be compared to the use of the alternative biomarker sets defined in the machine learning procedure. By knowing the usage gap between the two, recommendations can be made on future use in a clinical setting. All in all, recommendations resulting from the process mining procedure are compiled in the last step in Figure 2, covering the "Deployment" stage within the CRISP-DM model.

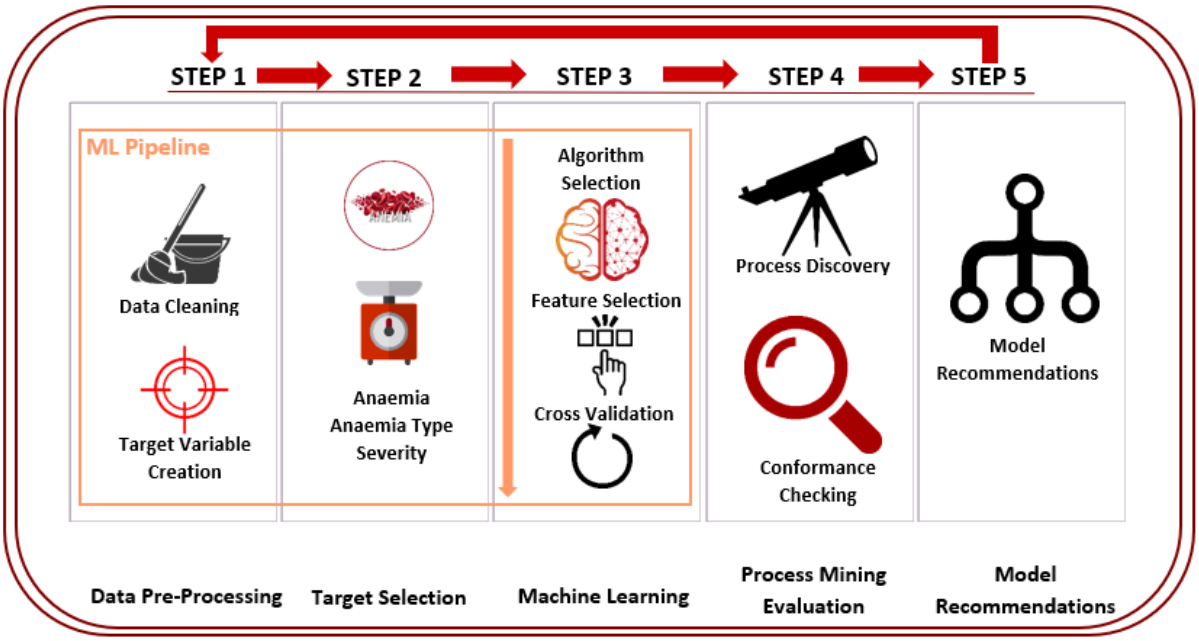

FIGURE 2: PROJECT PLAN

2 ANAEMIA BACKGROUND

This section gives an overview of the business background and anaemia biomarkers commonly used to diagnose the condition. It starts with a definition of anaemia according to WHO standards, together with a description of the NHG standard for anaemia diagnostics in the Netherlands (Section 2.1). Based on this information, business objectives can be defined that rely on various success criteria (Section 2.1.3). Continuing from there, section 2.2 investigates the production of red blood cells, which is necessary to understand the development of anaemia. Section 2.3, on the other hand, covers a variety of biomarkers used to diagnose anaemia. A deeper understanding of these biomarkers helps increase insight in sections 5 and 6, covering the selection of some biomarkers for use in machine learning models. Finally, section 2.4 assesses the current situation in anaemia diagnostics. All in all, this section covers the first two tasks of the business understanding step in the CRISP-DM model. These tasks were defined as:

1. Determine Business Objectives
2. Assess Situation

2.1 ANAEMIA DEFINITION & DIAGNOSTIC STANDARD

2.1.1 WHO ANAEMIA DEFINITION

Anaemia is a condition that affects people on a global scale. While governments and healthcare organizations around the world have put efforts into defining what anaemia is, this thesis focuses on the definition put forth by the World Health Organization (WHO) in their guideline "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity" [54]. According to the WHO, anaemia is characterized by a low oxygen-carrying capacity caused by a low number of red blood cells. Even though there are many types of anaemia, all of them have a low haemoglobin (Hb) count in common. Haemoglobin is the protein within red blood cells that carry oxygen throughout the body and are present in almost all vertebrates [54]. Because they also contain iron, a low haemoglobin count is often connected to iron deficiency.

Additionally, not all age groups have the same ranges for haemoglobin concentrations. Children, for instance, are known to have a lower Hb level per litre of blood than adults. Differences also need to be made between the sexes, as well as between pregnant and non-pregnant women. Where the cut-offs are set can be seen in Table 1.

TABLE 1: SEVERITY OF ANAEMIA MEASURED IN HAEMOGLOBIN IN MMOL/L [54]

Population	Mild	Moderate	Severe
Children < 5 years	6.21-6.77	4.35-6.20	<4.35
Children 5 – 11 years	6.83-7.08	4.97-6.83	<4.97
Children 11-14 years	6.83-7.39	4.97-6.83	<4.97
Non pregnant woman > 15	6.83-7.39	4.97-6.83	<4.97
Pregnant women	6.21-6.77	4.35-6.21	<4.35
Men > 15	6.83-8.01	4.97-6.83	<4.97

The reasons for this are missing data representing these variables and the altitude of the Netherlands, which is mostly at sea level. Even though levels need to be adjusted for smoking individuals and patients living in high altitudes, these differences are not considered in this thesis. Of course, while haemoglobin alone is enough to determine anaemia, more biomarkers need to be measured to determine the causes and type of anaemia. Which biomarkers need to be measured is often determined by experience and adherence to standard procedures captured in diagnostic standards like the NHG Anaemia Standard.

2.1.2 NHG ANAEMIA STANDARD

The NHG Standard for anaemia diagnostics from the "Nederlands Huisartsen Genootschap" is a guideline for the diagnostics and management of patients that have been diagnosed with a low

haemoglobin level [118]. Even though its drug policy is limited to iron deficiency anaemia and vitamin B12/folic acid deficiency anaemia, the guidelines explain how other types of anaemia can be diagnosed. The general workflow for making a diagnosis can be found in Figure 3.

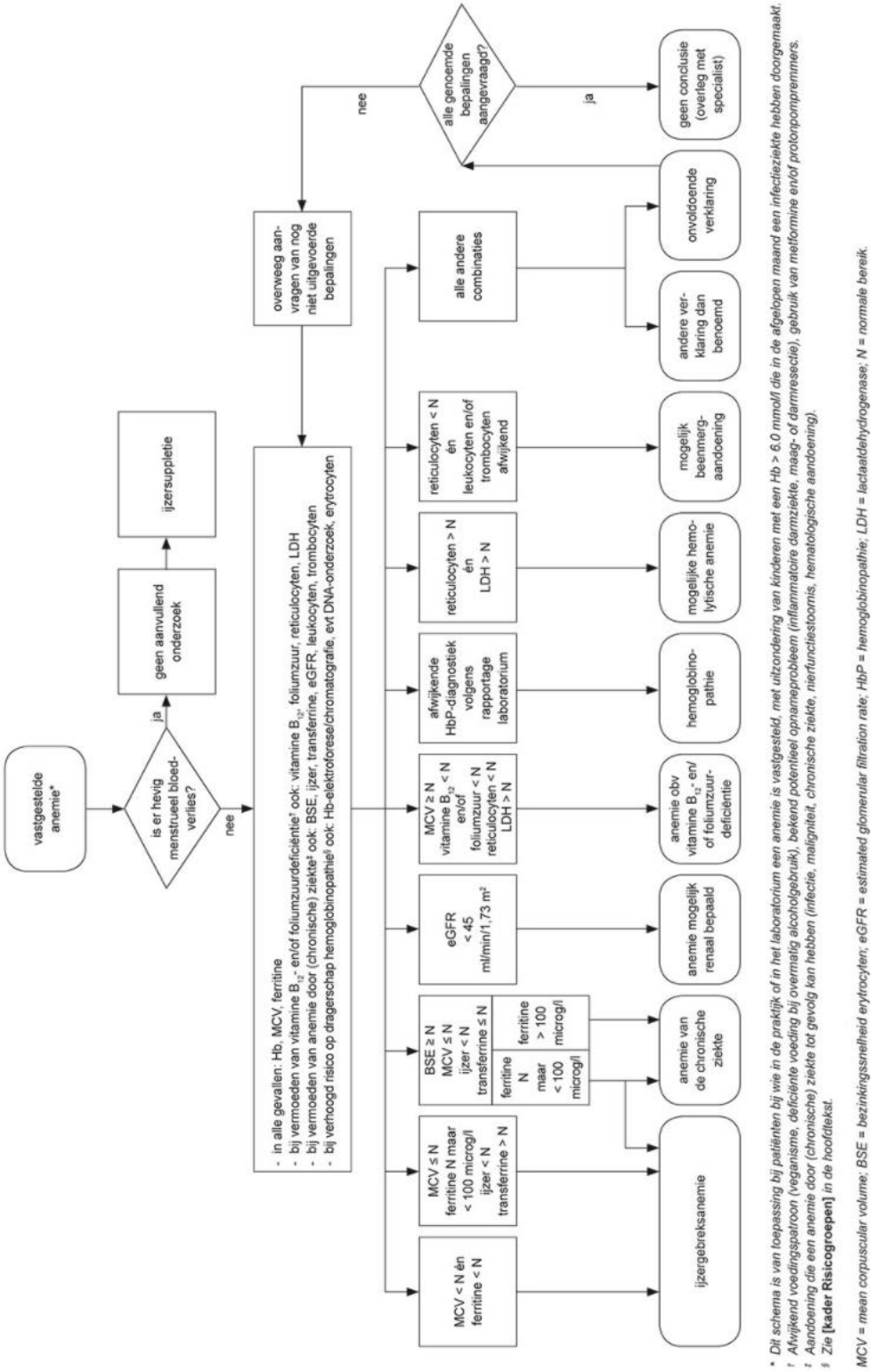

FIGURE 3: NHG DIAGNOSTIC STANDARD [118]

Figure 3 describes the diagnostics process for different kinds of anaemia once a patient with anaemic symptoms arrives. Different biomarkers are taken to determine the condition if the patient is not losing blood due to menstruation. In all cases, haemoglobin, the mean corpuscular volume (MCV), and ferritin within the blood are measured. If there is a suspicion of vitamin B12 deficiency or folic acid deficiency, additional measurements are taken to determine folic acid, reticulocytes, and lactate dehydrogenase (LDH) levels. On the other hand, if the suspicion arises that the patient has anaemia of chronic disease, additional measurements are taken for the erythrocyte sedimentation rate (BSE), iron, transferrin, the estimated glomerular filtration rate (eGFR), leukocytes, and thrombocytes. Finally, if a hemoglobinopathy is suspected, haemoglobin electrophoresis is taken, looking for different types within the blood to detect abnormal types.

Furthermore, DNA tests are also taken, and additional measurements regarding the erythrocytes are made. Which measurement levels can lead to an anaemia diagnosis can be seen in Table 2. An explanation for these biomarkers and other common ones follows in the subsequent sections.

TABLE 2: NHG DIAGNOSTIC STANDARD [118]

Biomarker Measurements	Diagnosis
MCV < Normal levels Ferritin < Normal levels	Iron deficiency anaemia
MCV < Normal levels Normal Ferritin values but < 100 mg/L Iron < Normal levels Transferrin > Normal levels	Iron deficiency anaemia
BSE >= Normal levels MCV <= Normal levels Iron < Normal levels Transferrin <= Normal levels Normal Ferritin values but < 100 mg/L	Iron deficiency anaemia/Anaemia of chronic disease
BSE >= Normal levels MCV <= Normal levels Iron < Normal levels Transferrin <= Normal levels Normal Ferritin values but > 100 mg/L	Anaemia of chronic disease
eGFR < 45 ml/1.73 m^2	Anaemia possibly renally determined
MCV >= Normal levels Vitamin B12 < Normal levels Folic acid < Normal levels Reticulocytes < Normal levels LDH > Normal levels	Vitamin B12/Folic acid deficiency anaemia
Abnormal hemoglobinopathy diagnostics according to lab report	Hemoglobinopathy
Reticulocytes > N LDH > N	Possible haemolytic anaemia
Reticulocytes < N Leukocytes and/or Thrombocytes divergent	Possible bone marrow disease
Every other combination	Other diagnosis/Insufficient explanation

2.1.2 BUSINESS OBJECTIVES

Through the use of diagnostic standards, like the one presented in section 2.1.2, diagnostic companies, hospitals, and doctors aim to maximize the following four attributes:

1. Accuracy in diagnosing anaemic patients
2. Efficiency of the diagnostic process
3. Transparency in the diagnostic process
4. Understandability of diagnostic process

Diagnostic standards are required to be understandable while at the same time ensuring high accuracy when diagnosing anaemic patients. For this particular use case, producing many true positive predictions takes up a higher priority than producing true negative predictions. In machine learning models of subsequent sections, the true positives are represented by the sensitivity value of the resulting prediction metrics. The understandability of the diagnostic process, on the other hand, is measured by the simplicity of the diagnosis flowcharts, represented by the number of biomarkers used while diagnosing a patient. Efficiency and transparency are the other two attributes that should be maximized. They are more intangible and do not fall within the scope of this thesis. Because of this, these attributes are not considered in the evaluation of the results.

2.2 PRODUCTION OF RED BLOOD CELLS AND ANEMIA

To better understand anaemia and its various forms, it is required to look at red blood cells themselves and their production. Red blood cells also referred to as erythrocytes, carry oxygen throughout the body and remove carbon dioxide by bringing it to the lungs to be exhaled. The production of these red blood cells is called erythropoiesis and mainly occurs in the bone marrow. While erythrocytes can typically live for around 120 days, 200 billion new ones are produced every day. Bone marrow, where red blood cells are produced, is a spongy tissue within the bones containing immature cells called stem cells [30]. Of the two types of bone marrow, red bone marrow contains hematopoietic cells, which can mature into red blood cells, white blood cells, or platelets. While red blood cells are responsible for oxygen delivery, white blood cells (leukocytes) can fight infections.

Furthermore, platelets (thrombocytes) are essential for blood clotting after an injury. Yellow bone marrow, on the other hand, uses mesenchymal stem cells to produce fat, cartilage, as well as bone [16]. A problem with the bone marrow can lead to dysfunctional red blood cells, leading to anaemia. Other conditions can result in the underproduction of red blood cells, leading to a state where the body cannot replace all of its dying red blood cells. This can also lead to anaemia. Aged red blood cells, which are more fragile than younger ones, get eaten up by macrophages, a type of white blood cell. During a process called phagocytosis, the iron contained within the cell returns either to the bone to produce new blood cells or to the liver or other tissue for storage.

Until a committed stem cell matures into a functional red blood cell, seven days can pass. The undeveloped stem cell is called an erythroblast, which is a nucleated cell without any haemoglobin. During its maturation, the cell develops haemoglobin, and its nucleus becomes smaller than before. This development continues until the seventh day when it loses its nucleus and enters the bloodstream [141]. If the red blood cell count is too low at any point in time, the body often makes up for it by releasing reticulocytes (immature red blood cells) into the bloodstream. Because of this, a high number of reticulocytes within the blood can be an indicator of anaemia.

The production of red blood cells is influenced by various micronutrients, like vitamin B-12 or folate, which contribute to the maturation of the cell [141]. Missing vitamin B-12 or folate can lead to erythrocytes larger than normal (macrocytic), potentially leading to anaemia. Missing vitamin B6 (pyroxide) or vitamin B2 (riboflavin), on the other hand, can lead to red blood cells that are too small (microcytic), also causing anaemia. Vitamins B6 and B2 are essential for the synthesis of globin, which is a protein. Finally, iron needs to be incorporated into the red blood cell haemoglobin, and a shortage may also lead to microcytic erythrocytes. However, a shortage of iron may be substituted by zinc, which increases protoporphyrin within the erythrocyte. Because of this, an elevated level of zinc protoporphyrin may indicate some form of anaemia.

2.3 TYPES OF ANAEMIA

The most common cause of anaemia today is a shortage of iron, which is estimated to cause approximately half of all anaemia cases. This figure may differ depending on the region, sex, and age of the patient. To understand the causes of anaemia, however, it may be worthwhile to investigate the types of anaemia and their characteristics.

2.3.1 IRON DEFICIENCY ANAEMIA

Iron deficiency anaemia is characterized by dyspnea (Shortness of breath), dizziness, an increased heart workload, tachycardia, and fatigue since there is not enough oxygen delivered to the tissues. Iron within the body is essential in haemoglobin production, which makes up a significant portion of the cell volume in erythrocytes (Red blood cells). A pigment called protoporphyrin 9 reacts with iron and gets converted into Heme, which is then used to produce haemoglobin [25]. The body can make no functional haemoglobin without iron, and an absence or decrease in iron can lead to smaller erythrocytes with a decreased cell volume. For determining iron deficiency anaemia, a blood test can be utilized to measure the so-called "Mean Corpuscular Volume" (MCV). The MCV is calculated by dividing the product of blood volume and haematocrit (proportion of blood that is cellular) with the number of erythrocytes in that blood volume. Whenever the MCV is below 80 femtolitres [157], the red blood cells are described as microcytic, resulting in the inability to deliver enough oxygen to the tissues. Causes of iron deficiency anaemia can be diverse, including blood loss, ulcers, menorrhagia (heavy menstruation), and a low-iron diet. Possible treatments for this condition are the supplementation of iron or the transfusion of blood [167].

2.3.2 PERNICIOUS ANAEMIA (B-12/FOLIC ACID DEFICIENCY)

Pernicious anaemia usually involves a deficiency in B-12, folic acid, or both. B-12 is present in leafy vegetables and certain types of meat, and a deficiency in B-12 is mainly caused by autoimmune conditions. Furthermore, it can also occur in elderly individuals, who often have a decreased intrinsic factor production [28]. B-12 naturally wants to bind to this intrinsic factor within the stomach, secreted by parietal cells. In some people, antibodies have been produced that bind to the intrinsic factor in place of B-12, blocking B-12 from binding. As a result, B-12 cannot be absorbed into the bloodstream. Since it is needed for the maturation and condensation of erythrocytes, a lack of B-12 can lead to disproportionately sizeable red blood cells [32][87].

On the other hand, folic acid is also responsible for the maturation of the red blood cell, so a lack thereof will result in large erythrocytes as well. A deficiency in either of them will hinder functional haemoglobin production, making the red blood cells so large that they can get stuck within the capillaries and possibly undergo haemolysis, which is a rupturing of the red blood cell. To diagnose this condition, one can use the MCV again. If the MCV is above 100, the red blood cells are described as macrocytic [10]. For treatment, intramuscular injections are a possibility, which enable absorption through an alternative route.

2.3.3 HEREDITARY SPHEROCYTOSIS

Hereditary Spherocytosis is a genetic condition with some mutation that disrupts spectrin or ankyrin production. These plasma membrane proteins are located in the red blood cell, and their absence can decrease the cell membrane's flexibility. As a result, the red blood cells change to become spherical, contrary to their usual biconcave structure, hence the name "spherocytosis" [87]. This transformation can throw off the MCV and lead to an inefficient transport of oxygen to the tissues, just like the other forms of anaemia. Furthermore, red blood cells can get stuck within the spleen's capillaries, potentially leading to an enlarged spleen (splenomegaly) [160]. Diagnosis of this anaemia type can be made by doing the osmotic fragility test, which tests the red blood cell membranes' permeability to salt and water [142]. Because of the unavailability of a cure, current treatment methods seek to manage the severity of the condition. Severe cases may require partial or total removal of the spleen.

2.3.4 G6PDH DEFICIENCY

Glucose 6-phosphate dehydrogenase (G6PDH) functions in a complex interplay with glutathione, helping it return to its non-reduced form. In that form, glutathione can catch free radicals, which are

toxic to the body. Examples of such radicals are the superoxide anion, the hydroxide free radical, or hydrogen peroxide, also called reactive oxygen species. To turn back into its non-reduced form, glutathione depends on a molecule called NADP, which gets produced within the red blood cell and requires the enzyme G6PDH. Without G6PDH, no NADPH can be produced, which leads to an accumulation of damaging reactive oxygen species and can finally lead to damaged haemoglobin within the red blood cell [22]. The damaged haemoglobin binds to the red blood cell's inner cell membrane, making the membrane less flexible. This binding can cause haemolytic anaemia, where red blood cells are destroyed, reducing the red blood cell count within the body. Like with the other types of anaemia, the result is the inability to efficiently transport oxygen to the tissues. When testing for this form of anaemia, one would test the haemoglobin binding to the cell membrane for Heinz bodies.

2.3.5 HAEMORRHAGIC ANAEMIA

Haemorrhagic anaemia is mainly caused by blood loss, which reduces the overall count of red blood cells within the body, leading to a reduced oxygen-carrying capacity. Some people have peptic ulcers or Helicobacter pylori which can cause bleeding, while other possible causes can be stab wounds, gunshot wounds, or aortic aneurysms. To treat this form of anaemia, blood transfusions and fluids are usually given to the patient. If necessary, damaged vessels may need to be fixed surgically. Severe blood loss can lead to haemorrhagic shock, a condition in which tissue perfusion cannot sustain aerobic metabolism [44].

2.3.6 SICKLE CELL ANAEMIA

This form of anaemia is a missense mutation that causes the three-dimensional structure of the red blood cell to change. In adults, haemoglobins usually have two alpha and two beta chains of amino acids. While the sixth amino acid in the beta chain is usually glutamic acid (Glu), a missense mutation can cause it to be converted into valine. This mutation causes the haemoglobin to display different physical properties since valine is a hydrophobic amino acid, while Glu is a hydrophilic or polar amino acid. Because of this, the molecules in the haemoglobins start polymerizing and connecting, resulting in a sickle shape of the red blood cell [92]. It has to be mentioned that patients with sickle cell anaemia do not only have sickle-shaped cells. The change in the three-dimensional structure only occurs if the haemoglobins are not bound to oxygen. Conversely, a sickle red blood cell can regain its biconcave form when it gets its oxygen back. This process is called sickling [92]. The danger of sickle cell anaemia is that the red blood cells can get stuck in the capillaries due to the cell's changing shape, leading to haemolysis. This can potentially lead to a vaso-occlusive crisis, where the cells get stuck in different body parts, such as the spleen [46]. Spleen removal might be necessary in this case, which can cause problems for elderly individuals since the spleen is essential to destroy encapsulated bacteria. Treatment options for this anaemia type include transfusions, additional oxygen, pain relievers, and fluids because of the possible blood loss or hydroxyurea. Patients with sickle cell anaemia showed to have a resistance to malaria [90].

2.3.7 APLASTIC ANAEMIA

Aplastic anaemia is a form of anaemia that does affect not only red blood cells but also white blood cells and platelets. The production of these cells in the bones requires the conversion of hemocytoblasts to myeloid stem cells. When this process is disrupted due to the destruction of bone marrow, low counts of red and white blood cells and platelets can cause various symptoms [124]. Next to the typical symptoms shown by people with other types of anaemia, low counts of white blood cells can lead to more infections, and low counts of platelets can disrupt blood clotting. Worse clotting of the blood can lead to bruises that are widespread over the body. 65% of aplastic anaemia is idiopathic. It can be caused by various kinds of drugs, as well as viruses and radiation. To treat the symptoms, continual transfusions can help manage the symptoms, while a more long-term solution would require a bone marrow transplant [123].

2.3.8 THALASSEMIA

Thalassemia is more common in people with Mediterranean ancestry. It is a genetic condition where beta or alpha chains within the haemoglobin are missing. There are two types of thalassemia, alpha-thalassemia and beta-thalassemia. The alpha variant causes the patient only to have one alpha chain in the haemoglobin, while the beta variant causes the absence of beta chains [73]. Because of this, cell

volume drops, leading to microcytosis and a low MCV. Constant perfusions, transfusions, iron supplements, and oxygen may be necessary for treatment. Alternatively, a bone stem cell transplant may be a more long-term solution for more functional haemoglobin production.

2.4 COMMON ANAEMIA BIOMARKERS

Because the different types of anaemia can have different causes in different locations within the body, each set of symptoms require their own biomarkers to measure. While many such biomarkers have an influence and are influenced by red blood cells, the most important ones are listed and explained in this section. Knowledge of these biomarkers helps choose biomarkers to create predictive models, which is covered in subsequent sections.

2.4.1 MEAN CORPUSCULAR VOLUME (MCV) AND HAEMATOCRIT

The MCV is used to measure the size of red blood cells. A value below 83 femtolitres indicates microcytic blood cells, which are erythrocytes that are too small. On the other hand, a value above 101 femtolitres indicates macrocytic blood cells, so too large erythrocytes [35]. A small value may be an indicator of iron deficiency but cannot be used on its own to diagnose it. This is because a small value may also be caused by thalassemia or anaemia caused by inflammation. A high value indicates deficiencies in folate or B-12, which may be caused by dietary reasons or poor absorption of those nutrients. Other than that, a low MCV could also be an indicator for alcohol abuse, liver disease, or hypothyroidism, which is an overactive thyroid. The MCV is calculated by taking the product of blood volume and the proportion of cellular (haematocrit) blood and dividing it by the red blood cell count in that volume. The haematocrit in this equation is the proportion of volume in the blood that contains red blood cells. When measuring the haematocrit alone, its typical values range from 0.4-0.5 L/L for men and 0.36-0.46 L/L for women. These values need to be increased for people that smoke. For people living high altitudes, on the other hand, these values need to be decreased. The MCV and the Mean Corpuscular Haemoglobin Concentration (MCHC) give identical results.

2.4.2 RETICULOCYTES

Reticulocytes are young and immature red blood cells with no nucleus and get released into the bloodstream to replace lost blood. They get produced in the bone marrow and eventually form fully mature red blood cells. Even though they have no nucleus, some remnant genetic material (RNA) can still be found within these cells [70]. In case the erythrocyte count within the blood sinks, for example, during anaemia, the reticulocytes are released into the bloodstream before they can fully mature to compensate for the loss. Under normal circumstances, a steady number of red blood cells is maintained within the blood, but a range of conditions, including anaemia, can influence the red blood cell count, leading to an increase of reticulocytes within the blood. Because of this, a higher number of reticulocytes can be an indicator of these conditions. On the other hand, a low number of reticulocytes can indicate insufficient blood cell production within the bone marrow, which can also indicate a range of conditions, such as aplastic anaemia [69].

2.4.3 RED CELL DISTRIBUTION WIDTH

The red cell distribution width is the variation for the width of red blood cells, with typical values ranging from 11 to 15%. A high measurement value usually means a larger variation in red blood cell size. This biomarker can be used to differentiate between a variety of different anaemia types. It is often used together with the mean corpuscular volume to determine the number of causes a case of anaemia might have. For example, a deficiency in vitamin B12 causes microcytosis with a typical red cell distribution width, while other anaemia types cause the patient to have an increased distribution width. Iron deficiency anaemia, for instance, shows an initial increase, while in a case where anaemia is caused by both iron deficiency and vitamin B12 deficiency, there be both small and big cells present within the bloodstream, causing an increase in the red cell distribution width [49].

2.4.4 FOLATES

Folate, also known as vitamin B9, is of great importance for haematopoiesis (red blood cell production) because it transfers one-carbon units in amino acid interconversions. This interconversion is essential for cell division which makes folate especially important for infants and pregnant women.

Because humans cannot produce folate, it needs to be a nutritional component of the diet. A deficiency of folate within the body can occur when not enough folate is ingested by dietary means or when folate absorption is hindered [88]. Because of its importance for cell division, a deficiency can impair the production of red blood cells, leading to megaloblastic anaemia and large, immature red blood cells. Other effects a deficiency might have are glossitis, diarrhoea, brain defects, fetal neural tube, and depression [26]. Even though serum folate concentrations start to fall after three weeks of folate deficiency, folate concentrations within erythrocytes stay stable for longer since they are better suited to represent folate storage. Because of this, to assess folate, serum and red blood cell concentrations should be measured in combination. The normal range starts at ten nmol/L for serum folate, while for red blood cell folate, the normal range starts at 340 nmol/L.

2.4.5 VITAMIN B12

Vitamin B12, also known as cyanocobalamin, is only available in animal-based foods like meat, eggs, milk, and cheese. Low intake of these foods can be due to unavailability of these products, high costs, or cultural reasons and can lead to a deficiency. Aside from these reasons, a deficiency can also be caused by other reasons such as malabsorption or a decreased internal factor production due to ageing. A low red blood cell folate concentration can indicate a vitamin B12 deficiency since missing B12 leads to a decreased folate metabolism. Without vitamin B12, folate remains trapped as N5-methyltetrahydrofolate without a possibility to get back into the folate pool. Furthermore, a vitamin B12 deficiency may lead to low plasma concentrations, an increased plasma homocysteine concentration, and increased urinary or serum methylmalonic acid concentration [139]. Plasma is the yellowish liquid component of blood that holds the blood cells and carries cells and proteins throughout the body. It holds several proteins, glucose, red blood cells, hormones, clotting factors, oxygen, carbon dioxide, and water. The usual range of plasma concentrations for vitamin B12 is above 180pmol/L [119].

2.4.6 VITAMIN A

Vitamin A's relationship to anaemia is essential because of its influence on the metabolism of iron. Even though its importance for anaemic patients was not recognized until the 70s, a study done by Hodges et al. described iron-type anaemia in vitamin-A-depleted subjects [140]. A deficiency in vitamin A is diagnosed by measuring the serum retinol concentrations. A value less than 0.7 µmol/L indicates a deficiency. Even though the effect of vitamin A on iron metabolism can be an influencing factor in the onset of iron anaemia, the role of infections should not be overlooked. This is because inflammation depresses plasma retinol concentrations, as well as the mobilization of iron. Vitamin A deficiency can often be observed in women and preschool children.

2.4.7 VITAMIN B2, B6

Vitamin B2, also known as riboflavin, is a water-soluble vitamin found in dairy products and in smaller quantities and other products. Just as with vitamin A, a deficiency in B2 may affect iron metabolism and may even impair iron absorption, increase the intestinal loss of iron, and hinder iron's role in the synthesis of iron. Because of these reasons, iron deficiency anaemia may also be an indicator of a vitamin B2 deficiency. A correction of low vitamin B2 has improved the response of patients with iron deficiency anaemia to iron therapy. Anaemia caused by low riboflavin is usually characterized by low production of red blood cells and an increase of reticulocytes within the bloodstream [143].
Vitamin B6, also known as pyridoxine, can be found in plant-based as well as animal-based food. A deficiency of vitamin B6 can cause microcytic anaemia but is rarely found [129].

2.4.8 IRON STORAGE, IRON DEPLETION, AND ZINC PROTOPORPHYRIN

Iron within the body that is not carried by the haemoglobin within the red blood cells is stored in ferritin complexes within all cells. These complexes are primarily present in the bone marrow, liver, and spleen. In total, the usual amount of iron within the body ranges from 4 to 5 grams, out of which 2.5 grams are contained within the haemoglobin. If the storage capacity of ferritin is exceeded, the excess iron forms into a complex with phosphate and hydroxide forms, which is called hemosiderin. If this hemosiderin exceeds normal levels, the excess iron gets stored in the liver and the heart, leading to

impaired functioning of these organs and, in some cases, death. In patients with insufficient iron, the body accesses these iron stored temporarily. Without enough iron, the body has difficulties making the protein haemoglobin, leading to iron-deficiency anaemia [17].
Iron depletion occurs when the iron is insufficiently available within the tissues, caused by low iron intake. This can be observed in patients with celiac disease, ulcerative colitis, or Crohn's disease. Alternatively, iron depletion can occur even with enough iron within the tissues by disrupting the physiological systems responsible for iron transport or absorption. In this case, usual treatment approaches, like supplementing iron, may not achieve their intended effects.
If iron is not available during the production of haemoglobin, zinc is inserted into the protoporphyrin molecule instead. A higher amount of zinc than expected within the red blood cells may indicate iron deficiency anaemia [72].

2.4.9 FERRITIN

Ferritin is a protein that stores iron and is produced in almost all organisms. The ferritin found in blood plasma is an indicator of the total amount of iron within the body, which makes it helpful in diagnosing conditions like iron deficiency anaemia. Typical ranges in men lie between 20 and 250 ng/mL and between 10 and 120 ng/ML in women. A low ferritin concentration characterizes a deficiency of iron in iron deficiency anaemia. Because these values can be raised significantly by inflammation during infection, a low value does hold more information than typical values for this biomarker. This is because a value can appear normal even while the patient may have low ferritin. On the other hand, an elevated level can indicate overaccumulation of iron in the body, which can be caused by disorders such as hemochromatosis [40].

2.4.10 SERUM IRON, IRON-BINDING CAPACITY, AND TRANSFERRIN SATURATION

Serum Iron is a test to measure the iron ions that are bound to transferrin within the blood. Transferrin is essential for transporting iron throughout the body, and the normal range for adult men lies between 60 to 170 µg/dL, while the range for women lies between 50 to 170 µg/dL. Measurement of transferrin concentrations is usually complementing the serum iron measurements [153]. The transferrin concentration measures how many iron ions fill the transferrin molecules. While a high value may indicate iron overload, a low value may indicate a shortage of iron. However, the concentration of iron may be depressed significantly by inflammation, which can lead to a decrease to somewhere between 5 and 15% during infection, whereas the normal range lies between 20 and 50%. The transferrin saturation is expressed as the iron-binding capacity of the plasma (TIBC) with average concentrations at around 60 µmol/L. While there is no difference between the sexes, a decline in concentration can be observed in older individuals. The TIBC is high in patients with iron deficiency but low in patients with haemolytic anaemia, iron overload, or undernutrition [61].

2.4.11 SERUM SOLUBLE TRANSFERRIN RECEPTORS

Serum Soluble Transferrin Receptors (sTfR) help distinguish between anaemia of chronic disease and anaemia caused by insufficient iron within the body. When only measuring ferritin within the blood, values may be significantly influenced by inflammations and may falsely indicate that iron stores are adequate within the body. Because sTfR is insensitive to inflammation, it can be used to measure the iron status within the body irrespective of inflammation present [34]. When iron binds to transferrin, a complex is formed that binds to receptors on the cell surface. Once the iron enters the cell, the transferrin receptors are cleaved from the cell and become soluble transferrin receptors found in the bloodstream [159]. The normal concentration of an adult lies between 2.8 and 8.5 mg/L. There are no differences between the sexes or ages between 18 and 80, but patients living in high altitudes and black patients might have increased concentrations. Low serum ferritin and iron concentrations indicate high sTfR concentrations, while low serum ferritin concentrations indicate high sTfR concentrations.

2.4.12 WHITE CELL COUNT

White blood cell count as a biomarker is important because an increase from 15.000 to 25,000 per microliter can indicate inflammation, influencing the interpretation of serum ferritin concentrations.

While the normal range of white blood cells is between 4,500 and 11,000 per microliter, leucocytes increase into the range mentioned earlier [174].

2.4.13 G6PD

Measuring G6PD helps determine G6PD deficiency anaemia in individuals who are affected by this condition. Since G6PD deficiency anaemia is a genetic condition, certain groups of people are more affected than others. In particular, people from West Africa, Central Africa, the Mediterranean, the Middle East, and Southeast Asia are affected the most. It has to be noted that all these regions are places where malaria is or has been prevalent. Since this condition is an X-linked recessive disorder, so carried on the X-chromosome, males are more often affected by female subjects. Aside from blood tests, genetic testing can be used to diagnose it [29].

2.4.14 HAPTOGLOBIN

Haptoglobin is used to bind with free haemoglobin from lysed erythrocytes in vivo to reduce its toxicity. During haemolysis, while red blood cells are ruptured, the number of free haemoglobin within the blood rises, causing haptoglobins to bind to them and depleting their total count. Because of this, a decreased number of haptoglobin may be an indicator of haemolysis. Since haemolysis occurs in haemolytic anaemias, a decrease in haptoglobin can be used to diagnose the condition, especially with a decreased blood cell count, haemoglobin, and haematocrit, and increased reticulocyte count. If the count of haptoglobins is normal, but the reticulocyte count is increased, it can be assumed that the destruction of red blood cells occurs in the liver or the spleen. This can indicate extravascular haemolytic anaemia, drug-induced haemolysis, or red cell dysplasia (Abnormal growth or development of the red blood cell). This is because the destruction of the cell in the spleen or liver does not cause a release of its content into the bloodstream [120].

2.4.15 FIBRINOGEN

Fibrinogen is a glycoprotein produced in the liver and converted to fibrin-based blood clots after a vascular injury. It promotes revascularization and wound healing and is a “positive” acute phase protein, which means that its levels increase during inflammation within the body, tissue injury, or various cancers. It is mainly produced by liver hepatocyte cells but is also produced in smaller amounts by endothelium cells. Studies have shown that hyperfibrinogenaemia was present in 75 to 85% of patients with haemoglobin concentrations above 110 to 120 g/litre [39] despite its observed independence to haemoglobin. Hyperfibrinogenaemia is a rare inherited disorder where blood does not clot because of reduced fibrinogen concentrations.

2.4.16 ERYTHROPOIETIN

Erythropoietin (EPO) is a hormone produced by the kidneys, which promotes the production of erythrocytes within the bone marrow. Aside from this, it also is responsible for initializing the synthesis of haemoglobin, the molecule that carries the oxygen within the red blood cell. It is a protein with an attached glycoprotein (sugar), and a low level may indicate anaemia. This is because the cells within the kidneys that are responsible for producing erythropoietin are sensitive to the oxygen within the blood, and a low level of oxygen triggers its production to make up for the low number of red blood cells or haemoglobin within the blood. To a lesser extent, it is also produced in the liver, which is taking care of approximately 10% of total erythropoietin production. Anaemia can occur when the kidneys can no longer produce adequate amounts of erythropoietin to create new red blood cells, which can be caused by chronic kidney disease [98].

2.4.17 D-DIMER

D-dimer is a fibrin degradation product, a protein found in the blood after a blood clot has degraded by fibrinolysis, the process responsible for preventing problematic blood clot growth. Even though it is typically used in diagnosing thrombosis, elevated levels can also indicate sickle cell anaemia, even though it is unclear if it is caused by activation of coagulation or the many vaso occlusive complications of this condition [33]. Furthermore, a highly elevated haematocrit has been shown to activate coagulation activities, in which D-dimer plasma levels are bound to increase [33]. Finally, an increase in D-dimer could also be observed in patients with thalassemia, with thrombosis as a well-described complication. Particularly after splenectomy, markers of thrombin generation, like, for example, prothrombin F1 plus two and D-dimer, were found to be increased [163].

2.5 ASSESSMENT OF THE CURRENT SITUATION

As could be seen in section 2.4, there are many more biomarkers available than there are currently present in the diagnostic flowchart of the NHG. While the flowchart is simplified for increased understandability and efficiency, the reality is often more complex and could be reflected better by a larger number of biomarkers. Alternatively, the same number of biomarkers with different biomarker selections may also provide better prediction performance for different types of anaemia. In sections 5 and 6, different prediction models are created to assess the impact of these selections on sensitivity and other metrics to see whether a different selection could provide this improvement.

3 LITERATURE REVIEW

This literature review aims to give insight into how machine learning can contribute to better decision-making in process mining. It provides an overview of what learning problems can be addressed by machine learning techniques/approaches/methods. Most importantly, it highlights the importance of machine learning for process mining, focusing on the healthcare industry. This literature review differentiates itself from other similar work, like, for example, [36], [135], or [169], by having its focus explicitly on the machine learning aspects of process mining while narrowing down its scope to healthcare at large. Ultimately, gaps should be identified within the current research to define possible research questions that can be addressed in the future. Using this knowledge, a use case for this research thesis is defined, covering the use of both machine learning and process mining within this work. While there is no step in the CRISP-DM model that represents this part of the thesis, it is nonetheless important in understanding the techniques used within this thesis and the interplay between these techniques.

3.1 RESEARCH QUESTIONS & SEARCH STRATEGY

Process mining is a relatively new scientific discipline that aims at bridging the gap between Business Process Management and Workflow Management [183]. Business process modelling techniques are typically disconnected from actual processes, while data-oriented analysis focuses on simple data mining and machine learning techniques. Process mining can provide the missing link between these disciplines and performance/compliance-related issues. It aims to confront event data, observed behaviour, and process models to optimise processes [182].

Event logs can generally be used in 3 different ways. Through process discovery, new models can be produced without any form of a-priori information, while conformance checking helps compare an event log with its model. On the other hand, process enhancement improves existing process models by using information about the actual process collected in some event logs [184]. These process mining techniques can improve processes in various application domains, like in industrial settings or the healthcare industry. Use cases could, for example, involve fraud detection [104], stroke care [145], hospital care flows [146], or auditing [105]. This literature review focuses on the healthcare industry, where process mining is mainly used in the control-flow perspective [15]. Standard process mining techniques or algorithms in healthcare are trace clustering, fuzzy miner, and heuristic miner due to their ability to handle noise and incomplete data.

In this section, the methodology used while doing the literature search is presented using the Kitchenham methodology [180]. It introduces the research questions for the review and describes the search strategy and the necessary steps.

3.1.1 RESEARCH QUESTIONS

This literature review has the aim of answering the following research questions:

1. When combined with machine learning, what applications do process mining techniques/methods/approaches have in the healthcare industry?
2. When combined with process mining, what machine learning techniques/methods/approaches can improve prescriptive analysis?
3. What learning problems can be solved by machine learning within the process mining domain?

3.1.2 SEARCH STRATEGY

The libraries used for this literature review are Scopus and FindUT, which cover a substantial number of publishers. The search is done using the Utwente network, making it possible to access all the papers accessible to the university. For the findings, only peer-reviewed papers are being considered. Other works, like books or dissertations, are excluded from the search. Furthermore, all papers not in the English language are not considered as well. The research is done using a keyword approach. The key-word combination used in this literature review is the following:

"Process Mining" AND "Machine Learning"

Using this keyword pair, 134 papers get returned in Scopus, and 119 papers get returned in FindUT. The findings of the 253 papers are cleaned up as follows:

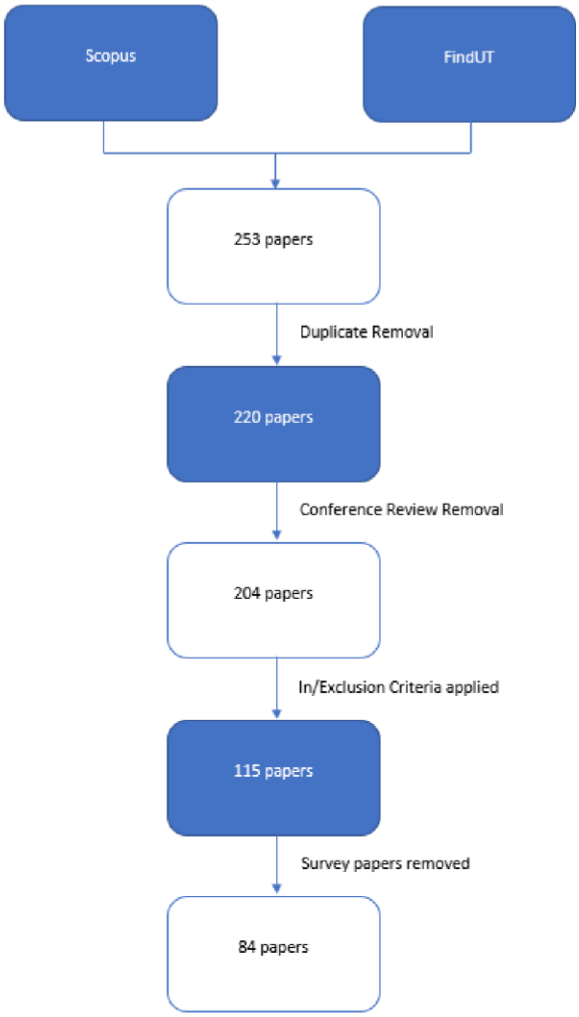

FIGURE 4: DATA EXTRACTION PROCESS

Step 1: Duplicates are removed that are present in both Scopus and FindUT. 220 unique papers are left as a result, which means that 33 were duplicates. This ratio is approximately 13%.

Step 2: Conference reviews are removed in the next step, which leaves 204 papers remaining. The 16 conference reviews in the results are approximately 6% of the initial 253 papers.

Step 3: To answer the research questions, inclusion criteria are utilized to keep papers with the highest relevancy. The inclusion criteria are the following:

IC1: The paper directly addresses the research questions for this literature review
IC2: The paper is focused on the healthcare industry

IC3: The paper focuses on the application of machine learning

Step 4: By using exclusion criteria, irrelevant papers are filtered out as well. The exclusion criteria are the following:

EC1: The paper cannot be downloaded from the internet

EC2: The paper has neither machine learning nor process mining as the main topic and only refers to them as a side topic or references these knowledge domains

EC3: The paper focuses exclusively on either machine learning or process mining and only references the other knowledge domain in a way that is not relevant to the research questions

By applying EC1, 16 papers got removed from the findings, leaving 186 papers to analyse. After using the other research questions and filtering out papers not relevant for this research, 115 papers are left. From the original 253 papers, approximately 54% are removed.

Step 5: In the final step, survey papers and other literature reviews and tutorials are removed. By doing this, 31 papers are removed, leaving a final 84 papers left to use for analysis. Out of the initial 253 papers, the final selection represents approximately 34% of these initial papers. Necessary information on all the papers can be found in the appendix.

A literature matrix is created using these papers, consisting of the necessary information about all the papers and answers to the research questions presented before. This literature matrix is used to analyse the results and get detailed findings that answer the aforementioned research questions.

3.2 DATA OVERVIEW

Visualizations are created to describe various aspects of the dataset to get an initial overview of the papers. In Appendix D, for example, one can find a bar graph on the number of papers published per year. As can be seen from that graph, most papers were published in 2018, with an ever-decreasing number the further one goes back in time. This finding indicates, for one, that the field of process mining, especially in combination with process mining, is relatively new. The second thing seen in the graph is the increasing academic interest in this topic, with a steady increase in research papers until 2018. Before 2012, there were barely any papers that attempted to handle this topic. By looking at Figure 5, one can deduce that there is potential for more research in the future.

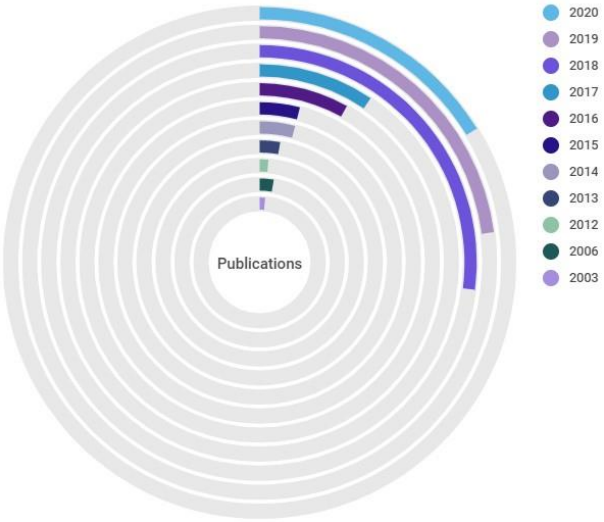

FIGURE 5: NUMBER OF PUBLICATIONS PER YEAR

The second figure describing the data is visualized in Figure 6, which includes an overview of the frequency of keywords in the papers. The two most frequent keywords are "process mining" and "machine learning," which is to be expected since the search terms for the literature extraction are "process mining" and "machine learning" as well. When looking at the other keywords, it becomes apparent that many papers used the keywords "classification" and "clustering," which most likely stems from the popularity of these techniques in the machine learning domain. Keywords also high in the list are "predictions" and "predictive process monitoring," which already indicate the most common use

cases of machine learning in process mining. Other keywords like "process discovery" or "decision trees" can be associated with either process mining or machine learning.

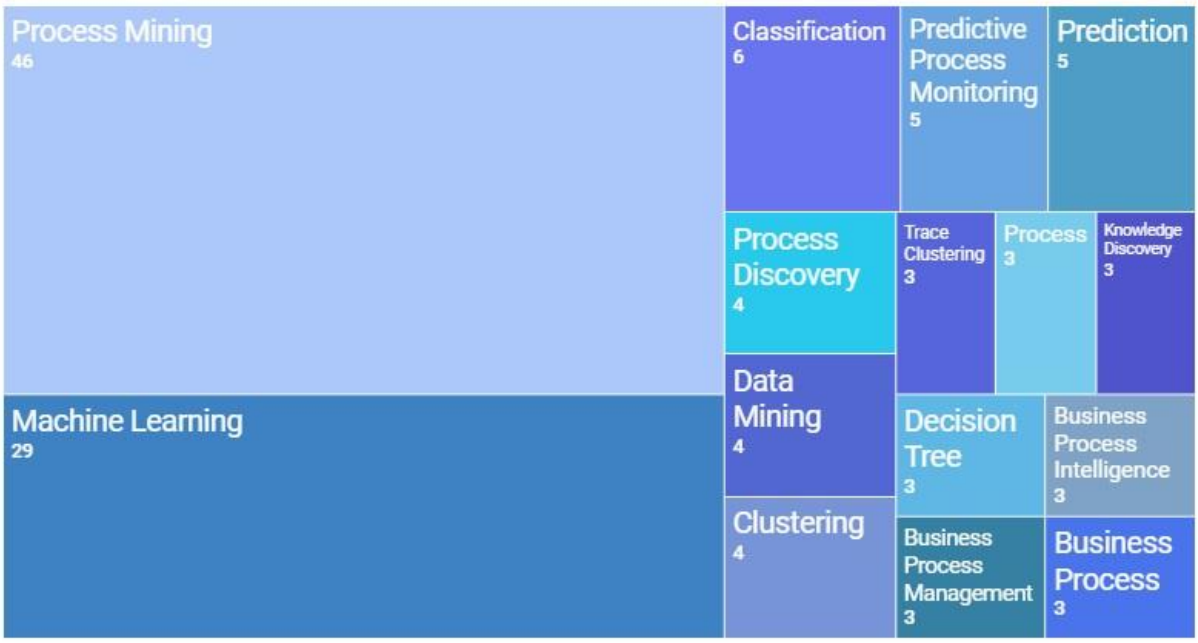

FIGURE 6: KEYWORD FREQUENCY

The third figure relevant for understanding the literature matrix can be found in Figure 7. This figure gives an overview of how individual researchers have contributed to process mining and machine learning publications. As can be seen from the graph, La Rosa seems to have made the most significant contribution, with a count of 5 papers. Malerba, Yakovlev, Appice, and Van der Aalst contributed to 3 papers and other noteworthy people who published about the combination of machine learning and process mining.

Finally, the last figure describing the literature matrix can be found in Figure 27, which illustrates the number of citations for the top thirty papers. It can be seen that only 12 papers have a citation count over 20, while only five papers have more than 50 citations and one more than 100 citations. In this regard, the most influential paper is "Decision mining in ProM," which was written by W. van der Aalst, who also appeared in the previous list of researchers with the most contributions.

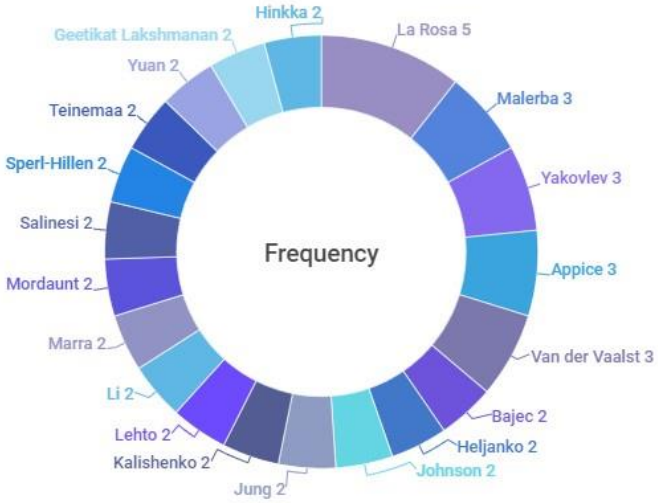

FIGURE 7: NUMBER OF PUBLICATIONS FOR MOST ACTIVE RESEARCHERS

Aside from the researchers contributing to machine learning and process mining, an analysis is made on the datasets used in their works. This analysis can be seen in Figure 28. Interestingly enough, while most papers do not focus on the healthcare domain and tried to generalize their work, 19 out of 74

papers decided to take a healthcare dataset to validate their results. This is 25.7% of all papers analysed in this literature review.

Even though this insight is fascinating, a definite conclusion on the usefulness of machine learning and process mining in healthcare cannot be made. Since many papers decided to use publicly available datasets, the choice of the datasets used might have been determined by availability instead of applicability. When looking at the rest of the datasets used in the papers, one can observe that finance and manufacturing play a significant role in the validation, as these are essential domains. Finally, some papers do not use any datasets or make use of synthetic datasets. The choice of using no dataset is mainly present in papers with a heavy theoretical focus.

3.3 RESULTS RESEARCH QUESTIONS

3.3.1 RQ1: WHAT APPLICATIONS DO PROCESS MINING TECHNIQUES/METHODS/APPROACHES HAVE IN THE HEALTHCARE INDUSTRY WHEN COMBINED WITH MACHINE LEARNING?

Of the 74 papers reviewed for this literature review, only 12 focus on process mining and machine learning in healthcare. The utilization of machine learning and process mining techniques for this domain is varied and utilized for many applications. In this section, a summary is given on how the analysed papers implemented these techniques. More details on them can be found in Table 25.

TABLE 3: RQ1 RESULTS

Pathway Discovery	Resources/Planning	Anomaly detection	Patient Treatments	Other
Healthcare Pathway discovery: [6]	Resource handling about hospital staff planning: [6], [110]	Failure Prediction: [52]	Treatment recommendations: [188]	To reveal communication intentions around the topic of healthcare: [31]
Patient location tracking: [130], [53]		To prevent medical fraud: [168]	Monitoring of patient conditions: [82]	Patient/Trace Clustering: [164]
				Detecting hidden healthcare subprocesses: [1]

The first application in this context is discovering pathways within the analysed event logs. [6], for example, builds a process mining pipeline to discover healthcare pathways using hospital records. Machine learning can then explore different pathways features that can give further insight into the data. Like [130] or [53], other papers focus on patterns of treatment processes in terms of patient movement, which allows for identifying movement trajectories in [130] and the ability to track patients in real-time in [53]. Machine learning is used in this context to classify patients and staff with similar characteristics.

Aside from pathways discovery, machine learning and process mining can also impact resource planning, which can be seen in paper [6], for example. [6] defines a probabilistic regression model that can estimate the recovery time using the information extracted from the process mining pipeline they created. By better understanding when a hospital can release their patients, scheduling can be made more accurate. [110] takes a different approach. By introducing a framework that uses a machine-learning, real-time, and data-driven prediction approach for system performance, it becomes possible to handle complex decision-making processes around hospital staff planning.

The third major use case of process mining and machine learning within the healthcare domain is detecting data anomalies. The research was done to predict failures and to detect fraud within this context. [52] for example, uses an inductive machine learning algorithm to generate decision rule sets that can help predict and eliminate treatment failures. [168] on the other hand, is focused on revealing fraud patterns by using an unsupervised hierarchical method as a pre-screening tool to aid in assessing potential fraud.

Patient treatment optimization and support are other applications that benefit from both process mining and machine learning, which can be achieved by recommending treatment options and monitoring patient conditions. [188] is an example of the former, where action rules, generated by the action rule mining technique, are used to identify treatments that co-occur with specific outcomes under some conditions. After using this technique, uplift trees are used to discover subgroups of cases for which the treatment has a high causal effect on the outcome. [82] on the other hand is an example for patient monitoring, which uses an approach where they use process mining to extract healthcare information from several contextual views and utilize machine learning to then monitor the conditions of the patients. To summarize, the four major applications identified for process mining and machine learning in healthcare are pathway discovery, resource planning, data anomaly detection, and patient treatment optimization and support. Aside from these categories, some papers could not be categorized into any application type. [31] is about revealing communication intentions around the topic of healthcare, [164] is about applying trace clustering techniques on a set of patients, and [1] is about detecting hidden healthcare subprocesses. All in all, these papers do not fit into any of the categories as mentioned earlier. Nonetheless, they are essential for a complete picture of the current research in this domain.

3.3.2 RQ2: WHAT MACHINE LEARNING TECHNIQUES/METHODS/APPROACHES CAN CONTRIBUTE TO AN IMPROVED PRESCRIPTIVE ANALYSIS WHEN COMBINED WITH PROCESS MINING?

In this section, a summary is given of the findings concerning the second research question. Just like with the first question, more detailed answers can be found in Table 25. Furthermore, like in question 1, a categorization is done to assign each paper to a prescriptive analysis contribution group. The first group identified is the "anomaly detection" group. Papers within this group focus on improving prescriptive analysis by detecting anomalies in various ways. [94] for example proposes an approach for detecting malware, where process mining techniques contribute to identifying patterns in the traces to characterize its behaviour. Other papers, like [82] or [83], focus on preventing data traffic overload. Both analyse wireless networks and their traffic by using process mining and machine learning techniques. For this purpose, wavelet neural networks, graph mining, time series analysis, and more are used. While malware detection and traffic analyses are very specific use cases, papers like [6] or [48] take a more general approach by focusing on deviations from the expected behaviour for certain variables. [6] for example, focuses on deviations in treatment durations and pathways.

Other uses of process mining and machine learning in anomaly detection are concept drift detection and the discovery of atypical business processes. [85] for instance, uses a concept drift algorithm to detect the challenges concept mining can cause for the ongoing value creation in process mining. For best results, incremental learning should be combined with the retraining of data in case of concept drift. [187] decides to focus more on a-typical processes by developing a business decision support system that uses a machine learning model. Finally, the last application in anomaly detection is the assistance in root cause analysis, which is handled in [101] and [41]. While [101] focuses on how feature selection results can help in computer-assisted root cause analysis, [41] focuses on using machine learning to identify weaknesses in business processes automatically.

TABLE 4: RQ2 RESULTS

Anomaly detection	Predictions	Performance improvements	Decision making support	Improved data understanding
To detect malware: [94]	By making predictions about future event: [112], [4], [117], [20], [65], [126], [110], [130], [38], [81], [5], [78], [97], [150], [154], [63], [64], [116], [76], [48]	By advising on how to improve process inefficiencies: [41], [170]	To assist in auditing: [103]	To identify prospective customers: [97]
To discover a-typical decision-making processes: [187]	By pre-empting the user with information that he needs in decision making: [99]	By reducing training times of samples: [106]	By helping in resource planning: [65], [6], [110], [185]	By giving more detailed information about the underlying process: [19], [107]
To prevent data traffic overload: [82], [83]	To predict the goals of the user: [47], [13]	To improve mining outcomes: [186], [31], [179], [74], [77], [132]	Inductive machine learning models can generate decision rules for better decision making: [52]	To better identify data dynamics on a microlevel: [158]
To help detect deviations from expected behaviour: [6], [48]		By automating: [171], [176], [147], [66]		By obtaining macrolevel characteristics: [158]
To detect concept drift: [85]				
By assisting in root cause analysis: [101], [41]				

The second major categorization, where process mining and machine learning can support prescriptive analysis, is "predictions." By making predictions about future events or KPIs, decision-making processes can be improved. This is handled in many papers, including [112], [4], [117], [20], [65], [126], [110], [130], [38], [81], [5], [78], [97], [150], [154], [63], [64], [116], [76], and [48]. An example can be given by looking at paper [64], which uses a black-box approach to predict quantitative process performance indicators for ongoing process instances. These performance indicators include remaining cycle time, cost, or probability of deadline violation. By knowing this information, better operational decisions can be made. These benefits also extend to customers, where machine learning and process mining can pre-empt needed information, like in [99]. This paper focuses on IT support ticket resolution, where process mining and machine learning can reduce the required number of inputs by pre-empting information needed for decision making. Finally, predictions can also help determine the user's intent, which is the focus of [47] and [13]. Both papers research intention mining, which is the ability to predict a user's goals. Knowing their intentions, the decision-making of network administrators, for example, can be improved.

Categorization number three for this research question is "performance improvement." By advising businesses on improving business inefficiencies and improving mining outcomes, performance can be increased. Process inefficiencies, for example, are covered by [41] and [170]. [170] for instance, analyses anomalies occurring in the execution of business processes. By doing that, inefficiencies can be detected and analysed. To this end, denoising autoencoders are used that are evaluated on several datasets. Mining outcome improvements are the aim of papers [186], [31], [106], [179], [74], [77], and [132]. These improvements can be exemplified by [74], which tries to maximize their fail detection algorithm's accuracy using the Hidden Semi Markov Model. [106] is another example

where the training time for their Small Sample Learning method should be minimized. Another way process mining and machine learning can improve performance is by automating, which is the focus of [171], [176], [147], and [66]. [171] for instance, attempts to define constraints in an automated manner to be used in runtime monitoring approaches. This automation solves the problem of missing domain knowledge when defining constraints and can make them more useful.

The next categorization for this research question is "decision making support." Using machine learning and process mining techniques can help generate decision rules for better decision-making or assist in auditing. An example of the former can be found in [52], which uses an inductive machine learning algorithm to generate decision rules and identify areas for improvement. An example of the latter is [103], which combines machine learning techniques with a human auditor's expertise by creating a continuous audit environment. This auditing technique is done through the use of three-way decision rules. Another application in decision-making support for this categorization is in resources planning, which is handled by [65], [6], [110], and [185][93][93]. For example, this application can be seen in [65], where they train a range of predictive models to predict various performance indicators. This way, issues can be identified early on, and resources can be reallocated from one case to another to avoid running overtime.

The last categorization where machine learning and process mining can help prescriptive analysis is "Improved data understanding." By giving more detailed information about the underlying process, decision-making can be improved, as in [19] and [107]. [107] analyses running processes to identify patterns that can be used as a model for decision support when new objects need to be assigned to an existing pattern. This assignment is done by comparing different representative attributes like average time to the organization's actual behaviour. [19] on the other hand, uses black-box machine learning to attain more detailed information about underlying processes. The data can also enrich a business's understanding by identifying prospective customers, which is done in [97]. [97] creates a process mining framework that contains machine learning and can use predictive modelling to do customer prospecting. Furthermore, using machine learning and process mining, data dynamics can be identified on a micro and macro level, which is the topic of concern in [158]. The micro-level dynamic can then contribute to better simulations, and the macro-level characteristics can be used to get more insight into valuable variables, like departmental load or queuing parameters.

To summarize, this section identifies five ways machine learning and process mining can contribute to improved prescriptive analysis. This contribution can be made by detecting anomalies, making predictions, improving performance, decision-making support, and understanding the data and its underlying processes.

3.3.3 RQ3: WHAT LEARNING PROBLEMS CAN MACHINE LEARNING TECHNIQUES/METHODS/APPROACHES SOLVE WITHIN THE PROCESS MINING DOMAIN?

In this section, a summary is given of the results concerning research question 3. More details on these results can be found in Appendix A, as well. Like in the last two sections, a categorization is applied to all papers to understand their contribution. This categorization is done with a focus on learning problems that can be solved with machine learning.

TABLE 5:RQ3 RESULTS

Predictions:	Process Analysis	Anomaly detection	Data Handling	User analysis	Clustering and Classification	Other:
Prediction about future event: [112], [4], [117], [20], [65], [126], [110], [130], [38], [81], [5], [78], [97], [150], [154], [63], [64], [116], [76], [48]	How can process models be analysed automaticall y?: [66]	How to detect anomalies in internal processes?: [170], [168]	How to filter chaotic activities from event logs?: [125]	How can user behaviour patterns be used in machine learning to improve process mining outcomes?: [186]	How can business process instances be classified?: [100], [166], [101], [176], [130], [94]	How can robust rule sets be induced?: [91], [52]
	How to detect subprocesses ?: [1], [168]	How to detect failures?: [74]	How to automatical ly extract data from big data logs?: [77]	How can user intent be mined?: [47], [13]	How to detect subgroups for process mining results?: [188]	How can constraints be defined in an automated manner?: [171]
	How to handle concurrency ?: [165]	How can defect rates be estimated using a combination of process mining and black-box machine learning techniques?: [19]	How to use small datasets for training?: [106]		How can Traces be grouped/clustered?: [3], [166], [79], [186], [107], [82], [27], [132], [185], [189], [53], [71]	How can constraints be recommended using machine learning?: [147]
			How to let classifiers deal with uncertainty: [103]		How to use unstructured data for training: [187]	How to evaluate detected patterns?: [5]
	What is the most efficient routing of a case? (Decision Mining): [96], [91], [148]	How can runtime monitoring approaches be used to detect possible deviations from the expected behaviour?: [171]			How can data be clustered to find better event paths?: [158]	How to measure variability to decide whether to use imperative or declarative miner?: [18]
					How can unstructured data be labelled to use in process mining?: [136], [55]	

The first category identified is "predictions," which includes all papers that attempt to solve learning problems to predict future events. These papers include [112], [4], [117], [20], [65], [126], [110], [130], [38], [81], [5], [78], [97], [150], [154], [63], [64], [116], and [76]. [97] for example, deals with customer prospecting, which is achieved by using predictive modelling to predict future events based on historical data. The predictions are made using decision trees and k-nearest-neighbours. Another example of predictive learning problems is [117], which uses stacked inception CNN modules to predict the next activity. [154] on the other hand, uses the WoMan framework for workflow

management to predict future activities. While the goal of making predictions is similar in all these papers, each one uses different machine learning techniques to achieve its results.

The second category identified is "process analysis." When aiming to get a better insight into business processes, machine learning can be used by doing decision mining, which can help attain a case's most efficient routing. Decision mining is the main topic in papers [96], [91], and [148]. [96] uses decision trees to detect data dependencies that affect the routing of the case, while [91] improves on that by first aligning log and model to deal with deviating behaviour and complex control flow constructs. [148] on the other hand, tries to utilize the rules obtained from decision mining to identify the form of choices using individual relationships of rules found from the workflow. The second learning problem to be solved in "process analysis" is concurrency covered in [165]. [165] creates an algorithm of workflow process mining based on machine learning that can handle concurrence and recurrence of business processes. Aside from this, another learning problem is concerned with detecting subprocesses, which are handled in [1] and [168]. Finally, [66] attempts to analyse process models automatically to get improved insights.

The third learning problem category defined is "anomaly detection." Within this category, learning problems are related to failure detection and other deviations from the expected behaviour. [19] for example uses a combination of process mining and black-box machine learning techniques to estimate defect rates, while [74] uses structural learning models. [171] on the other hand, makes use of runtime monitoring approaches to detect deviations. Other papers that focus on anomalies in internal processes are [170] and [168].

Another learning problem category deals with problems related to "data handling." [125] for example is occupied with filtering chaotic activities from event logs, which uses indirect entropy filters for best results. [77] on the other hand, is concerned with the extraction of data from big data logs, which formalizes data extraction and task identification as a problem of extracting attributes as process components. Sequence kernel techniques are used to solve this problem. The use of small datasets for model training is addressed in [106], which uses small sample learning (SLL) to tackle machine learning issues where only quantitatively insufficient datasets are available. Finally, [103] is about uncertainty, which can often not be handled by classifiers. A more intelligent classifier can be created to deal with the uncertainty using the rough set theory and the fuzzy sets theory.

The fifth learning problem category identified for this research question is "user analysis," which deals with user intent and behaviour patterns. An example of the former can be found in [47] and [13]. [47] uses the map miner method to mine user intent, while [13] uses multiagent systems in unsupervised learning. An example of user behaviour patterns can be found in [186], where these patterns are incorporated into sequence clustering for workflow model discovery.

The last learning problem category identified in this section is "clustering and classification." Papers in this category deal with the classification of process instances and clustering of traces. Trace grouping and clustering is the concern of [3], [166], [79], [186], [107], [82], [27],

[132], [185], [189], [53], and [71]. [3] for instance, uses trace clustering as a pre-processing step to split up the event log into clusters of similar traces. On the other hand, papers like [71] use clustering after the use of process mining. [71] in particular, uses process mining to identify action patterns that can be clustered to find group patterns. Like [132], other papers aim to improve human understanding of trace clustering solutions, a post hoc application of supervised learning with support vector machines on cluster results.

On the other hand, classification is either used to label unstructured data or to classify process instances. Papers that are about unstructured data are [136] and [55]. [136] proposes a framework to mine processes from CRM data by leveraging the unstructured part of the data. This processing is done using Latent Dirichlet Allocation(LDA), an unsupervised machine learning technique. [55] on the other hand, uses a modified prototypes clustering approach to obtain the labelled data. Classification of business process instances is addressed in [100], [101], [176], [130], and [94]. [100]

for example, classifies business process instances using Gates Recurrent Unit(GRU) and Long Short-Term Memory(LSTM) neural networks. This classification is done in a supervised fashion using unlabelled data. [101] classifies the process instances based on structural features derived from event logs, while [176] uses supervised classification as part of the machine learning layer of the framework they propose in their paper. [130] utilizes classification by classifying patients in healthcare data based on their movement trajectories. This classification is done using several machine learning methods, like SVC, Random Forest, or KNN. Finally, [94] uses classification algorithms to identify malware families.

To summarize, six different learning problem categories are created that contain papers aiming to solve that problem. Learning problems concerning process mining and machine learning involve predictions, process analyses, anomaly detection, data handling, user analysis, and clustering and classification. Besides these categories, several papers could not be classified into any category. [171] and [147], for instance, are about the recommendation and definition of constraints, while [91] and [52] are about the induction of robust rule sets. [5] attempts to evaluate detected patterns, while [18] proposes a way to measure variability to decide whether to use an imperative or declarative miner.

3.4 PROCESS MINING IN HEALTH CARE

Even though there are many possible use cases of process mining and machine learning in healthcare, its potential in the literature is not taken advantage of to its fullest extent. This potential can be seen when comparing the results from the first research question with the other research questions' results. While RQ1 is covered in 11 out of the 74 investigated, RQ2 is handled by 49 papers, and RQ3 is handled by 64. This gap is a stark contrast which shows that there is much room for the techniques in RQ2 and RQ3 to be applied in the healthcare sector. This gap can also be seen in Figure 8.

FIGURE 8: THE PROPORTION OF PAPERS THAT ANSWERED THE RESEARCH QUESTIONS

When looking at Table 3, Table 4, and Table 5, one can see that several categories from RQ2 and RQ3 are also relevant for healthcare. Anomaly detection is one of them, for example, which gets addressed

in all three research questions. In research question 1, it is the focus of papers [52] and [168], which deal with failures and fraud. Other categories that are important for healthcare are clustering and classification", "predictions," "decision-making support," "process analysis," and "user analysis." "Clustering and classification" are appearing in [164] and [130], for example, while "predictions" are used in [52] as well. "Decision support" is relevant in [6] and [151], which deal with resource planning in hospitals, while "process analysis" is appearing in [6] to discover healthcare pathways and in [1] to detect subprocesses. Finally, a "user analysis" is done in [31], where communications intentions around the topic of healthcare are revealed. Even though these categories are applied in multiple papers, many more process mining and machine learning techniques can be used in this context. In Table 4 (RQ2), for example, one can see nine papers handling anomaly detection, and in Table 5 (RQ3), one can see five papers handling anomaly detection, while in Table 3 (RQ1), which deals with healthcare, there are only two papers to be found covering it. This indicates that there is still much room for the application of different anomaly detection techniques in healthcare. The same applies to the other categories. While they are covered in healthcare literature, many more techniques could potentially be used in this context.

Categories from RQ2 and RQ3 that were not considered in healthcare literature are "data handling," "performance improvements," and "Improved data understanding."
Nonetheless, even though they did not appear in literature for the healthcare domain, it does not mean that these categories' techniques are not relevant or helpful in this regard. For example, data handling could help deal with irregular sparse time series data expected in healthcare organizations like hospitals. Improving data understanding could also be relevant to patients who need to understand healthcare-related devices that they need to use without a doctor's assistance. "Performance improvements," on the other hand, could be focused on improving classifiers to get better predictions, for example. A visualization of the categories considered in healthcare literature can be seen in Figure 9. Figure 10, on the other hand, gives an overview of how many papers were analysed per category.

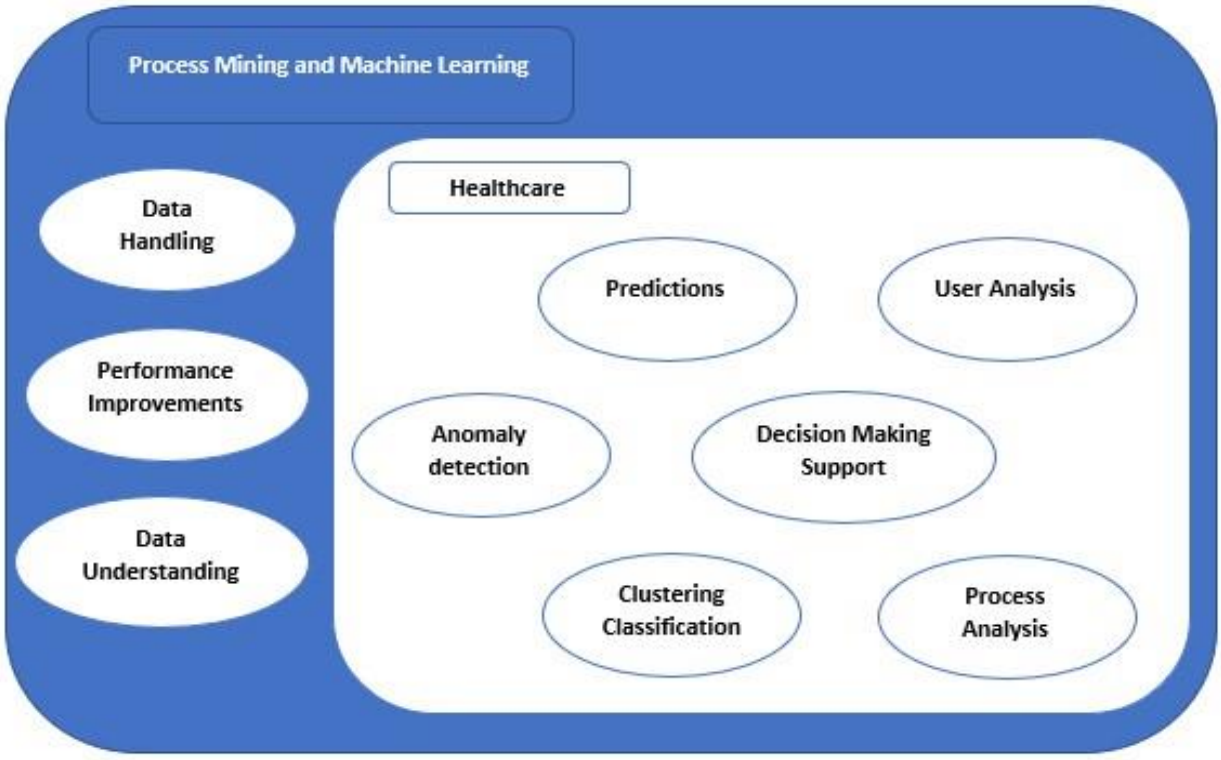

FIGURE 9: CATEGORIES OF RQ2 AND RQ3 APPLIED TO HEALTHCARE RESEARCH

3.5 MACHINE LEARNING AND PROCESS MINING

Another critical insight gained by reviewing the relevant papers is the interaction between machine learning and process mining. The use of machine learning techniques turned out to be taking place not

just at the same time as the process mining techniques, but sometimes before and sometimes after the use of process mining. This insight is visualized in Appendix J. As shown in the graph, most papers use machine learning techniques in conjunction with process mining, but many others use machine learning before process mining, for example, for pre-processing or after process mining, for analyses. The term "Together" in the graph is meant to describe papers that use machine learning techniques during the execution of running cases. Eleven papers do not use machine learning during process mining but instead presented their ideas to complement process mining techniques. These papers are categorized under "Additional."

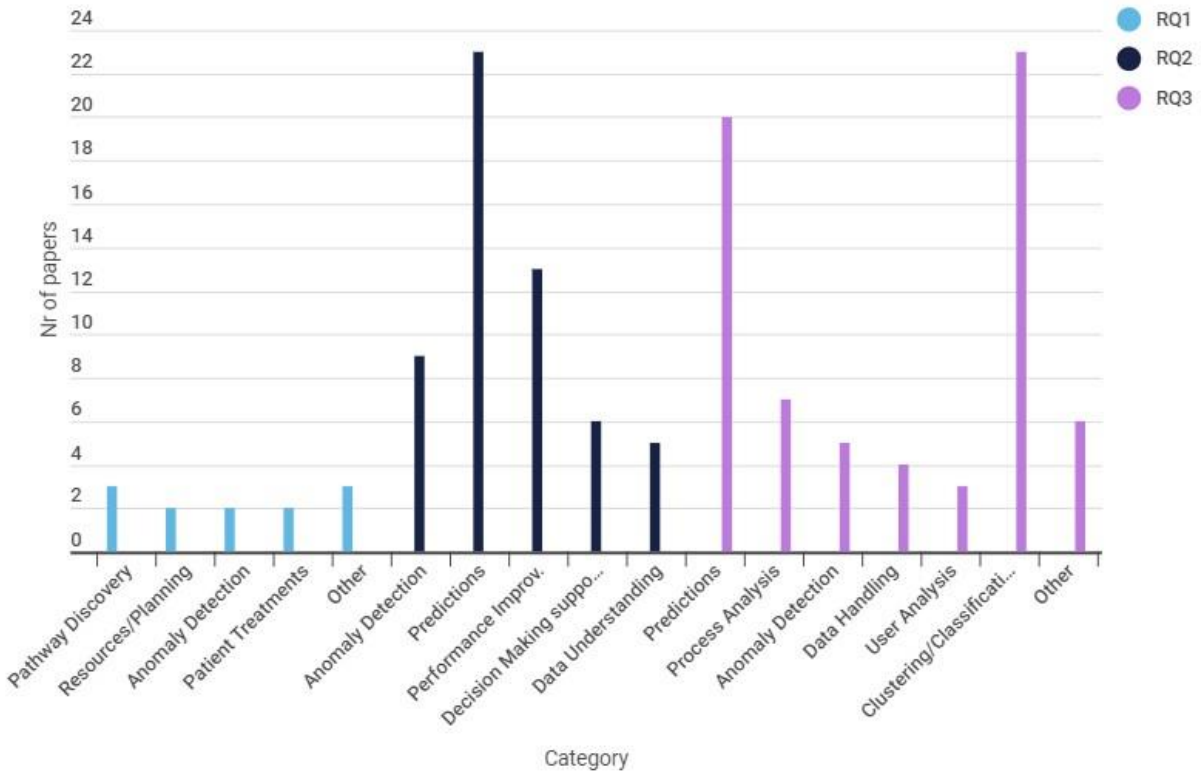

FIGURE 10: NUMBER OF PAPERS FOR THE DIFFERENT CATEGORIES IN EACH RQ

TABLE 6: MACHINE LEARNING AND PROCESS MINING

Together	[96], [91], [112], [79], [47], [4], [117], [165], [107], [99], [20], [65], [126], [110], [38], [82], [81], [41], [14], [5], [13], [74], [78], [97], [132], [64], [185], [116], [76], [48]
Pre	[3], [158], [166], [100], [101], [171], [79], [136], [186], [164], [31], [1], [55], [151], [27], [13], [187], [77], [125], [189], [168], [18]
Post	[158], [176], [147], [6], [188], [66], [13], [179], [150], [103], [94], [132], [71]
Additional	[52], [19], [148], [130], [85], [106], [83], [170], [63], [93], [53]

Which way the papers made use of machine learning techniques is summarized in Table 6. An example of a paper that used process mining and machine learning "together" is [117], which attempts to predict the following activity during process mining. This prediction is made using stacked inception CNN

modules, which is a machine learning technique. Another example is [20], which predicts the future behaviour of running instances in the process. Finally, this category's last example is paper [64], which predicts quantitative performance indicators for running process instances. As seen from the example and from reading the paper descriptions in Table 24, most papers in this category focus on prediction tasks. One can conclude that machine learning is most often used for prediction activities during process mining.

When looking at the "Pre" category, most papers use machine learning techniques to pre-process the datasets before using process mining. An example of this is [186], which clusters users' behaviour patterns to improve process mining results. [151] is another example that uses machine learning techniques before the process mining task. That paper makes use of a software process classifier for the labelling of missing activity attributes through the use of a naïve Bayes approach. Finally, the last example for this category is paper [3], which uses trace clustering as a pre-processing step to split up the event log into clusters of similar traces. Most papers in this category use machine learning for data pre-processing to make the data suitable for process mining or enhance results.

For the third category, "Posts," there are many examples that could be used. In these papers, machine learning is mainly used to improve the understanding of the process mining results or give recommendations. This can be seen in [71], which uses a cluster analysis to identify students' learning strategies after utilizing process mining techniques. Another example is [179], which identifies features that induce mistakes most commonly. This identification is made using post-hoc explainers. An example of how machine learning can give recommendations can be found in [188], where machine learning techniques are used to discover subgroups of cases where hospital treatments have a high causal effect on desired outcomes. These subgroups are then used to give treatment recommendations. This technique is used after their introduced action rule mining approach.

The last category, "Additional," contains papers that did not fit in any previous categories and only used machine learning techniques complementary to process mining. An example of this is [19], which presents machine learning and process mining techniques as part of a holistic toolbox that can be used in combination. It does, however, not specify whether they should be used simultaneously or whether machine learning should be used before or after the process mining techniques. This also applies to the other papers in this category. Hence it is called "Additional."

3.6 USE OF PROCESS MINING & MACHINE LEARNING IN THE REMAINDER OF THIS THESIS

Within the remainder of this thesis, both process mining and machine learning techniques are used to answer the research questions defined in section 1. In particular, machine learning techniques are used to create prediction models for anaemia, type of anaemia, and severity, while process mining is used to increase the data understanding of these results. In Figure 9, this would fit into the "Data Understanding" category, which did not get much attention yet in the literature. More specifically, process mining is used to evaluate the current diagnostic process for anaemia based on the biomarker sets produced in the machine learning step. Accordingly, this thesis starts with machine learning first and then do the process mining analysis. In Table 6, this would fit into the "Pre"-category. More details on the activities and the purpose of these techniques within the context of this thesis can be found in the next section, which is the methodology section.

3.7 THREATS TO VALIDITY

The biggest threat for this literature review in terms of validity is the potential incompleteness of the literature matrix. This threat was mitigated by defining solid inclusion and exclusion criteria that increase the likelihood of a valid representation in the field of process mining and machine learning. Furthermore, by only considering English literature, potentially relevant papers in other languages were ignored.

3.8 CONCLUSION

This section is going to summarize the findings from the results section.

While the first research question has the lowest number of relevant papers for its answer, there are enough to categorize the results to see which applications is most important in contemporary healthcare literature. The categories identified for this sake are "pathway discovery," "resource/planning," "anomaly detection," and "patient treatments."

For the second research question, machine learning techniques/methods/approaches relevant to process mining are identified and categorized to get a better overview. The categories identified in the literature are "anomaly detection," "predictions," "performance improvements," "decision making support," and "Improved data understanding."

For the last question, machine learning problems are identified in the literature relevant to the process mining domain. Categories are defined for them as well, which include "predictions," "process analysis," "anomaly detection," "data handling," "user analysis," and "clustering and classification."

This literature review's main conclusion is that there is still much room for machine learning and process mining applications in healthcare. While many techniques of these domains are already in use in healthcare, many others are still unused and could benefit healthcare organizations. Some machine learning and process mining techniques, on the other hand, did not receive any attention yet in healthcare-related literature. These techniques are related to either data handling, the improvement of data understand, or performance improvements of existing techniques. Focusing on them in the future could fill the research gap that is currently present in this regard.

4 METHODOLOGY

This section explains the methodology used to answer the research questions defined in section 1. It starts with section 4.1, describing how the research questions are applied to the CRISP-DM process, which is used as a methodological tool for answering these questions. Section 4.2 illustrates the machine learning process used to create prediction models, which acts as the backbone of this research. Information on the dataset can be found in section 4.3, highlighting the data characteristics that should be considered when creating a machine learning pipeline. That section increases the understanding of the dataset and belongs therefore into the "Data Understanding" step within the CRISP-DM model. Finally, section 4.4 defines objectives for the process mining techniques used within this thesis.

4.1 RESEARCH QUESTIONS APPLIED TO CRISP-DM

The CRISP-DM is known as an open standard that describes an approach commonly used by data science researchers. The standard has helped define these research questions and guides this thesis throughout the master thesis. This section highlights the importance of its various steps in the context of this master thesis.

For each research question defined in the introduction section, a cycle within the CRISP-DM model is completed. This work already started with an introduction into anaemia in section 2, which belonged to the "Business Understanding" step. As mentioned before, this section defines process mining objectives in section 4.4, which completes the last tasks in the "Business Understanding" Step. Furthermore, it includes an analysis of the dataset in section 4.3, which completes the "Data Understanding" step in CRISP-DM. From there, section five handles research question 5, which covers the "Data Preparation", "Modelling", as well as "Evaluation" steps. At the end of the section, an update is made on the business and data understanding gained by answering this question. By doing this, the CRISP-DM effectively circles back to the "Business Understanding" step, which can also be observed in Figure 11.

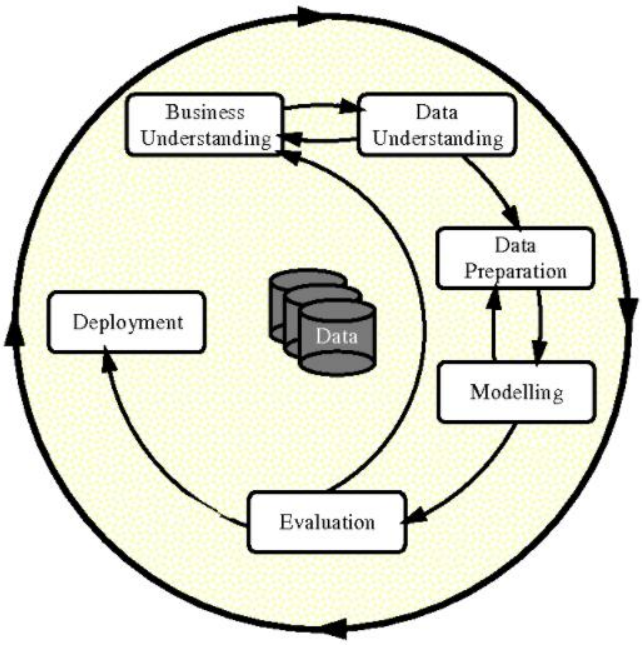

FIGURE 11: CRISP-DM STANDARD [21]

From there, section 6 handles research question 2, which covers the same CRISP-DM steps as research question 1: "Data Preparation", "Modelling", and "Evaluation". Different pre-processing techniques are used in this cycle to examine their effects on the same machine learning techniques used in the previous cycle. At the end of the section, just as in the previous cycle, updates on the business and data understanding is given, starting a new and last cycle. During the last cycle, research question 3 is handled, which covers, again, the "Data Preparation", "Modelling", and "Evaluation" steps in section 7. This time, however, process mining is used as a modelling technique, which

achieves an increased data understanding based on the results in the previous two cycles. More details on both the machine learning and process mining processes can be found in sections 4.2 and 4.4. Finally, after all three cycles have been completed, recommendations are given for the current state in anaemia diagnostics, resulting in proposed changes for the anaemia diagnostic model currently used. This is done in section 8, the discussion section, which handles the "Deployment" step of the CRISP-DM model. A visualization of the three proposed cycles in the CRISP-DM model can be found in Table 7.

TABLE 7: RESEARCH QUESTIONS APPLIED TO CRISP-DM STANDARD

CRISP_DM Cycle	Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
CYCLE 1	Section 2 Section 4.4	Section 4.3	Section 5	Section 5	Section 5	-
CYCLE 2	Section 5.9	Section 5.9	Section 6	Section 6	Section 6	-
CYCLE 3	Section 6.5	Section 6.5	Section 7	Section 7	Section 7	Section 8

4.2 CREATION OF A MACHINE LEARNING PIPELINE

The machine learning process for the anaemia dataset can be summarized into distinct steps illustrated in Figure 12. This machine learning pipeline is used to answer research questions 1 and 2 and is therefore used for the first and second cycle of the CRISP-DM model, as defined in Table 7. However, before the process can start, the data needs to be read and merged because it was provided in 2 different files. R is used as a programming language, and as an IDE, RStudio lends its help. After the data has been read, all rows with no biomarker readings need to be deleted. This is due to data imputation, which requires at least one biomarker reading per entry to infer the other missing values. For the anaemia dataset, this reduces the size of the entire dataset by 32 entries. Afterwards, data types need to be changed, and empty columns deleted. Furthermore, column names that do not follow the standard R naming conventions need to be renamed. Once this is completed, entries are deleted with missing values in the "Sex" and "Haemoglobin" column, as they are essential for defining a target variable.

After all of these steps are finalized, a target variable can be defined. Since the original dataset did not contain one, a custom one needs to be created that carries information on whether the patient in the entry has anaemia or not. A factor variable expresses this with the two factors "Anaemia" and "No Anaemia". This variable is created by considering the WHO anaemia definition in their report "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity", as discussed in the anaemia section [54]. Table 1 illustrates the rules used for the creation of the target variable. Considering these definitions, patients were classified as "anaemic" if they fell within any of the severity ranges within Table 1.

To understand how much the diagnostic standard features contribute to the prediction of different types of anaemia, predictions are also be made using each specific anaemia type as a target variable. To define these target variables, doctor notes containing diagnosis information noted down during patient examinations were used. In total, the dataset contained above 100,000 notes, which needed to be processed to extract the diagnosis information. This processing was done by using natural language processing libraries in R, which helped classify each data entry that contained a doctor note. Out of all the doctor notes, a third had to be excluded because of not determinable anaemia. The rest could be used for classification, however.

After target variables for anaemia and anaemia type have been defined, an additional target variable for the severity must be created. Severity, in this case, is defined as the extent of abnormality in the haemoglobin values of a patient. Technically, severity is a misnomer, primarily because of iron deficiency anaemia, which represents the end stage of iron deficiency and is, therefore, a severe form of iron deficiency in any case. However, the definitions for the predictions made in this section only

refer to the ones created by the WHO in their paper "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity [54]. In their paper, the WHO divides the severity of anaemia into three distinct groups, mild anaemia, moderate anaemia, and severe anaemia. This distinction is made by taking into account haemoglobin values and distinguishing between sexes, age, and pregnant and non-pregnant women. The values for their categorization can be found in Table 1.

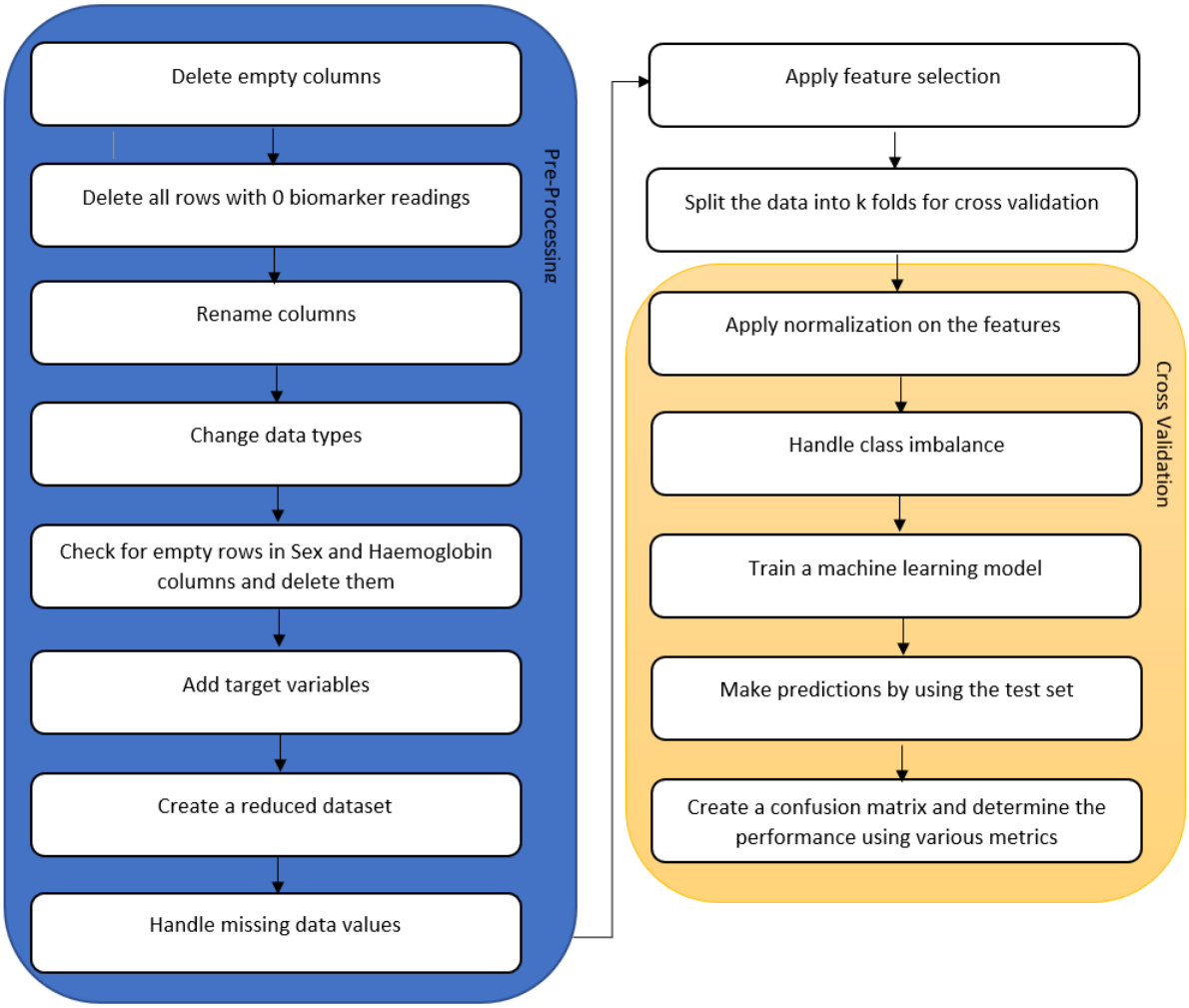

FIGURE 12: MACHINE LEARNING PIPELINE

To evaluate the machine learning results, an expert is used to assess the medical implications of the biomarker selection. The use of expert opinions for evaluation is used in a variety of domains, including healthcare. An example study utilizes expert opinions to extrapolate long-term survival for children and young adults with relapsed or refractory acute lymphoblastic leukaemia [152]. Another example uses experts to provide more stable treatment effect point estimates in analysing low-powered subgroups in clinical trials [149]. In the context of this thesis, biomarkers used in the diagnostic standard are compared to the biomarkers selected by the feature selection technique. By comparing the performance all feature sets have for different machine learning models, suggestions about the current state of diagnostics can be made. The expert, in this case, is going to help in assessing the implications of this analysis.

4.3 DATASET CHARACTERISTICS

The dataset used for this thesis is provided by Medlon BV, a diagnostic company situated in Enschede, Netherlands [175]. It consists of 747,466 entries, with each entry representing the measurements of various biomarkers for one patient. Due to privacy reasons, the dataset was anonymised and there was no possibility to trace back the identities of the patients. The dataset contains measurement data of 144

biomarkers related to anaemia diagnoses, such as haemoglobin, ferritin, or iron. Additional information on patient ID, measurement date, birth date, and patient age are provided. Since the dataset was delivered in 2 parts, one containing the patient information and one the biomarker data, both splits needed to be merged first. This split was done by matching order IDs, which were available in both datasets.

Because of various characteristics of the dataset, pre-processing methods need to be applied before the data can be used in machine learning models. The effect of different preprocessing methods on model performance can be found in section 6, which handles the second research question. One particular characteristic of the dataset is a large number of missing values in many features. This characteristic is illustrated in Figure 13, which graphs the percentage of missing values for the biomarkers from the diagnostic standard. In total, there are 97,542,249 missing values within the dataset, which represent around 79% of the total entries. Because of this large number of missing values, pre-processing techniques for imputation are used, as described in section 6.

A second reason pre-processing methods are needed is the class imbalance within the dataset since the data contains information about whether patients are diagnosed as anaemia positive or anaemia negative. In total, the data contains 113,833 entries, or 16.45% of all entries, where the patient could be diagnosed with anaemia, and 577,956 entries, or 83.55%, where patients were diagnosed to be anaemia negative. Since the anaemia diagnosis information was not provided in the original dataset, the haemoglobin values were used to define anaemia according to the WHO definitions presented in the data pre-processing section [54].

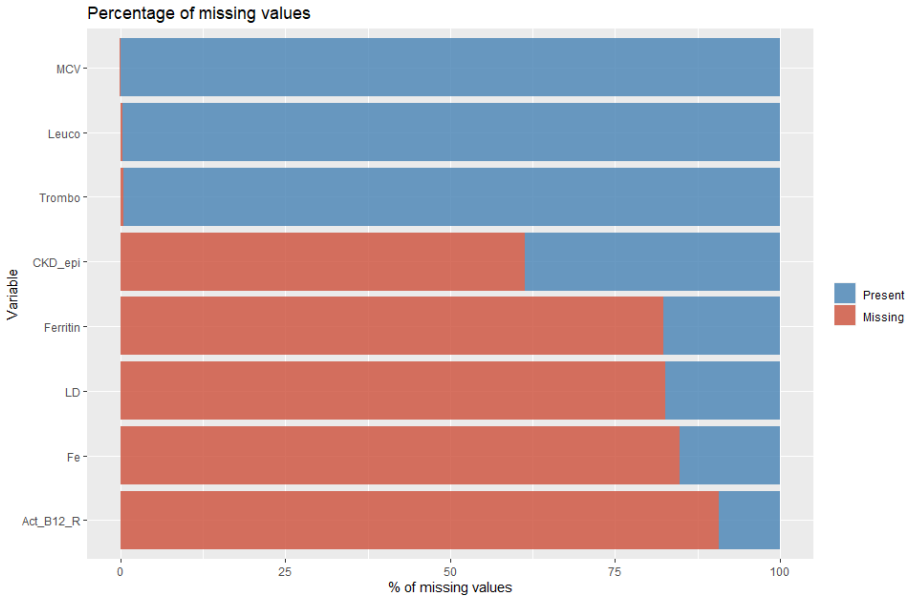

FIGURE 13: PERCENT OF MISSING VALUES FOR DIAGNOSTIC STANDARD FEATURES

Another notable piece of information is the male/female ratio, which lies at around 38/62. While the dataset does provide information about the age ranges of the patients, it does not contain data on whether female patients are pregnant or not. This limitation is important because of the significant effect pregnancies have on anaemia diagnosis [59]. The same can be said about the smoking habits of patients or about whether they live in elevated areas, but since this information was not recorded in the dataset, this thesis does not take these factors into account.

Furthermore, as shown in Figure 14, the data follows the normal distribution for the recorded haemoglobin values. As can be seen in these graphs, not all age groups follow the normal distribution, but this can be explained by the sample size of some of the age brackets. For example, the group of patients younger than five counts 348 entries, representing only 0.05% of the total entries. Other histograms on the distribution of different age groups can be found in the appendix.

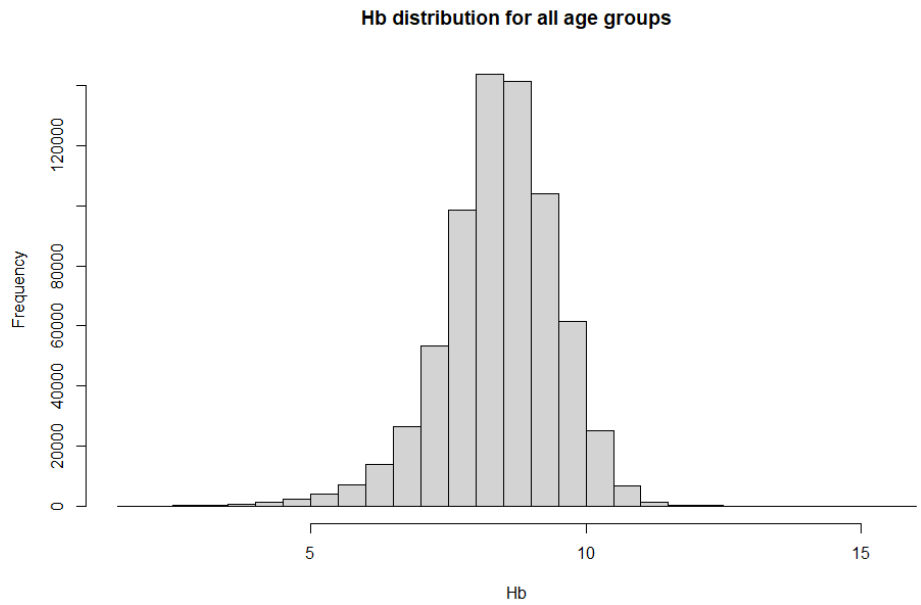

FIGURE 14: RESEARCH QUESTIONS APPLIED TO THE CRISP-DM STANDARD

4.4 PROCESS MINING OBJECTIVES

Once the machine learning results for the standard biomarkers and the other biomarker sets have been compared, process mining evaluates the actual diagnostic process. A comparison is made between the best biomarker set and the biomarkers used in the data. By doing this, recommendations can be made on the choice of biomarkers used in the process, and suggestions can be given on possible model improvements in the diagnostic standard. This is done to answer research question 3 and is covered in the third cycle of the CRISP-DM model, as defined in Table 7.

As described before, process mining can be used for process discovery, conformance checking, and model enhancement. The focus in this thesis is going to be on the first two, discovery and conformance checking. In accordance with the paper written by Ailenei et al. (2011), use cases have been defined that guides the use of process mining to analyse the diagnostic process [57]. These can be found in Table 8.

TABLE 8: USE CASES FOR PROCESS MINING IN THIS THESIS

Process Discovery	Structure of the process	Determine the actual process for anaemia diagnostics in practice.
Process Discovery	Most frequent paths in the process	Determine biomarkers used on the most frequent paths.
Process Discovery	Distribution of cases over paths	Discover common and uncommon biomarkers and behaviour in the case distribution over paths.
Conformance Checking	Exceptions from the expected path	Discover exceptional behaviour in the diagnostics process
Conformance Checking	Compliance with diagnostic standard	Check for differences between the actual process and the anaemia diagnostic standard

The purpose of process mining in this thesis is to increase data understanding, as defined in section 3.6. In particular, conformance to the diagnostic standard needs to be checked, which is necessary to make recommendations for possible changes. Since non-conformance to the diagnostic standard would make it meaningless, recommendations to said standard would make little sense if it were not used in

practice. This conformance is checked through process mining by evaluating the biomarker use across all paths for the diagnostic process. If the biomarkers in the diagnostic standard are, on average, used more often than other biomarkers, conformance is achieved. The relationship of process mining to the machine learning procedure becomes apparent when looking at Figure 2.

5 ANAEMIA PREDICTIONS

As described in section 4, this section shows the results of the machine learning experiments. This section aims to answer the first research question defined in the introduction section. Furthermore, it is completing the first cycle of the CRISP-DM model by preparing the data, making prediction models, and evaluating the results. Section 5.1 describes the different feature selection techniques used to train the models and justify why these techniques were chosen. Continuing from there, section 5.2 handles the cross-validation procedure, which is important to ensure statistical validity to the findings. After this, section 5.3 discusses the order in which the re-processing techniques are to be applied, while section 5.4 is about hyperparameter optimization, and the metrics used to evaluate the classifications. Finally, section 5.5 contains all the findings that were produced for predicting anaemia, as well as an analysis of their performance. 5.6 and 5.7 make additional comparisons for anaemia type and anaemia severity, evaluating which biomarkers selection techniques are best for predicting these two. Finally, section 5.8 contains conclusions, and a summary of the business and data understanding gained. The machine learning pipeline from the methodology section is used to execute all machine learning models.

5.1 BIOMARKER SELECTION FOR PREDICTION MODELS

To identify alternative biomarkers, different feature selection techniques had to be chosen for the machine learning process. First prediction models were produced using the features used in the Dutch diagnostic standard, which covers various anaemia types. Continuing from there, alternative feature sets were created by using different feature selection techniques. These techniques include an embedded method using random forest, LASSO regression, and Pearson's correlation coefficient.

Aside from facilitating data understanding, feature selection aimed to reduce training times and storage requirements [58]. Furthermore, it helped in excluding redundant features and features of low importance. There are three different types of feature selection methods: filter, wrapper, and embedded methods. While filter methods select biomarkers independently from classification, embedded methods use feature selection as an integral part of the classification model. On the other hand, Wrapper methods select or iteratively eliminate variables, all while utilizing the prediction metrics of the prediction model [9].

The embedded method using random forest was implemented using the ranger package in R[131]. To this end, the permutation variable importance approach is used, which considers a feature necessary if there is a positive effect on the prediction performance once included in the classification model [111]. Research has shown that accuracy and Kappa increase with this embedded method [138], which was why it was chosen for this thesis. This thesis's second feature selection technique is LASSO, which stands for "Least Absolute Shrinkage and Selection Operator". LASSO is a linear model that uses L1 regularization, and it was chosen for feature selection due to its ability to handle high-dimensional data [144]. As a last feature selection technique, Pearson's Correlation coefficient was used [37]. It is used to assess the strength and direction of linear relationships between different features and act as a baseline for comparison to the other feature selection techniques described in this section.

5.2 CROSS-VALIDATION PROCEDURE

To deal with the variance of the machine learning model, k-fold cross-validation was used for training. Cross-validation as a whole is one of the most widely used resampling methods to determine the actual error of predictions made by the models. Some types of cross-validation are single hold-out random subsampling, k-fold random subsampling, k-fold cross-validation, leave-one-out cross-validation, and jackknife [56]. This thesis made use of k-fold cross-validation, which partitions the data into k equally sized folds, which are subsequently used for k iterations of validation on the held-out fold and training on the rest of the folds. With each iteration, another fold is being held out, which leads to the use of the entire dataset [24]. Advantages and reasons to use this method are reduced pessimistic bias by using the whole data, no overlapping folds, no dependencies between iterations, and a guarantee that all folds are used [89].

For choosing the number of folds used in cross-validation, the data size was taken into account. To strike a balance between variance and bias, enough data points must be present in the validation set to test the trained model on unseen data, while, at the same time, the training set needs to be large enough to be able to learn from. For very large datasets, the proportion of data used in the validation set can be lower, as long as it contains enough variation to represent the underlying distribution [161]. In this case, the target variable for predicting anaemia had no missing values, which meant that, after downsampling, over 100,000 data points were left. With 10-fold cross-validation, this would mean that 10,000 entries would be used for the validation, which contains enough variation. However, other target variables, like those classifying haemolysis or bone marrow disease, only contained a few thousand data points for the minority class. This would mean that, after downsampling, the number of observations usable would be significantly lower for these predictions. For these cases, 5-fold cross-validation was chosen as a result. Aside from the fold numbers used during training, it was decided to perform repeated k-fold cross-validation to handle the variation in the data splitting.

5.3 ORDER OF PRE-PROCESSING TECHNIQUES

Regarding the pre-processing techniques, much consideration was given to the order of missing value handling, class imbalance handling, normalization, and feature selection. Reasons for considering the order of these techniques are possible data leakage between train and test data, introduced bias and computational limitations. One example of this was whether feature selection or missing value handling should be applied first. While two of the three feature selection techniques required the absence of missing values, doing the missing value handling first exceeded the computational limitations available for this research. In particular, the computation time of multiple imputation methods did become exceedingly long for high dimensional data. Because of this, it was decided to utilize a single imputation method using the variable median for each missing value in the dataset. The feature selection was then applied afterwards. A complete comparison of different missing value handling techniques can be found in section 6.4.

Another decision to be made affected the use of the class imbalance handling techniques. Possibilities were to apply them inside or outside of the cross-validation procedure. A problem when using sampling techniques outside of cross-validation is overfitting caused by resampled data instances. By duplicating entries before splitting into train and validation sets, duplicate entries end up being in multiple folds. If using 10-fold cross-validation on a dataset with a 9:1 class imbalance, there will be on average one instance of each minority class entry in every fold, which leads to overfitting of the training model.

Additionally, oversampling a dataset with a significant class imbalance can lead to bias since the generated artificial data may not represent the actual distribution of the variables. This is not an issue when under-sampling. However, using fewer data leads to decreased performance. For this master thesis, initial predictions were made using under sampling due to the relatively big data size. With around 700,000 entries and an 80/20 class imbalance, downsampling resulted in a balanced dataset with over 100,000 data points, which is plenty for the production of prediction models.

Nonetheless, because some target variables, in particular the ones classifying bone marrow disease, haemolysis, and vitamin B12/folic acid deficiency anaemia, have a much more significant class imbalance (99:1), a comparison was made between the machine learning models using the downsampling approach and machine learning models using other techniques, like oversampling, SMOTE, and ROSE. This comparison can be found in section 6.3. However, just downsampling was used to predict anaemia in general, where the target variable has an 80:20 imbalance. The same was done for the target variables classifying anaemia of chronic disease, iron deficiency anaemia, and severity of anaemia, which all had plenty of data points left after downsampling.

5.4 HYPERPARAMETER OPTIMIZATION & PERFORMANCE METRICS

For hyperparameter tuning, a grid search approach was used. For the KNN models, the number of neighbours (K) was tuned to find the best performing model among them. The K range used for this sake ranged from one to thirty. For the Naïve Bayes algorithm, tuning was done on the Laplace smoothing parameter, which handles the problem of zero probability in Naïve Bayes. While there are not many categorical variables in the dataset, the present ones have a low number of entries, making it possible for categories to be in the testing set that were not present in the training set. This is handled by introducing the smoothing parameter α into the classifier formula, which prevents the probability from becoming zero. For the last machine learning algorithm, the random forest, two hyperparameters were used for the grid search. The first one was the number of trees, and the second one was the number of variables considered at each split. For the number of trees, models were trained with either 16, 32, 64, or 128 trees, while for the variable number considered at each split, models were trained with 2, 3 or 4 variables considered. Experiments with higher tree numbers were also done, but due to diminishing returns and high computational costs, the maximum number of trees used during training was 128.

For the evaluation of the models, a variety of metrics were used to assess their performance. The two most important metrics used for selecting the best hyperparameter settings during cross-validation were Kappa or Sensitivity. The specific metrics choice per parameter is justified in subsequent sections. Further metrics recorded were accuracy, specificity, recall, precision, detection rate, detection prevalence, F1, balanced accuracy, and the 95%-confidence interval for the accuracy. For the cross-validation, the standard deviation for the Kappa and Sensitivity were recorded to check for overfitting. Complete metrics results for all models in this section can be found in the appendix.

5.5 PREDICTING ANAEMIA

As mentioned in section 5.1, multiple feature selection techniques were used to select the best biomarker set. The techniques used were an embedded method using random forest, the LASSO algorithm, and simple correlation. The eleven most important features selected through these techniques can be seen in Table 9, where feature set 1 represents the features selected using the embedded random forest method, while feature set 2 contains features from the LASSO method and set 3 features from the filter method using correlation.

The standard feature set contains eleven features, which represent the features currently used in anaemia diagnostics. It was expected that the random forest models for both feature selection and model training would lead to overfitting, which was a reason to try these feature sets with other machine learning models like KNN and naive Bayes. Regarding the number of features to be selected, a cut-off point needed to be chosen to give accurate predictions and not exceed computational limitations due to high complexity. To make the feature sets one, two and three comparable to the predictions made from the standard feature set, the top eleven features were selected for each feature selection technique. A complete overview of all selected features for each prediction can be found in the appendix.

The features selected in the feature selection for the anaemia predictions can be found in Table 9. As can be seen from the results, some differences can be found in the features used in the diagnostic process. In particular, features that were not considered in any of the feature sets are Trombo, Leuco, and Foliumz_R. Features resulting from the feature selection that are not part of the standard feature set are Transf_verz, Ret_He, TYBC, MCHC, Kreat, RDW, GF_MDRD, Na, and Ureum. Using these feature sets, predictions were made using various machine learning algorithms to evaluate their performance.

TABLE 9: FEATURE SET FOR PREDICTING ANAEMIA

Standard Feature Set	Feature Set 1	Feature Set 2	Feature Set 3
Fe	Ery	Ery	Ery
Ferritin	MCV	TYBC	RDW
MCV	Transf_verz	MCV	MCHC
Act_B12_R	Ret_He	MCHC	Reti_He
LD	TYBC	Transf	MCH_n
Leuco	MCH_n	Ret_He	GFR_MDRD
Trombo	Transf	Kreat	Kreat
CKD_epi	Reti_n	Ferritin	Ferritin
Foliumz_R	MCHC	RDW	Na
Reti_n	LD	Transf_verz	Ureum
Transf	Ferritin	Fe	CKD-epi

For the evaluation of the machine learning models, different metrics were used to assess the performance. In particular, for predicting Anaemia, Kappa was used to select the hyperparameter setting after the cross-validation. Kappa was used to evaluate inter-rater reliability, which is valuable in a medical setting because of its ability to take chance agreement into account [108]. Hyperparameters that were tuned are the number of neighbours (K) in KNN, the number of trees (Trees) and the number of variables randomly sampled (mtry) in Random Forest, as well as the Laplace smoothing in Naïve Bayes. Other metrics to evaluate the results are the accuracy, the sensitivity, the specificity, and the 95%-confidence interval for the accuracy. Especially the sensitivity is of great importance in this research context, as it represents the proportion of correctly diagnosed anaemia cases. This is important for the use case of anaemia diagnostics since not detected anaemia-positive patients bear more weight than falsely diagnosed patients without anaemia. Because of this, more weight should be given to the sensitivity, which makes it more important than the specificity or the accuracy. A complete collection of metric results for each prediction with metrics not considered for this analysis can be found in the appendix.

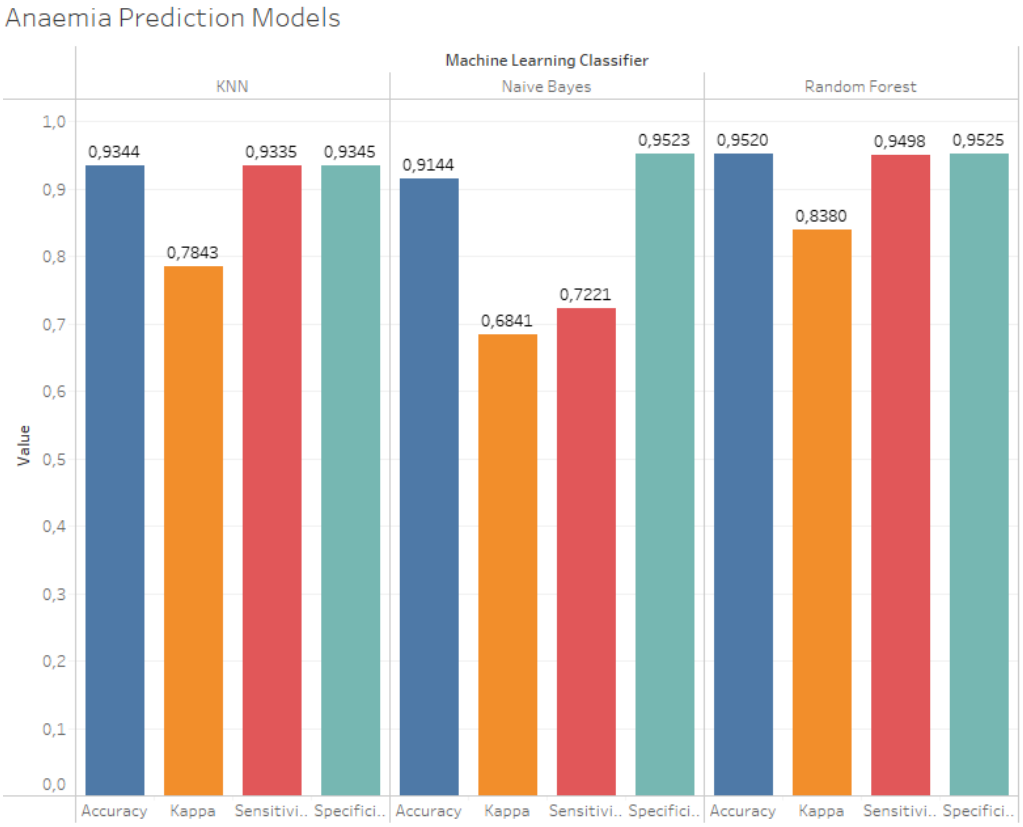

FIGURE 15: ANAEMIA PREDICTION MODEL RESULTS

When looking at the results in Figure 15, it becomes apparent that the Random Forest, when considering the accuracy, has the best performance out of the three machine learning classifiers. This holds not only for the standard feature set but also for the other feature sets using the feature selection methods. Furthermore, it can be seen that the feature sets resulting from these methods generally perform better than the feature set resulting from the standard diagnostic procedure, with feature set one performing the best across all metrics. When evaluating the Kappa for these predictions, it can be seen that the random forest method has the best interrater reliability, with values that exceed 0.8, indicating substantial agreement [108]. Results from the Naïve Bayes and KNN, on the other hand, only indicate moderate agreement. While the sensitivity for the random forest and KNN results is high, with values over 90%, Naïve Bayes results are significantly lower, ranging from 50% to 70%. Aside from this, the other metrics also seem to be lower for the Naïve Bayes compared to the other classifiers. This difference might be because of the algorithm's independence assumption between features, which is not realistic for biomarker data. This is because of the complex interplays between systems and processes within the body that all influence each other, leading to biomarkers influencing each other.

While it could be concluded that the biomarkers in the defined feature sets are superior to those in the diagnostic standard, this conclusion would disregard the nature of the purpose of the standard. The standard focuses on identifying the type of anaemia in each patient, while these predictions are for predicting the occurrence of anaemia in general. So even if the diagnostic standard biomarkers may not be ideal for predicting anaemia in general, they might still be optimal for diagnosing specific types of anaemia, which is what the standard ultimately is about. Because of this, more predictions need to be made using the type of anaemia as a target variable, which is covered in the following few sections.

5.6 PREDICTING TYPE OF ANAEMIA

While a multinomial classification was considered at first for predicting anaemia type, this idea was disregarded. The reason for this is that many data entries contained doctor comments that would

classify the patient into multiple types of anaemia. Because of the interplay of many processes within the body, some biomarker results could not be interpreted as one type of anaemia and were therefore put into multiple categories. This ruled out multinomial classification, which required each data entry to be assigned to one classification. Another possibility would have been to do a multi-label classification, which handles this issue, but libraries in R for this purpose are not mature yet, which led to the abandonment of this idea. Instead, the anaemia type prediction is treated as a binary problem, where predictions are made for each anaemia type separately.

5.6.1 TARGET VARIABLE DEFINITION

For iron deficiency anaemia, doctor notes were searched for the keyword “ijzergebrek”, the Dutch word for iron deficiency. The requirement was that the word would be preceded or followed by the keyword “anemie”, the Dutch word for anaemia. Adding the word anaemia was necessary for the search since iron deficiency could also occur without anaemia. This is because anaemia is only developing in the last stages of iron deficiency, indicating a severe state of iron deficiency [54]. The case of all keywords was ignored during the search to consider lower and uppercase versions of the search terms.

After predicting iron deficiency anaemia, the following type of anaemia handled was anaemia of chronic disease. The expectation for this type of anaemia was that it would result in similar results as iron deficiency anaemia since many data entries were classified simultaneously as iron deficiency and anaemia of chronic disease. Furthermore, the features in the diagnostic standard used to diagnose this type of anaemia were the same as for iron deficiency anaemia, which can be seen in the first column of Figure 3. For the dataset extraction, the doctor's comments were searched for the keyword “chronische ziekte”, which is the Dutch for chronic disease. Just as with iron deficiency anaemia, this should be followed or preceded by the keyword “anemie”, with is the Dutch expression for anaemia. Searching for both these keywords together was necessary because of the occurrence of chronic kidney disease in the dataset, which was not searched for in the context of the intended predictions.

Vitamins B12 and folic acid deficiency anaemia are predicted together because of their similar effects and causes, often developing because of a nutritional deficiency. Because of this, both of these anaemia types were combined for the predictions, even though there were not many instances where both types of anaemia were diagnosed together. The keywords used for the search in the comment’s column were “B12”, preceded or followed by “anemie”, or “foliumzuur”(folic acid), also preceded or followed by “anemie”. Contrary to the previous two anaemia types, this type was significantly less common, with only a few thousand occurrences, while the last two types had around 25,000 classifications. Because of the smaller dataset and the fact that downsampling was used for the class imbalance handling, the validation set ended up being smaller. Because of this, it was decided to train the models using 5-fold cross-validation instead of 10-fold cross-validation. This way, the proportion of data used for the validation set is larger, ensuring that enough variation is available to represent the underlying distribution.

For the prediction of bone marrow disease, standard features used are Leuco (white blood cells), Trombo (platelets), as well as Reti_n (Immature red blood cells). For the classification within the dataset, the comment column should contain the keywords “beenmergaandoening” or “beenmerg”, which are the Dutch words for bone marrow disease and bone marrow, respectively. Bone marrow is essential in anaemia diagnostics since it is the place of production of the red blood cells. If there is an issue with the bone marrow, it could result in insufficient production of red blood cells or dysfunctional red blood cells. Because bone marrow is also the place of production for white blood cells and platelets, biomarkers related to these cells are also important.

Finally, predictions were made for haemolysis, which is the rupturing of the red blood cells and releasing their contents into the bloodstream. When this happens in large quantities, a patient may develop anaemia if more red blood cells are destroyed than can be replaced by production within the

body. For the search through the doctor comments, the keywords “hemolysis” and “afbraak” were used, the latter being the Dutch word for breakdown, destruction, or decomposition.

5.6.2 FEATURE SELECTION FOR TYPE OF ANAEMIA

The feature sets identified for all types of anaemia can be found in the appendix, together with the full set of metrics recorded for each prediction model. Contrary to the anaemia prediction models from 5.5, only the features belonging to a specific datatype were selected for the standard feature sets. For iron deficiency anaemia, iron, ferritin, MCV, and transferrin are used for the diagnosis. For the alternative feature sets, the same feature selection techniques as in 5.5 are used. Eleven biomarkers were chosen for the alternative biomarker sets to discover a larger variety of biomarkers not present in the diagnostic standard.

5.6.3 PREDICTION RESULTS

The best prediction models for each anaemia type can be seen in Figure 16, together with metrics values for the accuracy, Kappa, sensitivity, and specificity of each prediction model. More complete results with all prediction model performances can be found in the appendix.

It was surprising for the iron deficiency anaemia results to see a higher performance across most metrics compared to the anaemia predictions in the last section, even though a smaller sample size was used with relatively similar biomarkers. When analysing the difference between the standard set performance and the metric values of the alternative biomarkers, it becomes apparent that the alternative features perform better than the diagnostic standard features. In particular, Random Forest performed best in total, with the highest values for each metric. Just as with the anaemia predictions, feature set 1 performs slightly better than the other feature set, at least when comparing the sensitivity and the Kappa values. In particular, feature set 1 in Table 31 had the highest values in sensitivity, with 99.02%, and 0.8326 for the Kappa, which can make this model very usable for anaemia diagnostics. However, the difference here is smaller than in the previous predictions, and more features were used as well. Like the anaemia predictions, the Naïve Bayes classifier performed worst across most metrics again, but with a more negligible difference than before.

When making predictions for anaemia of chronic disease, it became clear that the performance would also be similar to the one from the iron-deficiency anaemia predictions since similar features were being used. Because of that, the models covering anaemia of chronic disease have the same characteristics as with iron deficiency anaemia, with feature set 1 and the random forest classifier being superior to the other models created. The performance is, however, slightly worse than the metric values from the iron-deficiency anaemia predictions.

The machine learning results for vitamin B12/Folic acid deficiency anaemia show a significant performance improvement using feature selection techniques. As with the other models investigated, random forest performs slightly better than the other machine learning classifiers, especially when considering the sensitivity values. Furthermore, feature set one also performed better than the other feature sets, exemplifying the usefulness of random forest for the prediction of anaemia and its types. Until now, an apparent deviation from the other models is the low Kappa, which is significantly worse than the other predictions made in this section. Across all experiments for this anaemia type, Kappa does not exceed 15%. The high class imbalance could explain this in the test data. The class imbalance method before training the model was only applied on the training set and not on the testing set. The class imbalance handling was to generalize the model so that it would not discriminate against samples from the minority class in an unseen dataset. Since the trained model did take this into account, there was no need to apply the class imbalance on the testing set. This results, for data with a high class-imbalance, in a high no-information rate. As an example, we can take data with a 99:1 class imbalance. While this imbalance was taken into account while training the data, it is kept during testing since this is also how it would be sampled in the real world. If the testing class would now classify all of the points as the majority class, an accuracy of 99% could be achieved, even if the

models guessed the classifications at random. This leads to a high chance of agreement even if the specificity is high. In this case, it makes the most sense to focus on specificity as the primary metric and use it to decide the best hyperparameter configuration during the cross-validation.

Just as with the vitamin B12/folic acid deficiency anaemia, Kappa performs poorly for the bone marrow disease models. Because of the relative rarity of bone marrow disease compared to other conditions like iron deficiency, a significant class imbalance could be observed for these prediction models, which could have caused the low Kappa value. Furthermore, the pattern apparent in the previous prediction seems to be confirmed for this model, with the best sensitivity being produced by the random forest model and the feature set one. An anomaly can be observed in the prediction made by the Naïve Bayes algorithm, which has significantly lower values for the Kappa and specificity and accuracy, at least for feature sets two and three. When investigating the machine learning results from the cross-validation, it becomes clear that some models have classified all points as the minority class, leading to lousy results for these classifications. Average overall runs in the cross-validation, metrics have lower performance as a result. Based on this, it would be worth checking the results for different class imbalance handling and missing value handling techniques to see if the situation improves for these classifications. This is done in sections 6.3 and 6.4.

The metrics resulting from the hemolysis predictions show a low Kappa and highest sensitivity for feature set 1. Contrary to previous experiments, however, KNN seems to be performing better than the random Forest models when considering the accuracy and the specificity. On the other hand, the sensitivity values seem to remain the best in the random forest models. Just like with the prediction for bone marrow disease, the Naïve Bayes metrics show significantly lower values, this time for feature sets one and two.

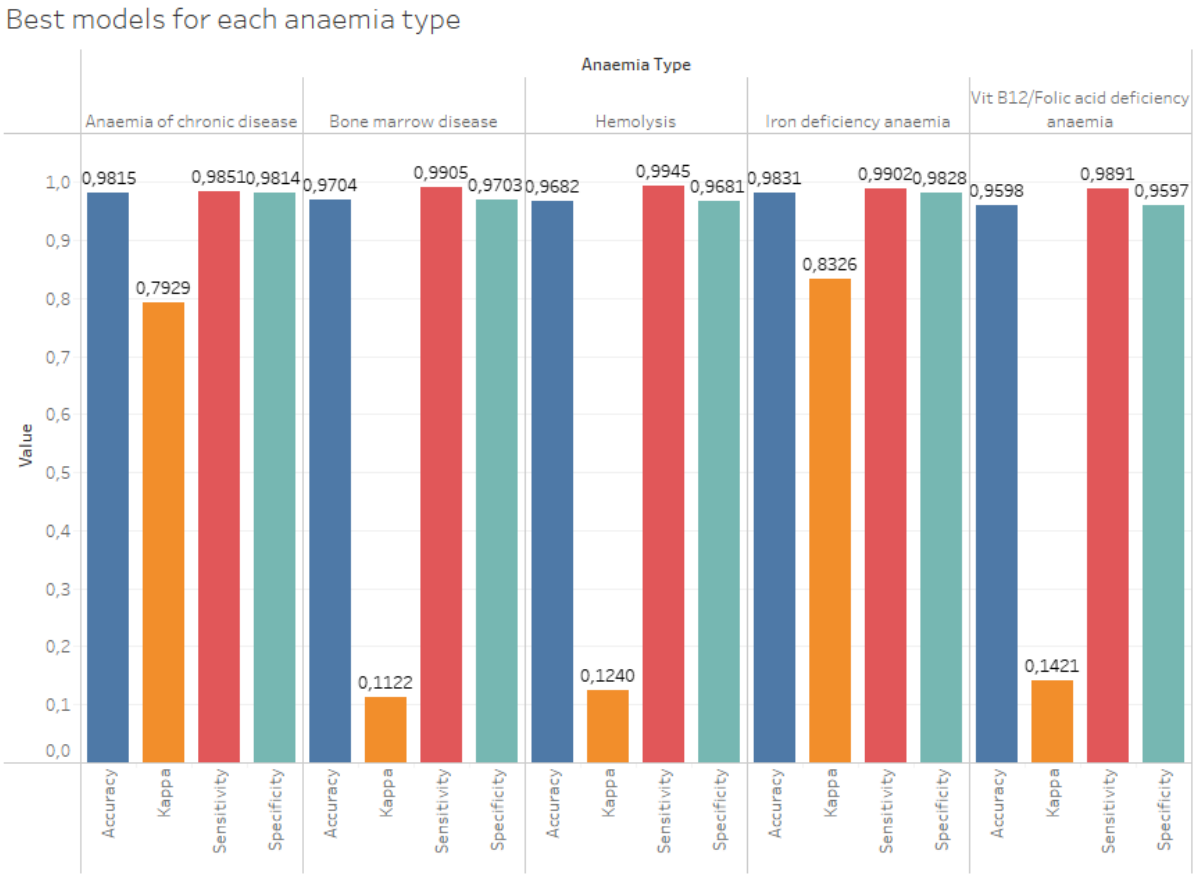

FIGURE 16: ANAEMIA TYPE PREDICTION MODEL RESULTS

5.7 PREDICTING SEVERITY OF ANAEMIA

After anaemia type predictions were covered in the last few sections, this section focuses on the prediction of anaemia severity. Severity, in this case, is defined as the extent of abnormality in the haemoglobin values of a patient. Technically, severity is a misnomer, primarily because of iron deficiency anaemia, which represents the end stage of iron deficiency and is, therefore, a severe form of iron deficiency in any case. However, the definitions for the predictions made in this section only refer to the ones created by the WHO in their paper "Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity [54]. In their paper, the WHO divides the severity of anaemia into three distinct groups, mild anaemia, moderate anaemia, and severe anaemia. This distinction is made by taking into account haemoglobin values and distinguishing between sexes, age, and pregnant and non-pregnant women. The values for their categorization can be found in Table 1.

For selecting the feature, only feature selection method one was used, which was the embedded method using random forest. This method was chosen because of its superiority over most metrics compared to the other feature sets. For the standard feature set, all features from the diagnostic flowchart were used. When looking at the results, feature set one contains two features present in the standard feature set. These biomarkers are Ferritin and MCV. Aside from these, the rest of the biomarkers selected in feature set one were not used in the diagnostic standard. These features are Ery, MCHC, RDW, Transf_verz, Ret_He, MCH_n, TYBC, and Transf. It can be noted that many of these new features did play a role in predicting anaemia and type of anaemia as well, highlighting their importance for anaemia prediction in general.

TABLE 10: FEATURE SELECTION FOR SEVERITY PREDICTION

Standard Feature Set	**Feature Set 1**
Fe	Ery
Ferritin	MCV
MCV	MCHC
Act_B12_R	RDW
LD	Transf_verz
Leuco	Ret_He
Trombo	Ferritin
CKD_epi	MCH_n
Foliumz_R	TYBC
Reti_n	Transf

The same algorithms were used for the predictions themselves as in the previous sections, KNN, random forest, and naïve Bayes. For each of these classifiers, a multinomial classification was run due to more than two classes in the target variable (Mild, moderate, severe). In Figure 17, metric results can be found for the random forest classifier for each severity class. It can be observed that feature set 1 produces better Kappa and sensitivity values for each anaemia severity when compared to the standard biomarker set values. Furthermore, predictions seem to perform better in severe anaemia cases compared to mild and moderate cases. This improvement may be because severe cases have biomarker values that stand out more, whereas mild cases could more easily be mistaken for non-anaemic. The random forest results were picked in Figure 17 because of their superior performance compared to the KNN and Naïve Bayes models.

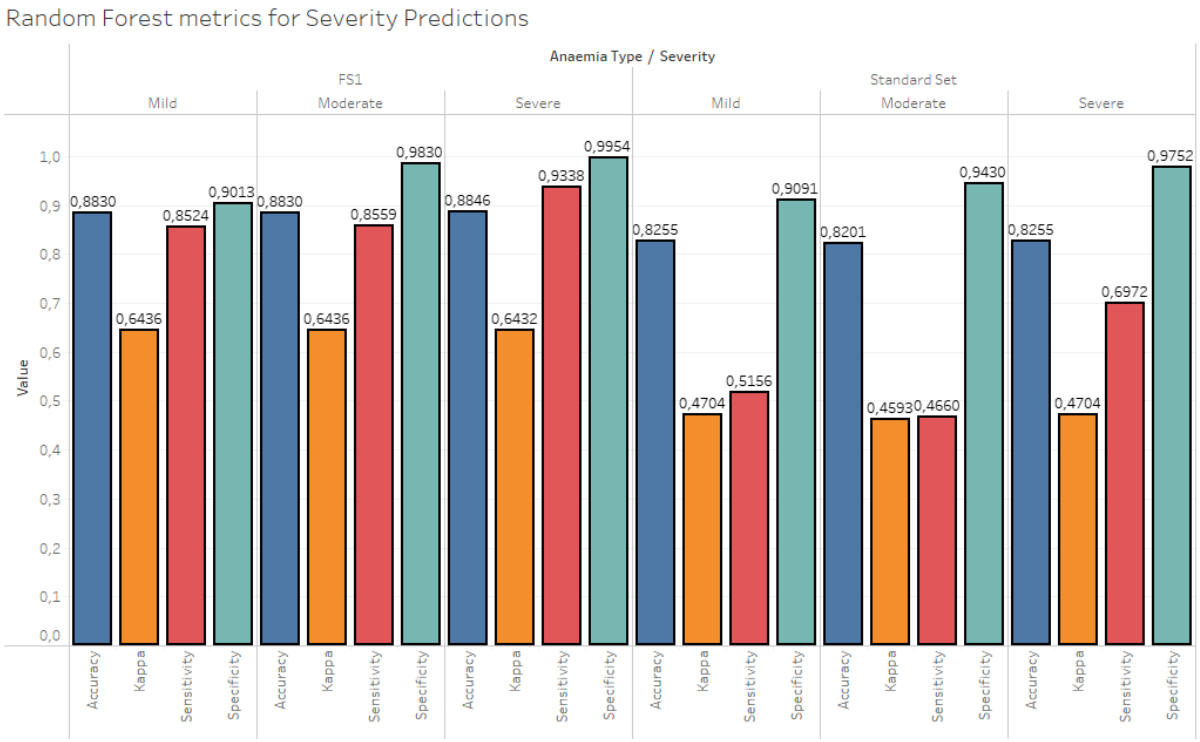

FIGURE 17: ANAEMIA SEVERITY PREDICTION MODEL RESULTS

5.8 UPDATE TO BUSINESS AND DATA UNDERSTANDING

For the prediction models created in this section, it can be concluded that feature selection technique one performed best in combination with the random forest classifier, which delivered superior performance to the other techniques used within the experiments. This indicates that other biomarkers could improve the diagnostic results when used in practice for the diagnostic standard biomarkers. These results, however, need to be validated by using alternative techniques for the missing value imputation and class imbalance handling. Because this section used the median to impute the missing values for variables with a substantial percentage of missing data, resulting in very low variation in their values, the effects of this can be exemplified by looking at the variable Eo, for example. With roughly 90% missing values, median imputation resulted in feature values where 90% were identical. Of course, this results in bias, leading to inflated prediction performance metrics. To address this issue, the next section looks at alternative techniques that can alleviate these issues.

6 EFFECTS OF PRE-PROCESSING TECHNIQUES ON MODEL PERFORMANCE & FEATURE SELECTION

6.1 SAMPLING TECHNIQUES USED FOR CLASS IMBALANCE HANDLING

There are several challenges connected to the use of imbalanced classes in machine learning models. Research has shown that they can lead to worse performance, decreased generalization capabilities for complex data, and small disjunct problems [2]. However, it should also be stated that other factors can also deteriorate the performance, such as sample size and class separability [11]. To handle imbalanced classes within this thesis, three methods are compared, including Synthetic Minority Over-sampling Technique (SMOTE), Random Over-Sampling Examples (ROSE), and as a cost-sensitive method using weights.

The ROSE method uses a smoothed bootstrap approach that generates artificial balanced samples [121] and proved to perform well for use on healthcare data [95]. On the other hand, SMOTE is using an approach where the minority class is oversampled while the majority class is under-sampled at the same time. This approach has proven to result in better results than under- or oversampling alone [127], even though its effectiveness on high dimensional data is reduced [137].

6.2 IMPUTATION TECHNIQUES

Once the target variable has been defined, missing values need to be handled. Since some of the imputation methods used are computationally expensive, however, some preliminary feature selection needs to be done. To do this, an expert opinion is used to select biomarkers that can be taken out from the data set. The use of expert opinions for feature selection is used in a variety of domains, including in industrial settings [23] and healthcare [156]. In medical data mining, expert opinions can improve the sensitivity of classifiers [172]. Furthermore, research has shown that feature selection can improve the precision of prediction models if done before missing value imputation [43].

After the preliminary feature selection is completed, missing value handling is used to impute the data. The reason for dealing with missing values is their substantial impact on precision in machine learning models [128]. Furthermore, missing values reduce statistical power, cause bias in parameter estimation, reduce representativeness of samples, and complicate the analysis of the thesis [75]. The type of missing data observed in the analysed dataset is “Missing at Random”(MAR) since the probability of an entry having a missing value could depend on the observed value but not on the missing value itself [56]. For handling missing values, two imputation methods are going to be compared. The first is the nonparametric missing value imputation method using random forest from the “missRanger” package, and the second is the predictive mean matching method from the “MICE” package [115].

6.3 COMPARING CLASS IMBALANCE HANDLING TECHNIQUES

This section compares the use of alternative class imbalance handling techniques and their effect on model performance for various metrics. The techniques used were downsampling, upsampling, SMOTE and ROSE. For the sake of the model comparison, predictions were made for hemolysis, bone marrow disease, and vitamin B12/folic acid deficiency anaemia. The reason these target variables were chosen was because of their high class imbalance. These comparisons aimed to determine whether using other class imbalance handling techniques would result in performance gains for sensitivity and Kappa. The metric results can be found in figures 1-3. Complete metric information can be found in the appendix.

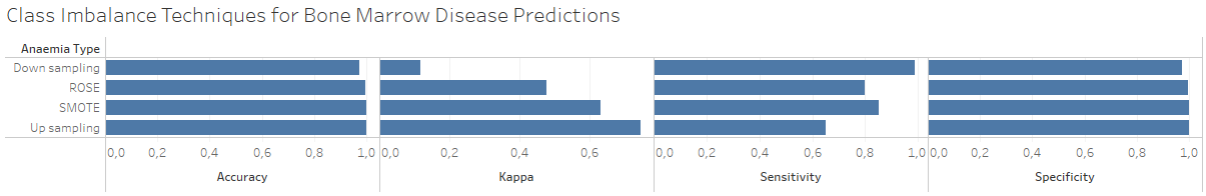

FIGURE 18: METRICS FOR BONE MARROW DISEASE USING RANDOM FOREST

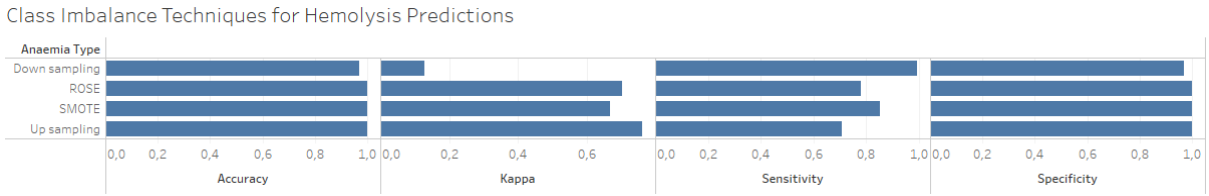

FIGURE 19: METRICS FOR HAEMOLYSIS USING RANDOM FOREST

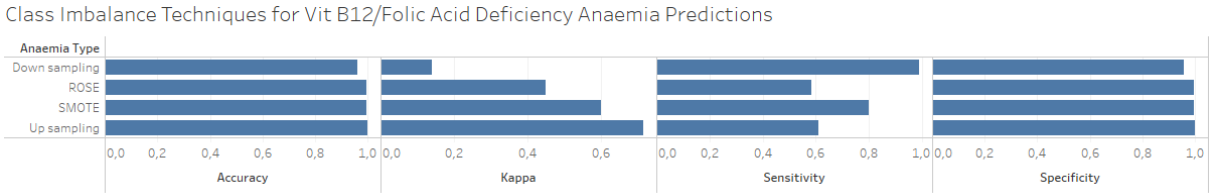

FIGURE 20: METRICS FOR VITAMIN B12/FOLIC ACID DEFICIENCY ANAEMIA USING RANDOM FOREST

As shown in tables 1-3, downsampling produces results with a significantly higher sensitivity than the rest of the models. This means that positive anaemia cases could be predicted correctly more often by using downsampling, even though many data entries are discarded during the use of this technique. The high variation present can explain this for upsampling, SMOTE, and ROSE, expressed in the "SD Sens" column. Because of a high standard deviation for these sampling techniques, overfitting seems likely. This is supported by the fact that there is a high standard deviation for the Kappa metric.

The reason for the overfitting in models using upsampling, SMOTE, and ROSE could be the creation of synthetic data. Because of the high class imbalance, instances of the minority class get copied many times to have the same number of majority and minority class instances. During cross-validation, this creates folds that each contains many copies of the duplicate entries, making it easy to predict the data instances in the training set but making it harder to generalize the results on unseen data. Based on this, it can be concluded that sampling techniques like upsampling, SMOTE, and ROSE, can result in overfitting in case of a significant class imbalance.

Another metric that should be analysed is Kappa, which significantly improves the models using upsampling, SMOTE, and ROSE. All three of these models have in common that they produced results with very high specificity across all models, contributing to the higher Kappa. The classification model using downsampling, on the other hand, produced results with a very low Kappa value, even though there is a high observed agreement for this model. The model that uses this class imbalance technique produced lower specificity values, resulting in a decreased overall accuracy. Because of this, the accuracy is below the no information rate, even though these models have high sensitivity values.

While the conclusions in this section now refer to cases with high to extreme class imbalance, it would make sense to compare them with a case where the class imbalance is less extreme. To do this, the exact class imbalance technique comparisons were performed on a target variable with a less extreme

class imbalance. The target variable used for this context was the iron-deficiency anaemia column, as shown in Figure 21.

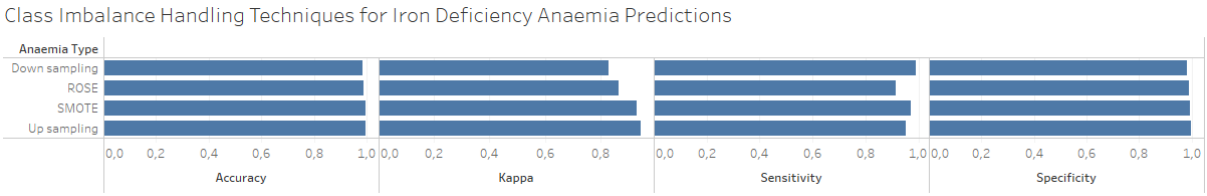

FIGURE 21: METRICS FOR IRON DEFICIENCY ANAEMIA USING RANDOM FOREST

As can be seen from the data, the classification models using upsampling, SMOTE, and ROSE have a lower standard deviation for the Kappa and Sensitivity values compared to the models created before. The reason for this is the similar class imbalance for each of these variables, which leads to a larger dataset after the downsampling. Furthermore, possibly due to the less extreme class imbalance, Kappa seems to be performing well, even for the model using down sampling. Since down sampling is computationally less expensive, it is the preferred class imbalance handling technique for iron deficiency anaemia predictions from here on out. Aside from iron deficiency anaemia predictions, predictions involving anaemia of chronic disease, anaemia occurrence in general, and severity make use of down sampling from now on.

6.4 COMPARING MISSING VALUE HANDLING TECHNIQUES

Previous prediction models in section 5 did use a dataset with imputed values as well, but the imputations were done using the Median. The problem with using the Median lies in the large number of missing values for many variables in the dataset. Because of this, many variables in the previous section ended up with reduced variance and the shrinkage of the standard error, which can invalidate confidence intervals by making their ranges shorter. To solve these problems, a comparison was made using alternative missing value handling techniques, as described in section 6.2. This section compares different missing value handling techniques and compares their impact on feature selection.

After the missing values have been imputed, SMOTE was used to handle the class imbalance present in the dataset. From there, feature selection was performed using the embedded random forest method used in the last section. The results for this can be found in Tables 5 and 6. While the features in

Table 11 refer to the features selected for the missForest dataset, Table 12 contains features selected for the MICE dataset. The complete form of the table with all features included can be found in the appendix.

TABLE 11: FEATURES AFTER USING MISSFOREST

Anaemia	**Iron deficiency anaemia**	**Anaemia of chronic disease**	**Vitamin B12/Folic acid deficiency anaemia**	**Bone marrow disease**	**Haemolysis**	**Severity**
Ery	Transf_verz	TYBC	LD	Ery	Ery	Ery
MCH_n	MCH_n	Transf_verz	RDW	Trombo	RDW	MCHC
Transf_verz	MCHC	Fe	Ery	RDW	Mg	VitB1
Mg	Ferritin	Ery	Reti_n	Mg	Reti_n	RDW
RDW	Ery	Transf	TYBC	NTproBNP	NTproBNP	MCH_n
MCHC	RDW	Ferritin	Ret_He	Leuco	LD	Transf_verz
MCV	Fe	MCHC	Haptoglo	TN_T_HS	Myelo	Myelo
NTproBNP	TYBC	Mg	Transf_verz	Trombo_cit	TN_T_HS	Meta
Ureum	Transf	NTproBNP	Transf	Ureum	MCHC	MCV
Fe	MCV	RDW	MCHC	Neutro	VitB1	NTproBNP
Kreat	NTproBNP	Ureum	MCV	Testost_lum	Meta	Fe

TABLE 12: FEATURES AFTER USING MICE

Anaemia	Iron deficiency anaemia	Anaemia of chronic disease	Vitamin B12/Folic acid deficiency anaemia	Bone marrow disease	Haemolysis	Severity
Ery	Ferritin	TYBC	MCV	Trombo	Reti_n	Ery
MCH_n	Transf_verz	Transf	Ery	Leuco	LD	MCHC
Ret_He	MCH_n	Fe	MCH_n	Ery	Ery	MCH_n
RDW	MCHC	Transf_verz	RDW	RDW	RDW	RDW
Testost_lum	Fe	Ery	TYBC	Segment	Haptoglo	Fe
Transf_verz	RDW	Ret_He	Ferritin	Neutro	Ret_He	Transf_verz
MCHC	Transf	Ferritin	Transf	MCV	MCV	MCV
Kreat	TYBC	RDW	MCHC	Ureum	VitB1	Testost_lum
Ureum	MCV	MCHC	Transf_verz	Ret_He	Mg	Mg
Fe	Ery	Ureum	Foliumz_R	Mg	Ureum	VitB1
MCV	Testost_lum	Kreat	Ret_He	Transf	TE	Ureum

6.5 UPDATE TO BUSINESS AND DATA UNDERSTANDING

As could be seen from the previous two sections, prediction performance and feature selection heavily depend on the technique used to generate these results. Using different feature selection and class imbalance handling techniques, problems from section 5 could be addressed and finally solved. 6.2 showed that interrater reliability could dramatically improve when using upsampling, SMOTE, or ROSE instead of downsampling. Furthermore, choosing different imputation techniques had a significant effect on the biomarkers selections as well. By choosing techniques that generate less bias, more useful features could be identified that can be used for the process mining in the next section.

7 EVALUATING THE DIAGNOSTIC PROCESS

This section answers the third research question. To this end, sections 7.1 and 7.2 handle the sub-questions by utilizing process mining techniques in Disco. In particular, process discovery and conformance checking are used to gain further insight into the actual diagnostic process. The process mining objectives defined in the methodology section are answered, and the data understanding is increased. Section 7.2 compares the actual process and the biomarker sets identified in section 6. By first checking the model's compliance to the diagnostic standard as specified in the methodology section, it is possible to compare these biomarkers with the use of the alternative biomarker sets.

7.1 DISCOVERING THE REAL DIAGNOSTIC PROCESS

This section handles the process discovery objectives defined in the methodology section. To gain further insight into the data through process mining, some pre-processing needed to be done first. In particular, data types needed to be changed to fit the dataset to the requirements defined in the methodology section. While the previous section had an interest in predicting accurate values for each biomarker, this section only requires knowledge about whether a biomarker was used or not. Since the specific measurement values are not crucial for answering the questions for this section, biomarker datatypes were changed to logical. If a biomarker was used for a particular patient measurement, the specific measurement value was therefore changed from a numeric number to a "TRUE" value. Biomarkers that were not used during a measurement had their NAs replaced with "FALSE". By doing this, the biomarker features were effectively transformed into categorical binary variables that indicate whether a biomarker was used or not. Continuing from there, the data was exported to Disco, which then served as a tool for creating various process mining models.

For each model, the patient ID column served as a case ID. As an activity, a combination of the measurement as well as the anaemia column was used. The measurement variable, in this case, defines the measurement number of a patient for a particular data entry. If a patient has their first anaemia biomarker measurement, the measurement column value would be one. For a patient that has a second measurement, the measurement value would be two.

On the other hand, the anaemia column indicates whether the patient has a haemoglobin value that would fall within the definition of anaemia. This column could have values that would either be TRUE or FALSE. Finally, the timestamp variable used was the "OrderDate" column, which contains the time and patient measurement data. The resulting process model can be seen in simplified form in Figure 22.

As can be seen in Figure 22, most patients tend to keep their anaemia between measurements. Nonetheless, there are also cases where patients become anaemic after not being anaemic in a previous measurement or cases where the patients stop being anaemic after previously being anaemic. This can be seen in the paths between measurement one and two, for example. While there are roughly 11.627 patients that kept being anaemic between measurements, there were almost 8,656 cases where patients stopped being anaemic. Similarly, there were roughly 120,000 patients that stayed anaemia negative, while there were 10,765 patients that became anaemic in the timeframe between measurements. The fact that proportionally more patients stop being anaemic shows the success of treatment measures between measurements.

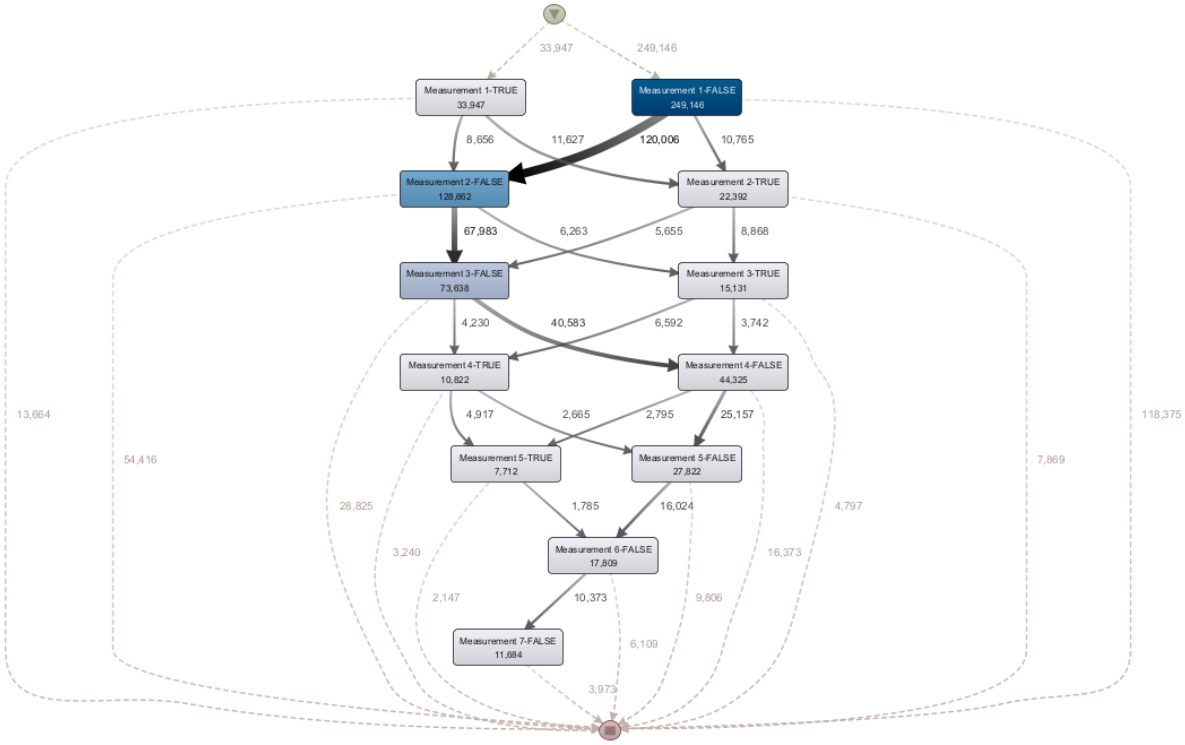

FIGURE 22: A PROCESS MODEL FOR ANAEMIC PATIENTS

Value	Frequency	Relative frequency
Measurement 1	283,107	40.97 %
Measurement 2	151,068	21.86 %
Measurement 3	88,783	12.85 %
Measurement 4	55,161	7.98 %
Measurement 5	35,548	5.14 %
Measurement 6	23,595	3.41 %
Measurement 7	15,886	2.3 %
Measurement 8	10,872	1.57 %
Measurement 9	7,511	1.09 %
Measurement 10	5,246	0.76 %
Measurement 11	3,632	0.53 %
Measurement 12	2,609	0.38 %
Measurement 13	1,868	0.27 %
Measurement 14	1,381	0.2 %
Measurement 15	991	0.14 %
Measurement 16	749	0.11 %
Measurement 17	557	0.08 %
Measurement 18	447	0.06 %
Measurement 19	365	0.05 %
Measurement 20	292	0.04 %

FIGURE 23: MEASUREMENT FREQUENCY FOR FIRST 20 MEASUREMENTS

Figure 23 illustrates the number of measurements more clearly. As shown in Figure 23, 40.97% of all patients only had one anaemia measurement, which equates to roughly 283,107. The rest of the patients had at least two measurements, with a maximum of 60 for the patient with the highest number of measurements. When considering sex, the data contains 425,733 entries (61.6%) where the patients were females and 265,349 entries where the patients were male (38.4%). The age column also indicates that older age groups get more measurements than younger age groups.

Furthermore, as shown in Figure 24, the median case duration took up 21 weeks, the time between the first and last measurement of a case. Additionally, the data contained measurements taken in the time range between the third of January 2011 to the 29th of April 2021. The cases in this time frame are divided between 3095 variants.

Events	691,082
Cases	283,107
Activities	116
Median case duration	21 wks
Mean case duration	27.3 mths
Start	03.01.2011 00:00:00
End	29.04.2021 00:00:00

FIGURE 24: CASE STATISTICS FOR PROCESS MINING MODEL

To determine exceptional behaviour, the most interesting cases had to be filtered out of the process model. Accordingly, filters were applied only to keep those cases where there is a change in the patient's diagnosis. This could, for example, be patients that were once diagnosed with anaemia but had a different diagnosis in later measurements. Likewise, other interesting cases involve patients who did not have anaemia but were later diagnosed with the condition. After applying these filters, 2825 variants were left. To get a better overview of these variants, only cases were investigated whose sequence of activities was shared by at least 114 other cases. This left over 1% of the variants, covering 95% of all cases and 84% of all events. In total, there were 19 variants and 12066 cases left, all of them illustrating the exceptional behaviour as defined before.

When looking at the variants describing the exceptional behaviour, one can see that the most common variants start with a FALSE diagnosis in the first measurement, followed by a TRUE diagnosis in the second measurement, followed by a FALSE diagnosis again in the third measurement. In total, about 22% of all cases could be assigned to this variant. It can be noted that most of the variants start with a FALSE diagnosis in the first measurement, indicating that there are often, at first, no abnormalities in the haemoglobin values of the patient. It could, for example, be that anaemia gets discovered on accident on these occasions, for instance in routine check-ups. It could also be that other biomarkers indicate an abnormality other than the haemoglobin value, which created the necessity for subsequent measurements by the doctor. Aside from this, there are also occasions where the diagnostic process starts with an anaemia TRUE diagnosis in the first measurement, but this occurs less often. It also has to be noted that among all of the 19 variants,there are none that include less than three events, indicating that there are usually more than two measurements for these abnormal cases.

7.2 COMPARING BIOMARKER USE IN THE DIAGNOSTIC PROCESS

To determine how the diagnostic standard biomarkers are used in the process, features contained in the standard were changed to a “Resource” in Disco. By doing this, it became possible to see the combination of biomarkers used for each case. The most common biomarkers for all cases can be seen in Table 13.

TABLE 13: MOST COMMON STANDARD BIOMARKER COMBINATIONS

Biomarker combination	Frequency	Relative Frequency
Leuco-MCV-Trombo	317,844	45.99%
CKD_epi-Leuco-MCV-Trombo	135,219	19.57%
Act_B12_R- CKD_epi-Leuco-MCV-Trombo	31,648	4.58%
CKD_epi-Fe-Ferritin-LD-Leuco-MCV-Reti_n-Transf-Trombo	28,354	4.1%
Fe-Ferritin-LD-Leuco-MCV-Transf-Trombo	18,848	2.73%
Fe-Ferritin-LD-Leuco-MCV-Reti_n-Transf-Trombo	13,636	1.97%
LD-Leuco-MCV-Trombo	13,164	1.9%

As can be seen from Table 13, a few biomarker combinations were used significantly more often than others. However, to determine for which cases each of the common combinations was used, the process model has to be filtered once more. Data entries were filtered based on anaemia type to clarify which biomarker combinations were used for the diagnosis of each type. The biomarkers that were used for the diagnosis of iron deficiency anaemia can be seen in Table 14. The seven biomarker combinations were identified for each anaemia type that occurred most often during the diagnosis process. By comparing the biomarkers used to the biomarkers mandated in the Dutch anaemia standard, one can determine whether the process complies with the diagnostic workflow model.

TABLE 14: MOST COMMON STANDARD BIOMARKER COMBINATIONS FOR DIAGNOSING IRON DEFICIENCY ANAEMIA

Act_B12_R	CKD_epi	Fe	Ferritin	Foliumz_R	LD	Leuco	MCV	Reti_n	Transf	Trombo	ID
	X	X	X		X	X	X	X	X	X	**1**
		X	X		X	X	X		X	X	**2**
			X			X	X			X	**3**
		X	X		X	X	X	X	X	X	**4**
	X		X			X	X			X	**5**
		X	X			X	X		X	X	**6**
X	X	X	X		X	X	X	X	X	X	**7**

TABLE 15: IRON DEFICIENCY ANAEMIA BIOMARKER FREQUENCIES (STANDARD BIOMARKER)

ID	Frequency	Relative Frequency
1	5,577	20.56%
2	3,768	13.89%
3	3,582	13.2%
4	2,923	10.77%
5	1,780	6.56%
6	1,453	5.36%
7	1,147	4.23%

The biomarkers that are of special interest for this type of anaemia are MCV, ferritin, Fe, and Transf, since they are the biomarkers mandated by the diagnostic anaemia standard. As can be seen from Table 14, measurements for MCV, ferritin, Fe, and Trombo were made for all of the seven most common combinations. From the results of the process mining model, it could be determined that MCV was used in 27,122 events (99.99% of all events), while ferritin was used in 23,652(87.18%) and Fe in 17,212 events (63.48%). Lastly, Transf has been measured in 19,315 events (71.27%). Based on these numbers, it can be concluded that the mandated anaemia standard biomarkers are used in most measurements. Tables containing the most common biomarker combinations for other anaemia types can be found in the appendix.

In Table 16, additional frequencies for the diagnostic standard biomarker are shown for each anaemia type. Each column in the table represents the frequencies of the standard biomarkers for the particular anaemia type. The iron deficiency anaemia column, for example, shows frequency values for Fe, Ferritin, MCV, as well as Transf, since the NHG diagnostic standard mandated these biomarkers. For this column, it can be seen that 63.48% of all events in the iron deficiency anaemia positive cases involved the use of Fe, while 87.18% involved Ferritin, 99.99% the MCV and 71.27% Transf. If one compares this column to the rightmost column in the table, one can see the frequencies for all events in the dataset, not just for the iron-deficiency anaemia cases. It can, for example, be observed that 15.24% of all cases in the whole dataset involve the measurement of Fe, while for iron deficiency anaemia, this number laid at 63.48%. This indicates that Fe is used more often for patients with iron deficiency anaemia than for the totality of all patients. This is to be expected since the diagnostic standard mandated the use of Fe for iron deficiency anaemia patients, which means that, at least for this biomarker for this type of anaemia, there is compliance to the diagnostic standard from the NHG. Not only that, but when observing the other values in the table, one can see that the mandated biomarkers for the diagnosis of other anaemia types were used as prescribed for every biomarker, indicating full compliance with the diagnostic model.

TABLE 16: STANDARD BIOMARKER FREQUENCIES PER ANAEMIA TYPE

	Iron Deficiency Anaemia	**Anaemia of chronic disease**	**Vit B12/Folic acid deficiency anaemia**	**Bone marrow disease**	**Hemolysis**	**No filter**
Act_B12_R	-	-	100%	-	-	**9.27%**
CKD_epi	-	-	-	-	-	**38.74%**
Fe	63.48%	87.8%	-	-	-	**15.24%**
Ferritin	87.18%	89.09%	-	-	-	**17.75%**
Foliumz_R	-	-	36.53%	-	-	**4.34%**
LD	-	-	89.57%	-	97.56%	**17.42%**
Leuco	-	-	-	100%	-	**99.75%**
MCV	99.99%	99.99%	100%	-	-	**99.99%**
Reti_n	-	-	58.83%	46.79%	61.95%	**11.38%**
Transf	71.27%	97.11%	-	-	-	**14.51%**
Trombo	-	-	-	99.55%	-	**99.54%**

While these frequencies are helpful in their own right, they should be compared to the frequencies of biomarkers identified in the machine learning models. Since these biomarkers proved to be superior to those used in the standard, it would be advisable to use them more often, or at least as often as the biomarkers defined in the diagnostic standard. Comparing these frequencies is the purpose of Table 17 and Table 18. Each of the tables shows the order of importance for each biomarker in the "Importance Order" column. The importance values for these variables resulted from the MICE imputation in section 6 and showed the ten most important biomarkers for the best prediction performance. The "Biomarker" column, on the other hand, shows the order of these biomarkers according to the usage frequency in the actual process. One can see in Table 17, for example, that Ferritin was ranked as the most important feature by the feature selection method in the previous section, while in practice, its rank drops from one to seven, with six biomarkers that are used more often than Ferritin. This change in position is contained in the "Position Change for importance order" column for each anaemia type. The "Frequency" column, on the other hand, shows the frequency of the biomarkers in the "Biomarker" column. In Table 17, for example, MCV was used in 99.9% of cases where the patient had iron deficiency anaemia.

Table 18 shows the same comparisons, just with focus on the missForest biomarker selection from section 6. Tables for the other types of anaemia for both biomarker sets can be found in the appendix. When investigating the contents of those tables, it becomes clear that the missForest biomarker sets selected more rarely used biomarkers than the MICE biomarker selections. Recommendations, based on those numbers, are given in section 7.

TABLE 17: MICE BIOMARKER FOR IRON DEFICIENCY ANAEMIA COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
99.99%	MCV	Ferritin	-5
99.66%	MCHC	Transf_verz	-7
99.46%	RDW	MCH_n	-3
99.36%	Ery	MCHC	+2
87.18%	Ferritin	Fe	-5
76.97%	MCH_n	RDW	+3
71.27%	Transf	Transf	0
71.07%	TYBC	TYBC	0
71.05%	Transf_verz	MCV	+8
63.48%	Fe	Ery	+6
0.01%	Testost_lum	Testost_lum	0

TABLE 18: MISSFOREST BIOMARKER FOR IRON DEFICIENCY ANAEMIA COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
99.99%	MCV	Ery	-4
99.66%	MCHC	MCH_n	-4
99.46%	RDW	Transf_verz	-4
99.42%	Kreat	Mg	-7
99.36%	Ery	RDW	+2
99.36%	MCH_n	MCHC	+4
76.97%	Transf_verz	MCV	+6
71.05%	Fe	NTproBNP	-1
11.37%	NTproBNP	Ureum	-1
7.55%	Ureum	Fe	+2
1.46%	Mg	Kreat	+7

8 DISCUSSION

This section discusses the implications of the findings found in the previous section. It starts with an evaluation of the anaemia biomarkers selected for anaemia predictions in section 8.1. This includes a discussion on the importance of biomarker selection for the prediction of anaemia type and severity. From there, 8.2 proposes possible changes within the diagnostic process and the diagnostic standard described in the anaemia section.

8.1 IMPORTANCE OF PREDICTION MODELS

As shown in the findings from section 6, the embedded method selected biomarkers with the highest importances for the MICE and missForest dataset. The biomarker set for predicting anaemia from the MICE dataset contained the following metrics: Ery, MCH_n, Ret_He, RDW, Testost_lum, Transf_verz, MCHC, Kreat, Ureum, and Fe. The procedure delegated by the diagnostic standard, on the other hand, contained the following biomarkers: Fe, Ferritin, MCV, Act_B12_R, LD, Leuco, Trombo, CKD_epi, Foliumz_R, as well as Reti_n. This can also be seen in Table 19.

TABLE 19: BEST FEATURE SET COMPARISON TO STANDARD FEATURE SET

Standard FS	FS1	Common Biomarkers	Exclusive Standard FS	Exclusive FS1
Fe	Ery	Fe	Ferritin	Ery
Ferritin	MCH_n		MCV	MCH_n
MCV	Ret_He		Act_B12_R	Ret_He
Act_B12_R	RDW		LD	RDW
LD	Testost_lum		Leuco	Testost_lum
Leuco	Transf_verz		Trombo	Transf_verz
Trombo	MCHC		CKD_epi	MCHC
CKD_epi	Kreat		Foliumz_R	Kreat
Foliumz_R	Ureum		Reti_n	Ureum
Reti_n	Fe			

When comparing these feature sets, it becomes apparent that they only have one biomarker in common, Fe. Since FS1 proved to perform better than the Standard FS in section 5, it can be concluded that the biomarkers currently used in the diagnostic workflow are not ideal for predicting anaemia as a condition. Even though that may be the case, this comparison is not entirely appropriate since the standard FS was intended to diagnose the type of anaemia and not the occurrence of anaemia in general. Nonetheless, this finding gives insight into which biomarkers are the best predictors of anaemia. In the unlikely case that there is no haemoglobin value available, the features in FS1 proved to be the next best predictors for diagnosing anaemia. For diagnostics, this might not be very interesting, which is why predictions were made for the type of anaemia as well. These were discussed in section 5.6.

8.2 POSSIBLE CHANGES WITHIN THE DIAGNOSTIC PROCESS & RECOMMENDATIONS

Changes to current diagnostic processes can be proposed by changing the diagnostic process mandated by the NHG and through practical recommendations on different biomarker use. Since the author of this work has no medical background, an evaluation was made by consulting a medical professional and getting feedback on the biomarker selections. By assessing about whether certain biomarker rankings are to be expected or not, conclusions can be made on future use and possible research opportunities. Assessments were given for both biomarker selections (MICE and missForest), as well as for each anaemia type. One such assessment can be observed in Table 20, which covers the rankings for the biomarkers best suited to perform anaemia predictions. “Ranking 1” refers to the MICE selection, while “ranking 2” refers to the missForest selection. For the eleven most important biomarkers in each selection an assessment was given on whether their high position is unexpected

(O) or expected (X). It can be seen that the selection using the missForest dataset (Ranking 2) contains one variable where the high position is assessed as unexpected, while there are two unexpected biomarkers for "ranking 1". Assessments for the other biomarkers from Table 17 and Table 18 can be found in the appendix.

TABLE 20: EXPERT ASSESSMENT FOR ANAMIA BIOMARKERS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
Ery	X	Ery	X
Transf_verz	X	MCH_n	X
Ret_He	X	Ret_He	X
Transf	X	RDW	X
TYBC	X	Testost_lum	X
Fe	X	Transf_verz	X
Mg	O	MCHC	X
MCH_n	X	Kreat	X
NTproBNP	O	Ureum	O
aTTG_n	X	Fe	X
Ferritin	X	MCV	X

When comparing the unexpected biomarkers with the results from section 7, it becomes clear that there is a large overlap between biomarkers that are used very rarely and biomarkers whose ranking was assessed as unexpected. Even though there may be an overlap, biomarkers where the high ranking was not expected might be assessed this way because of their rare use in practice and not because of their importance. This might be subject for future research. All of the unexpected biomarker in the top eleven for each anaemia type can be found in Table 21.

TABLE 21: BIOMARKER THAT WERE ASSESSED TO BE UNEXPECTED

Anaemia	Iron Deficiency Anaemia	Anaemia of Chronic DIsease	Vit B12/Folic Acid Deficiency Anaemia	Bone Marrow Disease	Haemolysis	Severity
Ureum	Testost_lum	Haptoglo	Transf	aTTG_n	Mg	VitB1
Mg		TN_T_HS	TYBC	NTproBNP	NTproBNP	Transf_verz
NTproBNP		Myelo	Transf_verz	Mg	Myelo	Fe
		Ureum	NTproBNP	TYBC	TN_T_HS	aTTG_n
			Ferritin	TN_T_HS	VitB1	PSA
			TN_T_HS	Transf	Meta	Transf
			Transf	RDW	Ret_He	TYBC
			Ret_He	Ureum	Ureum	NTproBNP
				Ret_He	TE	Fe
						Ureum
						MCV
						Testost_lum
						Mg

All biomarkers where the high ranking was expected can be found in table 22. These are the biomarkers that are considered important by the feature selection techniques, that are also expected to be important based on the expert opinion. Of course, only the top eleven biomarkers for each anaemia type were considered, so by taking into account more biomarkers the list could be expanded. For the sake of this thesis, however, only eleven biomarkers were considered due to the increased complexity the expanded list would produce. Based on the business objectives in section 2, understandability of a diagnostic model should be maximized, which also applies to this case. Even though more biomarkers

could deliver performance improvements, it would get exceedingly hard to incorporate all of these biomarkers into a standard that can be followed easily by the business user, while producing transparency at the same time.

TABLE 22: BIOMARKER THAT WERE ASSESSED TO BE EXPECTED

Anaemia	Iron Deficiency Anaemia	Anaemia of Chronic Disease	Vit B12/Folic Acid Deficiency Anaemia	Bone Marrow Disease	Haemolysis	Severity
Ery	Transf_verz	Transf_verz	MCV	Trombo	Ery	Ery
Transf_verz	Ferritin	Transf	Ery	Leuco	RDW	MCH_n
Ret_He	Transf	TYBC	MCH_n	Ery	Reti_n	MCHC
Transf	TYBC	Fe	aTTG_n	Trombo_cit	LD	RDW
TYBC	MCH_n	Ret_He	LD	Segment	MCHC	
Fe	Fe	NTproBNP	RDW	Neutro	Haptoglo	
MCH_n	Ret_He	Ferritin	MCHC	MCV	MCV	
aTTG_n	MCHC	Ery	Foliumz_R			
Ferritin	MCV	Ferritin				
Ery	aTTG_n	RDW				
Transf_verz	RDW	Kreat				
RDW	Ery					
Testost_lum						
MCHC						
Kreat						
MCV						

9 CONCLUSION & FUTURE WORK

9.1 CONCLUDING REMARKS

This thesis has defined a process for assessing diagnostic standards like the one from the NHG. By using machine learning techniques for feature selection, biomarkers were selected for the prediction of anaemia, type of anaemia, and severity of anaemia. Through the creation of prediction models, it could be shown that the use of these biomarkers led to superior performance compared to the models using the diagnostic standard biomarkers. Based on these findings, alternative predictions were made using a variety of pre-processing techniques, showing the effect each of them had on the prediction performance. By using process mining, the biomarker selection could be compared to the real process, enabling a deeper understanding of the data. Not only did it allow for conformance checking to the diagnostic standard, but also for an analysis on the utilization of the biomarkers. By doing this, biomarkers could be defined that could be used in addition to the biomarkers mandated in the anaemia standard.

9.2 IMPLICATIONS FOR SOCIETY

This thesis has implications for society by creating and evaluation approach used for assessing the impact of diagnostic standards. Specifically, it was shown that current diagnostic pathway models trade predictive performance for the simplicity they create. By using a complete set of features, anaemia could be identified more accurately and help treat anaemic patients. Possible changes are to either include more biomarkers into existing diagnostic models or create models per anaemia type, for instance. This would help in keeping the simplicity while at the same time involving a complete set of biomarkers.

9.3 CONTRIBUTION TO SCIENCE

The contribution to science in this thesis has its foundation in the methodology used to create the results. In particular, the novel use of process mining could help visualise the process and serve as a basis for machine learning analysis. Aside from that, this thesis aims at giving a comprehensive comparison of techniques for the pre-processing of predictive machine learning models. This was done by investigating the biomarkers used in the real process and comparing them to the feature sets created in the machine learning models for section 5 and 6. By comparing several techniques for class imbalance and missing value handling, methods that proved to be superior for a variety of metrics could be identified. Finally, predictions were made for a range of target variables, including the type of anaemia, anaemia occurrence, and severity of anaemia, which was not done in this complete form yet and with these specific sets of techniques used.

9.4 LIMITATIONS & FUTURE RESEARCH POSSIBILITIES

This thesis has several limitations that influenced the research direction and conclusions defined for this research thesis. Firstly, one of the most significant limitations had its origin in the available computational power for creating the models. Libraries and methods were chosen to accommodate a dataset with many dimensions to deal with exceedingly long waiting times. Future research possibilities lie in verifying the findings with the use of more efficient algorithms or methods. Furthermore, additional comparisons may be made as well. In particular, alternative machine learning classifiers, like SVM or neural networks, and other pre-processing techniques could improve the results.

While the dataset provided could be used extensively, the lack of information about whether a patient is pregnant, or whether a patient smokes or not, can also be regarded as a limitation. Since these factors have an influence on the haemoglobin values of a patient, it might be worthwhile to investigate whether the biomarker selections would be different for these groups of people. Furthermore, with additional data, like other biomarkers or target variables, more experiments could also be executed.

Of course, while the process mining helped in gaining deeper data understanding, this is not the only use case for combining process mining with machine learning techniques. As could be shown in the literature review in this thesis, there are many more possibilities for a conjunct use of techniques from these two fields.

10 ACRONYMS

TABLE 23: BIOMARKER SHORTCUTS

Shortcut	Meaning
25-OH-D3	Vitamin D3 - Calcidiol
AF	Alkaline Phosphatase
ALAT	Alanine-Aminotransferase (liver)
ASAT	Aspartat-Aminotransferase (liver, heart, skeletal muscle, kidneys, brain, and red blood cells)
Act-B12_R	Active vitamin B12
Alb	Serum Albumin
AlbKr	Albumin/creatinine ratio (urine)
Anti-TTG-IgA	Anti-Tissue-Transglutaminase IgA
Baso	Basophils
Bili_dir	Bilirubin Direct
Bili_tot	Bilirubin Total
CK	Creatine Kinase
CKD-epi	e-GFR based on the CKD-epi formula
CRP	C-reactive protein
CRP_POC	C-reactive protein (point of care)
Ca	Calcium
Ca2+_geion	Ionised Calcium
Ca2+_geion_iS1	Ionised Calcium (measured on I-stat)
Chol	Cholesterol
Cl	Chloride
D_Dimeer	D_Dimer
Eo	Eosinophil
Erytrobl	Erythroblasts
FT4	Free tetraiodothyronine
Fe	Iron
Ferritin	Ferritin
Fibrinog	Fibrinogen
Foliumz	Folic acid
GFR-MDRD	Estimated Glomerular Filtration Rate based on the Modification of Diet in Renal Disease formula
Gluc (2 times with different values?)	Glucose
HDL_chol	High-Density-Lipoprotein-Cholesterol
Haptoglo	Haptoglobin
Hb	Haemoglobin
Ht	Haematocrit
K	Potassium
Komm_A	Consult anaemia

Kreat	Creatinine
Kreat_U	Creatinine Urine
LD	Lactate dehydrogenase
LDL_chol	Low Density Lipoprotein - Cholesterol
Lactaat	Lactate
Leuco	Leucocytes
Lipase	Lipase
Lymfo	Lymphocytes
MCH	Mean corpuscular haemoglobin
MCHC	Mean corpuscular haemoglobin concentration
MCV	Mean corpuscular volume
Meta	Metamyelocyte
Mg	Magnesium
Mono	Monocytes
Myelo	Myelocyte
NTproBNP	N-terminal pro B-type natriuretic peptide
Na (2 times with different values?)	Sodium
Neutro	Neutrophils
P	Phosphate
PSA	Prostate-specific antigen
PT	Prothrombin time
PTH	Parathyroid hormone
PT_INR	Prothrombin time_international normalized ratio
Prolact_R	Prolactin
Promyelo	Promyelocytes
RDW	Red cell distribution width
RF	Rheumatoid factor
Reti	Reticulocytes
Segment	Segmented neutrophils
Staaf	Band neutrophils
TE	Total protein
TE_Kreat ratio	Total protein/creatinine ratio (urine)
TE_U	Total protein (urine)
TN-T_HS	High-sensitive troponin T
TSH	Thyroid-stimulating hormone
TYBC	Total Iron Binding Capacity (TYBC)
Testosteron	Testosterone
Transf	Transferrin
Transf_verz	Transferrin Saturation
Triglyce	Triglycerides

Trombo citr	Platelets in citrate
Trombo	Platelets
Uraat	Uric acid
Ureum (2 times with different values?)	Urea
VitB1	Vitamin B1
VitB6	Vitamin B6
GGT	Gamma-Glutamyltransferase

11 REFERENCES

[1] A. Alharbi, A. Bulpitt and O. Johnson, "Towards unsupervised detection of process models in healthcare", in MIE 2018, 2017.

[2] A. Ali, S. M. Shamsuddin, and A. Ralescu, “Classification with class imbalance problem: A review,” Int. J. Advance Soft Compu. Appl, vol. 5, no. 3, Nov. 2013.

[3] A. Appice and D. Malerba, "A Co-Training Strategy for Multiple View Clustering in Process Mining", IEEE Transactions on Services Computing, vol. 9, no. 6, pp. 832-845, 2016. Available: 10.1109/tsc.2015.2430327.

[4] A. Appice, N. Di Mauro and D. Malerba, "Leveraging shallow machine learning to predict business process behavior", in 2019 IEEE International Conference on Services Computing (SCC), Milan, Italy, 2019.

[5] A. Appice, P. Ardimento, D. Malerba, D. Marra, M. Mottola and G. Modugno, "Training in a Virtual Learning Environment: A Process Mining Approach", in IEEE Conference on Evolving and Adaptive Intelligent Systems, Bari, Italy, 2020.

[6] A. Kempa-Liehr et al., "Healthcare pathway discovery and probabilistic machine learning", International Journal of Medical Informatics, vol. 137, p. 104087, 2020. Available: 10.1016/j.ijmedinf.2020.104087.

[7] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, k-Nearest Neighbor Classification. 2009.

[8] A. Smola and S. V. N. Vishwanathan, Introduction to Machine Learning. Cambridge: press syndicate of the university of cambridge, 2008.

[9] A. Suppers, A. van Gool, and H. Wessels, “Integrated chemometrics and statistics to drive Successful Proteomics Biomarker Discovery,” Proteomes, vol. 6, no. 2, p. 20, 2018.

[10] A. Tefferi, “Anemia in Adults: A Contemporary Approach to Diagnosis,” Mayo Clinic Proceedings, vol. 78, no. 10, pp. 1274–1280, 2003.

[11] A. Wong, M. S. Kamel, and Y. Sun, “Classification of imbalanced data: a review,” International Journal of Pattern Recognition, vol. 23, no. 4, pp. 687–719, 2009.

[12] B. A. Tha, M. I. Hasan, and M. A. Desai, “Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier,” 2010 International Conference on Advances in Recent Technologies in Communication and Computing, 2010.

[13] B. Hajer, B. Arwa, H. Lobna and G. Khaled, "Intention Mining Data preprocessing based on Multi-Agents System", Procedia Computer Science, vol. 176, pp. 888-897, 2020. Available: 10.1016/j.procs.2020.09.084.

[14] B. Joo, S. Shim and H. Bae, "Utilization of sequential data for machine learning in process control", in PMC 2020, 2020.

[15] B. Kitchham, D. Budgen and P. Brereton, Evidence-Based Software Engineering and Systematic Reviews. hapman & Hall/CRC, 2015.

[16] “Blood,” Texas Heart Institute, 30-Sep-2020. [Online]. Available: https://www.texasheart.org/heart-health/heart-information-center/topics/blood/).

[17] C. A. FINCH, M. HEGSTED, T. D. KINNEY, E. D. THOMAS, C. E. RATH, D. HASKINS, S. FINCH, and R. G. FLUHARTY, “Iron metabolism; the pathophysiology of iron storage,” Blood, Nov-1950. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14791580.

[18] C. Back, S. Debois and T. Slaats, "Entropy as a Measure of Log Variability", Journal on Data Semantics, vol. 8, no. 2, pp. 129-156, 2019. Available: 10.1007/s13740-019-00105-3.

[19] C. Flath and N. Stein, "Towards a data science toolbox for industrial analytics applications", Computers in Industry, vol. 94, pp. 16-25, 2018. Available: 10.1016/j.compind.2017.09.003.

[20] C. Klinkmüller, N. van Beest and N. Weber, "Towards reliable predictive process monitoring", in International Conference on Advanced Information Systems Engineering, Tallinn, Estonia, 2018, pp. 163-181.

[21] “Crisp-dm: Towards a standard process model for data mining.” [Online]. Available: http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[22] C. Ruwende and A. Hill, “Glucose-6-phosphate dehydrogenase deficiency and malaria,” Journal of Molecular Medicine, vol. 76, no. 8, pp. 581–588, 1998.

[23] C. Uribe and C. Isaza, “Expert knowledge-guided feature selection for data-based industrial process monitoring,” Revista Facultad de Ingeniería Universidad de Antioquia, no. 65, 2012.

[24] D. Berrar, “Cross-Validation,” Encyclopedia of Bioinformatics and Computational Biology, pp. 542–545, 2019.

[25] D. Chiabrando, S. Mercurio, and E. Tolosano, “Heme and erythropoieis: more than a structural role,” Haematologica, vol. 99, no. 6, pp. 973–983, 2014.

[26] Dietary reference intakes for thiamin, riboflavin, niacin, vitamin B_6, folate, vitamin B_{12}, pantothenic acid, biotin, and choline. Washington, D.C.: National Academy Press, 1998.

[27] D. Jlailaty, D. Grigori and K. Belhajjame, "Multi-level clustering for extracting process-related information from email logs", in 2017 11th International Conference on Research Challenges in Information Science (RCIS), Brighton, UK, 2017.

[28] E. Andres and Serraj, “Optimal management of pernicious anemia,” Journal of Blood Medicine, p. 97, 2012.

[29] E. Beutler, “G6PD deficiency,” Blood, vol. 84, no. 11, pp. 3613–3636, 1994.

[30] E. E. Bouhassira, “Concise Review: Production of Cultured Red Blood Cells from Stem Cells,” STEM CELLS Translational Medicine, vol. 1, no. 12, pp. 927–933, 2012.

[31] E. Epure, D. Compagno, C. Salinesi, R. Deneckere, M. Bajec and S. Žitnik, "Process models of interrelated speech intentions from online health-related conversations", Artificial Intelligence in Medicine, vol. 91, pp. 23-38, 2018. Available: 10.1016/j.artmed.2018.06.007.

[32] E. Lahner and B. Annibale, “Pernicious anemia: New insights from a gastroenterological point of view,” World Journal of Gastroenterology, vol. 15, no. 41, p. 5121, 2009.

[33] “Elevated Fibrin D-Dimer Fragment in Sickle Cell Anemia: Evidence for Activation of Coagulation during the Steady State as well as in Painful Crisis,” Pathophysiology of Haemostasis and Thrombosis, vol. 19, no. 2, pp. 105–111, 1989.

[34] E. P. Kuiper-Kramer, C. M. Huisman, J. van Raan, and H. G. van Eijk, “Analytical and Clinical Implications of Soluble Transferrin Receptors in Serum,” Clinical Chemistry and Laboratory Medicine, vol. 34, no. 8, 1996.

[35] E. PMC, Europe PMC. [Online]. Available: https://europepmc.org/article/nbk/nbk545275.

[36] E. Rojas, J. Munoz-Gama, M. Sepúlveda and D. Capurro, "Process mining in healthcare: A literature review", Journal of Biomedical Informatics, vol. 61, pp. 224-236, 2016. Available: 10.1016/j.jbi.2016.04.007.

[37] E. S. Pearson, “The test of significance for the correlation coefficient,” Journal of the American Statistical Association, vol. 26, no. 174, pp. 128–134, 1931.

[38] E. Tello-Lael, J. Roa, M. Rubiolo and U. Ramirez-Alcocer, "Predicting activities in business processes with LSTM recurrent neural networks", in 2018 ITU Kaleidoscope: Machine Learning for a 5G Future (ITU K), Santa Fe, Argentina, 2019.

[39] E. W. Rice, “Plasma Fibrinogen, Cholinesterase Activity, and Anemia: Utility of Fibrinogen in Multiphasic Screening and in Assessing the Activity of Diseases,” Clinical Chemistry, vol. 23, no. 4, pp. 741–742, 1977.

[40] “Ferritin (Blood),” Ferritin (Blood) - Health Encyclopedia - University of Rochester Medical Center. [Online]. Available: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=ferritin_blood.

[41] F. Veit, J. Geyer-Klingeberg, J. Madrzak, M. Haug and J. Thomson, "The proactive insights engine: Process mining meets machine learning and artificial intelligence", in CEUR Workshop Proceedings, 2017.

[42] G. Biau, “Analysis of a Random Forests Model,” Journal of Machine Learning Research, vol. 13, pp. 1063–1095, 2012.

[43] G. Doquire and M. Verleysen, “Feature selection with missing data using mutual information estimators,” Neurocomputing, vol. 90, pp. 3–11, 2012.

[44] G. G. R. H. D. W.-G. ME; “Clinical review: hemorrhagic shock,” Critical care (London, England). [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15469601/.

[45] G. Hoffmann, J. Aufenanger, M. Födinger, J. Cadamuro, A. von Eckardstein, M. Kaeslin-Meyer, and W. Hofmann, “Benefits and limitations of laboratory diagnostic pathways,” Diagnosis, vol. 1, no. 4, pp. 269–276, 2014.

[46] G. J. Lonergan, Author Affiliations1From the Department of Radiology and Nuclear Medicine, M. Zulfiqar, and V. C. Andreu-Arasa, “Sickle Cell Anemia,” RadioGraphics, 01-Jul-2001. [Online]. Available: https://pubs.rsna.org/doi/full/10.1148/radiographics.21.4.g01jl23971.

[47] G. Khodabandelou, C. Hug and C. Salinesi, "Mining Users' Intents from Logs", International Journal of Information System Modeling and Design, vol. 6, no. 2, pp. 43-71, 2015. Available: 10.4018/ijismd.2015040102.

[48] G. Lakshmanan, D. Shamsi, Y. Doganata, M. Unuvar and R. Khalaf, "A markov prediction model for data-driven semi-structured business processes", Knowledge and Information Systems, vol. 42, no. 1, pp. 97126, 2013. Available: 10.1007/s10115-013-0697-8.

[49] G. L. Salvagno, F. Sanchis-Gomar, A. Picanza, and G. Lippi, “Red blood cell distribution width: A simple parameter with multiple clinical applications,” Critical Reviews in Clinical Laboratory Sciences, vol. 52, no. 2, pp. 86–105, 2014.

[50] “Global anaemia prevalence and number of individuals affected,” World Health Organization, 09-Jul-2008. [Online]. Available: https://www.who.int/vmnis/anaemia/prevalence/summary/anaemia_data_status_t2/en/.

[51] G. Malato, “Feature selection in machine learning using lasso regression,” Medium, 05-May-2021. [Online]. Available: https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a.

[52] G. Meyer et al., "A Machine Learning Approach to Improving Dynamic Decision Making", Information Systems Research, vol. 25, no. 2, pp. 239263, 2014. Available: 10.1287/are.2014.0513.

[53] G. Spini, M. van Heesch, T. Veugen and S. Chatterjea, "Private Hospital Workflow Optimization via Secure k-Means Clustering", Journal of Medical Systems, vol. 44, no. 1, 2019. Available: 10.1007/s10916-0191473-4.

[54] “Haemoglobin concentrations for the diagnosis of anaemia ...” [Online]. Available: https://www.who.int/vmnis/indicators/haemoglobin.pdf.

[55] H. Al-Ali, A. Cuzzocrea, E. Damiani, R. Mizouni and G. Tello, "A composite machine-learning-based framework for supporting low-level event logs to high-level business process model activities mappings enhanced by flexible BPMN model translation", Soft Computing, vol. 24, no. 10, pp. 7557-7578, 2019. Available: 10.1007/s00500-019-04385-6.

[56] H. Kang, “The prevention and handling of the missing data,” Korean Journal of Anesthesiology, vol. 64, no. 5, p. 402, 2013.

[57] I. Ailenei, A. Rozinat, A. Eckert, and W. M. van der Aalst, “Definition and Validation of Process Mining Use Cases,” Business Process Management Workshops, pp. 75–86, 2012.

[58] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003.

[59] “Indicator Metadata Registry Details,” World Health Organization. [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4552.

[60] I. Rish, “An empirical study of the naive bayes classifier,” IJCAI-01 workshop on ”Empirical Methods in AI,” 2001.

[61] “Iron and Total Iron-Binding Capacity,” Iron and Total Iron-Binding Capacity - Health Encyclopedia - University of Rochester Medical Center. [Online]. Available: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=iron_total_iron_binding_capacity.

[62] I. Sarker, M. Faruque, H. Alqahtani, and A. Kalim, “K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services,” ICST Transactions on Scalable Information Systems, p. 162737, 2018.

[63] I. Teinemaa, M. Dumas, A. Leontjeva and F. Maggi, "Temporal stability in predictive process monitoring", Data Mining and Knowledge Discovery, vol. 32, no. 5, pp. 1306-1338, 2018. Available: 10.1007/s10618-018-0575-9.

[64] I. Verenich, M. Dumas, M. La Rosa and H. Nguyen, "Predicting process performance: A white-box approach based on process models", Journal of Software: Evolution and Process, vol. 31, no. 6, 2019. Available: 10.1002/smr.2170.

[65] I. Verenich, S. Mõškovski, S. Raboczi, M. La Rosa and F. Maggi, "Predictive process monitoring in apromore", in Information Systems in the Big Data Era: CAiSE Forum 2018, Springer, Switzerland, 2018, pp. pp. 244-253.

[66] I. Zakarija, F. Škopljanac-Mačina and B. Blašković, "Automated simulation and verification of process models discovered by process mining", Automatika, vol. 61, no. 2, pp. 312-324, 2020. Available: 10.1080/00051144.2020.1734716.

[67] J. A. Akrimi, A. R. Ahmad, and L. E. George, “Review of Machine Learning Techniques in Anemia Recognition,” International Journal of Science and Research (IJSR), vol. 2, no. 3, Mar. 2013.

[68] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” IJCSI International Journal of Computer Science Issues, vol. 9, no. 5, Sep. 2012.

[69] J. D. Bessman, Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. LexisNexis UK, 1990.

[70] J. F. Koepke and J. A. Koepke, “Reticulocytes,” Clinical & Laboratory Haematology, vol. 8, no. 3, pp. 169–179, 1986.

[71] J. Fu, D. Zapata and E. Mavronikolas, "Statistical Methods for Assessments in Simulations and Serious Games", ETS Research Report Series, vol. 2014, no. 2, pp. 1-17, 2014. Available: 10.1002/ets2.12011.

[72] J. Hastka, J. J. Lasserre, A. Schwarzbeck, and R. Hehlmann, “Central role of zinc protoporphyrin in staging iron deficiency,” Clinical Chemistry, vol. 40, no. 5, pp. 768–773, 1994.

[73] J. Herbert L. Muncie and J. S. Campbell, “Alpha and Beta Thalassemia,” American Family Physician, 15-Aug-2009. [Online]. Available: https://www.aafp.org/afp/2009/0815/p339.html.

[74] J. Jaramillo and J. Arias, "Automatic classification of event logs sequences for failure detection in WfM/BPM systems", in 2019 IEEE Colombian Conference on Applications in Computational Intelligence, Barranquilla, Colombia, 2019.

[75] J. Kaiser, “Dealing with Missing Values in Data,” Journal of Systems Integration, pp. 42–51, 2014.

[76] J. Krumeich, D. Werth and P. Loos, "Prescriptive Control of Business Processes", Business & Information Systems Engineering, vol. 58, no. 4, pp. 261-280, 2015. Available: 10.1007/s12599-015-0412-2.

[77] J. Li, H. Wang and X. Bai, "An intelligent approach to data extraction and task identification for process mining", Information Systems Frontiers, vol. 17, no. 6, pp. 1195-1208, 2015. Available: 10.1007/s10796-0159564-3.

[78] J. Rodrigues, P. Sousa and J. Rodrigues, "Real-time business process recommendations", in Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao, 2018.

[79] J. Saint, D. Gašević and A. Pardo, "Detecting Learning Strategies Through Process Mining", in 13th European Conference on Technology Enhanced Learning, UK, 2018.

[80] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: from early developments to recent advancements,” Systems Science & Control Engineering, vol. 2, no. 1, pp. 602–609, 2014.

[81] K. Jorbina et al., "Nirdizati: A web-based tool for predictive process monitoring", in CEUR Workshop Proceedings, Volume 1920, 2017.

[82] K. Krinkin and E. Kalishenko, "Traffic prediction in wireless mesh networks using process mining algorithms", in 2012 11th Conference of Open Innovations Association (FRUCT), St. Petersburg, Russia, 2012.

[83] K. Krinkin, E. Kalishenko and S. Prakash, "Process mining approach for traffic analysis in wireless mesh networks", in International Conference on Next Generation Wired/Wireless Networking, 2012, pp. 260-269.

[84] K. M. Al-Aidaroo, A. A. Bakar, and Z. Othman, “Medical Data Classification with Naive Bayes Approach,” Information Technology Journal, vol. 11, no. 9, pp. 1166–1174, 2012.

[85] L. Baier, J. Reimold and N. Kuhl, "Handling Concept Drift for Predictions in Business Process Mining", in 2020 IEEE 22nd Conference on Business Informatics, 2020.

[86] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 3, pp. 261–277, 2001.

[87] L. Da Costa, J. Galimand, O. Fenneteau, and N. Mohandas, “Hereditary spherocytosis, elliptocytosis, and other red cell membrane disorders,” Blood Reviews, vol. 27, no. 4, pp. 167–178, 2013.

[88] L. H. Allen, “Causes of Vitamin B12and Folate Deficiency,” Food and Nutrition Bulletin, vol. 29, no. 2_suppl1, 2008.

[89] L. Liu and Özsu M. Tamer, Encyclopedia of database systems. New York: Springer, 2009.

[90] L. Luzzatto, “SICKLE CELL ANAEMIA AND MALARIA,” Mediterranean Journal of Hematology and Infectious Diseases, vol. 4, no. 1, 2012.

[91] L. Măruşter, A. Weijters, W. Van Der Aalst and A. Van Den Bosch, "A Rule-Based Approach for Process Discovery: Dealing with Noise and Imbalance in Process Logs", Data Mining and Knowledge Discovery, vol. 13, no. 1, pp. 67-87, 2006. Available: 10.1007/s10618-005-0029-z.

[92] L. Pauling, H. A. Itano, S. J. Singer, and I. C. Wells, “Sickle Cell Anemia, a Molecular Disease,” Science, vol. 110, no. 2865, pp. 543–548, 1949.

[93] M. Bello, G. Nápoles, R. Sánchez, R. Bello and K. Vanhoof, "Deep neural network to extract high-level features and labels in multi-label classification problems", Neurocomputing, vol. 413, pp. 259-270, 2020. Available: 10.1016/j.neucom.2020.06.117.

[94] M. Bernardi, M. Cimitile, D. Distante, F. Martinelli and F. Mercaldo, "Dynamic malware detection and phylogeny analysis using process mining", International Journal of Information Security, vol. 18, no. 3, pp. 257-284, 2018. Available: 10.1007/s10207-018-0415-3.

[95] M. Calvo, I. Cano, C. Hernandez, V. Ribas, F. Miralles, J. Roca, and R. Jane, “Class Imbalance Impact on the Prediction of Complications during Home Hospitalization: A Comparative Study,” 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.

[96] M. De Leoni and W. Van Der Aalst, "Data-aware process mining:
Discovering decisions in processes using alignments", in SAC '13: Proceedings of the 28th Annual ACM Symposium on Applied Computing, 2013, pp. 1454–1461.

[97] M. Diapouli, S. Kapetanakis, M. Petridis and R. Evans, "Behavioural analytics using process mining in on-line advertising", in ICCBR 2017, 2017.

[98] M. D. Siamak N. Nabili, “What Is Erythropoietin (EPO)? Test, Definition & Side Effects,” MedicineNet, 03-Dec-2019. [Online]. Available: https://www.medicinenet.com/erythropoietin/article.htm.

[99] M. Gupta, A. Asadullah, S. Padmanabhuni and A. Serebrenik, "Reducing user input requests to improve IT support ticket resolution process", Empirical Software Engineering, vol. 23, no. 3, pp. 1664-1703, 2017. Available: 10.1007/s10664-017-9532-2.

[100] M. Hinkka, T. Lehto, K. Heljanko and A. Jung, "Classifying Process Instances Using Recurrent Neural Networks", in BPM 2018: Business Process Management Workshops, 2018, pp. 313-324.

[101] M. Hinkka, T. Lehto, K. Heljanko and A. Jung, "Structural feature selection for event logs", in Business Process Management Workshops, 2018, pp. 20-35.

[102] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, “Machine Learning Algorithms for Anemia Disease Prediction,” Lecture Notes in Electrical Engineering, pp. 463–469, 2018.

[103] M. Jans and M. Hosseinpour, "How active learning and process mining can act as Continuous Auditing catalyst", International Journal of Accounting Information Systems, vol. 32, pp. 44-58, 2019. Available: 10.1016/j.accinf.2018.11.002.

[104] M. Jans, J. van der Werf, N. Lybaert and K. Vanhoof, "A business process mining application for internal transaction fraud mitigation", Expert Systems with Applications, vol. 38, no. 10, pp. 13351-13359, 2011. Available: 10.1016/j.eswa.2011.04.159.

[105] M. Jans, M. Alles and M. Vasarhelyi, "The case for process mining in auditing: Sources of value added and areas of application", International Journal of Accounting Information Systems, vol. 14, no. 1, pp. 1-20, 2013. Available: 10.1016/j.accinf.2012.06.015.

[106] M. Käppel, S. Schönig and S. Jablonski, "Leveraging Small Sample Learning for Business Process Management", Information and Software Technology, vol. 132, p. 106472, 2021. Available: 10.1016/j.infsof.2020.106472.

[107] M. Kopka and M. Kudělka, "Analysis of SAP Log Data Based on Network Community Decomposition", Information, vol. 10, no. 3, p. 92, 2019. Available: 10.3390/info10030092.

[108] M. L. McHugh, “Interrater reliability: the kappa statistic,” Biochemia Medica, pp. 276–282, 2012.

[109] M. Mallappallil, J. Sabu, A. Gruessner, and M. Salifu, “A review of big data and medical research,” SAGE Open Med. , vol. 8, 2020.

[110] M. Mesabbah, W. AboHamad and S. McKeever, "A Hybrid Process Mining Framework for Automated Simulation Modelling for Healthcare", in 2019 Winter Simulation Conference (WSC), National Harbor, MD, USA, 2019.

[111] M. N. Wright, A. Ziegler, and I. R. König, “Do little interactions get lost in dark random forests?,” BMC Bioinformatics, vol. 17, no. 1, 2016.

[112] M. Polato, A. Sperduti, A. Burattin and M. Leoni, "Time and activity sequence prediction of business process instances," Computing, vol. 100, no. 9, pp. 1005-1031, 2018. Available: 10.1007/s00607-018-0593-x.

[113] M. Sheppard and M. Cartwright, “Predicting with sparse data,” Proceedings Seventh International Software Metrics Symposium.

[114] M. Shouman, T. Turner, and R. Stocker, “Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients,” International Journal of Information and Education Technology, pp. 220–223, 2012.

[115] “Multivariate imputation by chained equations [r package mice version 3.13.0],” The Comprehensive R Archive Network, 27-Jan-2021. [Online]. Available: https://cran.r-project.org/web/packages/mice/index.html.

[116] M. Unuvar, G. Lakshmanan and Y. Doganata, "Leveraging path information to generate predictions for parallel business processes", Knowledge and Information Systems, vol. 47, no. 2, pp. 433-461, 2015. Available: 10.1007/s10115-015-0842-7.

[117] N. Di Mauro, A. Appice and T. Basile, "Activity Prediction of Business Process Instances with Inception CNN Models", in AI*IA 2019 – Advances in Artificial Intelligence, 2019, pp. pp 348-361.

[118] NHG-werkgroep Bouma M, “Anemie,” NHG. [Online]. Available: https://richtlijnen.nhg.org/standaarden/anemie.

[119] NHS Choices. [Online]. Available: https://www.gloshospitals.nhs.uk/our-services/services-we-offer/pathology/tests-and-investigations/vitamin-b12-and-serum-folate/.

[120] N. K. Shinton, R. W. Richardson, and J. D. Williams, “Diagnostic value of serum haptoglobin,” Journal of Clinical Pathology, vol. 18, no. 1, pp. 114–118, 1965.

[121] N. Lunardon, G. Menardi, and N. Torelli, “ROSE: a Package for Binary Imbalanced Learning,” The R Journal, vol. 6, no. 1, p. 79, 2014.

[122]N. Milman, “Anemia—still a major health problem in many parts of the world!,” Annals of Hematology, vol. 90, no. 4, pp. 369–377, 2011.

[123]N. S. Young, “Acquired Aplastic Anemia,” JAMA, vol. 282, no. 3, p. 271, 1999.

[124]N. S. Young, P. Scheinberg, and R. T. Calado, “Aplastic anemia,” Current Opinion in Hematology, vol. 15, no. 3, pp. 162–168, 2008.

[125]N. Tax, N. Sidorova and W. van der Aalst, "Discovering more precise process models from event logs by filtering out chaotic activities", Journal of Intelligent Information Systems, vol. 52, no. 1, pp. 107-139, 2018. Available: 10.1007/s10844-018-0507-6.

[126]N. Tax, S. van Zelst and I. Teinemaa, "An experimental evaluation of the generalizing capabilities of process discovery techniques and black-box sequence models", in 19th International Conference, BPMDS 2018, Dordrecht, 2018, pp. 165-180.

[127]N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[128]O. F. Ayilara, L. Zhang, T. T. Sajobi, R. Sawatzky, E. Bohm, and L. M. Lix, “Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry,” Health and Quality of Life Outcomes, vol. 17, no. 1, 2019.

[129]“Office of Dietary Supplements - Vitamin B6,” NIH Office of Dietary Supplements. [Online]. Available: https://ods.od.nih.gov/factsheets/VitaminB6-HealthProfessional/#:~:text=Vitamin%20B6%20deficiency%20is%20associated,function%20%5B1%2C2%5D.

[130]O. Metsker et al., "Modelling and analysis of complex patient-treatment process using graphminer toolbox", in Computational Science – ICCS 2019, 2019.

[131]“Package ranger,” CRAN. [Online]. Available: https://cran.r-project.org/web/packages/ranger/index.html.

[132]P. De Koninck, J. De Weerdt and S. vanden Broucke, "Explaining clusterings of process instances", Data Mining and Knowledge Discovery, vol. 31, no. 3, pp. 774-808, 2016. Available: 10.1007/s10618-016-0488-4.

[133]P. Misra and A. S. Yadav, “Impact of Pre-processing Methods on Healthcare Predictions,” SSRN Electronic Journal, 2019.

[134]P. Theerthagiri, I. J. Jacob, A. U. Ruby, and Y.Vamsidhar, “Prediction of COVID-19 Possibilities using KNN Classification Algorithm,” 2020.

[135]R. Ahmed, M. Faizan and A. Burney, "Process Mining in Data Science: A Literature Review", in 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), 2019.

[136]R. Banzinger, A. Basukoski, and T. Chaussalet, "Discovering Business Processes in CRM Systems by leveraging unstructured text data", in IEEE 20th International Conference on High Performance Computing and Communications, Exeter, UK, 2018.

[137]R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” BMC Bioinformatics, vol. 14, no. 1, 2013.

[138]R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” Journal of Big Data, vol. 7, no. 1, 2020.

[139]R. C. Oh and D. L. Brown, “Vitamin B12 Deficiency,” American Family Physician, 01-Mar-2003. [Online]. Available: https://www.aafp.org/afp/2003/0301/p979.html?mod=article_inline.

[140]R. E. Hodges, H. E. Sauberlich, J. E. Canham, D. L. Wallace, R. B. Rucker, L. A. Mejia, and M. Mohanram, “Hematopoietic studies in vitamin A deficiency,” The American Journal of Clinical Nutrition, vol. 31, no. 5, pp. 876–885, 1978.

[141]Red Blood Cell Production By Evan M. Braunstein, By, E. M. Braunstein, and Last full review/revision Sep 2020| Content last modified Sep 2020, “Red Blood Cell Production - Hematology and Oncology,” MSD Manual Professional Edition. [Online]. Available: https://www.msdmanuals.com/professional/hematology-and-oncology/approach-to-the-patient-with-anemia/red-blood-cell-production.

[142]R. L. Crisp, L. Solari, D. Vota, E. García, G. Miguez, M. E. Chamorro, G. A. Schvartzman, G. Alfonso, D. Gammella, S. Caldarola, C. Riccheri, D. Vittori, B. Venegas, A. Nesse, and H. Donato, “A prospective study to assess the predictive value for hereditary spherocytosis using five laboratory tests (cryohemolysis test, eosin-5′-maleimide flow cytometry, osmotic fragility test, autohemolysis test, and SDS-PAGE) on 50 hereditary spherocytosis families in Argentina,” Annals of Hematology, vol. 90, no. 6, pp. 625–634, 2010.

[143]“Riboflavin – Vitamin B2,” The Nutrition Source, 11-Aug-2020. [Online]. Available: https://www.hsph.harvard.edu/nutritionsource/riboflavin-vitamin-b2/.

[144]R. Jain and W. Xu, “HDSI: High DIMENSIONAL selection with interactions algorithm on feature selection and testing,” PLOS ONE, vol. 16, no. 2, 2021.

[145]R. Mans et al., "Process mining techniques: an application to stroke care", Studies in Health Technology and Informatics, vol. 6, no. 136, pp. 573578, 2021. [Accessed 31 January 2021].

[146]R. Mans, M. Schonenberg, M. Song, W. van der Aalst and P. Bakker, "Application of process mining in healthcare–a case study in a dutch hospital", in BIOSTEC 2008: Biomedical Engineering Systems and Technologies, 2008, pp. 425-438.

[147]R. Oberhauser and S. Stigler, "Microflows: Leveraging process mining and an automated constraint recommender for microflow modeling", in BMSD 2017: Business Modeling and Software Design, 2018, pp. pp 2548.

[148]R. Sarno, P. Sari, D. Sunaryono, B. Amaliah and I. Mukhlash, "Mining decision to discover the relation of rules among decision points in a nonfree choice construct", in 2014 International Conference on Information, Communication Technology and System, Surabaya, Indonesia, 2014.

[149]R. Thirard, R. Ascione, J. Blazeby, and C. A. Rogers, "Integrating expert opinions with clinical trial data to analyse low-powered subgroup analyses: a Bayesian analysis of the VeRDiCT trial," 2020.

[150]R. Umer, T. Susnjak, A. Mathrani and S. Suriadi, "On predicting academic performance with process mining in learning analytics", Journal of Research in Innovative Teaching & Learning, vol. 10, no. 2, pp. 160176, 2017. Available: 10.1108/jrit-09-2017-0022.

[151]R. Zhu et al., "Automatic Real-Time Mining Software Process Activities From SVN Logs Using a Naive Bayes Classifier", IEEE Access, vol. 7, pp. 146403-146415, 2019. Available: 10.1109/access.2019.2945608.

[152]S. Cope, D. Ayers, J. Zhang, K. Batt, and J. P. Jansen, "Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia," BMC Medical Research Methodology, vol. 19, no. 1, 2019.

[153]"Serum Iron," ucsfhealth.org, 06-Oct-2020. [Online]. Available: https://www.ucsfhealth.org/medical-tests/serum-iron-test.

[154]S. Ferilli and S. Angelastro, "Activity prediction in process mining using the WoMan framework", Journal of Intelligent Information Systems, vol. 53, no. 1, pp. 93-112, 2019. Available: 10.1007/s10844-019-00543-2.

[155]S. J. Hickey, "Naive Bayes Classification of Public Health Data with Greedy Feature Selection," Communications of the IIMA, vol. 13, no. 2, 2013.

[156]S. Khalid and D. Prieto-Alhambra, "Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research," Current Epidemiology Reports, vol. 6, no. 3, pp. 364–372, 2019.

[157]S. Killip, J. M. Bennett, and M. D. Chambers, "Iron Deficiency Anemia," American Family Physician, 01-Mar-2007. [Online]. Available: https://www.aafp.org/afp/2007/0301/p671.html.

[158]S. Kovalchuk, A. Funkner, O. Metsker and A. Yakovlev, "Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification", Journal of Biomedical Informatics, vol. 82, pp. 128-142, 2018. Available: 10.1016/j.jbi.2018.05.004.

[159]"Soluble Transferrin Receptor," Lab Tests Online. [Online]. Available: https://labtestsonline.org/tests/soluble-transferrin-receptor.

[160]S. Perrotta, P. G. Gallagher, and N. Mohandas, "Hereditary spherocytosis," The Lancet, vol. 372, no. 9647, pp. 1411–1426, 2008.

[161]S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," arXiv preprint, 2018.

[162]S.-R. Pasricha, "Anemia: a comprehensive global estimate," Blood, vol. 123, no. 5, pp. 611–612, 2014.

[163]S. S. Adam, N. S. Key, and C. S. Greenberg, "D-dimer antigen: current concepts and future prospects," Blood, vol. 113, no. 13, pp. 2878–2887, 2009.

[164]S. Tabatabaei, X. Lu, M. Hoogendoorn and H. Reijers, "Trace Clustering on Very Large Event Data in Healthcare Using Frequent Sequence Patterns", in BPM 2019, Vienna, Austria, 2019, pp. 198-215.

[165]S. Zhang, N. Gu, J. Lian and S. Li, "Workflow process mining based on machine learning", in International Conference on Machine Learning and Cybernetics, Xi'an, China, 2003.

[166]T. Becker and W. Intoyoad, "Context Aware Process Mining in Logistics", in Procedia CIRP 63, 2017, pp. 557 – 562.

[167]T. D. Johnson-Wimbley and D. Y. Graham, "Diagnosis and management of iron deficiency anemia in the 21st century," Therapeutic Advances in Gastroenterology, vol. 4, no. 3, pp. 177–184, 2011.

[168]T. Ekin, G. Lakomski and R. Musal, "An unsupervised Bayesian hierarchical method for medical fraud assessment", Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 12, no. 2, pp. 116124, 2019. Available: 10.1002/sam.11408.

[169]T. Erdogan and A. Tarhan, "Systematic Mapping of Process Mining Studies in Healthcare", IEEE Access, vol. 6, pp. 24543-24567, 2018. Available: 10.1109/access.2018.2831244.

[170]T. Nolle, S. Luettgen, A. Seeliger and M. Mühlhäuser, "Analyzing business process anomalies using autoencoders", Machine Learning, vol. 107, no. 11, pp. 1875-1893, 2018. Available: 10.1007/s10994-018-5702-
8.

[171]T. Krismayer, R. Rabiser, and P. Grunbacher, "Mining constraints for event-based monitoring in systems of systems", in 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE), 2017.

[172]Tsang-Hsiang Cheng, Chih-Ping Wei, and V. S. Tseng, "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches," 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), 2006.

[173]T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," Biomedical Signal Processing and Control, vol. 52, pp. 456–462, 2019.

[174]"WBC Count," ucsfhealth.org, 06-Oct-2020. [Online]. Available: https://www.ucsfhealth.org/medical-tests/wbc-count.

[175]"Welkom bij Medlon, medische diagnostiek," Medlon. [Online]. Available: https://www.medlon.nl/.

[176]W. Es-Soufi, E. Yahia and L. Roucoules, "On the use of process mining and machine learning to support decision making in systems design", in Product Lifecycle Management, 2016.

[177]W. I. L. Van der Aalst, "Process Mining: Overview and Opportunities," ACM Trans. Manag. Inform. Syst, vol. 99, no. 99, Feb. 2012.

[178]W. M. P. van der Aalst, "Process Mining in the Large: A Tutorial," Business Intelligence, pp. 33–76, 2014.

[179]W. Rizzi, C. Di Francescomarino and F. Maggi, "Explainability in predictive process monitoring: When understanding helps improving", in Business Process Management Forum, 2020.

[180]W. van der Aalst and A. Rozinat, "Decision mining in ProM", in BPM'06: Proceedings of the 4th international conference on Business Process Management, 2006, pp. 420-425.

[181]W. Van der Aalst, "Process Mining," Communications of the ACM, vol. 55, no. 8, pp. 76–83, Aug. 2012.

[182]W. VAN DER AALST, Process Mining: Data Science in Action. [Place of publication not identified]: SPRINGER, 2016.

[183]W. Van der Aalst, "Process Mining in the Large: A tutorial", in eBISS 2013: Business Intelligence, 2014, pp. 33-76.

[184]W. Van der Aalst, "Process Mining: Overview and Opportunities", ACM Transactions on Management Information Systems, vol. 3, no. 2, pp. 117, 2012.

[185]W. Zhao, H. Liu, W. Dai and J. Ma, "An entropy-based clustering ensemble method to support resource allocation in business process management", Knowledge and Information Systems, vol. 48, no. 2, pp. 305-330, 2015. Available: 10.1007/s10115-015-0879-7.

[186]X. Liu, H. Liu, and C. Ding, "Incorporating user behavior patterns to discover workflow models from event logs", in IEEE 20th International Conference on Web Services, Santa Clara, CA, USA, 2013.

[187]Y. Nakayama, M. Mori, Y. Naruse and H. Morikawa, "The process discovery approaches for decision making in sales activities", in 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS), Toyama, Japan, 2018.

[188]Z. Bozorgi, I. Teinemaa, M. Dumas, M. La Rosa and A. Polyvyanyy, "Process mining meets causal machine learning: Discovering causal rules from event logs", in 2020 2nd International Conference on Process Mining (ICPM), 2020.

[189]Z. Tariq, N. Khan, D. Charles, S. McClean, I. McChesney and P. Taylor, "Understanding Contrail Business Processes through Hierarchical Clustering: A Multi-Stage Framework", Algorithms, vol. 13, no. 10, p. 244, 2020. Available: 10.3390/a13100244.

[190]Z. Zhang, "Introduction to machine learning: k-nearest neighbors," Annals of Translational Medicine, vol. 4, no. 11, pp. 218–218, 2016.

12 APPENDICES

12.1 DATA EXTRACTION FROM SCOPUS AND FINDUT

TABLE 24: DATA EXTRATION

Sr. No	Title	Authors	Year	Source Title	Cited By	Keywords
1	Decision mining in ProM	Rozinat, A., Van Der Aalst, W.M.P.	2006	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	172	business process intelligence, process mining, petri nets, decision trees
2	Data-aware process mining: Discovering decisions in processes using alignments	De Leoni, M., Van Der Aalst, W.M.P.	2013	Proceedings of the ACM Symposium on Applied Computing	97	process discovery, machine-learning techniques, business process data-flow perspective
3	A rule-based approach for process discovery: Dealing with noise and imbalance in process logs	Măruşter, L., Weijters, A.J.M.M., Van Der Aalst, W.M.P., Van Den Bosch, A.	2006	Data Mining and Knowledge Discovery	57	rule induction, process mining, knowledge discovery, petri nets
4	A Co-Training Strategy for Multiple View Clustering in Process Mining	Appice, A., Malerba, D.	2016	IEEE Transactions on Services Computing	32	clustering, co-training, multiple view learning, process mining
5	Time and activity sequence prediction of business process instances	Polato, M., Sperduti, A., Burattin, A., Leoni, M.	2018	Computing	31	process mining, prediction, remaining time, machine learning
6	A machine learning approach to improving dynamic decision making	Meyer, G., Adomavicius, G., Johnson, P.E., (...), Sperl-Hillen, J.A.M., O'Connor, P.J.	2014	Information Systems Research	25	dynamic decision making, process control, data mining, process mining, machine learning, simulation, healthcare
7	Towards a data science toolbox for industrial analytics applications	Flath, C.M., Stein, N.	2018	Computers in Industry	23	predictive analytics, manufacturing, process mining
8	Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification	Kovalchuk, S.V., Funkner, A.A., Metsker, O.G., Yakovlev, A.N.	2018	Journal of Biomedical Informatics	19	clinical pathways, discrete-event simulation, process mining, data mining, acute coronary syndrome, electronic health records, classification
9	Context-Aware Process Mining in Logistics	Becker, T., Intoyoad, W.	2017	Procedia CIRP	13	logistics, process mining, context awareness
10	Classifying Process Instances Using Recurrent Neural Networks	Hinkka, M., Lehto, T., Heljanko, K., Jung, A.	2019	Lecture Notes in Business Information Processing	8	process mining, prediction, classification, machine learning, deep learning, recurrent neural networks, long short-term memory, gated recurrent unit, natural language processing
11	Structural feature selection for event logs	Hinkka, M., Lehto, T., Heljanko, K., Jung, A.	2018	Lecture Notes in Business Information Processing	8	automatic business process discovery, process mining, prediction, classification, machine learning, clustering, feature selection
12	Mining constraints for event-based monitoring in systems of systems	Krismayer, T., Rabiser, R., Grunbacher, P.	2017	ASE 2017 - Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering	8	constraint mining, event-based monitoring, systems of systems
13	Detecting Learning Strategies Through Process Mining	Saint, J., Gašević, D., Pardo, A.	2018	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	6	learning analytics, process mining, first order markov models, temporal dynamics, self-regulated learning

14	Discovering Business Processes in CRM Systems by leveraging unstructured text data	Banziger, R., Basukoski, A., Chaussalet, T.	2019	Proceedings - 20th International Conference on High-Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/Smart City/DSS 2018	5	process mining, process discovery, customer relationship management, crm, business process management, latent dirichlet allocation
15	On the use of process mining and machine learning to support decision making in systems design	Es-Soufi, W., Yahia, E., Roucoules, L.	2016	IFIP Advances in Information and Communication Technology	5	collaborative design process, process mining, supervised classification, process patterns, decision-making
16	Mining users' intents from logs	Khodabandelou, G., Hug, C., Salinesi, C.	2015	International Journal of Information System Modeling and Design	5	intentional process models, machine learning, process mining, hidden markov models
17	Mining decision to discover the relation of rules among decision points in a non-free choice construct	Sarno, R., Sari, P.L.I., Sunaryono, D., Amaliah, B., Mukhlash, I.	2014	Proceedings of 2014 International Conference on Information, Communication Technology and System, ICTS 2014	5	decision mining, decision tree, non-free choice, process mining
18	Incorporating user behavior patterns to discover workflow models from event logs	Liu, X., Liu, H., Ding, C.	2013	Proceedings - IEEE 20th International Conference on Web Services, ICWS 2013	5	probabilistic suffix tree, workflow model discovery, coclustering
19	Leveraging shallow machine learning to predict business process behavior	Appice, A., Di Mauro, N., Malerba, D.	2019	Proceedings - 2019 IEEE International Conference on Services Computing, SCC 2019 - Part of the 2019 IEEE World Congress on Services	4	process prediction, feature construction, machine learning
20	Activity Prediction of Business Process Instances with Inception CNN Models	Di Mauro, N., Appice, A., Basile, T.M.A.	2019	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	4	business process monitoring, sequence prediction, deep learning
21	Trace Clustering on Very Large Event Data in Healthcare Using Frequent Sequence Patterns	Lu, X., Tabatabaei, S.A., Hoogendoorn, M., Reijers, H.A.	2019	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	4	trace clustering, process mining, frequent sequential patterns, machine learning
22	Process models of interrelated speech intentions from online health-related conversations	Epure, E.V., Compagno, D., Salinesi, C., (...), Bajec, M., Žitnik, S.	2018	Artificial Intelligence in Medicine	4	intention mining, text mining, natural language processing, machine learning, process mining, speech acts, speech intentions, conversational processes, conversation analysis
23	Workflow process mining based on machine learning	Zhang, S.-H., Gu, N., Lian, J.-X., Li, S.-H.	2003	International Conference on Machine Learning and Cybernetics	4	worknow process mining, worknow modeling, machine learning
24	Towards unsupervised detection of process models in healthcare	Alharbi, A., Bulpitt, A., Johnson, O.A.	2018	Studies in Health Technology and Informatics	3	process mining, unsupervised learning, hidden markov models, electronic health records, event abstraction, mimic-III
25	Microflows: Leveraging process mining and an automated constraint recommender for microflow modeling	Oberhauser, R., Stigler, S.	2018	Lecture Notes in Business Information Processing	3	business process modeling, workflow management systems, microservices, service orchestration, agent systems, semantic technology, declarative programming, recommenders, recommendation engines, business process mining, business process modeling notation
26	Healthcare pathway discovery and probabilistic machine learning	Kempa-Liehr, A.W., Lin, C.Y.-C., Britten, R., (...), Mordaunt, D., O'Sullivan, M.	2020	International Journal of Medical Informatics	2	healthcare pathway, process mining, electronic health record, probabilistic programming

27	Analysis of SAP log data based on network community decomposition	Kopka, M., Kudělka, M.	2019	Information (Switzerland)	2	decision support, process log data, network construction, visualization (visual data mining), community detection (network clustering), pattern and outlier analysis, recursive procedure (cluster quality)
28	Reducing user input requests to improve IT support ticket resolution process	Gupta, M., Asadullah, A., Padmanabhuni, S., Serebrenik, A.	2018	Empirical Software Engineering	2	software process, machine learning, process mining , service level agreement, ticket resolution time
29	Towards reliable predictive process monitoring	Klinkmüller, C., van Beest, N.R.T.P., Weber, I.	2018	Lecture Notes in Business Information Processing	2	behavioural classification, predictive process monitoring, process mining, machine learning
30	Predictive process monitoring in apromore	Verenich, I., Mõškovski, S., Raboczi, S., (...), La Rosa, M., Maggi, F.M.	2018	Lecture Notes in Business Information Processing	2	process mining, predictive monitoring, business process, machine learning
31	An experimental evaluation of the generalizing capabilities of process discovery techniques and Blackbox sequence models	Tax, N., van Zelst, S.J., Teinemaa, I.	2018	Lecture Notes in Business Information Processing	2	process mining, behavioural generalization, next activity prediction, process discovery, sequence modeling
32	Process mining meets causal machine learning: Discovering causal rules from event logs	Bozorgi, Z.D., Teinemaa, I., Dumas, M., La Rosa, M., Polyvyanyy, A.	2020	Proceedings - 2020 2nd International Conference on Process Mining, ICPM 2020	1	process mining, causal ml, uplift modeling
33	A composite machine learning-based framework for supporting low-level event logs to high-level business process model activities mappings enhanced by flexible BPMN model translation	Al-Ali, H., Cuzzocrea, A., Damiani, E., Mizouni, R., Tello, G.	2020	Soft Computing	1	business process management systems, business process intelligence, low-level event logs, high-level business process model activities, bpmn, model translation
34	A Hybrid Process Mining Framework for Automated Simulation Modelling for Healthcare	Mesabbah, M., Abo-Hamad, W., McKeever, S.	2019	Proceedings - Winter Simulation Conference	1	
35	Automatic Real-Time Mining Software Process Activities from SVN Logs Using a Naive Bayes Classifier	Zhu, R., Dai, Y., Li, T., (...), Yuan, J., Huang, Y.	2019	IEEE Access	1	activity classifier, machine learning, software process activity, svn log
36	Modelling and analysis of complex patient treatment process using graphminer toolbox	Metsker, O., Kesarev, S., Bolgova, E., (...), Yakovlev, A., Kovalchuk, S.	2019	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	1	graph mining, process mining, community detection, process modeling, cardiology, complex process analysis
37	Predicting activities in business processes with LSTM recurrent neural networks	Tello-Leal, E., Roa, J., Rubiolo, M., RamirezAlcocer, U.M.	2018	10th ITU Academic Conference Kaleidoscope: Machine Learning for a 5G Future, ITU K 2018	1	lstm, event log, process mining, business process
38	Traffic prediction in wireless mesh networks using process mining algorithms	Krinkin, K., Kalishenko, E.	2018	Conference of Open Innovation Association, FRUCT	1	wireless mesh networks, routing, process mining, traffic overload
39	Development of an Emental health infrastructure for supporting interoperability and data analysis	Rabbi, F., Lamo, Y.	2018	CEUR Workshop Proceedings	1	healthcare systems, internet of things, process mining, machine learning, hl7 fhir

40	Multi-level clustering for extracting processrelated information from email logs	Jlailaty, D., Grigori, D., Belhajjame, K.	2017	Proceedings - International Conference on Research Challenges in Information Science	1	email analysis, process model, process mining, process information, clustering
41	Nirdizati: A web-based tool for predictive process monitoring	Jorbina, K., Rozumnyi, A., Verenich, I., (...), La Rosa, M., Raboczi, S.	2017	CEUR Workshop Proceedings	1	process mining, predictive process monitoring, machine learning
42	The proactive insights engine: Process mining meets machine learning and artificial intelligence	Veit, F., Geyer-Klingeberg, J., Madrzak, J., Haug, M., Thomson, J.	2017	CEUR Workshop Proceedings	1	process mining, process intelligence, machine learning, artificial intelligence
43	Utilization of sequential data for machine learning in process control	Joo, B., Shim, S., Bae, H.	2016	ICIC Express Letters	1	sequence analysis, machine learning, probability density function (pdf), k-means clustering, process analysis
44	Handling Concept Drift for Predictions in Business Process Mining	Baier, L., Reimold, J., Kuhl, N.	2020	Proceedings - 2020 IEEE 22nd Conference on Business Informatics, CBI 2020	0	
45	Training in a Virtual Learning Environment: A Process Mining Approach	Appice, A., Ardimento, P., Malerba, D., (...), Marra, D., Mottola, M.	2020	IEEE Conference on Evolving and Adaptive Intelligent Systems	0	process mining, machine learning, classification
46	Automated simulation and verification of process models discovered by process mining	Zakarija, I., Škopljanac-Mačina, F., Blašković, B.	2020	Automatika	0	process mining, iot, model checking, inductive machine learning, big data, mas
47	Leveraging Small Sample Learning for Business Process Management	Käppel, M., Schönig, S., Jablonski, S.	2020	Information and Software Technology	0	small sample learning, bpm, process prediction, machine learning, process mining
48	Intention mining data preprocessing based on multi-agents' system	Hajer, B., Arwa, B., Lobna, H., Khaled, G.	2020	Procedia Computer Science	0	process mining, intention mining, multi-agents' system, machine learning, unsupervised learning, data-preprocessing
49	Explainability in predictive process monitoring: When understanding helps improving	Rizzi, W., Di Francescomarino, C., Maggi, F.M.	2020	Lecture Notes in Business Information Processing	0	predictive process monitoring, process mining, machine learning, explainable artificial intelligence
50	Automatic classification of event logs sequences for failure detection in WfM/BPM systems	Jaramillo, J., Arias, J.	2019	2019 IEEE Colombian Conference on Applications in Computational Intelligence, ColCACI 2019 - Proceedings	0	
51	The process discovery approaches for decision making in sales activities	Nakayama, Y., Mori, M., Naruse, Y., Morikawa, H.	2018	Proceedings - 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2018	0	sales activity, decision making, process mining, decision mining, process discovery, hidden markov model
52	Real-time business process recommendations | [Recomendações em processos de negócio em tempo-real]	Rodrigues, J., Sousa, P., Rodrigues, J.	2018	Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao	0	process mining, decision mining; supervised learning, process, operational support, event logs
53	Behavioural analytics using process mining in on-line advertising	Diapouli, M., Kapetanakis, S., Petridis, M., Evans, R.	2017	CEUR Workshop Proceedings	0	process mining, process-oriented workflows, classification, online display advertising
54	Process mining approach for traffic analysis in wireless mesh networks	Krinkin, K., Kalishenko, E., Prakash, S.P.S.	2012	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	0	wireless mesh networks, adaptive routing, process mining, traffic hotspots, power aware metric
55	On predicting academic performance with process mining in learning analytics	Rahila Umer, Teo Susnjak, Anuradha Mathrani, Suriadi Suriadi	2018	Journal of Research in Innovative Teaching & Learning	15	prediction, moocs, machine learning, learning analytics, process mining, Education data mining

56	How active learning and process mining can act as Continuous Auditing catalyst	Mieke Jans, Marzie Hosseinpour	2019	International Journal of Accounting Information Systems	9	continuous auditing, internal control testing, process mining, data mining, active learning
57	Dynamic malware detection and phylogeny analysis using process mining	Mario Luca Bernardi, Marta Cimitile, Damiano Distante, Fabio Martinelli, Francesco Mercaldo	2019	International Journal of Information Security	18	malware detection, malware evolution, malware phylogeny, security, process mining, linear temporal logic, declare
58	Activity prediction in process mining using the WoMan framework	Stefano Ferilli, Sergio Angelastro	2019	Journal of Intelligent Information Systems	6	process mining, activity prediction, process model
59	An intelligent approach to data extraction and task identification for process mining	Jiexun Li, Harry Jiannan Wang, Xue Bai	2015	Information Systems Frontiers	16	business process management, computational experiments, data extraction, process mining, task identification, text mining
60	Explaining clusterings of process instances	Pieter De Koninck, Jochen De Weerdt, Seppe K L M vanden Broucke	2017	Data Mining and Knowledge Discovery	10	process discovery, trace clustering, human understanding, instance-level explanations, support vector machines
61	Analyzing business process anomalies using autoencoders	Nolle T, Luettgen S, Seeliger A, Muhlhauser M	2018	Machine Learning	16	deep learning, autoencoder, anomaly detection, process mining, business intelligence
62	Temporal stability in predictive process monitoring	Irene Teinemaa, Marlon Dumas, Anna Leontjeva, Fabrizio Maria Maggi	2018	Data Mining and Knowledge Discovery	13	predictive process monitoring, early sequence classification, stability
63	Predicting process performance: A Whitebox approach based on process models	Ilya Verenich, Marlon Dumas, Marcello La Rosa, Hoang Nguyen	2019	Journal of Software Evolution and Process	8	process mining, predictive process monitoring, flow analysis
64	Discovering more precise process models from event logs by filtering out chaotic activities	Niek Tax, Natalia Sidorova, Wil M P van der Aalst	2019	Journal of Intelligent Information Systems	20	information systems, business process intelligence, process mining, knowledge discovery
65	An entropy-based clustering ensemble method to support resource allocation in business process management	Weidong Zhao, Haitao Liu, Weihui Dai, Jian Ma	2016	Knowledge and Information Systems	17	business process management, resource allocation, multi-criteria recommendation, process mining, clustering ensemble
66	Understanding Contrail Business Processes through Hierarchical Clustering: A MultiStage Framework	Zeeshan Tariq, Naveed Khan, Darryl Charles, Sally McClean, Ian McChesney, Paul Taylor	2020	Algorithms	1	processmining, trace clustering, machine learning, knowledge discovery, process analytics
67	Deep neural network to extract high-level features and labels in multi-label classification problems	Marilyn Bello, Gonzalo Nápoles, Ricardo Sánchez, Rafael Bello, Koen Vanhoof	2020	Neurocomputing	1	deep neural networks, multi-label classification, high-level features, high-level labels, associationbased pooling
68	Private Hospital Workflow Optimization via Secure kMeans Clustering	Gabriele Spini, Maran van Heesch, Thijs Veugen, Supriyo Chatterjea	2020	Journal of Medical Systems	43	secure multi-party computation, hospital, workflow optimization, privacy, real-time locating system, clustering, kmeans
69	Leveraging path information to generate predictions for parallel business processes	Merve Unuvar, Geetika T Lakshmanan, Yurdaer N Doganata	2016	Knowledge and Information Systems	24	prediction, path representation, decision tree, training, prediction model, accuracy, complexity
70	An unsupervised Bayesian hierarchical method for medical fraud assessment	Tahir Ekin, Greg Lakomski, Rasim Muzaffer Musal	2019	Statistical Analysis and Data Mining	1	medical fraud, health care fraud, medical audits, unsupervised data mining, Bayesian hierarchical methods

71	Prescriptive Control of Business Processes: New Potentials Through Predictive Analytics of Big Data in the Process Manufacturing Industry	Julian Krumeich, Dirk Werth Dr., Peter Loos Prof. Dr.	2016	Business & Information Systems Engineering	27	predictive analytics, complex event processing, prescriptive analytics, event-driven business process management, big data, process industry
72	A markov prediction model for data-driven semi-structured business processes	Geetika T Lakshmanan, Davood Shamsi, Yurdaer N Doganata, Merve Unuvar, Rania Khalaf	2015	Knowledge and Information Systems	67	markov chain, data driven, decision tree, business process, prediction
73	Entropy as a Measure of Log Variability	Christoffer Olling Back, Søren Debois, Tijs Slaats	2019	Journal on Data Semantics	53	process mining, hybrid models, process variability, process flexibility, information theory, entropy, knowledge work
74	Statistical Methods for Assessments in Simulations and Serious Games	Jianbin Fu, Diego Zapata, Elia Mavronikolas	2014	ETS Research Report Series	8	simulation-based assessment, game-based assessment, cognitive diagnostic model, data mining, review

12.2 RESEARCH QUESTION ANSWERS

TABLE 25: RESEARCH QUESTION ANSWERS

Sr. No	RQ1 Answer	RQ2 Answer	RQ3 Answer
1			Decision Trees can be used to detect data dependencies that affect the routing of a case. This detection process is called decision mining.
2			Improvement of paper 1, which was about decision mining. Before decision trees are used, log and model are first aligned to deal with deviating behaviour and complex control flow constructs
3			Machine Learning is used to induce noise-robust rule sets for causal relations and parallel/exclusive relations.
4			To prevent spaghetti results from the process models, trace clustering is used as a pre-processing step. This pre-processing is done to split up the logs into clusters of similar traces. A Multiview aware approach to trace clustering is taken, which is using a co-training strategy.
5		This paper introduces a method of predicting the remaining time of running cases and uses two case studies to exemplify this.	For the prediction tasks in this paper, Support Vector Regression and the Naive Bayes Classifier is used
6	In this paper, an iterative approach for improving treatment strategies is introduced. This improvement is achieved by predicting and eliminating treatment failures, using information from an electronic medical record system in a healthcare management organization.	The approach introduced is utilized to enhance dynamic decision-making in a healthcare organization and for a manufacturing task. It applies the concluded information to improve the decision-making strategies.	The paper's solution uses an inductive machine learning algorithm to generate a set of decision rules and identify areas for improvement.
7		The predictive analytics system proposed in this paper, which contains both process mining and black-box machine learning techniques, can give more detailed information about the underlying process. The combination of both machine learning and process mining makes this toolbox more valuable.	This paper's learning problem is to determine defect rates in a manufacturing process using process mining and black-box machine learning techniques. It is concluded that black box techniques and process mining can work in unison.
8	Clustering is used to group patients and to make a pathway analysis for them using process mining	The paper's approach allows for automatically identifying patient dynamics on a micro level to perform more realistic simulations and obtain macro-level characteristics, such as departmental load, queuing parameters, or patient experiences.	Learning problem: How can patient flow be simulated using a combination of data, text, and process mining techniques. Patients are clustered based on treatment strategies, then used to find these patients' pathways using process mining. For a more in-depth analysis of individual clusters, decision trees were used.
9			Clustering analysis is used in a logistics domain to classify single items in large datasets. This clustering

			is for adding context awareness to unstructured data to improve the likely results of process mining.
10			This paper uses Gated Recurrent Unit(GRU) and Long Short-Term Memory(LSTM) neural networks to classify business process instances. This classification is done in a supervised fashion on unlabelled data. Using NLP also resulted in improved classification model training time.
11		A discussion is made on how feature selection results can be used in computer-assisted root cause analysis. Furthermore, they look at properties of different structural feature types in the context of feature selection	Business process instances are classified based on structural features derived from event logs. A discussion is added on how many features should be used for optimal classification performance. Nine different feature selection techniques are proposed and tested. The best feature selection technique is using the k-means clustering algorithm
12		The approach in this paper helps to elevate the problem of missing domain knowledge when defining constraints.	This paper's learning problem is to define constraints in an automated manner for runtime monitoring approaches.
13			Learning strategies for higher education are detected using a process mining approach based on first-order Markov models combined with unsupervised machine learning techniques to perform Intra and inter-strategy analysis. Learning sessions were clustered for this.
14			The problem in this paper is to make use of process mining techniques for unstructured data. It proposes a framework to mine processes from CRM data while at the same time dealing with the unstructured part of the data. This is done by using Latent Dirichlet Allocation (LDA), which is unsupervised machine learning. This technique can automatically detect and assign labels to activities.
15		This paper introduces a double-layer framework that identifies the most important patterns that need execution in its design context. At the same time, it gives the most important parameters for every activity in the process pattern.	One layer in the framework is the machine learning layer that supports supervised classification.
16			Research Problem: How can the intend of the user be mined by using the Map Miner Method?
17			Aside from using decision mining to analyse rules that affect case routing, it can also identify choices using relationships and rules found from the workflow. The paper shows that decision point rules in non-free choice constructs have similarities because implicit dependencies make limited choices in some cases. If there are the same rules found between two decision points in the process mining process, it might be that the decision points are in a non-free choice relationship even if the model does not show a non-free choice construct.
18		The mining process's outcome can be improved by learning each user's behaviour patterns and incorporating the learning result into the clustering process.	Learning problems: How can user behaviour patterns be used in machine learning to improve process mining outcomes? This is done in the paper by incorporating the behaviour patterns into sequence clustering for workflow model discovery.
19			To predict business process behaviour, this paper makes use of shallow machine learning techniques. Their holistic approach combines feature construction, local and global learning, classification, and regression algorithms.
20			To predict the next activity, this paper used stacked inception CNN modules. They propose a neural network architecture that performs better than RNN architectures.
21	These papers perform a new trace clustering approach on a sample of patients. Using it, they determine frequent sequence patterns on a sample set, rank patients based on these patterns, and determine clusters in an automated manner. Frequent sequence patterns are used to discover a process map.		

22	This paper introduces a method to reveal process models of interrelated speech intentions from conversations automatically.	The results may give new perspectives for analysing health communication and behaviour and discourse in general.	A domain independent taxonomy of speech intention is used to analyse the speech intention, a corpus of Reddit conversations is released, and they use a supervised classifier to predict speech intentions. This classifier is trained using ten-fold cross-validation and an approach to transform conversations into logs of verbal behaviour. These logs can be used for process mining.
23			This paper introduces an algorithm that uses process mining based on machine learning to handle concurrent and recurrent business processes.
24	The method in this paper to detect hidden sub-processes in healthcare is applied to event data for 'Altered Mental Status' patients that were extracted from a US hospital database (MIMIC-III).		This paper introduces a method to detect hidden healthcare subprocesses in an unsupervised way. This detection process is done using the hidden markov model (Viterbi algorithm). Their approach includes the enrichment of the event logs with HMM-derived states and remodelling processes in healthcare using state transitions.
25		This paper extracts BPMN models via process mining, then used to train a recommender system to get the best constraints. The process mining is done on Microflow execution log files. With this microflow approach, only essential rigidity is specified via constraints, allowing for more agility in the business process models because the remaining unspecified areas are automatically determined and planned, which means that they remain dynamically adaptable.	Learning problem: How can the best constraints be recommended through the use of machine learning? This is done in the paper by using an artificial multilayer network using DeepLearning4J
26	This paper designs a process mining pipeline that uses hospital data to discover healthcare pathways and enrich them. This is done using ProM. This pipeline is applied to a case study to discover appendicitis from the records. Machine learning techniques based on probabilistic programming are used to analyse pathway features that influence the recovery time of a patient.	It is possible to deduce reasons for the patients' longer recovery times by analysing the discovered pathway models. A probabilistic regression model is created to estimate the recovery time that can be useful for hospital scheduling purposes.	
27		It is shown how process mining is used to analyse running processes by taking process logs and using a network(community or cluster) analysis in the constructed network from the SAP business process log. By finding patterns in the network, one can use them as a model for decision support for a**ssigning** a new object to an existing pattern with a possible comparison of representative attributes(number of roles, average time, etc.) and the **actu**al behaviour in an organization.	Learning problem: How can a network be constructed by identifying clusters within the SAP process logs? This is done by using the Louvain method, which is based on the nearest neighbour analysis.
28		Process mining and machine learning in this paper are used to analyse ticket resolution and reduce the number of inputs required by users. By doing that, the system can pre-empt the user with potential information needs and help make decisions.	Learning problem: How to reduce user inputs during the ticket resolution by using machine learning? This is done using an SVM classifier-based pre-emptive model to pre-empt users to request additional information while submitting the ticket.
29			This paper analyses methods for predictive monitoring, which is concerned with predicting the future behaviour of running instances in the process. They find that encodings that represent ongoing process instances should take the relative positioning of events into account. Furthermore, training prediction models at fixed positions have the risk of creating spurious correlations between unimportant features and the outcome. Lastly, they find that to avoid this risk. One should use it for tracing the complete traces instead of shortening them to prefixes.
30		This paper integrates Nirdizati, a predictive process monitoring tool, into Apromore, a web-based process analytics platform. This integration can help managers recognize issues early on and reallocate resources from one case to another to prevent one case from running over time. To do this, they use the predictive models of Apromore,	For each bucket of feature vectors, they train a predictive model using one of four supported machine learning techniques: decision tree, random forest, gradient boosting and extreme gradient boosting (XGBoost)

		which, when trained, can predict various KPIs of running process cases from a live event stream.	
31			This paper evaluates automated process discovery techniques and the black box sequence model, predicting the next event. By comparing a range of process discovery and sequence modelling techniques, they conclude that LSTM networks accurately describe previously unseen traces.
32	Treatment recommendations for patients are given by using the action rule mining technique.		This paper introduces an approach using the action rule mining approach to recommend treatment options with the highest probability of achieving a particular outcome. Afterwards, they use a causal machine learning technique to discover subgroups where the treatment has a high causal effect on the outcome. This discovery is made using uplift trees.
33		This paper's introduced framework includes a prediction approach that is real-time and relies on machine learning. Using this, system performance can be predicted using advanced activity blocks for the autogenerated models. This prediction is based on live stream data coming from patients.	This paper introduces a machine learning approach for mapping low-level event logs to high-level activities. This mapping is useful when high-level labels of the low-level events are not available and happens in 2 phases. The first phase includes automatic labelling and the second machine learning-based classification. While labelling, a k-prototypes clustering approach is used.
34	This paper represents an extension of the Auto Simulation Model Builder (ASMB) introduced in another paper. This extension is done to be used in complex healthcare systems to improve their decision making.	The introduced framework includes a comprehensive solution for handling resources for complex decision-making processes around hospital staff planning. It can be used for single healthcare units' management, as well as multiple healthcare unit management.	This paper's introduced framework includes a prediction approach that is real-time and relies on machine learning. Using this, system performance can be predicted using advanced activity blocks for the autogenerated models. This prediction is based on live stream data coming from patients.
35			This paper's problem is that event logs from software configuration management(SCM) systems cannot use traditional process mining techniques because of missing activity attributes. This paper attempts to solve this problem by introducing a software process classifier. The used approach extracts activities from the log based on semantic features. Furthermore, it includes a new technique to associate events dynamically using a naïve Bayes approach.
36	This paper attempts to develop tools for analysing highly variable processes, which are applied to the example of patients with cardiovascular disease. To do this, methods and algorithms are introduced that can deal with large volumes of unstructured data.		Patients are identified in this paper by analysing their movements and classifying them using machine learning techniques. Furthermore, they identify patterns in the treatment processes using machine learning to predict the path in the area of phase space. SVC, Random Forest, KNeigbor, Logistic Regression, Naïve Bayes, GB (gradient boosting), ensemble V1 (RandomForest + KNeigbor) and ensemble V2 (RandomForest + LogisticRegression) were used.
37			This paper uses LSTM networks to predict the execution of some instances. The event log for this comes from the IoT and industry 4.0 domain. According to the literature, this event log is used to train and test these algorithms, which have predictions that indicate an acceptable prediction.
38		This paper focuses on predicting traffic prediction in wireless mesh networks using process mining algorithms, which can help predict traffic overload.	A methodology is proposed where traffic traces are converted from NS3 software to MXML format in PoM. Once it is converted, machine learning/process mining techniques can be used to predict traffic overload. The algorithms that are being compared for this are using wavelet neural networks, clustering approaches, Graph mining, or time series analysis
39	A proposal is given for infrastructure in healthcare systems that can support interoperability and data analysis. This approach applies process mining techniques to extract the overall picture of healthcare information from various contextual views and different abstraction levels and utilizes machine learning techniques to monitor patients' conditions and raise alarms.		

40			This paper proposes a new method for mining process models from email logs that leverages unsupervised machine learning techniques. The method is a 3-step process wherein the second step, emails, is clustered by the process model. For this, hierarchical clustering is used.
41			This paper introduces the web-based application Nirdizati, which can make predictions for running cases in a business process. It is utilized to predict the end outcome, the next event(s), the remaining time, or the workload for a day of each case. The user can choose from a list of prediction methods and prepared algorithms that can predict various KPIs.
42		This paper presents a demo of the proactive insight engine, which uses machine earning to automatically identify weaknesses in business processes, reveal their root causes, and give thoughtful advice on improving process inefficiencies.	
43			This paper uses statistical methods in process analysis to calculate the activities probability distribution using machine learning. This calculation is done on synthetic datasets. Furthermore, statistical methods are also used to calculate the probability density function. On top of that, they used classification and prediction. For the classification, they used Support Vector Machine, and for the prediction, they used supervised learning with Artificial neural networks. In the end, they modelled the clustered process by K-means clustering.
44		This paper attempts to deal with concept drift by developing a concept drift detection algorithm applied to a use case to predict the delivery time for a purchase. By doing this, internal service processes can be improved. The researchers figured out that the best results are achieved by combining incremental learning with retraining in case of concept drift. Furthermore, data scientists should use the last collected batch data for the retraining for the data selection.	
45		This paper describes a process mining methodology to yield accurate predictions of a worker's virtual training session's outcome. The effectiveness of the proposed methodology has been validated against a real use case.	To evaluate the accuracy of the learned predictive patterns, they use standard machine learning metrics. Specifically, they compute accuracy and F-score to evaluate predictions on testing traces.
46		Automative analysis of process models.	This paper introduces a new approach for the automated analysis of process models using process mining. This automation is possible by using libalf library that includes various learning techniques and algorithms based on the finite automata theory. For this, they use inductive machine learning and apply it to a hotel's property management system(PMS). To make the process discovery, they utilize a k-tail algorithm, also known as Biermann's algorithm.
47		This paper introduces a concept for using small sample learning (SLL) in business process management in process mining. They conclude that the SLL technique performs better than the traditional technique on the event logs.	Since SLL is mostly used in computer vision and NLP, it is still very neglected for BPM. This is important since many organizations can frequently not provide big and sufficient datasets. This paper uses small sample learning, which attempts to solve small sample sizes, which can be problematic for most machine learning techniques.
48		This work deals with intention mining that is a very active and promising research area. Intention mining is the ability to predict a user's goals. Knowing the user's intention can support the decision-making of the network administrators.	This paper aims to define their approach for intention mining based on a multiagent system in unsupervised learning. They build this approach with four layers to discover intentional process models automatically from event logs and recommend a new intention. Their approach has the following characteristics: First, they use the multi-agent's system by creating autonomous agents to reduce the complexity of discovering an intentional process model and recommend a new model. Second, they divide the main goal into sub-goals represented in different layers, and each layer will be done by a specific agent using their processes and responsibilities to attend to their goals. Third, the

			different agents communicate and collaborate in order to achieve the primary goal.
49		This paper introduces a methodology to explain why predictive models can sometimes make wrong predictions, ultimately improving accuracy. It uses post hoc explainers for identifying the features that induce mistakes most commonly.	
50		The fail detection approach is evaluated on 460,000 event logs, which results in an accuracy of 86.7% for the hidden semi-Markov model.	This paper's learning problem is related to failure detection in BPM systems based on machine learning models from event logs. This fail detection is done by using the hidden Markov model and the semi-hidden Markov model. While both models are acceptable, the latter can consider the duration of the process in one state.
51		To discover atypical business processes, this paper develops a business decision support system using machine learning.	This paper introduces two pre-processing techniques for process mining. The first one uses unstructured data to do an activity estimation, such as daily reports. The second pre-processing technique is a process estimation that can express the irregularity in a stochastic manner. In process estimation, the hidden Markov models prove to have the best performance. 2 machine learning algorithms are used to extract data and perform comparative validation. The first one is a Bayesian network that can create a model even with a small dataset. The second one is the aforementioned hidden Markov model that can model a customer's state as a latent variable and predict the state change. This prediction is made while taking into account the time series.
52		This paper is concerned with predicting the next best event, which is achieved by identifying possible decision points, mining its data objects, and applying probabilistic supervised learning algorithms to make predictions.	The classifier must provide probabilistic insight into the possible action and rank them, excluding decision trees to predict the next best action. Furthermore, it needs to be fast since the predictions need to be made in real-time. Finally, it has to learn from new observations continuously. Based on these requirements, they chose to construct a hybrid naïve Bayes classifier.
53		This paper tries to benchmark machine learning algorithms and attribute pre-processing techniques commonly used in behavioural targeting. For evaluation, they use the conversion rate as a performance metric. This makes it possible to identify prospective customers in most cases. This paper predicts whether a new web user is likely to be converted based on other customers' profiles in the past.	To do the customer prospecting, data samples have to be selected that make a valuable analysis of the underlying patterns possible. This prospecting is done through predictive modelling, which uses historical data to make predictions about future events. Decision trees and KNNs were used over other techniques like logistic regression and collaborative filtering classification methods.
54		The goal of this paper is to predict traffic overload while changing the network topology. To do this, a combination of machine learning and process mining techniques is used to analyse the traffic coming from several moving nodes. These algorithms are wavelet neural networks, clustering, graph mining, and time series analysis.	
55		To improve the student experience in massive open online courses (MOOCs), this paper attempts to make early predictions using a combination of process mining and machine learning. These predictions help combat the high dropout rate in these courses and improve the students' learning experience.	To make predictions about MOOCs' student experience, this paper uses four machine learning classification techniques: logistic regression, naïve Bayes, random forest, and k nearest neighbour. This monitoring is done weekly to predict their overall performance.
56		This paper introduces a framework that is concerned with building a continuous auditing environment. This framework's focus is the transaction verification level and combines techniques from data mining and process mining. Furthermore, it also includes the auditor as a human expert, giving inputs for the machine learning classifications.	This paper uses three-way decision rules to allow the classifier to deal with uncertainty, including the rough set theory and the fuzzy sets theory. The active learning mechanism they propose combines machine learning techniques and human expertise in the form of the auditor. This auditor acts as an oracle to feed the machine learning classification algorithms.
57		This paper aims to detect malware and to study phylogeny, for which they use process mining. Process mining is used to identify relationships and patterns from app system call traces to characterize its behaviour.	To do dynamic malware detection and malware phylogeny tracking, process mining is used to extract a declarative model from system call traces of malware and safe application, which are then compared to characterize the malware variant. For

			classification, machine learning algorithms are used and compared.
58			To predict activities, this paper uses the WoMan framework for workflow management. Two strategies are used: the former is superior, while the latter is based on a consolidated naïve Bayes approach.
59		This paper introduces an approach to improve the accuracy of data extraction and task identification for process mining. To illustrate their approach, a business expense reimbursement use case is used.	This paper's problem is that data extraction from massive event log databases for process mining requires rich domain knowledge and advanced database skills, which are not always available and are labour-intensive. To solve this, this paper introduces an intelligent approach to data extraction and task identification, which formalizes them as a problem of extracting attributes as process components and relations among process components. This approach is made using sequence kernel techniques.
60		An approach is introduced to improve understanding of process mining clustering techniques that outperforms alternatives and can identify features that lead to shorter and more accurate explanations.	This paper attempts to improve human understanding of clustering techniques in process mining, which are usually analysed by visual inspection. An approach is introduced that applies machine learning in a post hoc fashion through supervised learning with support vector machines. This machine learning is applied to the cluster results to learn rules to understand why a specific instance is included. To do this, specific control flow-based variable features are used.
61	.	This paper introduces an approach to analyse detected anomalies to see which event within one execution of the process causes the anomaly. This approach is an extension of their previous research in 2016 but is extended with improved performance and an evaluation of better datasets	This paper attempts to detect anomalies in internal processes to identify possible fraud or inefficiencies. This detection is done using autoencoders, which detect and analyse anomalies. The method does not need any previous knowledge and can be trained on noisy datasets which already include anomalies.
62		This paper is concerned with improving the temporal stability of predictive monitoring approaches. This type of monitoring is essential in environments where users have to decide and respond to the predictions they receive. Traditionally, accuracy is the metric that is usually optimized. This paper finds that using the XGBoost and LSTM neural networks has the highest temporal stability, which can be enhanced by optimizing the inter-run stability of random forests and XGBoost classifiers. Furthermore, time series smoothing can also improve temporal stability, even if it comes at the cost of lower accuracy.	This paper attempts to evaluate process monitoring methods concerning their temporal stability. Temporal stability can tell how much successive predictions differ from each other. The more they differ, the more volatile the classifier is, while less difference will lead to a soother time series. They found that the best results could be achieved by evaluating seven predictive monitoring methods using the XGBoost and LSTM neural networks.
63		This paper attempts to introduce a white-box approach to predicting quantitative performance indicators for running process instances. This is done by first predicting the activities level indicators and then aggregating them at the process instance through flow analysis techniques. By doing this, one can predict the remaining cycle time, cost, or the probability of deadline violation. Most other approaches utilize a Blackbox approach, while this paper introduces a white-box approach.	To predict the performance indicators of a running case, this paper uses the XGBoost 34.
64		This paper introduces an approach to filter out chaotic activities from event logs, decreasing the process models' quality discovered by process discovery techniques. Traditionally, this is done by filtering out infrequent events, but this paper shows that doing that does not solve the problems caused by chaotic activities.	Four techniques are proposed from information theory and Bayesian statistics to filter out chaotic activities from event logs. Experiments on 17 datasets found that the new methods outperform the traditional filtering out of infrequent events. However, the performance of these methods is highly dependent on the characteristics of the used event log.
65		The proposed method for improving resource allocation is used in real scenario-based experiments, outperforming several other methods. This approach can help in making better operational decisions while at the same time finding more appropriate resources. Furthermore, it helps to maintain a reasonable workload for the employees.	This paper attempts to improve resource allocation in process management by using an entropy-based clustering ensemble method. Process mining is used to analyse performance elements and competence metrics. The entropy clustering ensemble method is used to capture multiple types of preference patterns to gain each task's unique preference.
66		This paper introduces a multi-stage hierarchical framework to cluster processes with many events based on business logic. This framework helps	A hybrid approach combines rule-based mining techniques with a new form of agglomerative hierarchical clustering to cluster business processes.

		increase the understanding of the business users' cluster results since these results are often not business logic-driven. This paper first introduces the term contrail processes, which describe how complex business processes can be characterized using contrail-like models. Furthermore, they introduce an algorithm for hierarchical clustering to discover the clusters from these contrail processes.	The initial event log is first decomposed into high-level business classes before feature engineering is used. This feature engineering is then based on the business context features.
67		The deep neural network architecture proposed in this paper can extract high-level features and labels from datasets with no topological organization. The bidirectional neural network reduces the number of problem features but still maintains the network's discriminatory power.	To reduce complexity and the number of parameters in deep neural networks, one can pool layers. While these operators can deal with a single label and multi-label problems, they can also reduce the feature space. In case there is a multi-label problem, this should be done in the label space. However, for datasets with an explicit topological organization, the existing pooling operators are not enough. This paper introduces a deep learning architecture using bidirectional association-based pooling layers, which helps extract high-level features and labels in multilabel classification problems.
68	Using a real-time location system to track hospital staff and patients can optimize workflows in a hospital. Because of privacy regulations, this is hard to achieve, however. The hospital and the labour union in this paper have a two-party protocol, in which they cluster staff members by looking at how frequently they visit patients. Using a cryptographic technique called Secure multiparty computation, they provide a secure way to track.		Clustering of staff members.
69			The problem of this paper is to make predictions while some paths are executed in parallel. These predictions should be made on unstructured data. They propose a way to determine whether parallel paths are independent and model these paths as independent. Five different methods are compared, which help represent execution traces as path attributes for a prediction model. The methodology is tested on a marketing campaign, which uses decision trees for predictions. Finally, they compare the prediction and accuracy that the model produced.
70	They are preventing medical fraud.		The size and complexity of healthcare make it a target for fraud. Because of this, statistical methods can help medical auditors to find fraud within claims data. This paper uses the unsupervised Bayesian hierarchical method as a pre-screening tool to find hidden patterns in medical procedures.
71		This paper defines seven requirements necessary for a successful implementation of their architecture. Furthermore, it introduces a generic architecture of prescriptive Enterprise systems. This architecture comprises five system elements that can recognize intricate patterns in events within multi-sensor environments. Its implementation is done to compare them with past data for calculating predictions. These predictions are then utilized to give the ideal course of action while the process is ongoing to minimize or maximize KPIs. On the other hand, predictive process analytics is different from other descriptive methods because it aims to predict future process conditions by looking at historical process executions. The underlying goal is to recognize rules and patterns, which can forecast running process instances. Furthermore, prescriptive process analytics utilizes the trained prediction models to forecast process executions. Using these forecasts predicts the following running process instances. Because of this, one can say that	The second component of the architecture, "Induction of Complex Event Patterns and Their Detection in Process Instantiations," is utilized to automatically recognize patterns in the event log that influence process outcomes. This recognition is done by the use of pattern matching and recognition algorithms. These algorithms can determine the importance of events that impact the probability of running different process branches or consuming more or fewer resources or longer and faster cycle times.

		prescriptive analytics help to control business processes proactively.	
72		Predictions provided by their technique could create early alerts for workers about the probability of essential or unwanted outcomes in an ongoing case instance.	The learning problem here is that they want to predict the likelihood of a future event in semi-structured business processes: This is done by using an instance-specific probabilistic process model(PPM) that can be transformed into a Markov chain under non-restrictive circumstances.
73			This paper's learning problem is to find the overall entropy to use as a measure of variability to know whether to use imperative miners or declarative miners in process mining. The entropy is found by using knn. For the entropy rate, Lempel-Ziv and specific variants of k-block estimators are best.
74			The learning problem in this paper is to detect problem-solving strategies in simulation or game-based assessments. This detection can be done by using decision trees or a supervised neural network. If student strategies are unknown, they need to be identified from action patterns, which can be seen using process mining. Afterwards, one can use cluster analysis or unsupervised machine learning to identify the student strategies. This paper uses the hidden Markov model.

12.3 FUTURE RESEARCH POSSIBILITIES

TABLE 26: FUTURE RESEARCH IN LITERATURE REVIEW

Sr. No	Future work/Open problems
1	Future research may include support for more types of process models, such as EPCs, and the use of different algorithms from data mining. Furthermore, data mining techniques could also be applied beyond the analysis of decisions made in the paper. A free specification of the learning problem could be used on the data. An example is to mine association rules.
2	Future research may improve because this paper's implementation only discovers guards that are conjunctions/disjunctions of expressions of the form variable-operator-constant (e.g., $x > 4$). Furthermore, as they do not consider post-conditions, one can also try to mine these conditions
3	More real-world case studies and the adaption and discovery of different factors that may be essential characteristics of the logs
4	A weighting schema could determine the relative importance of a trace profile in the co-training strategy. Furthermore, model-based stopping criteria for the co-training process may also be defined. Aside from this, the strategies can also be parallelized.
5	The calibration of the parameters could be improved, and it should be checked if only using workhours in the prediction will lead to better results. Finally, the approach should be used in a real scenario.
6	More research is needed to determine how decision performance can be increased and how information systems can give actionable, interpretable, domain-appropriate, and concise conclusions to improve decision making.
7	
8	Additional systemization, integration with predictive models to simulate the behavioural and clinical evolution of cases, a generalization of the solution, and implementing a data-driven solution to deal with large and diverse datasets.
9	Future research should take into account additional context information coming from event logs. Furthermore, this context information should be evaluated in comparison to the frequency of the processes. Furthermore, an approach could be introduced that can identify the main processes in the first place, and additional clustering algorithms can be tested.
10	
11	
12	Experiments with other systems and larger event logs should be done. Additional information should also be provided to users, like domain knowledge about known cycles or issues.
13	The generalisability and replication value should be tested by providing a set of benchmark metrics for analysis. Furthermore, high and low performers could additionally be compared and the first half of the term and second half. Higher-order markov models and hidden markov models can also be used.
14	The event classification results need to be validated with human-tagged event logs. Then one can compare the model produced using these event logs with the model produced in this paper to verify the results. Furthermore, other text clustering techniques can be used as well, and further validation can be made by creating process models with the help of domain experts. This paper's approach can also be applied to other domains, like, for example, healthcare.
15	In the future, one can consider processes containing Xor/Or gateways, where only the one that fulfils the condition must be executed. Aside from this, performance indicators like development time or change propagation can further validate the proposal.
16	In future work, one can assist new users or users unfamiliar with the software by using their previous activities as a guideline. By knowing the intentions, one can recommend them fitting strategies and activities.
17	
18	
19	In the future, one can research other clustering solutions to improve the prediction performance. Furthermore, this paper's method can also be extended to stream learning to make predictions with models that can change over time when new traces come in. Prescriptive learning theories may improve the approach with guidelines that can describe what to do to reach the desired outcome or extend the model by including reactions to alerts based on the predictions.
20	A future direction could be to predict the next activity and its execution time simultaneously. Furthermore, the naïve inception module could be adjusted to a more complex one.

21	To improve F1 scores, other strategies for selecting sequence patterns can be considered. Aside from this, one can experiment with different sample sizes and their effect on the F1 score. One can also use the frequent sequence patterns to discover formal process models for the cluster pathways. Furthermore, the map can also be validated by working with domain experts and using it in other hospitals or other patients.
22	Future work still needs to be done to improve and validate the overall approach.
23	
24	The experiments in this paper can be done with larger and more complex cohorts of patients, and the processes can be compared. Furthermore, the approach needs to be validated by applying it to local healthcare data.
25	Support for BPMN 2.0 may be included in the future, integration of advanced integration and verification techniques semantic support for the discovery service, support for compensation and long-running processes, improved declarative and semantic support, a tuning of the recommender system, and additional empirical validation.
26	Future work may consist of looking at the business process monitoring tools used to mine healthcare pathways.
27	This paper's method can be made more universal by creating a compatible solution for any standard SAP solution. Furthermore, different methods for visualization can be used.
28	The detection model in the approach can be extended by identifying cases where unnecessary information is asked.
29	The first improvement that can be made is improving the use of sub trace encoding and the global prediction approach. The second improvement worth making is to use tools that can support the performance and reliability assessment. This can be done on real-world data.
30	
31	A more extensive collection of event logs and different process discovery techniques and sequence modelling methods may be used.
32	Future work may include identifying contextual variables like stakeholders or weather and validating the causal relations discovered with these variables. This could be done by using structural models, for example. Including feedback can do further validation from users or use external validation through the use of A/B testing.
33	
34	
35	Improvements can be made by having a more significant database sample which will improve the classifier's accuracy. Furthermore, one could add weights to define the importance of the indicators used in the event mapping. Finally, a standard method could be used to evaluate the results of unsupervised machine learning.
36	Future research may include studying various other graphs, the community's identification, improved calculation performance, using different machine learning methods, and visualizing the results. Furthermore, the toolbox can be used in other domains as well.
37	One could consider event logs with more traces and include two or more classes to predict the next activity.
38	
39	One could do more research in machine learning, data mining, NLP, and process mining, to investigate their impact on interoperability and healthcare analytics.
40	More powerful semantic similarity measurement tools may be used like Word2vec, relevant for the activity recognition phase. The researchers plan to create a recommendation system for emails that recommend possible activities to add to the users' to-do lists. Furthermore, efficiency and quality may be tested on a more extensive dataset.
41	
42	
43	In the future, non-linear relations within sequence data instances may be considered. The work may also be extended to be used for real-time sequential process analysis of actual data.
44	More sophisticated approaches need to be created to deal with concept drift since this paper only proposes 3. The work may also be adapted to handle the target variable's transformation from a regression problem to a multi-class classification problem. More use cases could also be used for further validation since this paper only looked at one use case. By doing this, a more generalized recommendation may be given.
45	New data collected in future training sessions may be used to test the introduced approach's effectiveness. It may also be extended to stream learning to mine patterns that may change over time. Another extension may be transfer learning to transfer the learned patterns with data collected in the training session with data collection in a new session.
46	
47	Their research initially focuses on implementing and evaluating the methods described above. Different methods, possibly for different BPM areas, will be tested for suitability in the next step. The long-term goal is to offer well-evaluated methods for the various requirements of BPM that perform better on small amounts of data than conventional methods.
48	
49	An improvement would be implementing more sophisticated abstraction mechanisms to discover complex conditions such as comparisons to numerical attributes and inequalities between categorical attributes. The experimentation may also be done with other post hoc explainers to compare and evaluate their results.
50	
51	
52	
53	Experimentation may be done on more sophisticated models o other algorithms to improve performance. Furthermore, individual settings can be tested, including the model's temporal dimension, the model's combination with a content-based approach, and more category and continuously based attributes that may specify the time spent on each category. More user and publisher information from third-party media providers may also be considered.
54	
55	An improvement could be made by using a more extensive dataset.
56	Future research could involve a pilot implementation of the presented framework as a validation.
57	
58	Other domains may be looked at to apply for this study's work, like Industry 4.0 ones. Furthermore, the prediction module may be embedded in other applications as well.
59	In future work, the ranking attribute for the attributes may be validated more and developed even further. Furthermore, additional datasets may be used to test the performance of the learning methods. Moreover, more experimentation could be done on process policies to test the portability of their approach.
60	More domain knowledge may be incorporated, instance-level explanations may be aggregated, and the practical application of SECPI needs to be validated.
61	

62	A more robust notion of temporal stability may be created for the future, which would not necessitate a penalty for changing the prediction when an event has a relevant signal. For developing an adaptive smoothing method, techniques for early sequence classification may be of use. Finally, temporal stability can be extended to other prediction tasks like multi-class predictions and regression.
63	Further research needs to be made to identify factors that can impact this paper's approach's performance. By changing the derivation of the flow analysis formula, more quantitative performance indicators can be predicted. An extension can also be made to be able to handle more complex models with overlapping loops. Finally, the approach may also be extended to predict cost-related properties.
64	This paper's researchers plan to create a hybrid activity filtering technique that utilizes this paper's techniques to predict how removing one activity can impact the final results. This filtering could be done by using supervised learning techniques from the domain of data mining.
65	Combining domain knowledge with process data could analyse other resource characteristics such as knowledge or skills to improve resource recommendations. Furthermore, one can look at how allocation policies are adjusted in a dynamic process environment.
66	
67	
68	Future work could focus on testing the solution on real-world data to assess the data analysis's impact. Furthermore, other machine learning techniques could also be implemented, leading to a comparative evaluation of the different possibilities.
69	
70	A semi-supervised extension can be done to get better accuracy. This should be possible if there is labelled medical data available. Furthermore, using Bayesian non-parametric methods does not have to use a fixed number of groups anymore. By using temporal models, one can see the changes in billings over time. Moreover, one can take into account the covariate information of providers and procedures. Using a Bayesian analysis, one can distribute the providers' hellsinger distance compared to his peers. Providers on the far-left side of the tail can then be considered to be outliers.
71	In this paper, the researchers have a product that looks into the concept's technical implementation. This is necessary since this paper abstracted the details of the technical realization.
72	The techniques in this paper may be applied to entire semi-structured business process management systems. Furthermore, this paper must research the uncertainty created when using the prediction technique and the measurement design algorithms. Finally, while this paper assumes that existing mining techniques are enough to mine process models in semi-structured businesses, this may not be true in real life. This means that more research needs to be done concerning process mining for semi-structured business processes.
73	A clear partitioning between logs into imperative and declarative classes needs to be done. A way needs to be found to estimate whether declarative or imperative mining produces better models. With a clearly defined error measure and one or more entropy estimators, one could, for example, use them as input features to a classification algorithm to determine which mining approach to use for a log.
74	1. Compare different scaling methods 2. Compare different statistical and data mining methods 3. Check the reliability, validity, and fairness issues of simulation or game-based assessments

12.4 NUMBER OF PUBLICATIONS PER YEAR

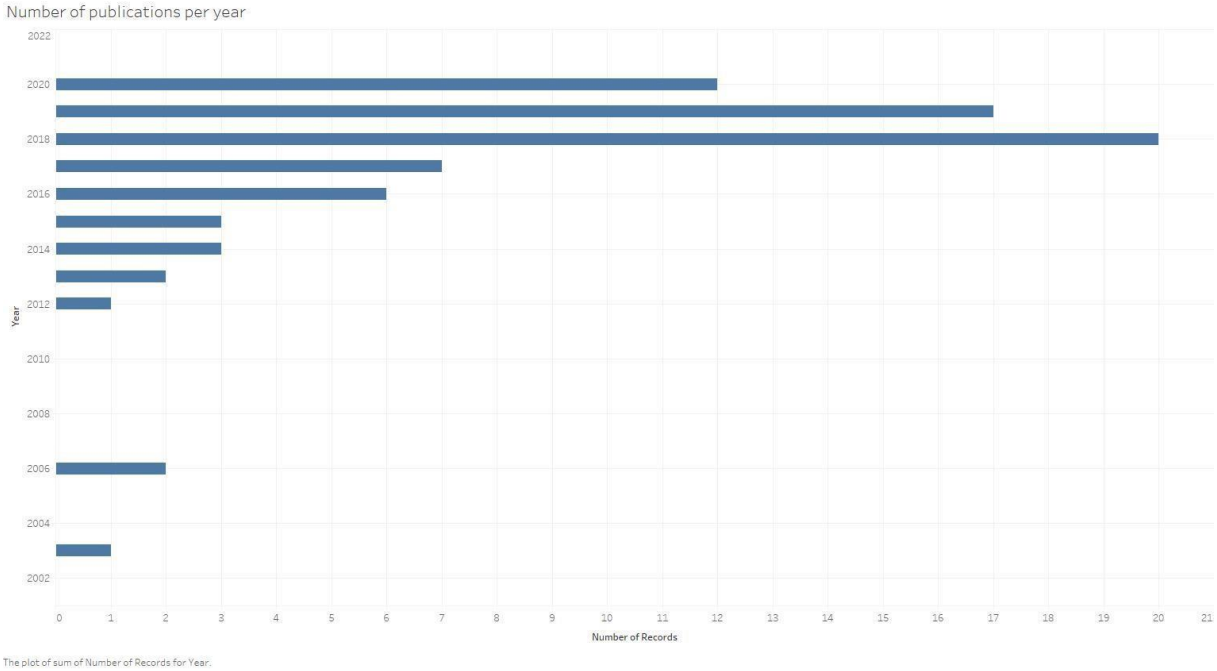

FIGURE 25: NUMBER OF PUBLICATIONS PER YEAR

12.5 RESEARCHER FREQUENCY

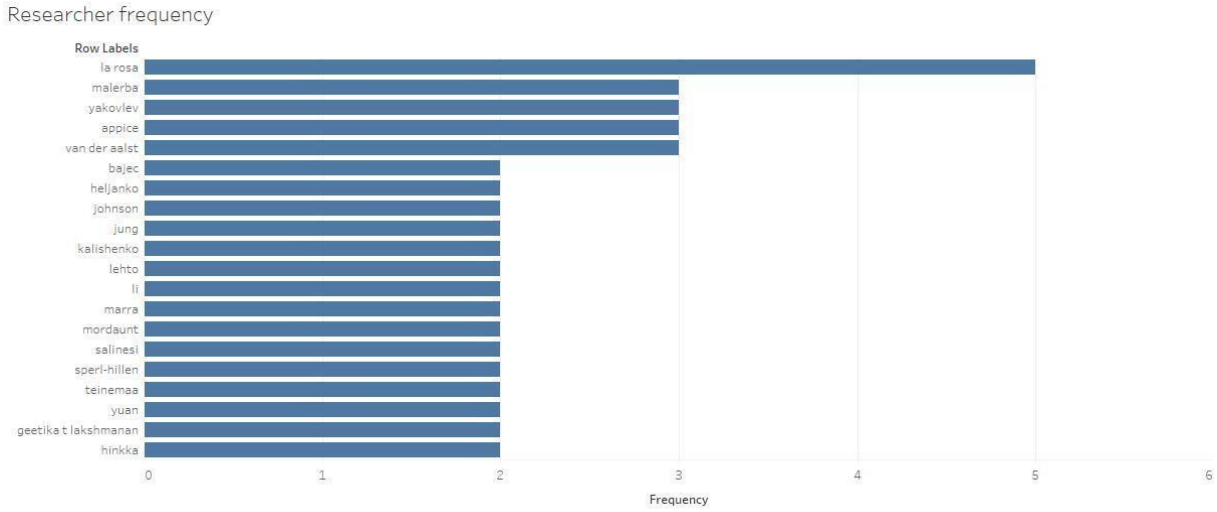

FIGURE 26: FREQUENCY OF RESEARCHER CONTRIBUTIONS

12.6 NUMBER OF CITATIONS

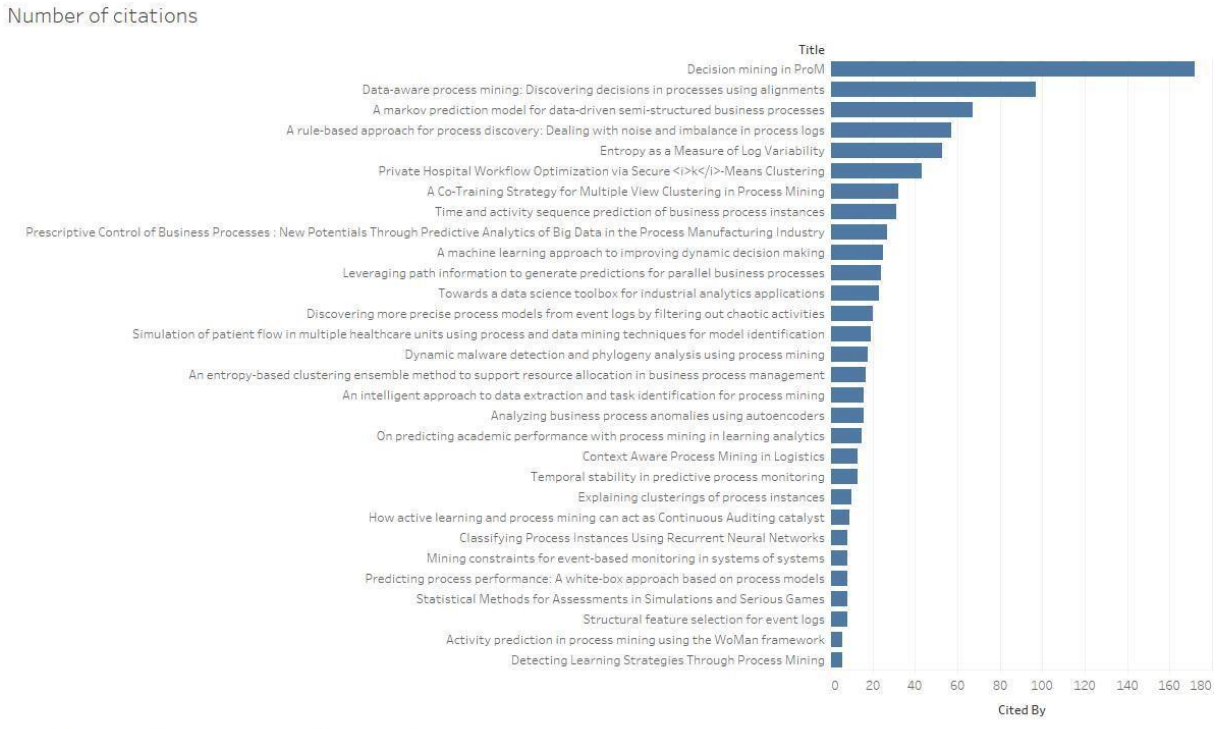

FIGURE 27: NUMBER OF CITATIONS FOR TOP 30 PAPERS

12.7 DATASET DOMAINS

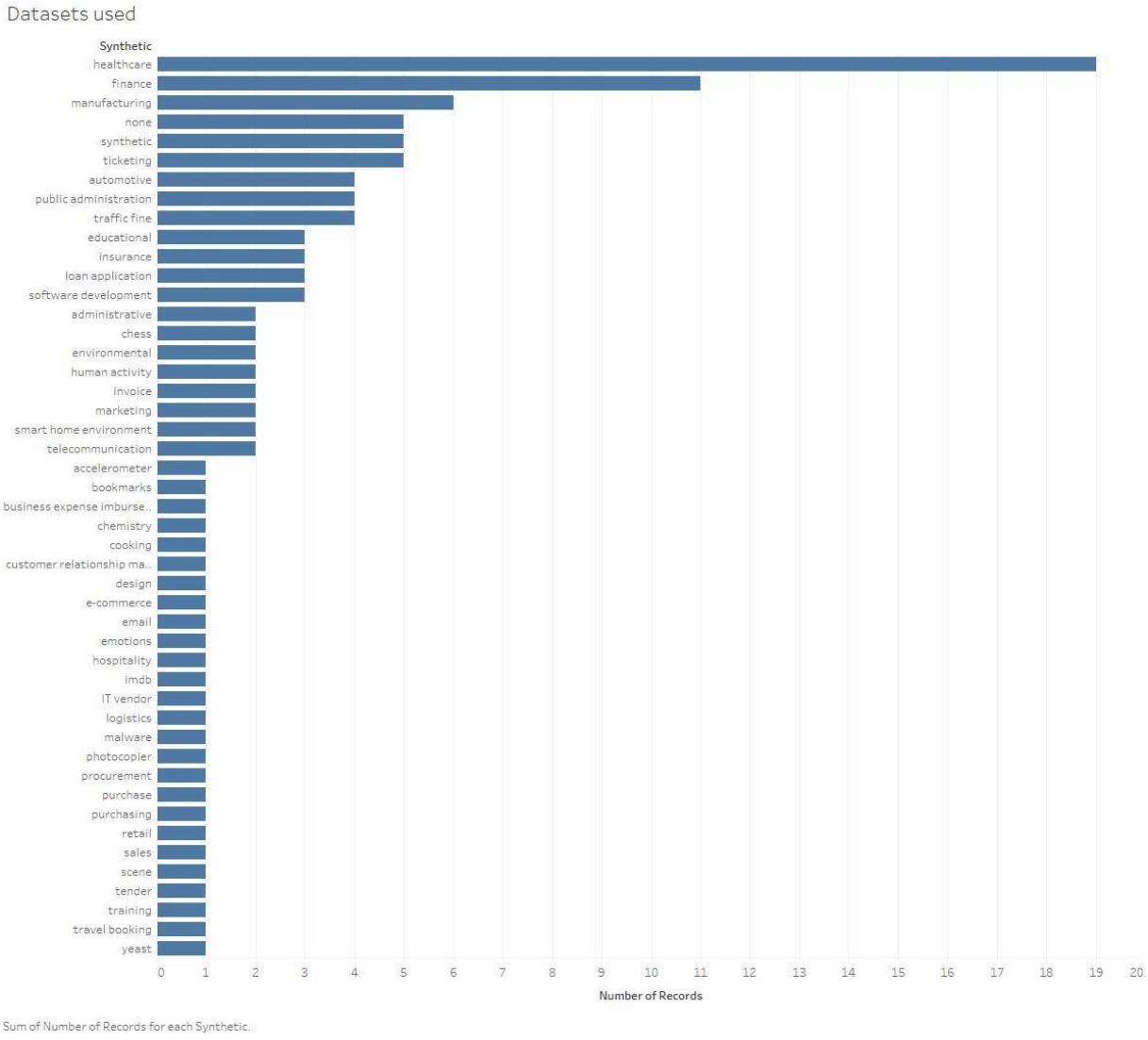

FIGURE 28: DOMAINS COVERED IN THE DATASETS

12.8 MACHINE LEARNING IN PROCESS MINING

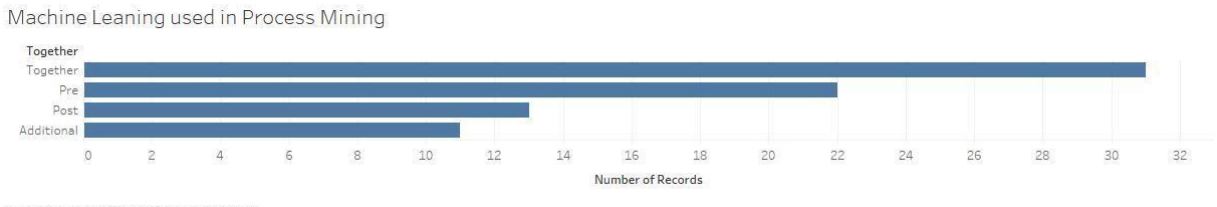

FIGURE 29: MACHINE LEARNING IN CONJUNCTION WITH PROCESS MINING

12.9 HB DISTRIBUTION FOR CHILDREN YOUNGER THAN 5

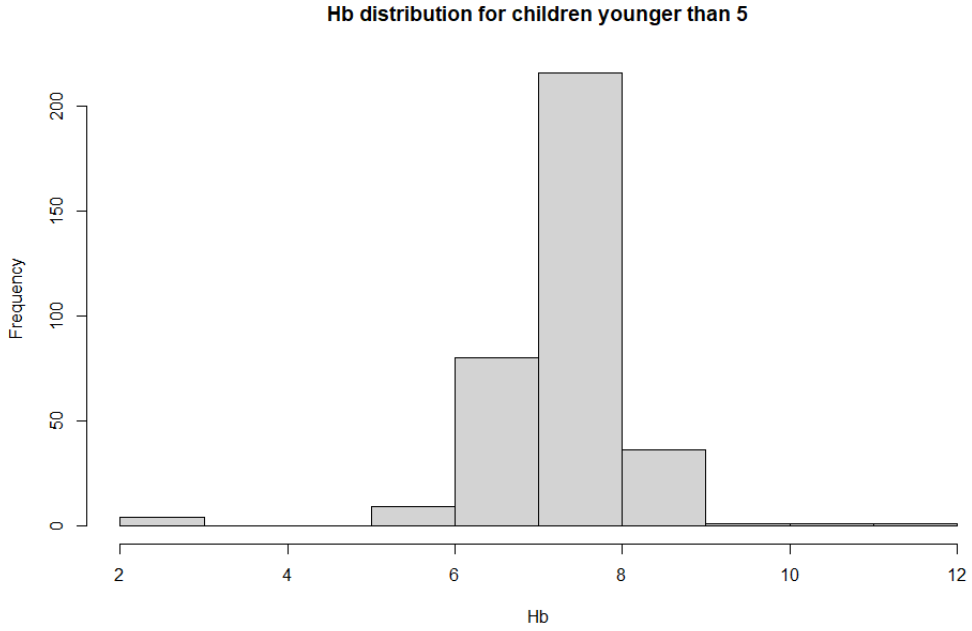

FIGURE 30: HB DISTRIBUTION FOR CHILDREN YOUNGER THAN 5

12.10 HB DISTRIBUTION FOR CHILDREN BETWEEN 5 AND 11

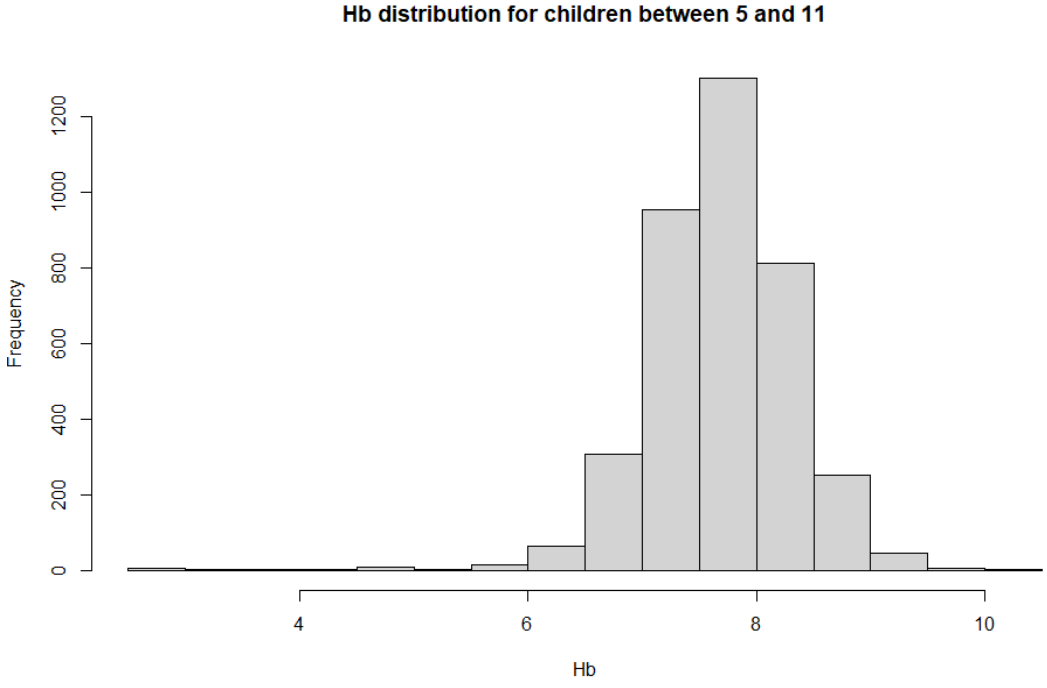

FIGURE 31: HB DISTRIBUTION FOR CHILDREN BETWEEN 5 AND 11

12.11 HB DISTRIBUTION FOR CHILDREN BETWEEN 11 AND 14

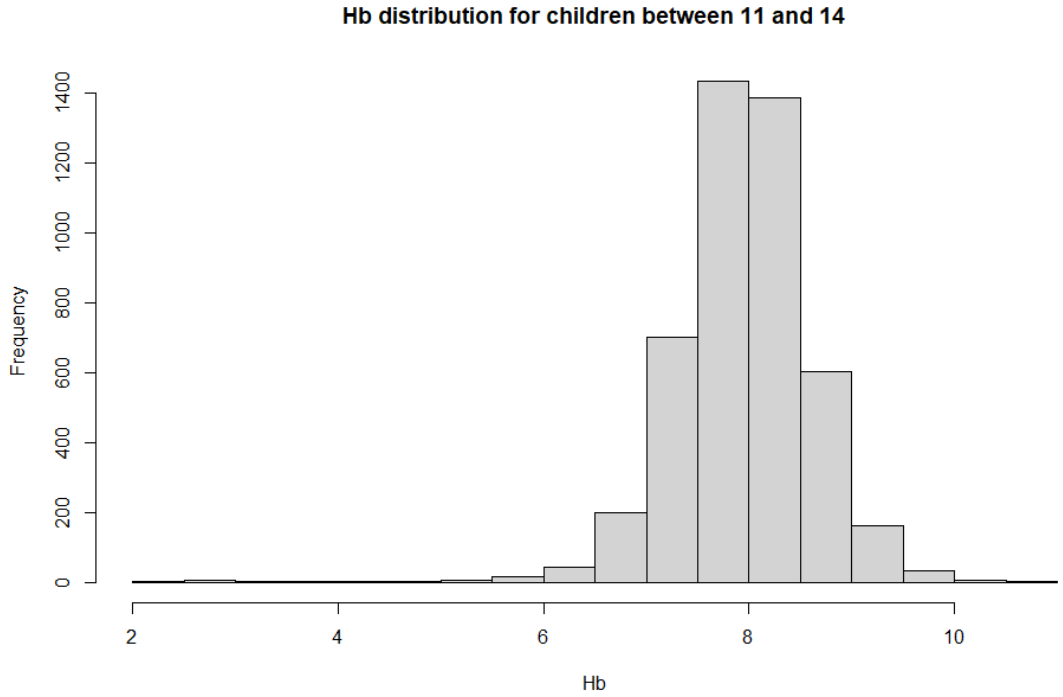

FIGURE 32: HB DISTRIBUTION FOR CHILDREN BETWEEN 11 AND 14

12.12 HB DISTRIBUTION FOR FEMALES ABOVE 14

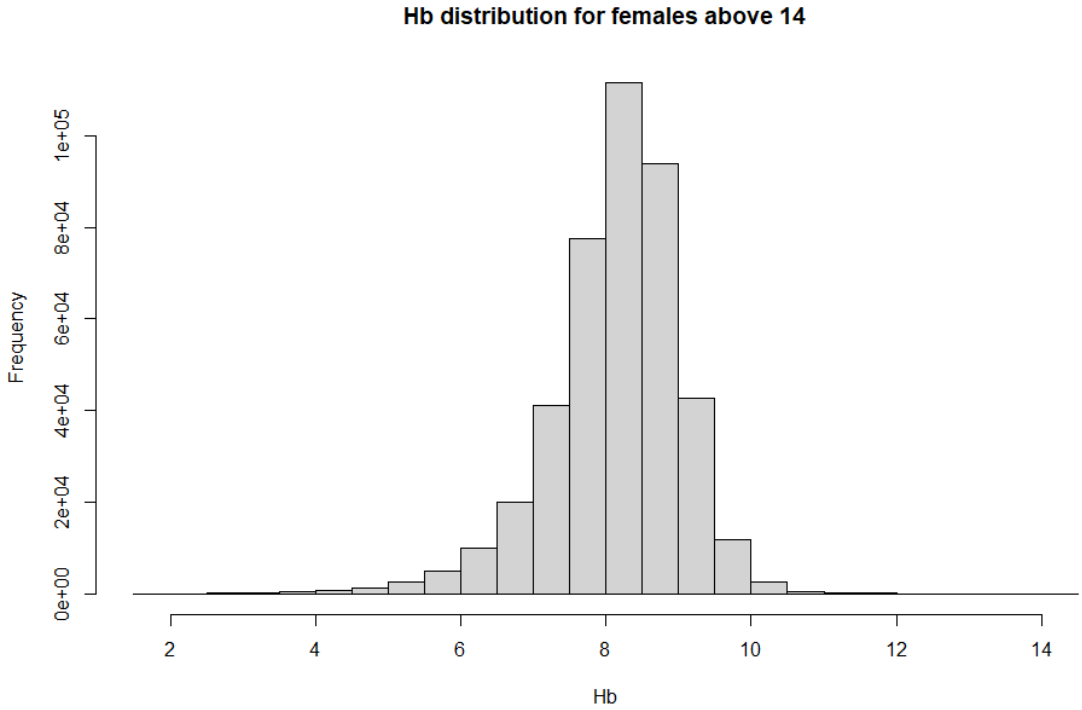

FIGURE 33: HB DISTRIBUTION FOR FEMALES ABOVE 14

12.13 HB DISTRIBUTION FOR MALES ABOVE 14

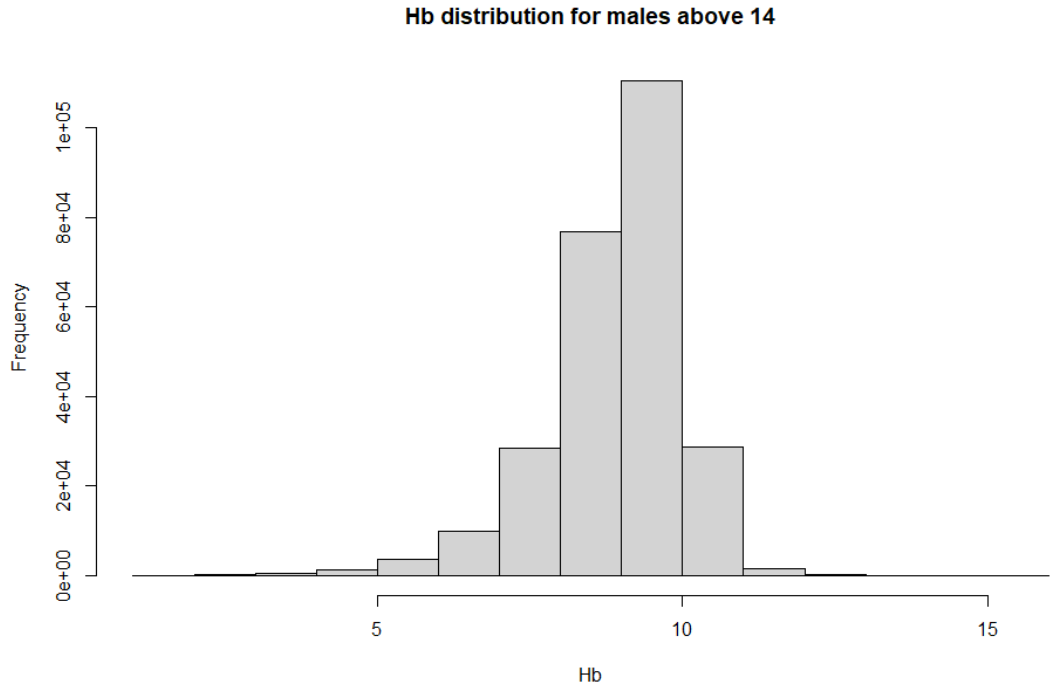

FIGURE 34: HB DISTRIBUTION FOR MALES ABOVE 14

12.14 MACHINE LEARNING ALGORITHMS

For predicting anaemia, three machine learning models were compared based on a selection of specific performance metrics. The models selected for this thesis are K-Nearest Neighbours(KNN), Random Forest, and Naive Bayes. The purpose of these models is to provide predictions that classify patients into anaemic or non-anaemic based on training with various biomarkers. Furthermore, additional comparisons were made between models predicting severity and type of anaemia.

KNN is a classifier that classifies an unknown sample by learning from the classification of its surrounding samples. To do this, the distance between the unknown sample and all of the surrounding samples are calculated. The data point is then classified based on the k neighbours with the shortest distance [7]. By principle, KNN utilizes the Euclidean distance, which can be calculated by using the following formula [190]:

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

The use of KNN models in healthcare for diagnostics has been exemplified by many studies [62], [134], [114], making its use justifiable for this thesis.

Aside from KNN, Random Forest is the second machine learning model that is used to predict anaemia. The method was developed by Breiman in 2001[86] and uses a collection of decision trees with a controlled variation where each tree represents a classifier to determine the classification of the unlabelled data instance. To do this, each classifier casts a vote for the predicted class. Using majority voting, the class label with the majority of votes is used to classify the data point [80]. Like the KNN classifier, decision trees are also used extensively in healthcare, such as breast cancer predictions [68] or outcome prediction in antibody incompatible kidney transplantation [173]. Furthermore, random forests adapt to sparsity, meaning that its convergence rate depends only on the number of solid features and not on noise variables within the data [60].

The last machine learning model used for the analysis of the anaemia dataset is Naive Bayes. Even though it is a classifier that assumes independence between features, which is not the case in the anaemia dataset, research has shown that it can also be used for functionally dependent features [84]. Indeed, the accuracy of the Naive Bayes is not directly correlated with the degree of feature

dependencies in the dataset [84]. On top of that, its use in many healthcare-related application areas makes it feasible for use in this diagnostic scenario [12], [155].

Because of the large amount of data and the high number of models to be compared, machine learning implementations capable of handling high dimensional data were chosen. For KNN, this was the "fastKNN" package in R. For random forest, the "ranger" package was used, and for the naive Bayes, the "fastNaiveBayes" package was chosen, both in R as well. Since most of these methods did not contain a cross-validation function with inherent normalization and class imbalance handling out of the box, cross-validation was performed manually. This was done using a foreach loop to run the code in parallel over multiple cores. The implementation was performed on Rstudio Server using an AMI in AWS to ensure high computational power. Normalization was implemented for the KNN and the Naïve Bayes but not for the random forest classifier. The reason for this is random forest's dependency on decision rules instead of distances like the Euclidean distance, for example.

Predictions for anaemia were made based on the machine learning pipeline described in the methodology section. After cleaning the anaemia dataset and defining a target variable, less critical biomarkers were deleted based on an expert opinion. Additionally, variables with a very low number of observations in the single digits or low double digits were deleted. Since the chosen machine learning algorithms require an absence of missing values, imputation methods are needed to address this issue. Because of computational reasons, the initial imputation was done using the median of each biomarker column. Because such single imputation methods can lead to bias, a comparison was made to other missing value handling techniques. The effect of these methods on the selected metrics can be found in section 6, which covers the second research question.

12.15 FEATURE SET 1 FOR PREDICTING ANAEMIA

TABLE 27: EMBEDDED FEATURE SELECTION METHOD FOR ANAEMIA PREDICTIONS

Biomarker	Importance
Ery	5.141328e-02
MCV	2.287826e-02
Transf_verz	2.283106e-02
Ret_He	1.965687e-02
TYBC	1.736411e-02
MCH_n	1.462395e-02
Transf	1.446468e-02
Reti_n	1.278921e-02
MCHC	1.253169e-02
LD	1.113416e-02
Ferritin	7.651413e-03
Fe	6.054338e-03
RDW	6.008566e-03
Kreat	5.217984e-03
GFR_MDRD	2.897521e-03
LDL_chol	1.108854e-03
Leuco	1.026843e-03
Triglyce	9.947921e-04
CKD_epi	9.762196e-04
Trombo	8.029662e-04
HDL_chol	7.739661e-04
TSH	7.428026e-04
Neutro	7.192642e-04
Ureum	5.995817e-04
Na	5.456610e-04
Gluc_n	4.928960e-04
gGT	4.893862e-04
Gluc	3.947306e-04
NTproBNP	3.065693e-04
HbA1cN	2.961608e-04
PSA	2.854002e-04
K	2.545288e-04
Kreat_U	2.537282e-04
Chol	1.991031e-04
Cho_HDLR	1.939349e-04
Bili_tot	1.885528e-04
Lymfo	1.706307e-04
CRP	1.436077e-04
FT4	1.224053e-04
Mono	9.403322e-05
X25_OH_D3	9.299313e-05
Act_B12_R	8.643755e-05
Ca	8.300130e-05
Foliumz_R	7.578650e-05
ALAT	6.733379e-05
Alb	6.342729e-05
AlbKr	6.197101e-05
ASAT	6.187937e-05
VitB1	3.760294e-05
Haptoglo	3.676144e-05
VitB6	3.204763e-05
AF	2.910689e-05
Segment	2.363910e-05
Bili_dir	2.035980e-05
Erytrobl	1.727545e-05
Mg	1.641673e-05
P	1.607209e-05
Eo	1.110604e-05
Lipase	1.053521e-05
Baso	1.024587e-05
Staaf	8.336615e-06
Uraat	7.598488e-06
TN_T_HS	7.317065e-06
Meta	5.721791e-06
Myelo	5.430860e-06
PSA_Ratio	4.644683e-06
TE	4.644330e-06
D_Dimeer	2.383262e-06
Promyelo	1.084669e-06
Cl	2.201531e-07
Krea_Ul	2.119032e-07
Testost_lum	1.572126e-07
PTH	1.419234e-07
aTTG_n	1.412579e-07
Bili_U	-9.433407e-08
Trombo_cit	-1.179242e-07

12.16 FEATURE SET 2 FOR PREDICTING ANAEMIA

TABLE 28: LASSO FEATURE SELECTION FOR ANAEMIA PREDICTIONS

Biomarker	Values
Ery	-2.63
MCV	-1.51
MCHC	-1.20
Transf	-1.10
Fe	-0.40
Na	-0.20
HDL_chol	-0.16
Triglyce	-0.12
LDL_chol	-0.10
Bili_tot	-0.04
Ca	-0.04
Kreat_U	-0.04
Lymfo	-0.04
Meta	-0.04
Mg	-0.04
Cl	-0.03
P	-0.03
ASAT	-0.02
Segment	-0.02
TSH	-0.02
VitB6	-0.02
X25_OH_D3	-0.01
Act_B12_R	-0.01
Alb	-0.01
AlbKr	-0.01
D_Dimeer	-0.01
Haptoglo	-0.01
PTH	-0.01
Trombo	-0.01
Trombo_cit	-0.01
X.Intercept..1	0.00
AF	0.00
Baso	0.00
CRP	0.00
Erytrobl	0.00
FT4	0.00
Gluc	0.00
Gluc_n	0.00
Lipase	0.00
Myelo	0.00
PSA	0.00
Promyelo	0.00
TE	0.00
TN_T_HS	0.00
aTTG_n	0.00
Bili_dir	0.01
Eo	0.01
Krea_Ul	0.01
PSA_Ratio	0.01
Staaf	0.01
Testost_lum	0.01
Uraat	0.01
ALAT	0.02
Cho_HDLR	0.02
Chol	0.02
K	0.02
Neutro	0.02
Foliumz_R	0.03
NTproBNP	0.03
Ureum	0.03
Reti_n	0.04
CKD_epi	0.05
gGT	0.05
HbA1cN	0.06
VitB1	0.06
Leuco	0.08
LD	0.14
Mono	0.19
GFR_MDRD	0.20
MCH_n	0.27
X.Intercept.	0.44
Transf_verz	0.62
RDW	0.66
Ferritin	0.68
Kreat	0.69
Ret_He	0.90
TYBC	1.62
Ery	-2.63
MCV	-1.51
MCHC	-1.20
Transf	-1.10
Fe	-0.40
Na	-0.20
HDL_chol	-0.16
Triglyce	-0.12
LDL_chol	-0.10
Bili_tot	-0.04
Ca	-0.04
Kreat_U	-0.04
Lymfo	-0.04
Meta	-0.04
Mg	-0.04
Cl	-0.03
P	-0.03
ASAT	-0.02
Segment	-0.02
TSH	-0.02
VitB6	-0.02
X25_OH_D3	-0.01
Act_B12_R	-0.01
Alb	-0.01
AlbKr	-0.01
D_Dimeer	-0.01
Haptoglo	-0.01
PTH	-0.01
Trombo	-0.01
Trombo_cit	-0.01
X.Intercept..1	0.00
AF	0.00
Baso	0.00
CRP	0.00
Erytrobl	0.00
FT4	0.00
Gluc	0.00
Gluc_n	0.00
Lipase	0.00
Myelo	0.00
PSA	0.00
Promyelo	0.00
TE	0.00
TN_T_HS	0.00
aTTG_n	0.00
Bili_dir	0.01
Eo	0.01
Krea_Ul	0.01
PSA_Ratio	0.01
Staaf	0.01
Testost_lum	0.01
Uraat	0.01
ALAT	0.02
Cho_HDLR	0.02
Chol	0.02
K	0.02
Neutro	0.02
Foliumz_R	0.03
NTproBNP	0.03
Ureum	0.03
Reti_n	0.04
CKD_epi	0.05
gGT	0.05
HbA1cN	0.06
VitB1	0.06
Leuco	0.08
LD	0.14
Mono	0.19
GFR_MDRD	0.20
MCH_n	0.27
X.Intercept.	0.44
Transf_verz	0.62
RDW	0.66
Ferritin	0.68
Kreat	0.69
Ret_He	0.90
TYBC	1.62

12.17 FEATURE SET 3 FOR PREDICTING ANAEMIA

TABLE 29: CORRELATION FEATURE SELECTION FOR ANAEMIA PREDICTIONS

Biomarker	Values
Ery	-0.4703589954
MCHC	-0.3769286891
MCH_n	-0.2082747220
GFR_MDRD	-0.1708567575
Na	-0.1370884121
CKD_epi	-0.1203250920
MCV	-0.1114449982
Fe	-0.0794375313
Alb	-0.0747385898
LDL_chol	-0.0571434163
Kreat_U	-0.0446995741
Ca	-0.0441571914
Triglyce	-0.0348870163
HDL_chol	-0.0300991720
Transf_verz	-0.0270093190
Mg	-0.0232606156
Chol	-0.0180820866
Cho_HDLR	-0.0149149342
TE	-0.0103263129
VitB6	-0.0054673600
TSH	-0.0040121567
ALAT	-0.0024342293
Krea_Ul	-0.0018458889
Cl	-0.0018365773
Testost_lum	-0.0008200138
Eo	-0.0003881692
Trombo_cit	0.0002635772
X25_OH_D3	0.0011266070
Act_B12_R	0.0013782446
PSA_Ratio	0.0014866931
FT4	0.0019788918
aTTG_n	0.0022484629
AlbKr	0.0044816966
Baso	0.0045719050
TN_T_HS	0.0074528987
D_Dimeer	0.0089268306
ASAT	0.0095183915
Promyelo	0.0096694984
Lymfo	0.0097262162
Uraat	0.0101735509
Bili_dir	0.0103847080
Myelo	0.0112983436
Erytrobl	0.0115566799
Bili_tot	0.0117652143
Lipase	0.0125711247
PSA	0.0129690271
Meta	0.0136204704
Staaf	0.0145624775
PTH	0.0153714482
Gluc_n	0.0162368358
Haptoglo	0.0165596365
P	0.0192086976
Mono	0.0200591181
Segment	0.0251021007
VitB1	0.0271589602
AF	0.0281202351
Foliumz_R	0.0297111046
gGT	0.0397964206
HbA1cN	0.0558314575
K	0.0617537799
Neutro	0.0635039411
Gluc	0.0654892992
Transf	0.0680971998
Leuco	0.0694651624
CRP	0.0713138474
TYBC	0.0721367457
NTproBNP	0.0881202776
LD	0.0884046044
Trombo	0.1044081439
Reti_n	0.1090177225
Ureum	0.1349786538
Ferritin	0.1377460868
Kreat	0.1565549983
Ret_He	0.2835374242
RDW	0.4290208011
Ery	-0.4703589954
MCHC	-0.3769286891
MCH_n	-0.2082747220
GFR_MDRD	-0.1708567575
Na	-0.1370884121
CKD_epi	-0.1203250920
MCV	-0.1114449982
Fe	-0.0794375313
Alb	-0.0747385898
LDL_chol	-0.0571434163
Kreat_U	-0.0446995741
Ca	-0.0441571914
Triglyce	-0.0348870163
HDL_chol	-0.0300991720
Transf_verz	-0.0270093190
Mg	-0.0232606156
Chol	-0.0180820866
Cho_HDLR	-0.0149149342
TE	-0.0103263129
VitB6	-0.0054673600
TSH	-0.0040121567
ALAT	-0.0024342293
Krea_Ul	-0.0018458889
Cl	-0.0018365773
Testost_lum	-0.0008200138
Eo	-0.0003881692
Trombo_cit	0.0002635772
X25_OH_D3	0.0011266070
Act_B12_R	0.0013782446
PSA_Ratio	0.0014866931
FT4	0.0019788918
aTTG_n	0.0022484629
AlbKr	0.0044816966
Baso	0.0045719050
TN_T_HS	0.0074528987
D_Dimeer	0.0089268306
ASAT	0.0095183915
Promyelo	0.0096694984
Lymfo	0.0097262162
Uraat	0.0101735509
Bili_dir	0.0103847080
Myelo	0.0112983436
Erytrobl	0.0115566799
Bili_tot	0.0117652143
Lipase	0.0125711247
PSA	0.0129690271
Meta	0.0136204704
Staaf	0.0145624775
PTH	0.0153714482
Gluc_n	0.0162368358
Haptoglo	0.0165596365
P	0.0192086976
Mono	0.0200591181
Segment	0.0251021007
VitB1	0.0271589602
AF	0.0281202351
Foliumz_R	0.0297111046
gGT	0.0397964206
HbA1cN	0.0558314575
K	0.0617537799
Neutro	0.0635039411
Gluc	0.0654892992
Transf	0.0680971998
Leuco	0.0694651624
CRP	0.0713138474
TYBC	0.0721367457
NTproBNP	0.0881202776
LD	0.0884046044
Trombo	0.1044081439
Reti_n	0.1090177225
Ureum	0.1349786538
Ferritin	0.1377460868
Kreat	0.1565549983
Ret_He	0.2835374242
RDW	0.4290208011

12.18 FEATURE SET METRICS FOR PREDICTING ANAEMIA

TABLE 30: FULL METRICS FOR ANAEMIA PREDICTIONS

CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc	SD Kappa	SD Accuracy	SD Specificity	SD Sensitivity
0.911	0.9233	0.8357	0.6822	0.9661	0.7512	0.8795	0.0031	9e-04	0.0013	**0.003**
0.8895	0.9377	0.6308	0.6662	0.9280	0.6479	0.7842	0.0057	0.0014	0.0026	**0.0104**
0.9319	0.9379	0.8899	0.7386	0.9774	0.8072	0.9139	0.0029	0.0008	0.0009	**0.0030**
0.9362	0.9345	0.9335	0.7375	0.9862	0.824	0.934	0.0029	8e-04	9e-04	**0.0022**
0.9164	0.9523	0.7221	0.7488	0.9456	0.7352	0.8372	0.0049	0.0014	0.0016	**0.0063**
0.9536	0.9525	0.9498	0.7974	0.9897	0.867	0.9511	0.0028	9e-04	0.0012	**0.0016**
0.9292	0.9269	0.9294	0.7146	0.9852	0.8079	0.9281	0.0034	0.0011	0.0014	**0.0056**
0.9088	0.9475	0.6994	0.7241	0.9412	0.7115	0.8234	0.0052	0.0014	0.0014	**0.0023**
0.9459	0.9440	0.9453	0.7689	0.9887	0.8480	0.9447	0.0026	0.0009	0.0012	**0.0022**
0.9018	0.9012	0.8916	0.640	0.9769	0.7451	0.8964	0.0041	0.0015	0.0025	**0.0048**
0.8822	0.9386	0.5810	0.6511	0.9192	0.6140	0.7598	0.0046	0.0010	0.0022	**0.0098**
0.9358	0.9326	0.9411	0.7334	0.9877	0.8244	0.9368	0.0030	0.0009	0.0010	**0.0023**

ML Algo	Feature Set	Acc	Kappa	CI95% lower
KNN	Standard	0.9089	0.6961	0.9067
Naïve Bayes	Standard	0.8872	0.5808	0.8848
Random Forest	Standard	0.9300	0.7649	0.9281
KNN	1	0.9344	0.7843	0.9325
Naïve Bayes	1	0.9144	0.6841	0.9123
Random Forest	1	0.952	0.838	0.9504
KNN	2	0.9273	0.764	0.9253
Naïve Bayes	2	0.9067	0.6559	0.9045
Random Forest	2	0.9442	0.8143	0.9425
KNN	3	0.8996	0.6846	0.8973
Naïve Bayes	3	0.8798	0.5430	0.8773
Random Forest	3	0.9340	0.7845	0.9321

12.19 FEATURE SET 1 FOR PREDICTING IRON DEFICIENCY ANAEMIA

TABLE 31: EMBEDDED FEATURE SELECTION FOR IRON DEFICIENCY ANAEMIA PREDICTIONS

Biomarker	Importance
Transf_verz	1.208929e-02
MCV	8.454839e-03
Ferritin	8.422607e-03
TYBC	5.225512e-03
MCH_n	5.157949e-03
MCHC	3.856283e-03
Ery	3.707565e-03
Transf	3.476938e-03
LD	2.444711e-03
RDW	2.125564e-03
Ret_He	2.013464e-03
Fe	1.908674e-03
Kreat	5.888113e-04
Reti_n	5.261386e-04
GFR_MDRD	4.538731e-04
CKD_epi	2.524402e-04
Leuco	1.667383e-04
LDL_chol	1.413463e-04
gGT	1.335995e-04
Neutro	1.259840e-04
Triglyce	1.244612e-04
Trombo	1.242916e-04
HDL_chol	9.020669e-05
Gluc_n	7.379240e-05
HbA1cN	6.764089e-05
TSH	6.611396e-05
Gluc	6.281039e-05
Na	5.872541e-05
K	5.220351e-05
Kreat_U	3.630917e-05
NTproBNP	3.328976e-05
Lymfo	2.852816e-05
Chol	2.744578e-05
Cho_HDLR	2.668286e-05
CRP	2.657341e-05
PSA	2.452861e-05
ALAT	1.770856e-05
FT4	1.568596e-05
Ureum	1.559884e-05
Bili_tot	1.456754e-05
Mono	1.260836e-05
Act_B12_R	1.072615e-05
AlbKr	9.996271e-06
Alb	7.203080e-06
X25_OH_D3	6.466019e-06
Foliumz_R	6.384476e-06
Ca	5.549036e-06
ASAT	5.196455e-06
VitB6	4.213581e-06
Lipase	4.150594e-06
Baso	3.782278e-06
Leuco_U	3.251344e-06
Uraat	2.844794e-06
Bili_dir	2.701409e-06
AF	2.483031e-06
Segment	2.017452e-06
P	1.341335e-06
D_Dimeer	1.323840e-06
Mg	1.260511e-06
VitB1	1.188362e-06
Ery_U	9.903356e-07
Eo	3.853909e-07
TN_T_HS	3.697129e-07
Staaf	3.421065e-07
Bili_U	1.806547e-08
Krea_Ul	1.804722e-08
Trombo_cit	1.801859e-08
Testost_lum	1.800618e-08
Cl	1.796068e-08
Promyelo	-3.828904e-11
aTTG_n	-8.957834e-09
Myelo	-1.793417e-08
Meta	-2.706449e-08
PSA_Ratio	-9.894927e-08
Erytrobl	-1.348078e-07
PTH	-1.890566e-07
Haptoglo	-7.114515e-07
TE	-1.242813e-06

12.20 FEATURE SET 2 FOR PREDICTING IRON DEFICIENCY ANAEMIA

TABLE 32: LASSO FEATURE SELECTION FOR IRON DEFICIENCY ANAEMIA PREDICTIONS

Biomarker	Values
Ery	-1.98
MCHC	-1.95
MCV	-1.89
Transf	-1.36
Ferritin	-0.76
Transf_verz	-0.40
Bili_dir	-0.19
HDL_chol	-0.19
Triglyce	-0.19
Na	-0.16
Leuco	-0.14
Ureum	-0.14
LD	-0.12
LDL_chol	-0.12
Staaf	-0.11
Fe	-0.10
NTproBNP	-0.07
P	-0.07
Chol	-0.06
ASAT	-0.04
Ca	-0.04
Lipase	-0.04
VitB6	-0.04
gGT	-0.04
CKD_epi	-0.03
Lymfo	-0.03
Segment	-0.03
Alb	-0.02
AlbKr	-0.02
FT4	-0.02
Gluc_n	-0.02
MCH_n	-0.02
Meta	-0.02
PSA_Ratio	-0.02
TE	-0.02
Baso	-0.01
Kreat_U	-0.01
Mono	-0.01
PTH	-0.01
TSH	-0.01
aTTG_n	-0.01
X.Intercept..1	0.00
Cl	0.00
Foliumz_R	0.00
Mg	0.00
TN_T_HS	0.00
Testost_lum	0.00
Trombo_cit	0.00
Uraat	0.00
X25_OH_D3	0.01
Act_B12_R	0.01
D_Dimeer	0.01
Eo	0.01
K	0.01
Myelo	0.01
Promyelo	0.01
Trombo	0.01
CRP	0.02
Gluc	0.02
Krea_Ul	0.02
AF	0.03
Erytrobl	0.03
Haptoglo	0.03
ALAT	0.05
PSA	0.05
Reti_n	0.05
Cho_HDLR	0.06
VitB1	0.06
Bili_tot	0.13
Neutro	0.13
HbA1cN	0.21
GFR_MDRD	0.22
Kreat	0.59
RDW	0.76
Ret_He	0.92
X.Intercept.	1.21
TYBC	3.36
Ery	-1.98
MCHC	-1.95
MCV	-1.89
Transf	-1.36
Ferritin	-0.76
Transf_verz	-0.40
Bili_dir	-0.19
HDL_chol	-0.19
Triglyce	-0.19
Na	-0.16
Leuco	-0.14
Ureum	-0.14
LD	-0.12
LDL_chol	-0.12
Staaf	-0.11
Fe	-0.10
NTproBNP	-0.07
P	-0.07
Chol	-0.06
ASAT	-0.04
Ca	-0.04
Lipase	-0.04
VitB6	-0.04
gGT	-0.04
CKD_epi	-0.03
Lymfo	-0.03
Segment	-0.03
Alb	-0.02
AlbKr	-0.02
FT4	-0.02
Gluc_n	-0.02
MCH_n	-0.02
Meta	-0.02
PSA_Ratio	-0.02
TE	-0.02
Baso	-0.01
Kreat_U	-0.01
Mono	-0.01
PTH	-0.01
TSH	-0.01
aTTG_n	-0.01
X.Intercept..1	0.00
Cl	0.00
Foliumz_R	0.00
Mg	0.00
TN_T_HS	0.00
Testost_lum	0.00
Trombo_cit	0.00
Uraat	0.00
X25_OH_D3	0.01
Act_B12_R	0.01
D_Dimeer	0.01
Eo	0.01
K	0.01
Myelo	0.01
Promyelo	0.01
Trombo	0.01
CRP	0.02
Gluc	0.02
Krea_Ul	0.02
AF	0.03
Erytrobl	0.03
Haptoglo	0.03
ALAT	0.05
PSA	0.05
Reti_n	0.05
Cho_HDLR	0.06
VitB1	0.06
Bili_tot	0.13
Neutro	0.13
HbA1cN	0.21
GFR_MDRD	0.22
Kreat	0.59
RDW	0.76
Ret_He	0.92
X.Intercept.	1.21
TYBC	3.36

12.21 FEATURE SET 3 FOR PREDICTING IRON DEFICIENCY ANAEMIA

TABLE 33: CORRELATION FEATURE SELECTION FOR IRON DEFICIENCY ANAEMIA PREDICTIONS

Biomarker	Values
MCHC	-0.3989298615
MCH_n	-0.3924401053
Transf_verz	-0.3495723704
MCV	-0.3054983945
Fe	-0.2091757363
Ery	-0.2010255509
Ferritin	-0.1195536509
GFR_MDRD	-0.0464649734
CKD_epi	-0.0412430548
LDL_chol	-0.0366832877
Na	-0.0353429760
Kreat_U	-0.0195956155
Triglyce	-0.0177893349
HDL_chol	-0.0143008218
Chol	-0.0138052770
Alb	-0.0125690090
gGT	-0.0106098117
LD	-0.0096369780
Cho_HDLR	-0.0095155182
Ca	-0.0061509065
Mg	-0.0058736367
TE	-0.0055975301
ALAT	-0.0049755593
Lymfo	-0.0042328576
ASAT	-0.0028115504
Lipase	-0.0027318920
FT4	-0.0026787392
Staaf	-0.0025016728
PSA	-0.0016659976
Bili_dir	-0.0015651039
Meta	-0.0013464314
Krea_Ul	-0.0010936928
Testost_lum	-0.0010844107
Myelo	-0.0009343844
PSA_Ratio	-0.0008545171
Promyelo	-0.0006335140
X25_OH_D3	-0.0003285426
D_Dimeer	-0.0003221276
Trombo_cit	-0.0002717499
Erytrobl	-0.0001864621
VitB6	-0.0001777860
TN_T_HS	0.0000671259
Uraat	0.0001689452
TSH	0.0002397512
Segment	0.0002884255
Act_B12_R	0.0003520605
Baso	0.0007614988
Mono	0.0009521310
Foliumz_R	0.0012194120
Eo	0.0015686221
AF	0.0017805546
Cl	0.0020783781
AlbKr	0.0022176111
CRP	0.0023865438
Bili_tot	0.0026335231
aTTG_n	0.0033725549
P	0.0034638745
Leuco	0.0056005129
PTH	0.0077712079
Haptoglo	0.0094641990
Gluc	0.0116617153
Neutro	0.0118483319
VitB1	0.0187492231
Gluc_n	0.0206288140
K	0.0220641167
Ureum	0.0233437560
Kreat	0.0274522419
NTproBNP	0.0373128933
HbA1cN	0.0539651426
Reti_n	0.0634602901
Trombo	0.1335874949
Ret_He	0.2277094191
RDW	0.4212293538
Transf	0.5320932253
TYBC	0.5373415327
MCHC	-0.3989298615
MCH_n	-0.3924401053
Transf_verz	-0.3495723704
MCV	-0.3054983945
Fe	-0.2091757363
Ery	-0.2010255509
Ferritin	-0.1195536509
GFR_MDRD	-0.0464649734
CKD_epi	-0.0412430548
LDL_chol	-0.0366832877
Na	-0.0353429760
Kreat_U	-0.0195956155
Triglyce	-0.0177893349
HDL_chol	-0.0143008218
Chol	-0.0138052770
Alb	-0.0125690090
gGT	-0.0106098117
LD	-0.0096369780
Cho_HDLR	-0.0095155182
Ca	-0.0061509065
Mg	-0.0058736367
TE	-0.0055975301
ALAT	-0.0049755593
Lymfo	-0.0042328576
ASAT	-0.0028115504
Lipase	-0.0027318920
FT4	-0.0026787392
Staaf	-0.0025016728
PSA	-0.0016659976
Bili_dir	-0.0015651039
Meta	-0.0013464314
Krea_Ul	-0.0010936928
Testost_lum	-0.0010844107
Myelo	-0.0009343844
PSA_Ratio	-0.0008545171
Promyelo	-0.0006335140
X25_OH_D3	-0.0003285426
D_Dimeer	-0.0003221276
Trombo_cit	-0.0002717499
Erytrobl	-0.0001864621
VitB6	-0.0001777860
TN_T_HS	0.0000671259
Uraat	0.0001689452
TSH	0.0002397512
Segment	0.0002884255
Act_B12_R	0.0003520605
Baso	0.0007614988
Mono	0.0009521310
Foliumz_R	0.0012194120
Eo	0.0015686221
AF	0.0017805546
Cl	0.0020783781
AlbKr	0.0022176111
CRP	0.0023865438
Bili_tot	0.0026335231
aTTG_n	0.0033725549
P	0.0034638745
Leuco	0.0056005129
PTH	0.0077712079
Haptoglo	0.0094641990
Gluc	0.0116617153
Neutro	0.0118483319
VitB1	0.0187492231
Gluc_n	0.0206288140
K	0.0220641167
Ureum	0.0233437560
Kreat	0.0274522419
NTproBNP	0.0373128933
HbA1cN	0.0539651426
Reti_n	0.0634602901
Trombo	0.1335874949
Ret_He	0.2277094191
RDW	0.4212293538
Transf	0.5320932253
TYBC	0.5373415327

12.22 FEATURE SET METRICS FOR PREDICTING IRON DEFICIENCY ANAEMIA

TABLE 34: FULL METRICS FOR IRON DEFICIENCY ANAEMIA PREDICTIONS

Feature Set	Acc	Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
Standard	0.9616216	0.6779443	0.9605230	0.9626980	0.9608523	0.9779295	0.6265162	0.9625762	0.6966237	**0.9693909**
Standard	0.9559955	0.6252591	0.9548239	0.9571451	0.9589047	0.8942926	0.5066923	0.9948288	0.6468087	**0.9265986**
Standard	0.9676215	0.7177617	0.9661796	0.9690182	0.9669452	0.9818067	0.5860775	0.9991076	0.7338200	**0.9743759**
1	0.9785281	0.7934570	0.9773404	0.9796694	0.9785827	0.9773843	0.6835405	0.9989068	0.8044402	**0.9779835**
1	0.9554373	0.6307746	0.9537618	0.9570689	0.9568297	0.9260562	0.5037634	0.9963584	0.6523541	**0.9414430**
1	0.9831361	0.8326403	0.9820773	0.9841480	0.9828001	0.9902359	0.7314043	0.9995307	0.8413323	**0.9865180**
2	0.9786153	0.7866787	0.9774300	0.9797543	0.9786769	0.9772501	0.6737668	0.9989541	0.7975633	**0.9779635**
2	0.9560336	0.6178193	0.9543681	0.9576551	0.9581064	0.9096815	0.4925978	0.9958038	0.6390284	**0.9338940**
2	0.9814642	0.8115347	0.9803570	0.9825248	0.9811109	0.9893357	0.7017150	0.9995122	0.8210089	**0.9852233**
3	0.9744634	0.7539471	0.9731739	0.9757070	0.9743011	0.9780891	0.6306331	0.9989921	0.7667879	**0.9761951**
3	0.9504197	0.5917861	0.9486581	0.9521377	0.9515274	0.9256507	0.4606545	0.9965181	0.6151158	**0.9385891**
3	0.9786153	0.7868171	0.9774302	0.9797541	0.9781753	0.9885052	0.6686801	0.9994768	0.7976599	**0.9833403**

ML Algo	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest

12.23 FEATURE SET 1 FOR PREDICTING ANAEMIA OF CHRONIC DISEASE

TABLE 35: EMBEDDED FEATURE SELECTION FOR ANAEMIA OF CHRONIC DISEASE PREDICTIONS

Biomarker	Importance
Transf_verz	1.183408e-01
TYBC	1.073906e-01
Transf	1.063581e-01
Fe	2.782266e-02
Ferritin	1.898816e-02
Ret_He	1.804284e-02
Ery	1.638819e-02
LD	1.282478e-02
Reti_n	9.969180e-03
MCHC	3.912879e-03
RDW	3.807653e-03
MCV	3.371110e-03
MCH_n	3.299248e-03
Kreat	1.461100e-03
Neutro	9.104993e-04
Trombo	9.014640e-04
GFR_MDRD	8.494890e-04
Leuco	8.076822e-04
CKD_epi	6.404034e-04
Triglyce	5.180862e-04
CRP	5.002084e-04
LDL_chol	4.922193e-04
Na	4.255894e-04
NTproBNP	4.127559e-04
HDL_chol	3.535458e-04
Ureum	2.379419e-04
gGT	2.132689e-04
Haptoglo	1.754440e-04
Gluc	1.279848e-04
K	1.042647e-04
Gluc_n	9.625030e-05
TSH	8.232609e-05
Kreat_U	7.287022e-05
Mono	6.901158e-05
Lymfo	6.833267e-05
Meta	5.345431e-05
Alb	4.958793e-05
HbA1cN	4.649699e-05
Foliumz_R	3.839490e-05
VitB6	2.849317e-05
Act_B12_R	2.759009e-05
Chol	2.730162e-05
Ca	2.441033e-05
Lipase	2.301471e-05
PSA	2.247728e-05
ASAT	2.007279e-05
FT4	1.773630e-05
Segment	1.269752e-05
Staaf	1.137527e-05
Cho_HDLR	1.105977e-05
VitB1	9.229634e-06
AF	9.011328e-06
Mg	8.005428e-06
Bili_tot	7.776898e-06
Erytrobl	7.773443e-06
TE	6.930074e-06
Promyelo	6.708579e-06
TN_T_HS	5.739794e-06
Myelo	5.735254e-06
Baso	5.358722e-06
X25_OH_D3	4.987649e-06
Eo	4.770214e-06
Bili_dir	2.974734e-06
Uraat	2.737365e-06
PTH	2.271952e-06
PSA_Ratio	1.442686e-06
P	1.082371e-06
D_Dimeer	4.832185e-07
Testost_lum	2.387392e-07
Krea_Ul	0.000000e+00
Trombo_cit	0.000000e+00
Cl	-6.981857e-10
aTTG_n	-3.610133e-07
AlbKr	-6.833284e-06
ALAT	-1.252043e-05

12.24 FEATURE SET 2 FOR PREDICTING ANAEMIA OF CHRONIC DISEASE

TABLE 36: LASSO FEATURE SELECTION FOR ANAEMIA OF CHRONIC DISEASE PREDICTIONS

Biomarker	Values
Transf	-20.82
MCV	-0.88
MCHC	-0.77
Ferritin	-0.72
Ery	-0.54
Bili_dir	-0.47
Fe	-0.40
NTproBNP	-0.39
Triglyce	-0.31
LDL_chol	-0.28
Trombo	-0.26
P	-0.23
MCH_n	-0.22
Gluc_n	-0.19
Na	-0.19
Leuco	-0.15
AF	-0.14
FT4	-0.13
Segment	-0.12
HDL_chol	-0.11
LD	-0.10
Uraat	-0.10
TSH	-0.09
K	-0.08
PTH	-0.08
gGT	-0.08
Ureum	-0.07
aTTG_n	-0.06
PSA_Ratio	-0.05
Chol	-0.04
Myelo	-0.04
Haptoglo	-0.03
Mg	-0.03
VitB6	-0.03
CRP	-0.02
Baso	-0.01
X.Intercept..1	0.00
X25_OH_D3	0.00
ASAT	0.00
Act_B12_R	0.00
Alb	0.00
AlbKr	0.00
Krea_Ul	0.00
Meta	0.00
Promyelo	0.00
Testost_lum	0.00
Trombo_cit	0.00
Cl	0.01
Mono	0.01
Foliumz_R	0.02
Staaf	0.02
VitB1	0.03
Lipase	0.04
D_Dimeer	0.06
PSA	0.06
TE	0.06
Ca	0.07
Gluc	0.07
Lymfo	0.08
Cho_HDLR	0.09
ALAT	0.10
CKD_epi	0.10
Eo	0.10
Bili_tot	0.12
Kreat_U	0.13
TN_T_HS	0.15
Neutro	0.23
Reti_n	0.25
HbA1cN	0.33
GFR_MDRD	0.44
Erytrobl	0.52
RDW	0.58
Transf_verz	0.62
X.Intercept.	0.96
Kreat	1.02
Ret_He	1.53
TYBC	23.41
Transf	-20.82
MCV	-0.88
MCHC	-0.77
Ferritin	-0.72
Ery	-0.54
Bili_dir	-0.47
Fe	-0.40
NTproBNP	-0.39
Triglyce	-0.31
LDL_chol	-0.28
Trombo	-0.26
P	-0.23
MCH_n	-0.22
Gluc_n	-0.19
Na	-0.19
Leuco	-0.15
AF	-0.14
FT4	-0.13
Segment	-0.12
HDL_chol	-0.11
LD	-0.10
Uraat	-0.10
TSH	-0.09
K	-0.08
PTH	-0.08
gGT	-0.08
Ureum	-0.07
aTTG_n	-0.06
PSA_Ratio	-0.05
Chol	-0.04
Myelo	-0.04
Haptoglo	-0.03
Mg	-0.03
VitB6	-0.03
CRP	-0.02
Baso	-0.01
X.Intercept..1	0.00
X25_OH_D3	0.00
ASAT	0.00
Act_B12_R	0.00
Alb	0.00
AlbKr	0.00
Krea_Ul	0.00
Meta	0.00
Promyelo	0.00
Testost_lum	0.00
Trombo_cit	0.00
Cl	0.01
Mono	0.01
Foliumz_R	0.02
Staaf	0.02
VitB1	0.03
Lipase	0.04
D_Dimeer	0.06
PSA	0.06
TE	0.06
Ca	0.07
Gluc	0.07
Lymfo	0.08
Cho_HDLR	0.09
ALAT	0.10
CKD_epi	0.10
Eo	0.10
Bili_tot	0.12
Kreat_U	0.13
TN_T_HS	0.15
Neutro	0.23
Reti_n	0.25
HbA1cN	0.33
GFR_MDRD	0.44
Erytrobl	0.52
RDW	0.58
Transf_verz	0.62
X.Intercept.	0.96
Kreat	1.02
Ret_He	1.53
TYBC	23.41

12.25 FEATURE SET 3 FOR PREDICTING ANAEMIA OF CHRONIC DISEASE

TABLE 37: CORRELATION FEATURE SELECTION FOR ANAEMIA OF CHRONIC DISEASE PREDICTIONS

Biomarker	Values
Transf	-4.987181e-01
TYBC	-4.964558e-01
Fe	-3.185683e-01
Ery	-2.968303e-01
MCHC	-2.139430e-01
Transf_verz	-1.879487e-01
Na	-1.229290e-01
MCH_n	-1.183068e-01
GFR_MDRD	-9.791298e-02
CKD_epi	-7.787880e-02
Alb	-5.753678e-02
LDL_chol	-3.000578e-02
HDL_chol	-2.610609e-02
Triglyce	-2.125054e-02
Kreat_U	-1.949270e-02
Mg	-1.680050e-02
Ca	-1.677502e-02
Chol	-1.441685e-02
MCV	-1.430714e-02
TSH	-8.648634e-03
Cho_HDLR	-8.161780e-03
VitB6	-7.498657e-03
Lymfo	-7.230289e-03
TE	-5.706465e-03
ALAT	-3.895565e-03
PSA_Ratio	-3.297657e-03
aTTG_n	-1.490747e-03
ASAT	-1.370121e-03
Krea_Ul	-1.226995e-03
Lipase	-1.099570e-03
Eo	-2.124795e-04
P	2.773927e-05
PTH	1.346271e-04
Cl	2.710057e-04
Act_B12_R	2.909028e-04
Testost_lum	4.148464e-04
Baso	9.396537e-04
X25_OH_D3	1.091678e-03
Trombo_cit	1.158794e-03
AlbKr	2.963065e-03
Promyelo	3.330166e-03
Bili_dir	3.975912e-03
TN_T_HS	5.183081e-03
Foliumz_R	6.085943e-03
Gluc_n	7.212909e-03
Erytrobl	7.246812e-03
Myelo	7.499908e-03
D_Dimeer	7.569666e-03
Bili_tot	8.736415e-03
Meta	8.909899e-03
FT4	8.967122e-03
Uraat	9.017017e-03
Staaf	1.285956e-02
K	1.396063e-02
AF	1.470816e-02
PSA	2.224811e-02
Segment	2.232084e-02
VitB1	2.534455e-02
HbA1cN	3.153464e-02
gGT	4.495785e-02
Mono	5.034384e-02
Gluc	5.176517e-02
Haptoglo	5.961658e-02
Ureum	7.742900e-02
Reti_n	8.175757e-02
NTproBNP	8.459446e-02
Kreat	9.931091e-02
LD	1.119452e-01
CRP	1.129797e-01
Leuco	1.204732e-01
Trombo	1.389782e-01
Neutro	1.395484e-01
RDW	2.562077e-01
Ret_He	2.982289e-01
Ferritin	3.142638e-01
Transf	-4.987181e-01
TYBC	-4.964558e-01
Fe	-3.185683e-01
Ery	-2.968303e-01
MCHC	-2.139430e-01
Transf_verz	-1.879487e-01
Na	-1.229290e-01
MCH_n	-1.183068e-01
GFR_MDRD	-9.791298e-02
CKD_epi	-7.787880e-02
Alb	-5.753678e-02
LDL_chol	-3.000578e-02
HDL_chol	-2.610609e-02
Triglyce	-2.125054e-02
Kreat_U	-1.949270e-02
Mg	-1.680050e-02
Ca	-1.677502e-02
Chol	-1.441685e-02
MCV	-1.430714e-02
TSH	-8.648634e-03
Cho_HDLR	-8.161780e-03
VitB6	-7.498657e-03
Lymfo	-7.230289e-03
TE	-5.706465e-03
ALAT	-3.895565e-03
PSA_Ratio	-3.297657e-03
aTTG_n	-1.490747e-03
ASAT	-1.370121e-03
Krea_Ul	-1.226995e-03
Lipase	-1.099570e-03
Eo	-2.124795e-04
P	2.773927e-05
PTH	1.346271e-04
Cl	2.710057e-04
Act_B12_R	2.909028e-04
Testost_lum	4.148464e-04
Baso	9.396537e-04
X25_OH_D3	1.091678e-03
Trombo_cit	1.158794e-03
AlbKr	2.963065e-03
Promyelo	3.330166e-03
Bili_dir	3.975912e-03
TN_T_HS	5.183081e-03
Foliumz_R	6.085943e-03
Gluc_n	7.212909e-03
Erytrobl	7.246812e-03
Myelo	7.499908e-03
D_Dimeer	7.569666e-03
Bili_tot	8.736415e-03
Meta	8.909899e-03
FT4	8.967122e-03
Uraat	9.017017e-03
Staaf	1.285956e-02
K	1.396063e-02
AF	1.470816e-02
PSA	2.224811e-02
Segment	2.232084e-02
VitB1	2.534455e-02
HbA1cN	3.153464e-02
gGT	4.495785e-02
Mono	5.034384e-02
Gluc	5.176517e-02
Haptoglo	5.961658e-02
Ureum	7.742900e-02
Reti_n	8.175757e-02
NTproBNP	8.459446e-02
Kreat	9.931091e-02
LD	1.119452e-01
CRP	1.129797e-01
Leuco	1.204732e-01
Trombo	1.389782e-01
Neutro	1.395484e-01
RDW	2.562077e-01
Ret_He	2.982289e-01
Ferritin	3.142638e-01

12.26 FEATURE SET METRICS FOR PREDICTING ANAEMIA OF CHRONIC DISEASE

TABLE 38: FULL METRICS FOR ANAEMIA OF CHRONIC DISEASE PREDICTIONS

Feature Set	Acc	Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
Standard	0.9637876	0.6528608	0.9627134	0.9648391	0.9636213	0.9680099	0.5124500	0.9986903	0.6701067	**0.9658156**
Standard	0.9547710	0.5631606	0.9535783	0.9559415	0.9593925	0.8377934	0.4493869	0.9933668	0.5847647	**0.8985929**
Standard	0.9703259	0.6995742	0.9689348	0.9716712	0.9702076	0.9733031	0.5639318	0.9989130	0.7140533	**0.9717553**
1	0.9771731	0.7504509	0.9763110	0.9780119	0.9778093	0.9610692	0.6311698	0.9984298	0.7618973	**0.9694393**
1	0.9581304	0.5913623	0.9569800	0.9592586	0.9617201	0.8672924	0.4722693	0.9945785	0.6114601	**0.9145063**
1	0.9814941	0.7928649	0.9803822	0.9825590	0.9813501	0.9851184	0.6766573	0.9994005	0.8022201	**0.9832342**
2	0.9583473	0.6179255	0.9571997	0.9594725	0.9581479	0.9633966	0.4762170	0.9984935	0.6373526	**0.9607723**
2	0.9670852	0.6360404	0.9660585	0.9680891	0.9731343	0.8139145	0.5447199	0.9925052	0.6526042	**0.8935244**
2	0.9814094	0.7923103	0.9802949	0.9824768	0.9811074	0.9890600	0.6740740	0.9995596	0.8016950	**0.9850837**
3	0.9674179	0.6764450	0.9663962	0.9684169	0.9675370	0.9644235	0.5398604	0.9985493	0.6922046	**0.9659803**
3	0.9573345	0.5768941	0.9561739	0.9584728	0.9622249	0.8335097	0.4657831	0.9932133	0.5975162	**0.8978673**
3	0.9794930	0.7754690	0.9783254	0.9806138	0.9792929	0.9845399	0.6538380	0.9993731	0.7857766	**0.9819164**

ML Algo	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest

12.27 FEATURE SET 1 FOR PREDICTING VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

TABLE 39: EMBEDDED FEATIURE SELECTION FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA PREDICTIONS

Biomarker	Importance
MCV	7.831370e-02
Ery	6.898089e-02
LD	5.612235e-02
Ret_He	5.143613e-02
Transf_verz	3.729151e-02
TYBC	3.203379e-02
Transf	2.790136e-02
RDW	2.655066e-02
Ferritin	2.218524e-02
Reti_n	2.041842e-02
MCH_n	1.449787e-02
MCHC	1.219130e-02
Fe	1.102708e-02
Foliumz_R	5.410928e-03
Act_B12_R	4.010321e-03
Kreat	2.123451e-03
VitB6	1.375463e-03
LDL_chol	1.245277e-03
Trombo	1.115513e-03
GFR_MDRD	9.245211e-04
CKD_epi	8.043545e-04
Leuco	7.391023e-04
HDL_chol	4.744250e-04
VitB1	4.379043e-04
Ureum	3.758317e-04
Triglyce	3.434084e-04
NTproBNP	3.115782e-04
Neutro	2.479453e-04
ALAT	2.470689e-04
gGT	2.305111e-04
K	1.991354e-04
Chol	1.602292e-04
Lymfo	1.388630e-04
Erytrobl	1.384030e-04
Gluc_n	1.334642e-04
Na	1.002010e-04
Eo	9.672274e-05
CRP	9.506434e-05
Gluc	9.332355e-05
Lipase	9.238128e-05
ASAT	8.839169e-05
Bili_dir	8.579597e-05
Cho_HDLR	8.509464e-05
Mono	8.014824e-05
Alb	7.077365e-05
HbA1cN	5.819920e-05
Kreat_U	5.380692e-05
Baso	5.336669e-05
AlbKr	5.017488e-05
Haptoglo	4.179707e-05
AF	2.471269e-05
P	2.314956e-05
TSH	2.082742e-05
Mg	1.877460e-05
Segment	1.761413e-05
TN_T_HS	1.369931e-05
PSA	1.345443e-05
Ca	6.689242e-06
PSA_Ratio	3.961727e-06
Myelo	1.377410e-06
Staaf	1.308181e-06
Cl	0.000000e+00
Krea_Ul	0.000000e+00
Promyelo	0.000000e+00
Testost_lum	0.000000e+00
Trombo_cit	0.000000e+00
aTTG_n	0.000000e+00
TE	-1.383299e-08
Meta	-1.362398e-06
Uraat	-9.526807e-06
D_Dimeer	-9.582829e-06
Bili_tot	-1.503700e-05
X25_OH_D3	-1.869895e-05
PTH	-2.595355e-05
FT4	-8.529101e-05

12.28 FEATURE SET 2 FOR PREDICTING VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

TABLE 40: LASSO FEATURE SELECTION FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA PREDICTIONS

Biomarker	Values
Ery	-2.43
VitB6	-1.23
MCHC	-1.07
Fe	-0.55
Act_B12_R	-0.45
Lipase	-0.36
HDL_chol	-0.32
LDL_chol	-0.30
Na	-0.19
Reti_n	-0.17
Foliumz_R	-0.14
Triglyce	-0.13
AF	-0.11
AlbKr	-0.11
Leuco	-0.10
Mg	-0.09
gGT	-0.08
Bili_dir	-0.06
TE	-0.05
Ca	-0.04
K	-0.04
TN_T_HS	-0.03
Eo	-0.02
PSA_Ratio	-0.02
Trombo_cit	-0.02
Chol	-0.01
Uraat	-0.01
X.Intercept..1	0.00
X25_OH_D3	0.00
ASAT	0.00
Baso	0.00
Bili_tot	0.00
CKD_epi	0.00
Cho_HDLR	0.00
Cl	0.00
Erytrobl	0.00
FT4	0.00
HbA1cN	0.00
Kreat_U	0.00
Lymfo	0.00
MCH_n	0.00
Mono	0.00
Myelo	0.00
P	0.00
Promyelo	0.00
Segment	0.00
Staaf	0.00
Testost_lum	0.00
Transf	0.00
Trombo	0.00
aTTG_n	0.00
D_Dimeer	0.01
Haptoglo	0.01
LD	0.01
Neutro	0.01
ALAT	0.02
Alb	0.02
Meta	0.02
NTproBNP	0.03
Gluc	0.04
Krea_Ul	0.04
PSA	0.04
Ureum	0.04
PTH	0.06
CRP	0.07
Gluc_n	0.07
TSH	0.23
Kreat	0.26
VitB1	0.27
GFR_MDRD	0.28
MCV	0.37
Ferritin	0.42
X.Intercept.	0.58
Ret_He	0.76
TYBC	0.88
Transf_verz	0.92
RDW	1.45
Ery	-2.43
VitB6	-1.23
MCHC	-1.07
Fe	-0.55
Act_B12_R	-0.45
Lipase	-0.36
HDL_chol	-0.32
LDL_chol	-0.30
Na	-0.19
Reti_n	-0.17
Foliumz_R	-0.14
Triglyce	-0.13
AF	-0.11
AlbKr	-0.11
Leuco	-0.10
Mg	-0.09
gGT	-0.08
Bili_dir	-0.06
TE	-0.05
Ca	-0.04
K	-0.04
TN_T_HS	-0.03
Eo	-0.02
PSA_Ratio	-0.02
Trombo_cit	-0.02
Chol	-0.01
Uraat	-0.01
X.Intercept..1	0.00
X25_OH_D3	0.00
ASAT	0.00
Baso	0.00
Bili_tot	0.00
CKD_epi	0.00
Cho_HDLR	0.00
Cl	0.00
Erytrobl	0.00
FT4	0.00
HbA1cN	0.00
Kreat_U	0.00
Lymfo	0.00
MCH_n	0.00
Mono	0.00
Myelo	0.00
P	0.00
Promyelo	0.00
Segment	0.00
Staaf	0.00
Testost_lum	0.00
Transf	0.00
Trombo	0.00
aTTG_n	0.00
D_Dimeer	0.01
Haptoglo	0.01
LD	0.01
Neutro	0.01
ALAT	0.02
Alb	0.02
Meta	0.02
NTproBNP	0.03
Gluc	0.04
Krea_Ul	0.04
PSA	0.04
Ureum	0.04
PTH	0.06
CRP	0.07
Gluc_n	0.07
TSH	0.23
Kreat	0.26
VitB1	0.27
GFR_MDRD	0.28
MCV	0.37
Ferritin	0.42
X.Intercept.	0.58
Ret_He	0.76
TYBC	0.88
Transf_verz	0.92
RDW	1.45

12.29 FEATURE SET 3 FOR PREDICTING VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

TABLE 41: CORRELATION FEATURE SELECTION FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA PREDICTIONS

Biomarker	Values
Ret_He	1.174369e-01
RDW	1.087947e-01
LD	9.977182e-02
MCV	8.012189e-02
TYBC	3.728306e-02
MCH_n	3.624507e-02
Transf	3.543422e-02
Kreat	2.888310e-02
Reti_n	2.681612e-02
VitB1	2.330253e-02
Ureum	1.835526e-02
NTproBNP	1.582115e-02
PTH	1.493526e-02
Myelo	1.268001e-02
Meta	1.226440e-02
Ferritin	1.125224e-02
Staaf	1.062703e-02
Gluc	9.793020e-03
K	9.424533e-03
Neutro	8.176493e-03
CRP	7.971963e-03
Trombo	7.237485e-03
Baso	6.915880e-03
Mono	5.463790e-03
Leuco	4.415043e-03
Promyelo	4.131696e-03
gGT	4.130719e-03
Segment	4.012240e-03
D_Dimeer	3.607544e-03
HbA1cN	3.581231e-03
TSH	2.691523e-03
Erytrobl	2.163937e-03
Lymfo	2.156265e-03
Gluc_n	1.593501e-03
Cl	1.518508e-03
P	1.269043e-03
AF	1.153785e-03
Krea_Ul	9.612366e-04
PSA	9.144253e-04
AlbKr	2.267832e-04
Eo	8.408822e-05
Testost_lum	-1.095601e-04
Bili_tot	-1.952672e-04
aTTG_n	-2.035712e-04
Trombo_cit	-2.184893e-04
Uraat	-2.665585e-04
FT4	-3.758249e-04
TN_T_HS	-4.233860e-04
Bili_dir	-5.734186e-04
PSA_Ratio	-1.339599e-03
Lipase	-1.422202e-03
ASAT	-2.241846e-03
ALAT	-2.393152e-03
Cho_HDLR	-2.928623e-03
Mg	-3.059630e-03
X25_OH_D3	-3.799061e-03
Transf_verz	-5.867256e-03
Triglyce	-6.154375e-03
Kreat_U	-6.245483e-03
VitB6	-6.411828e-03
Chol	-6.635939e-03
Ca	-6.685219e-03
HDL_chol	-7.997707e-03
Haptoglo	-8.367582e-03
Alb	-1.013087e-02
TE	-1.380802e-02
LDL_chol	-1.392759e-02
Na	-1.506427e-02
Fe	-1.630371e-02
CKD_epi	-2.225703e-02
Foliumz_R	-2.324265e-02
GFR_MDRD	-2.480384e-02
Act_B12_R	-4.343512e-02
MCHC	-5.822311e-02
Ery	-1.327810e-01
Ret_He	1.174369e-01
RDW	1.087947e-01
LD	9.977182e-02
MCV	8.012189e-02
TYBC	3.728306e-02
MCH_n	3.624507e-02
Transf	3.543422e-02
Kreat	2.888310e-02
Reti_n	2.681612e-02
VitB1	2.330253e-02
Ureum	1.835526e-02
NTproBNP	1.582115e-02
PTH	1.493526e-02
Myelo	1.268001e-02
Meta	1.226440e-02
Ferritin	1.125224e-02
Staaf	1.062703e-02
Gluc	9.793020e-03
K	9.424533e-03
Neutro	8.176493e-03
CRP	7.971963e-03
Trombo	7.237485e-03
Baso	6.915880e-03
Mono	5.463790e-03
Leuco	4.415043e-03
Promyelo	4.131696e-03
gGT	4.130719e-03
Segment	4.012240e-03
D_Dimeer	3.607544e-03
HbA1cN	3.581231e-03
TSH	2.691523e-03
Erytrobl	2.163937e-03
Lymfo	2.156265e-03
Gluc_n	1.593501e-03
Cl	1.518508e-03
P	1.269043e-03
AF	1.153785e-03
Krea_Ul	9.612366e-04
PSA	9.144253e-04
AlbKr	2.267832e-04
Eo	8.408822e-05
Testost_lum	-1.095601e-04
Bili_tot	-1.952672e-04
aTTG_n	-2.035712e-04
Trombo_cit	-2.184893e-04
Uraat	-2.665585e-04
FT4	-3.758249e-04
TN_T_HS	-4.233860e-04
Bili_dir	-5.734186e-04
PSA_Ratio	-1.339599e-03
Lipase	-1.422202e-03
ASAT	-2.241846e-03
ALAT	-2.393152e-03
Cho_HDLR	-2.928623e-03
Mg	-3.059630e-03
X25_OH_D3	-3.799061e-03
Transf_verz	-5.867256e-03
Triglyce	-6.154375e-03
Kreat_U	-6.245483e-03
VitB6	-6.411828e-03
Chol	-6.635939e-03
Ca	-6.685219e-03
HDL_chol	-7.997707e-03
Haptoglo	-8.367582e-03
Alb	-1.013087e-02
TE	-1.380802e-02
LDL_chol	-1.392759e-02
Na	-1.506427e-02
Fe	-1.630371e-02
CKD_epi	-2.225703e-02
Foliumz_R	-2.324265e-02
GFR_MDRD	-2.480384e-02
Act_B12_R	-4.343512e-02
MCHC	-5.822311e-02
Ery	-1.327810e-01

12.30 FEATURE SET METRICS FOR PREDICTING VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

TABLE 42: FULL METRICS FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

Feature Set	Acc	Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
Standard	0.8404075	0.02477221	0.8384673	0.8423335	0.8403473	0.8610915	0.01543347	0.9995202	0.03032264	**0.8507194**
Standard	0.9153523	0.03070853	0.9138737	0.9168135	0.9164223	0.5469150	0.01868907	0.9985662	0.03613773	**0.7316686**
Standard	0.8344397	0.02605036	0.8316513	0.8372000	0.8341742	0.9250030	0.01608892	0.9997385	0.03162346	**0.8795886**
1	0.9507769	0.11316372	0.9495163	0.9520148	0.9507476	0.9594623	0.06340537	0.9998514	0.11891203	**0.9551050**
1	0.9478668	0.09123277	0.9465720	0.9491391	0.9483444	0.8101377	0.05167981	0.9993064	0.09713696	**0.8792410**
1	0.9597900	0.14211876	0.9581596	0.9613741	0.9596872	0.9890991	0.07984382	0.9999591	0.14769912	**0.9743932**
2	0.9321324	0.08012023	0.9306698	0.9335733	0.9321583	0.9245266	0.04520387	0.9997197	0.08618268	**0.9283425**
2	0.9285823	0.06526669	0.9270853	0.9300579	0.9290495	0.7936240	0.03740812	0.9992312	0.07143678	**0.8613367**
2	0.9413841	0.09503449	0.9394424	0.9432814	0.9413026	0.9655366	0.05323639	0.9998750	0.10084241	**0.9534196**
3	0.9494532	0.10860201	0.9481770	0.9507069	0.9494764	0.9427521	0.06090287	0.9997910	0.11438917	**0.9461142**
3	0.9503847	0.09930750	0.9491205	0.9516264	0.9507753	0.8380293	0.05612876	0.9994088	0.10514132	**0.8944023**
3	0.9561835	0.12921445	0.9544861	0.9578351	0.9560687	0.9891006	0.07235371	0.9999614	0.13480952	**0.9725846**

ML Algo	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest

12.31 FEATURE SET 1 FOR PREDICTING BONE MARROW DISEASE

TABLE 43: EMBEDDED FEATURE SELECTION FOR BONE MARROW DISEASE PREDICTIONS

Biomarker	Importance
Ery	8.791000e-02
Trombo	6.466367e-02
Ret_He	4.340343e-02
LD	4.044490e-02
Transf	3.918772e-02
Leuco	3.835939e-02
TYBC	3.709240e-02
Transf_verz	3.013468e-02
MCV	3.012680e-02
RDW	1.752675e-02
Reti_n	8.514640e-03
Ferritin	8.063103e-03
Neutro	5.611277e-03
Fe	5.288826e-03
MCH_n	5.195737e-03
Mono	3.788272e-03
MCHC	3.206718e-03
Kreat	2.432325e-03
Lymfo	2.359753e-03
GFR_MDRD	2.282553e-03
Segment	1.337854e-03
Myelo	8.314583e-04
LDL_chol	8.160611e-04
Erytrobl	8.033101e-04
CKD_epi	7.818035e-04
Staaf	6.897509e-04
Gluc	5.095883e-04
Foliumz_R	4.562133e-04
Ureum	4.494856e-04
TSH	3.584593e-04
X25_OH_D3	3.358326e-04
HDL_chol	2.940443e-04
Gluc_n	2.678265e-04
Na	2.676024e-04
Meta	2.433743e-04
NTproBNP	1.976573e-04
AlbKr	1.893825e-04
ALAT	1.781256e-04
Alb	1.634695e-04
Chol	1.343356e-04
gGT	1.185776e-04
Act_B12_R	1.102191e-04
Triglyce	9.858176e-05
K	8.266207e-05
Cho_HDLR	8.053640e-05
CRP	7.888712e-05
AF	6.925576e-05
Eo	6.469779e-05
Promyelo	4.704403e-05
Mg	4.665798e-05
Lipase	4.524905e-05
VitB6	2.795230e-05
Haptoglo	2.715936e-05
Ca	2.451096e-05
TE	1.965068e-05
PSA	4.583672e-06
PSA_Ratio	2.511043e-06
PTH	2.457002e-06
Cl	0.000000e+00
Krea_Ul	0.000000e+00
Testost_lum	0.000000e+00
Trombo_cit	0.000000e+00
aTTG_n	0.000000e+00
Bili_tot	-3.077137e-07
TN_T_HS	-4.807735e-06
D_Dimeer	-4.828637e-06
Bili_dir	-4.833461e-06
VitB1	-7.448214e-06
P	-9.965575e-06
Kreat_U	-1.612471e-05
Uraat	-1.717278e-05
Baso	-2.427668e-05
HbA1cN	-3.961620e-05
ASAT	-3.978788e-05
FT4	-5.897952e-05

12.32 FEATURE SET 2 FOR PREDICTING BONE MARROW DISEASE

TABLE 44: LASSO FEATURE SELECTION FOR BONE MARROW DISEASE PREDICTIONS

Biomarker	Values
Ery	-2.82
Leuco	-1.59
Neutro	-1.38
Trombo	-1.06
MCHC	-0.80
Lipase	-0.43
Fe	-0.39
Na	-0.38
Ureum	-0.19
ALAT	-0.17
Bili_dir	-0.15
Triglyce	-0.14
Chol	-0.11
HDL_chol	-0.09
LDL_chol	-0.09
PSA_Ratio	-0.06
X25_OH_D3	-0.03
Ca	-0.03
X.Intercept..1	0.00
AF	0.00
ASAT	0.00
Bili_tot	0.00
Cho_HDLR	0.00
Cl	0.00
D_Dimeer	0.00
Eo	0.00
Erytrobl	0.00
Haptoglo	0.00
K	0.00
Krea_Ul	0.00
Kreat_U	0.00
MCV	0.00
Meta	0.00
Mg	0.00
PSA	0.00
Promyelo	0.00
Reti_n	0.00
TE	0.00
TN_T_HS	0.00
TSH	0.00
Testost_lum	0.00
Transf	0.00
Trombo_cit	0.00
Uraat	0.00
VitB6	0.00
aTTG_n	0.00
gGT	0.00
Myelo	0.01
P	0.02
Baso	0.03
NTproBNP	0.04
Alb	0.05
FT4	0.06
GFR_MDRD	0.06
Gluc_n	0.06
PTH	0.06
VitB1	0.09
CRP	0.10
Gluc	0.10
Foliumz_R	0.13
TYBC	0.13
Act_B12_R	0.14
AlbKr	0.19
Segment	0.19
HbA1cN	0.22
MCH_n	0.27
Mono	0.46
CKD_epi	0.51
Kreat	0.77
Lymfo	1.11
Transf_verz	1.27
Ferritin	1.37
Staaf	1.42
Ret_He	1.43
RDW	1.67
X.Intercept.	1.83
LD	2.61
Ery	-2.82
Leuco	-1.59
Neutro	-1.38
Trombo	-1.06
MCHC	-0.80
Lipase	-0.43
Fe	-0.39
Na	-0.38
Ureum	-0.19
ALAT	-0.17
Bili_dir	-0.15
Triglyce	-0.14
Chol	-0.11
HDL_chol	-0.09
LDL_chol	-0.09
PSA_Ratio	-0.06
X25_OH_D3	-0.03
Ca	-0.03
X.Intercept..1	0.00
AF	0.00
ASAT	0.00
Bili_tot	0.00
Cho_HDLR	0.00
Cl	0.00
D_Dimeer	0.00
Eo	0.00
Erytrobl	0.00
Haptoglo	0.00
K	0.00
Krea_Ul	0.00
Kreat_U	0.00
MCV	0.00
Meta	0.00
Mg	0.00
PSA	0.00
Promyelo	0.00
Reti_n	0.00
TE	0.00
TN_T_HS	0.00
TSH	0.00
Testost_lum	0.00
Transf	0.00
Trombo_cit	0.00
Uraat	0.00
VitB6	0.00
aTTG_n	0.00
gGT	0.00
Myelo	0.01
P	0.02
Baso	0.03
NTproBNP	0.04
Alb	0.05
FT4	0.06
GFR_MDRD	0.06
Gluc_n	0.06
PTH	0.06
VitB1	0.09
CRP	0.10
Gluc	0.10
Foliumz_R	0.13
TYBC	0.13
Act_B12_R	0.14
AlbKr	0.19
Segment	0.19
HbA1cN	0.22
MCH_n	0.27
Mono	0.46
CKD_epi	0.51
Kreat	0.77
Lymfo	1.11
Transf_verz	1.27
Ferritin	1.37
Staaf	1.42
Ret_He	1.43
RDW	1.67
X.Intercept.	1.83
LD	2.61

12.33 FEATURE SET 3 FOR PREDICTING BONE MARROW DISEASE

TABLE 45: CORRELATION FEATURE SELECTION FOR BONE MARROW DISEASE PREDICTIONS

Biomarker	Values
Ery	-1.181739e-01
Trombo	-7.116006e-02
TYBC	-5.467183e-02
Transf	-5.289823e-02
GFR_MDRD	-3.700315e-02
MCHC	-2.939339e-02
CKD_epi	-1.502463e-02
Na	-1.425416e-02
Neutro	-1.362455e-02
Haptoglo	-1.136092e-02
LDL_chol	-9.705305e-03
Mg	-6.802792e-03
Trombo_cit	-6.617769e-03
Kreat_U	-5.613673e-03
Alb	-5.121710e-03
HDL_chol	-4.968998e-03
Triglyce	-4.805886e-03
Ca	-4.287997e-03
Chol	-2.903337e-03
Cho_HDLR	-2.501622e-03
TSH	-1.623957e-03
ALAT	-1.580776e-03
VitB6	-1.175955e-03
Lipase	-9.432305e-04
Bili_dir	-7.161217e-04
PSA_Ratio	-5.131670e-04
aTTG_n	-3.361791e-04
TN_T_HS	-2.717997e-04
Krea_Ul	-1.915243e-04
Bili_tot	-1.209972e-04
AF	-8.423431e-05
Testost_lum	-8.143448e-05
Reti_n	1.644090e-05
ASAT	2.213128e-04
PTH	4.048435e-04
AlbKr	6.901800e-04
Cl	9.091553e-04
FT4	1.492238e-03
P	2.129158e-03
TE	2.267626e-03
D_Dimeer	2.579256e-03
X25_OH_D3	2.607523e-03
Uraat	3.089639e-03
Baso	3.223747e-03
Gluc_n	3.626507e-03
Act_B12_R	4.412442e-03
Leuco	4.569915e-03
CRP	4.695577e-03
HbA1cN	5.023767e-03
Eo	6.054129e-03
VitB1	6.389202e-03
Gluc	6.618237e-03
gGT	7.481870e-03
K	9.558266e-03
Segment	1.524126e-02
NTproBNP	1.646607e-02
PSA	1.702634e-02
Foliumz_R	1.803217e-02
Ureum	2.328727e-02
Erytrobl	2.660002e-02
Lymfo	2.911846e-02
MCH_n	3.220710e-02
Fe	4.010347e-02
Mono	4.112271e-02
Kreat	4.396376e-02
Myelo	4.691981e-02
Promyelo	4.846945e-02
MCV	5.480904e-02
Meta	6.092260e-02
Staaf	6.531121e-02
Ferritin	7.415347e-02
LD	8.612863e-02
Transf_verz	9.849676e-02
RDW	9.986480e-02
Ret_He	1.313658e-01
Ery	-1.181739e-01
Trombo	-7.116006e-02
TYBC	-5.467183e-02
Transf	-5.289823e-02
GFR_MDRD	-3.700315e-02
MCHC	-2.939339e-02
CKD_epi	-1.502463e-02
Na	-1.425416e-02
Neutro	-1.362455e-02
Haptoglo	-1.136092e-02
LDL_chol	-9.705305e-03
Mg	-6.802792e-03
Trombo_cit	-6.617769e-03
Kreat_U	-5.613673e-03
Alb	-5.121710e-03
HDL_chol	-4.968998e-03
Triglyce	-4.805886e-03
Ca	-4.287997e-03
Chol	-2.903337e-03
Cho_HDLR	-2.501622e-03
TSH	-1.623957e-03
ALAT	-1.580776e-03
VitB6	-1.175955e-03
Lipase	-9.432305e-04
Bili_dir	-7.161217e-04
PSA_Ratio	-5.131670e-04
aTTG_n	-3.361791e-04
TN_T_HS	-2.717997e-04
Krea_Ul	-1.915243e-04
Bili_tot	-1.209972e-04
AF	-8.423431e-05
Testost_lum	-8.143448e-05
Reti_n	1.644090e-05
ASAT	2.213128e-04
PTH	4.048435e-04
AlbKr	6.901800e-04
Cl	9.091553e-04
FT4	1.492238e-03
P	2.129158e-03
TE	2.267626e-03
D_Dimeer	2.579256e-03
X25_OH_D3	2.607523e-03
Uraat	3.089639e-03
Baso	3.223747e-03
Gluc_n	3.626507e-03
Act_B12_R	4.412442e-03
Leuco	4.569915e-03
CRP	4.695577e-03
HbA1cN	5.023767e-03
Eo	6.054129e-03
VitB1	6.389202e-03
Gluc	6.618237e-03
gGT	7.481870e-03
K	9.558266e-03
Segment	1.524126e-02
NTproBNP	1.646607e-02
PSA	1.702634e-02
Foliumz_R	1.803217e-02
Ureum	2.328727e-02
Erytrobl	2.660002e-02
Lymfo	2.911846e-02
MCH_n	3.220710e-02
Fe	4.010347e-02
Mono	4.112271e-02
Kreat	4.396376e-02
Myelo	4.691981e-02
Promyelo	4.846945e-02
MCV	5.480904e-02
Meta	6.092260e-02
Staaf	6.531121e-02
Ferritin	7.415347e-02
LD	8.612863e-02
Transf_verz	9.849676e-02
RDW	9.986480e-02
Ret_He	1.313658e-01

12.34 FEATURE SET METRICS FOR PREDICTING BONE MARROW DISEASE

TABLE 46: FULL METRICS FOR BONE MARROW DISEASE PREDICTIONS

Feature Set	Acc	Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
Standard	0.9340674	0.04631032	0.9326274	0.9354856	0.9341609	0.8861743	0.02566871	0.9997658	0.04984990	**0.9101676**
Standard	0.9660112	0.05132084	0.9649540	0.9670449	0.9668835	0.5093911	0.02896267	0.9990317	0.05474875	**0.7381373**
Standard	0.9197691	0.03779884	0.9175292	0.9219668	0.9198061	0.9003077	0.02118843	0.9997932	0.04138630	**0.9100569**
1	0.9662058	0.09529254	0.9651493	0.9672388	0.9662068	0.9655824	0.05194617	0.9999320	0.09857199	**0.9658946**
1	0.9557542	0.06673375	0.9545552	0.9569303	0.9559223	0.8685443	0.03656975	0.9997359	0.07015423	**0.9122333**
1	0.9703834	0.11216575	0.9689717	0.9717476	0.9703446	0.9905373	0.06133338	0.9999810	0.11539393	**0.9804409**
2	0.9613568	0.08278599	0.9602319	0.9624586	0.9613844	0.9468656	0.04513438	0.9998944	0.08613063	**0.9541250**
2	0.5780478	0.03652064	0.5772837	0.5788082	0.5776137	0.8019584	0.02104792		0.04009509	**0.6897861**
2	0.9635112	0.09187022	0.9619533	0.9650224	0.9634554	0.9923445	0.05002542	0.9999844	0.09520379	**0.9778999**
3	0.9587481	0.07887816	0.9575872	0.9598859	0.9587355	0.9654071	0.04296174	0.9999308	0.08224664	**0.9620713**
3	0.6403681	0.04326580	0.6395201	0.6412093	0.6399576	0.8508972	0.02446603	0.9996307	0.04681396	**0.7454274**
3	0.9628099	0.08937101	0.9612386	0.9643347	0.9627654	0.9861369	0.04865808	0.9999725	0.09268264	**0.9744512**

ML Algo	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest

12.35 FEATURE SET 1 FOR PREDICTING HAEMOLYSIS

TABLE 47: EMBEDDED FEATURE SELECTION FOR HAEMOLYSIS PREDICTIONS

Biomarker	Importance
LD	9.713684e-02
Reti_n	7.143360e-02
Ret_He	5.749570e-02
Ery	5.129959e-02
TYBC	3.858714e-02
Transf_verz	3.682025e-02
Transf	3.165701e-02
Haptoglo	2.434785e-02
RDW	2.348984e-02
MCV	1.571618e-02
Fe	9.054042e-03
Ferritin	8.759363e-03
MCHC	7.963187e-03
MCH_n	2.633694e-03
Trombo	2.017596e-03
NTproBNP	1.805605e-03
Kreat	1.282335e-03
LDL_chol	8.088110e-04
GFR_MDRD	7.402120e-04
CKD_epi	7.118616e-04
Leuco	6.475684e-04
Ureum	5.598032e-04
gGT	5.412321e-04
Foliumz_R	4.498584e-04
Gluc	4.385645e-04
CRP	3.620123e-04
Na	3.242559e-04
Neutro	3.011582e-04
Lipase	2.578645e-04
TSH	2.411756e-04
HDL_chol	2.411024e-04
HbA1cN	2.264694e-04
Triglyce	2.222783e-04
ASAT	2.162235e-04
Lymfo	1.920901e-04
Cho_HDLR	1.606494e-04
K	1.572612e-04
Alb	1.569167e-04
X25_OH_D3	1.333014e-04
Ca	1.314738e-04
PSA	1.046625e-04
Gluc_n	1.007246e-04
ALAT	8.685838e-05
Bili_tot	8.000725e-05
Eo	6.817220e-05
Act_B12_R	5.274302e-05
Erytrobl	5.212896e-05
Chol	4.951462e-05
Segment	4.008436e-05
AlbKr	2.495013e-05
Mono	2.446192e-05
Myelo	2.209080e-05
VitB1	1.925613e-05
AF	1.805167e-05
FT4	1.657027e-05
P	1.606749e-05
Kreat_U	1.381688e-05
Bili_dir	9.833898e-06
TE	8.023669e-06
TN_T_HS	6.058164e-06
Uraat	4.028742e-06
Meta	4.006316e-06
VitB6	3.910742e-06
PTH	7.757168e-08
Cl	0.000000e+00
Krea_Ul	0.000000e+00
PSA_Ratio	0.000000e+00
Promyelo	0.000000e+00
Testost_lum	0.000000e+00
Trombo_cit	0.000000e+00
aTTG_n	0.000000e+00
Staaf	-2.086777e-08
Mg	-2.193304e-06
D_Dimeer	-2.213899e-05
Baso	-5.275026e-05

12.36 FEATURE SET 2 FOR PREDICTING HAEMOLYSIS

TABLE 48: LASSO FEATURE SELECTION FOR HAEMOLYSIS PREDICTIONS

Biomarker	Values
Ery	-2.26
MCHC	-0.83
Haptoglo	-0.66
Foliumz_R	-0.36
Na	-0.25
HDL_chol	-0.19
TSH	-0.15
VitB6	-0.13
TE	-0.11
CRP	-0.10
Chol	-0.09
Leuco	-0.08
LDL_chol	-0.04
ASAT	-0.02
X.Intercept..1	0.00
X25_OH_D3	0.00
AF	0.00
ALAT	0.00
Act_B12_R	0.00
AlbKr	0.00
Baso	0.00
Bili_dir	0.00
Bili_tot	0.00
Cl	0.00
Erytrobl	0.00
FT4	0.00
Fe	0.00
GFR_MDRD	0.00
Gluc	0.00
Gluc_n	0.00
Krea_Ul	0.00
Lipase	0.00
Lymfo	0.00
MCH_n	0.00
MCV	0.00
Meta	0.00
Mono	0.00
Myelo	0.00
Neutro	0.00
P	0.00
PSA	0.00
PSA_Ratio	0.00
PTH	0.00
Promyelo	0.00
Segment	0.00
Staaf	0.00
TN_T_HS	0.00
TYBC	0.00
Testost_lum	0.00
Transf	0.00
Trombo_cit	0.00
Uraat	0.00
Ureum	0.00
VitB1	0.00
aTTG_n	0.00
Cho_HDLR	0.01
D_Dimeer	0.01
Mg	0.01
Alb	0.03
Eo	0.03
NTproBNP	0.03
Kreat_U	0.04
HbA1cN	0.05
CKD_epi	0.07
Trombo	0.07
Ca	0.08
K	0.08
Kreat	0.10
Triglyce	0.10
gGT	0.16
Transf_verz	0.40
Ferritin	0.45
RDW	1.16
Ret_He	1.44
X.Intercept.	3.10
Reti_n	3.64
LD	5.31
Ery	-2.26
MCHC	-0.83
Haptoglo	-0.66
Foliumz_R	-0.36
Na	-0.25
HDL_chol	-0.19
TSH	-0.15
VitB6	-0.13
TE	-0.11
CRP	-0.10
Chol	-0.09
Leuco	-0.08
LDL_chol	-0.04
ASAT	-0.02
X.Intercept..1	0.00
X25_OH_D3	0.00
AF	0.00
ALAT	0.00
Act_B12_R	0.00
AlbKr	0.00
Baso	0.00
Bili_dir	0.00
Bili_tot	0.00
Cl	0.00
Erytrobl	0.00
FT4	0.00
Fe	0.00
GFR_MDRD	0.00
Gluc	0.00
Gluc_n	0.00
Krea_Ul	0.00
Lipase	0.00
Lymfo	0.00
MCH_n	0.00
MCV	0.00
Meta	0.00
Mono	0.00
Myelo	0.00
Neutro	0.00
P	0.00
PSA	0.00
PSA_Ratio	0.00
PTH	0.00
Promyelo	0.00
Segment	0.00
Staaf	0.00
TN_T_HS	0.00
TYBC	0.00
Testost_lum	0.00
Transf	0.00
Trombo_cit	0.00
Uraat	0.00
Ureum	0.00
VitB1	0.00
aTTG_n	0.00
Cho_HDLR	0.01
D_Dimeer	0.01
Mg	0.01
Alb	0.03
Eo	0.03
NTproBNP	0.03
Kreat_U	0.04
HbA1cN	0.05
CKD_epi	0.07
Trombo	0.07
Ca	0.08
K	0.08
Kreat	0.10
Triglyce	0.10
gGT	0.16
Transf_verz	0.40
Ferritin	0.45
RDW	1.16
Ret_He	1.44
X.Intercept.	3.10
Reti_n	3.64
LD	5.31

12.37 FEATURE SET 3 FOR PREDICTING HAEMOLYSIS

TABLE 49: CORRELATION FEATURE SELECTION FOR HAEMOLYSIS PREDICTIONS

Biomarker	Values
Reti_n	4.339822e-01
LD	2.254020e-01
RDW	1.300577e-01
Ret_He	1.226502e-01
Ferritin	7.395144e-02
MCV	5.423842e-02
Transf_verz	4.863017e-02
NTproBNP	4.085598e-02
Staaf	2.990549e-02
Myelo	2.943310e-02
Meta	2.927199e-02
Promyelo	2.903096e-02
Kreat	2.835609e-02
Erytrobl	2.624822e-02
Leuco	2.524210e-02
Ureum	2.379633e-02
Mono	2.088198e-02
MCH_n	2.057716e-02
Fe	2.003106e-02
gGT	1.767760e-02
Trombo	1.552770e-02
Neutro	1.304731e-02
Segment	1.177423e-02
Foliumz_R	1.155823e-02
Lymfo	1.128420e-02
CRP	1.108829e-02
Bili_tot	1.088171e-02
Gluc	1.004096e-02
VitB1	8.141198e-03
Bili_dir	6.628410e-03
K	5.775506e-03
AF	5.578382e-03
Gluc_n	5.168387e-03
PTH	5.134824e-03
D_Dimeer	3.079949e-03
HbA1cN	2.654056e-03
Uraat	2.320945e-03
Act_B12_R	1.881405e-03
P	1.305478e-03
Triglyce	1.183426e-03
PSA_Ratio	1.173295e-03
ASAT	8.673274e-04
AlbKr	7.204491e-04
VitB6	6.253265e-04
Baso	6.029159e-04
Testost_lum	3.646752e-04
Cho_HDLR	2.600531e-04
Cl	1.423991e-04
TN_T_HS	9.482285e-05
Lipase	8.719324e-05
ALAT	5.399026e-05
Krea_Ul	-2.118014e-04
Eo	-2.297754e-04
aTTG_n	-3.686666e-04
FT4	-4.716287e-04
TSH	-5.932979e-04
PSA	-6.173489e-04
Mg	-2.045805e-03
X25_OH_D3	-2.432803e-03
Trombo_cit	-2.522999e-03
Kreat_U	-4.327401e-03
Chol	-4.897525e-03
Ca	-5.019279e-03
TE	-7.309661e-03
HDL_chol	-8.736416e-03
Alb	-9.223196e-03
LDL_chol	-9.888072e-03
CKD_epi	-1.461543e-02
TYBC	-1.801822e-02
Transf	-2.047285e-02
Na	-2.069373e-02
GFR_MDRD	-2.743731e-02
MCHC	-5.569976e-02
Ery	-1.180216e-01
Haptoglo	-3.015127e-01
Reti_n	4.339822e-01
LD	2.254020e-01
RDW	1.300577e-01
Ret_He	1.226502e-01
Ferritin	7.395144e-02
MCV	5.423842e-02
Transf_verz	4.863017e-02
NTproBNP	4.085598e-02
Staaf	2.990549e-02
Myelo	2.943310e-02
Meta	2.927199e-02
Promyelo	2.903096e-02
Kreat	2.835609e-02
Erytrobl	2.624822e-02
Leuco	2.524210e-02
Ureum	2.379633e-02
Mono	2.088198e-02
MCH_n	2.057716e-02
Fe	2.003106e-02
gGT	1.767760e-02
Trombo	1.552770e-02
Neutro	1.304731e-02
Segment	1.177423e-02
Foliumz_R	1.155823e-02
Lymfo	1.128420e-02
CRP	1.108829e-02
Bili_tot	1.088171e-02
Gluc	1.004096e-02
VitB1	8.141198e-03
Bili_dir	6.628410e-03
K	5.775506e-03
AF	5.578382e-03
Gluc_n	5.168387e-03
PTH	5.134824e-03
D_Dimeer	3.079949e-03
HbA1cN	2.654056e-03
Uraat	2.320945e-03
Act_B12_R	1.881405e-03
P	1.305478e-03
Triglyce	1.183426e-03
PSA_Ratio	1.173295e-03
ASAT	8.673274e-04
AlbKr	7.204491e-04
VitB6	6.253265e-04
Baso	6.029159e-04
Testost_lum	3.646752e-04
Cho_HDLR	2.600531e-04
Cl	1.423991e-04
TN_T_HS	9.482285e-05
Lipase	8.719324e-05
ALAT	5.399026e-05
Krea_Ul	-2.118014e-04
Eo	-2.297754e-04
aTTG_n	-3.686666e-04
FT4	-4.716287e-04
TSH	-5.932979e-04
PSA	-6.173489e-04
Mg	-2.045805e-03
X25_OH_D3	-2.432803e-03
Trombo_cit	-2.522999e-03
Kreat_U	-4.327401e-03
Chol	-4.897525e-03
Ca	-5.019279e-03
TE	-7.309661e-03
HDL_chol	-8.736416e-03
Alb	-9.223196e-03
LDL_chol	-9.888072e-03
CKD_epi	-1.461543e-02
TYBC	-1.801822e-02
Transf	-2.047285e-02
Na	-2.069373e-02
GFR_MDRD	-2.743731e-02
MCHC	-5.569976e-02
Ery	-1.180216e-01
Haptoglo	-3.015127e-01

12.38 FEATURE SET METRICS FOR PREDICTING HAEMOLYSIS

TABLE 50: FULL METRICS FOR HAEMOLYSIS PREDICTIONS

Feature Set	Acc	Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
Standard	0.9343701	0.05951153	0.9329292	0.9357892	0.9343220	0.9549645	0.03298443	0.9998870	0.06375822	**0.9446433**
Standard	0.9736466	0.12745471	0.9727128	0.9745567	0.9739835	0.8293071	0.07144689	0.9995914	0.13120682	**0.9016453**
Standard	0.9425197	0.06674709	0.9405952	0.9443996	0.9425081	0.9478290	0.03683024	0.9998728	0.07085966	**0.9451686**
1	0.9728069	0.13941841	0.9718548	0.9737353	0.9728201	0.9673375	0.07735543	0.9999211	0.14313690	**0.9700788**
1	0.4487179	0.04388282	0.4480333	0.4494053	0.4474759	0.9789395	0.02537151	0.9998905	0.04824148	**0.7132077**
1	0.9681669	0.12402510	0.9667054	0.9695812	0.9681045	0.9945483	0.06835156	0.9999869	0.12786879	**0.9813264**
2	0.9714245	0.13068552	0.9704499	0.9723754	0.9714903	0.9433904	0.07244423	0.9998635	0.13445553	**0.9574404**
2	0.6383915	0.05878220	0.6375078	0.6392685	0.6376303	0.9624213	0.03311039	0.9998855	0.06303775	**0.8000258**
2	0.9612167	0.10082900	0.9596140	0.9627729	0.9611757	0.9789817	0.05538365	0.9999496	0.10474106	**0.9700787**
3	0.9746055	0.14666384	0.9736841	0.9755030	0.9746505	0.9557392	0.08166367	0.9998929	0.15033397	**0.9651948**
3	0.9578377	0.09149587	0.9566669	0.9589855	0.9578795	0.9406240	0.05037561	0.9998539	0.09552119	**0.9492517**
3	0.9686561	0.12749030	0.9672069	0.9700581	0.9686039	0.9908727	0.07044680	0.9999774	0.13133582	**0.9797383**

ML Algo	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest	KNN	Naïve Bayes	Random Forest

12.39 PREDICTING SEVERITY OF ANAEMIA

TABLE 51: EMBEDDED FEATURE SELECTION FOR SEVERITY PREDICTIONS

Biomarker	Importance
Ery	2.123741e-01
MCV	6.429054e-02
MCHC	6.358413e-02
RDW	5.247058e-02
Transf_verz	4.850736e-02
Ret_He	4.650134e-02
Ferritin	4.065575e-02
MCH_n	3.680451e-02
TYBC	3.585071e-02
Transf	3.463419e-02
Fe	2.778423e-02
Reti_n	2.725675e-02
LD	2.327521e-02
Kreat	1.065983e-02
Trombo	7.099609e-03
GFR_MDRD	6.243524e-03
Leuco	3.164522e-03
Ureum	2.424377e-03
CKD_epi	2.249270e-03
Na	1.833244e-03
LDL_chol	1.443334e-03
Neutro	1.430503e-03
Erytrobl	1.243120e-03
Ca	1.102318e-03
gGT	9.902220e-04
Gluc	9.708548e-04
TSH	9.493726e-04
Triglyce	9.309784e-04
K	8.882345e-04
HDL_chol	7.230966e-04
Mono	6.364831e-04
Lymfo	6.166383e-04
HbA1cN	6.148590e-04
CRP	5.736769e-04
Gluc_n	5.367878e-04
NTproBNP	4.740275e-04
Alb	4.381716e-04
Chol	3.932641e-04
AF	2.413946e-04
ALAT	2.229200e-04
Segment	1.910780e-04
Bili_tot	1.836109e-04
Eo	1.727387e-04
Lipase	1.700831e-04
ASAT	1.672803e-04
FT4	1.585748e-04
Cho_HDLR	1.316448e-04
Kreat_U	1.259755e-04
Foliumz_R	6.452085e-05
Staaf	4.662769e-05
P	4.474132e-05
Bili_dir	4.308649e-05
AlbKr	4.245831e-05
X25_OH_D3	4.159512e-05
Myelo	3.816034e-05
Meta	3.005101e-05
VitB1	2.649690e-05
VitB6	2.611223e-05
Uraat	2.289123e-05
Haptoglo	1.965102e-05
PSA	1.274477e-05
Baso	9.821005e-06
Mg	9.483061e-06
D_Dimeer	2.809221e-06
Trombo_cit	1.050776e-06
Krea_Ul	0.000000e+00
Testost_lum	0.000000e+00
Cl	-3.518843e-07
TN_T_HS	-6.926672e-07
Act_B12_R	-7.270355e-07
PSA_Ratio	-1.078494e-06
PTH	-1.779987e-06
aTTG_n	-5.994606e-06
TE	-7.378757e-06
Promyelo	-8.834355e-06

12.40 FEATURE SET METRICS FOR PREDICTING MILD ANAEMIA

TABLE 52: FULL METRICS FOR PREDICTING MILD ANAEMIA

Feature Set	Acc	Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
Standard	0.7983065	0.38789681	0.7961824	0.8004181	0.8948495	0.4281470	0.28829157	0.9402647	0.34450020	**0.6614983**
Standard	0.8274753	0.35122914	0.8254743	0.8294626	0.9164482	0.3931192	0.31935723	0.9382393	0.35187002	**0.6547837**
Standard	0.8255255	0.47039589	0.8226768	0.8283469	0.9091342	0.5156144	0.36221812	0.9494503	0.42529597	**0.7123743**
1	0.8562134	0.56914544	0.8543536	0.8580583	0.8857977	0.7538202	0.39629448	0.9731135	0.51944387	**0.8198089**
1	0.8653939	0.50372179	0.8635843	0.8671882	0.9396581	0.4412329	0.42120300	0.9441853	0.43064747	**0.6904455**
1	0.8830065	0.64360861	0.8805886	0.8853924	0.9013329	0.8523860	0.46167651	0.9840205	0.59881908	**0.8768595**

ML Algo	KNN	Naive Bayes	Random Forest	KNN	Naive Bayes	Random Forest

12.41 FEATURE SET METRICS FOR PREDICTING MODERATE ANAEMIA

TABLE 53: FULL METRICS FOR PREDICTING MODERATE ANAEMIA

Kappa	CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
0.40093000	0.8088909	0.8130242	0.9529531	0.4110406	0.31315621	0.9687898	0.35536123	**0.6819969**
0.34803395	0.8258231	0.8298087	0.9790224	0.1691970	0.30151439	0.9576596	0.21057776	**0.5741097**
0.45933105	0.8172631	0.8230006	0.9429836	0.4659591	0.29840338	0.9714020	0.36368593	**0.7044713**
0.55855969	0.8494198	0.8531762	0.9755208	0.7184670	0.60486070	0.9851773	0.65673342	**0.65673342**
0.50527995	0.8638373	0.8674383	0.9730347	0.4589893	0.47040085	0.9718412	0.46404899	**0.7160120**
0.64360861	0.8805886	0.8853924	0.9829841	0.8558726	0.72352294	0.9924320	0.78411762	**0.9194284**

ML Algo	Feature Set	Acc
KNN	Standard	0.8109640
Naïve Bayes	Standard	0.8278227
Random Forest	Standard	0.8201453
KNN	1	0.8513053
Naïve Bayes	1	0.8656455
Random Forest	1	0.8830065

12.42 FEATURE SET METRICS FOR PREDICTING SEVERE ANAEMIA

TABLE 54: FULL METRICS FOR PREDICTING SEVERE ANAEMIA

CI95% lower	CI95% upper	Specificity	Sensitivity	Pos Pred Value	Neg Pred Value	F1	Balanced Acc
0.8113137	0.8154264	0.9730717	0.6098538	0.11270399	0.9977598	0.19018871	**0.7914628**
0.8258231	0.8298087	0.9841007	0.2191215	0.07213697	0.9955761	0.10831033	**0.6016111**
0.8226768	0.8283469	0.9751870	0.6972067	0.13550292	0.9982719	0.22683959	**0.8361969**
0.8554110	0.8591045	0.9905563	0.8596028	0.33792866	0.9992062	0.48501869	**0.9250795**
0.8635843	0.8671882	0.9854134	0.7117093	0.21547068	0.9983635	0.33034427	**0.8485613**
0.8822287	0.8870036	0.9953822	0.9338113	0.53217606	0.9996263	0.67764937	**0.9645968**

ML Algo	Feature Set	Acc	Kappa
KNN	Standard	0.8133766	0.40341446
Naïve Bayes	Standard	0.8278227	0.3480339
Random Forest	Standard	0.8255255	0.47039589
KNN	1	0.8572652	0.56889176
Naïve Bayes	1	0.8653939	0.50372179
Random Forest	1	0.8846323	0.64321137

12.43 FEATURE SET METRICS FOR PREDICTING HAEMOLYSIS WITH DIFFERENT CLASS IMBALANCE HANDLING TECHNIQUES

TABLE 55: CLASS IMBLANACE HANDLING TECHNIQUE EXPERIMENTS FOR HAEMOLYSIS PREDICTIONS

	Accuracy	Sensitivity	Specificity	Kappa	95%- CI	SD Kappa	SD Sens
Down sampling	0.9693	0.9921	0.9693	0.1280	(0.9683, 0.9703)	0.0119	0.0076
Up sampling	0.999	0.7082	0.9997	0.7609	(0.9987, 0.9992)	0.0276	0.0451
SMOTE	0.9980	0.8512	0.9984	0.6685	(0.9976, 0.9984)	0.0238	0.0293
ROSE	0.9985	0.7784	0.9990	0.7033	(0.9981, 0.9988)	0.0284	0.0277

12.44 FEATURE SET METRICS FOR PREDICTING BONE MARROW DISEASE WITH DIFFERENT CLASS IMBALANCE HANDLING TECHNIQUES

TABLE 56: CLASS IMBALANCE HANDLING TECHNIQUE EXPERIMENTS FOR BONE MARROW DISEASE PREDICTIONS

	Accuracy	Sensitivity	Specificity	Kappa	95%- CI	SD Kappa	SD Sens
Down sampling	0.9719	0.9872	0.9718	0.1159	(0.9709, 0.9728)	0.0113	0.0081
Up sampling	0.9992	0.6506	0.9998	0.7465	(0.9989, 0.9994)	0.0304	0.0373
SMOTE	0.9981	0.8528	0.9984	0.6335	(0.9977, 0.9985)	0.0377	0.0354
ROSE	0.9967	0.7993	0.9971	0.4776	(0.9962, 0.9971)	0.0283	0.0424

12.45 FEATURE SET METRICS FOR PREDICTING VIT B12/FOLIC ACID DEFICIENCY ANAEMIA WITH DIFFERENT CLASS IMBALANCE HANDLING TECHNIQUES

TABLE 57: CLASS IMBALANCE HANDLING TECHNIQUE EXPERIMENTS FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA PREDICTIONS

	Accuracy	Sensitivity	Specificity	Kappa	95%- CI	SD Kappa	SD Sens
Down sampling	0.9597	0.9889	0.9596	0.1397	(0.9586, 0.9608)	0.0048	0.0054
Up sampling	0.9983	0.6103	0.9997	0.7165	(0.9980, 0.9987)	0.0332	0.0401
SMOTE	0.9964	0.7986	0.99704	0.6004	(0.9958,0.9968)	0.0267	0.0326
ROSE	0.9951	0.5845	0.9965	0.4498	(0.9945,0.9957)	0.0281	0.0384

12.46 FEATURE SET METRICS FOR PREDICTING IRON DEFICIENCY ANAEMIA WITH DIFFERENT CLASS IMBALANCE HANDLING TECHNIQUES

TABLE 58: CLASS IMBALANCE HANDLING TECHNIQUE EXPERIMENTS FOR IRON DEFICIENCY ANAEMIA PREDICTIONS

	Accuracy	Sensitivity	Specificity	Kappa	95%- CI	SD Kappa	SD Sens
Down sampling	0.9828	0.9901	0.9825	0.8298	(0.9821,0.9836)	0.0073	0.0018
Up sampling	0.9952	0.9541	0.9972	0.9450	(0.9947,0.9958)	0.0026	0.0036
SMOTE	0.9939	0.9729	0.9949	0.9313	(0.9932,0.9945)	0.0028	0.0036
ROSE	0.9879	0.9132	0.9914	0.8652	(0.9870,0.9887)	0.0046	0.0055

12.47 FEATURE SET COMPARISON FOR MISSRANGER DATASET

TABLE 59: FEATURE SET USING THE MISSRANGER DATASET

Anaemia	Iron deficiency anaemia	Anaemia of chronic disease	Vit B12/Folic acid deficiency anaemia	Bone marrow disease	Hemolysis	Severity
Ery	Transf_verz	Transf_verz	MCV	Trombo	Ery	Ery
Transf_verz	Ferritin	Transf	Ery	Leuco	RDW	VitB1
Ret_He	Transf	TYBC	Transf	aTTG_n	Mg	Transf_verz
Transf	TYBC	Fe	TYBC	Ery	Reti_n	MCH_n
TYBC	MCH_n	Ret_He	Transf_verz	Trombo_cit	NTproBNP	MCHC
Fe	Fe	Haptoglo	NTproBNP	NTproBNP	LD	Fe
Mg	Ret_He	NTproBNP	MCH_n	Mg	Myelo	aTTG_n
MCH_n	MCHC	Ferritin	aTTG_n	TYBC	TN_T_HS	PSA
NTproBNP	MCV	TN_T_HS	Ferritin	Segment	MCHC	Transf
aTTG_n	aTTG_n	Ery	LD	TN_T_HS	VitB1	TYBC
Ferritin	RDW	Myelo	TN_T_HS	Transf	Meta	NTproBNP
Reti_n	NTproBNP	Ureum	Ret_He	Neutro	Ureum	RDW
LD	Testost_lum	Mg	Fe	PTH	Erytrobl	Ret_He
PSA	LD	LD	RDW	Ureum	LDL_chol	Mg
Ureum	Ureum	aTTG_n	Mg	Lymfo	Alb	MCV
MCHC	Ery	Alb	Foliumz_R	Ret_He	Anemia	Ferritin
TN_T_HS	Staaf	Promyelo	Myelo	MCV	Staaf	Ca

RDW	Myelo	Testost_lum	D_Dimeer	D_Dimeer	Krea_Ul	PTH
MCV	TN_T_HS	Reti_n	Ureum	Testost_lum	gGT	TN_T_HS
VitB1	HbA1cN	PSA	PSA	Mono	Gluc	Ureum
Krea_Ul	PTH	VitB1	Reti_n	PSA	Act_B12_R	Myelo
Kreat_U	Mg	Meta	MCHC	VitB1	Testost_lum	Meta
Myelo	Promyelo	D_Dimeer	Promyelo	Bili_dir	Bili_dir	Reti_n
Kreat	PSA	Erytrobl	LDL_chol	TE	Ret_He	Staaf
LDL_chol	Meta	MCH_n	PTH	Haptoglo	CKD_epi	LD
P	Haptoglo	Kreat	Alb	LD	Na	Testost_lum
Promyelo	Reti_n	HbA1cN	Act_B12_R	RDW	D_Dimeer	Alb
HbA1cN	Erytrobl	Staaf	Erytrobl	LDL_chol	MCV	Kreat_U
Testost_lum	Kreat	Segment	Meta	Transf_verz	Ferritin	Trombo_cit
Trombo_cit	Lipase	PTH	Staaf	gGT	ASAT	LDL_chol
Na	Uraat	MCV	VitB6	Ferritin	PSA	Promyelo
Uraat	VitB1	Uraat	TE	Reti_n	Eo	TE
PTH	LDL_chol	MCHC	VitB1	Alb	Baso	D_Dimeer
Lymfo	Krea_Ul	Mono	Trombo_cit	MCH_n	Foliumz_R	Krea_Ul
Alb	Leuco	Neutro	Testost_lum	AF	Lymfo	Na
Erytrobl	Segment	RDW	AlbKr	AlbKr	AF	P
Leuco	Alb	LDL_chol	Uraat	Myelo	Ca	HbA1cN
D_Dimeer	D_Dimeer	Leuco	GFR_MDRD	Fe	HbA1cN	VitB6
Segment	Neutro	GFR_MDRD	Krea_Ul	PSA_Ratio	GFR_MDRD	Lymfo
Meta	P	Cl	Kreat	HbA1cN	Gluc_n	Kreat
Staaf	Cl	Krea_Ul	Haptoglo	Gluc_n	Haptoglo	Leuco
Neutro	Mono	CRP	P	Ca	AlbKr	Erytrobl
GFR_MDRD	Trombo_cit	Trombo_cit	Lipase	Foliumz_R	Kreat_U	Uraat
CKD_epi	GFR_MDRD	Lymfo	Baso	Uraat	TSH	Neutro
Foliumz_R	Gluc	CKD_epi	HbA1cN	ASAT	Trombo_cit	Segment
TE	Lymfo	gGT	gGT	Staaf	P	Bili_dir
Haptoglo	Gluc_n	TE	Cl	Lipase	PSA_Ratio	Gluc_n
Gluc_n	Kreat_U	Kreat_U	Bili_dir	VitB6	PTH	Haptoglo
PSA_Ratio	PSA_Ratio	Trombo	AF	Gluc	Triglyce	AlbKr
Cl	VitB6	AF	Lymfo	Meta	ALAT	Mono
VitB6	CKD_epi	PSA_Ratio	Segment	Krea_Ul	VitB6	PSA_Ratio
HDL_chol	Na	Gluc	Kreat_U	Promyelo	MCH_n	Triglyce
Mono	Bili_tot	Ca	PSA_Ratio	Kreat_U	Uraat	TSH
Triglyce	Eo	Gluc_n	Leuco	MCHC	Fe	HDL_chol
Gluc	Triglyce	Na	Neutro	GFR_MDRD	X25_OH_D3	Gluc
Lipase	Baso	Triglyce	CKD_epi	Bili_tot	Leuco	GFR_MDRD
gGT	Bili_dir	Lipase	Eo	Erytrobl	Mono	Foliumz_R
Ca	CRP	HDL_chol	Mono	Baso	TE	AF
CRP	AF	P	CRP	Na	Cho_HDLR	Baso
Bili_tot	Trombo	VitB6	Ca	Kreat	Cl	CKD_epi
Trombo	Foliumz_R	AlbKr	Gluc_n	Triglyce	Lipase	Trombo
AF	gGT	Bili_dir	Trombo	CKD_epi	Segment	gGT
Bili_dir	HDL_chol	Baso	Bili_tot	P	HDL_chol	Cl
K	TE	Bili_tot	Triglyce	HDL_chol	Chol	Lipase
AlbKr	AlbKr	Foliumz_R	Na	Cl	Neutro	CRP
Eo	Ca	Eo	Gluc	CRP	TYBC	Bili_tot
Baso	K	Chol	ASAT	TSH	Bili_tot	Eo
Chol	FT4	ASAT	K	Act_B12_R	K	FT4
TSH	Chol	K	HDL_chol	Chol	Kreat	ASAT
Act_B12_R	ASAT	FT4	Chol	ALAT	Transf	K
FT4	Act_B12_R	Cho_HDLR	TSH	K	Transf_verz	Chol
ASAT	Cho_HDLR	Act_B12_R	FT4	Eo	FT4	Cho_HDLR
Cho_HDLR	ALAT	ALAT	ALAT	Cho_HDLR	aTTG_n	ALAT
ALAT	TSH	TSH	Cho_HDLR	FT4	Trombo	Act_B12_R
X25_OH_D3	X25_OH_D3	X25_OH_D3	X25_OH_D3	X25_OH_D3	Promyelo	X25_OH_D3

12.48 FEATURE SET COMPARISON FOR MICE DATASET

TABLE 60: FEATURE SET USING THE MICE DATASET

Anaemia	Iron deficiency anaemia	Anaemia of chronic disease	Vit B12/Folic acid deficiency anaemia	Bone marrow disease	Hemolysis	Severity
Ery	Ferritin	TYBC	MCV	Trombo	Reti_n	Ery
MCH_n	Transf_verz	Transf	Ery	Leuco	LD	MCHC
Ret_He	MCH_n	Fe	MCH_n	Ery	Ery	MCH_n
RDW	MCHC	Transf_verz	RDW	RDW	RDW	RDW
Testost_lum	Fe	Ery	TYBC	Segment	Haptoglo	Fe
Transf_verz	RDW	Ret_He	Ferritin	Neutro	Ret_He	Transf_verz
MCHC	Transf	Ferritin	Transf	MCV	MCV	MCV
Kreat	TYBC	RDW	MCHC	Ureum	VitB1	Testost_lum
Ureum	MCV	MCHC	Transf_verz	Ret_He	Mg	Mg
Fe	Ery	Ureum	Foliumz_R	Mg	Ureum	VitB1
MCV	Testost_lum	Kreat	Ret_He	Transf	TE	Ureum
Transf	Ret_He	MCH_n	Act_B12_R	TYBC	MCHC	Ferritin
TYBC	Krea_Ul	Haptoglo	Fe	NTproBNP	NTproBNP	Ret_He
Mg	Ureum	Neutro	Krea_Ul	MCH_n	Transf	TYBC
Krea_Ul	Kreat	VitB1	Testost_lum	Lymfo	TYBC	LDL_chol
Ferritin	LD	Testost_lum	Ureum	LDL_chol	Ferritin	Transf
VitB1	HbA1cN	Segment	LD	Mono	Kreat_U	Krea_Ul
LD	LDL_chol	Krea_Ul	Mg	TE	Alb	Kreat
GFR_MDRD	Mg	LD	VitB1	Krea_Ul	Trombo	NTproBNP
Kreat_U	GFR_MDRD	Alb	Kreat	Kreat	MCH_n	Alb
NTproBNP	NTproBNP	Mg	Alb	GFR_MDRD	Krea_Ul	Kreat_U
LDL_chol	Kreat_U	MCV	NTproBNP	LD	LDL_chol	Na
Na	VitB1	Trombo	Kreat_U	VitB1	Fe	aTTG_n
Uraat	Uraat	Na	LDL_chol	Ferritin	gGT	TE
Reti_n	Alb	Leuco	VitB6	Transf_verz	Transf_verz	GFR_MDRD
HbA1cN	Trombo	GFR_MDRD	TE	Fe	Na	Ca
Alb	Haptoglo	NTproBNP	GFR_MDRD	Alb	aTTG_n	Trombo
Segment	Segment	LDL_chol	Trombo	Kreat_U	Testost_lum	Reti_n
Trombo	Na	HbA1cN	Reti_n	PSA_Ratio	Kreat	LD
CKD_epi	CKD_epi	Reti_n	Na	CKD_epi	D_Dimeer	HbA1cN
Neutro	Neutro	CRP	Haptoglo	MCHC	HbA1cN	Haptoglo
Leuco	Reti_n	Kreat_U	Segment	Ca	TN_T_HS	Uraat
Haptoglo	Leuco	Uraat	Trombo_cit	TN_T_HS	Gluc	Segment
TE	aTTG_n	CKD_epi	Uraat	aTTG_n	Neutro	CKD_epi
P	Gluc_n	gGT	HbA1cN	Reti_n	Trombo_cit	Leuco
HDL_chol	Gluc	Gluc	Leuco	PTH	GFR_MDRD	Gluc
PTH	PTH	Cl	CKD_epi	Trombo_cit	ASAT	PSA
Cl	HDL_chol	D_Dimeer	D_Dimeer	gGT	PSA	D_Dimeer
Ca	Trombo_cit	TE	HDL_chol	D_Dimeer	Cl	Neutro
Gluc	Triglyce	HDL_chol	PTH	PSA	Ca	Triglyce
Triglyce	TE	Gluc_n	K	Gluc	Segment	P
gGT	CRP	Ca	Mono	HbA1cN	Foliumz_R	HDL_chol
Trombo_cit	P	Triglyce	Cl	Na	Leuco	PTH
Gluc_n	Cl	PTH	Ca	P	PTH	Trombo_cit
K	Ca	Trombo_cit	P	Uraat	Lymfo	gGT
aTTG_n	gGT	Mono	Neutro	Testost_lum	HDL_chol	Cl
D_Dimeer	Lymfo	AF	CRP	HDL_chol	Uraat	CRP
CRP	PSA_Ratio	Lymfo	Gluc	Haptoglo	Triglyce	TN_T_HS
PSA_Ratio	D_Dimeer	PSA	Gluc_n	K	CKD_epi	Lymfo
Lymfo	K	aTTG_n	aTTG_n	Gluc_n	Gluc_n	PSA_Ratio
PSA	PSA	P	gGT	AlbKr	K	Gluc_n
Cho_HDLR	Chol	K	Lymfo	Cl	AF	FT4
TSH	Bili_tot	PSA_Ratio	PSA	Myelo	Meta	K
Mono	Mono	Chol	ASAT	TSH	PSA_Ratio	TSH
AF	AF	Cho_HDLR	Triglyce	AF	P	Mono
FT4	ASAT	AlbKr	ALAT	Triglyce	TSH	AF
TN_T_HS	Cho_HDLR	Bili_tot	X25_OH_D3	Chol	CRP	ASAT
Chol	FT4	FT4	AF	Eo	Staaf	Foliumz_R
ASAT	Foliumz_R	ASAT	Meta	ASAT	FT4	VitB6
Foliumz_R	TN_T_HS	TN_T_HS	TSH	Baso	ALAT	Myelo
AlbKr	TSH	Staaf	FT4	Lipase	VitB6	Cho_HDLR

Bili_tot	AlbKr	TSH	TN_T_HS	Cho_HDLR	Chol	Erytrobl
VitB6	X25_OH_D3	VitB6	Chol	VitB6	Myelo	AlbKr
ALAT	ALAT	ALAT	Staaf	Act_B12_R	Cho_HDLR	Meta
Staaf	Act_B12_R	Bili_dir	Lipase	Meta	Promyelo	Staaf
Bili_dir	Meta	Lipase	Myelo	Bili_dir	Lipase	Bili_dir
Lipase	Staaf	Foliumz_R	Erytrobl	Foliumz_R	Bili_dir	Chol
Meta	Baso	Meta	Bili_dir	CRP	Erytrobl	Eo
X25_OH_D3	Lipase	Eo	Cho_HDLR	Erytrobl	Mono	Promyelo
Act_B12_R	VitB6	X25_OH_D3	Bili_tot	ALAT	Bili_tot	Bili_tot
Erytrobl	Bili_dir	Promyelo	PSA_Ratio	Promyelo	Act_B12_R	Baso
Eo	Eo	Baso	Promyelo	FT4	AlbKr	ALAT
Baso	Promyelo	Act_B12_R	AlbKr	X25_OH_D3	Eo	Lipase

12.49 BIOMARKER COMBINATIONS FOR ANAEMIA OF CHRONIC DISEASE

TABLE 61: MOST COMMON STANDARD BIOMARKER COMBINATIONS FOR DIAGNOSING ANAEMIA OF CHRONIC DISEASE

Act_B12_R	CKD_epi	Fe	Ferritin	Foliumz_R	LD	Leuco	MCV	Reti_n	Transf	Trombo	ID
	X	X	X		X	X	X	X	X	X	**1**
		X	X		X	X	X		X	X	**2**
		X	X		X	X	X	X	X	X	**3**
X	X	X	X		X	X	X	X	X	X	**4**
		X	X			X	X		X	X	**5**
					X	X	X		X	X	**6**
X	X	X	X	X	X	X	X	X	X	X	**7**

12.50 BIOMARKER FREQUENCIES FOR ANAEMIA OF CHRONIC DISEASE

TABLE 62: ANAEMIA OF CHRONIC DISEASE BIOMARKER FREQUENCIES (STANDARD BIOMARKER)

ID	Frequency	Relative Frequency
1	7,608	33.63%
2	4,753	21.01%
3	3,716	16.43%
4	1,114	4.92%
5	1,104	4.88%
6	1,010	4.46%
7	581	2.57%

12.51 BIOMARKER COMBINATIONS FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

TABLE 63: MOST COMMON STANDARD BIOMARKER COMBINATIONS FOR DIAGNOSING VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

Act_B12_R	CKD_epi	Fe	Ferritin	Foliumz_R	LD	Leuco	MCV	Reti_n	Transf	Trombo	ID
					X	X	X			X	**1**
X	X		X	X	X	X	X	X		X	**2**
		X	X		X	X	X		X	X	**3**
X	X	X	X	X	X	X	X	X	X	X	**4**
	X			X	X	X	X	X		X	**5**
	X	X	X		X	X	X	X	X	X	**6**
	X	X	X	X	X	X	X	X	X	X	**7**

12.52 BIOMARKER FREQUENCIES FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA

TABLE 64: VIT B12/FOLIC ACID DEFICIENCY ANAEMIA BIOMARKER FREQUENCIES (STANDARD BIOMARKER)

ID	Frequency	Relative Frequency
1	324	16.23%
2	348	12.42%
3	210	10.52%
4	147	7.36%
5	125	6.26%
6	119	5.96%
7	88	4.41%

12.53 BIOMARKER COMBINATIONS FOR BONE MARROW DISEASE

TABLE 65: MOST COMMON STANDARD BIOMARKER COMBINATIONS FOR DIAGNOSING BONE MARROW DISEASE

Act_B12_R	CKD_epi	Fe	Ferritin	Foliumz_R	LD	Leuco	MCV	Reti_n	Transf	Trombo	ID
		X	X		X	X	X		X	X	**1**
					X	X	X			X	**2**
	X	X	X		X	X	X	X	X	X	**3**
		X	X			X	X		X	X	**4**
		X	X		X	X	X	X	X	X	**5**
					X	X	X		X	X	**6**
	X			X	X	X	X	X		X	**7**

12.54 BIOMARKER FREQUENCIES FOR BONE MARROW DISEASE

TABLE 66: BONE MARROW DISEASE BIOMARKER FREQUENCIES (STANDARD BIOMARKER)

ID	Frequency	Relative Frequency
1	227	20.54%
2	138	12.49%
3	122	11.04%
4	105	9.5%
5	92	8.33%
6	72	6.52%
7	64	5.79%

12.55 BIOMARKER COMBINATIONS FOR HAEMOLYSIS

TABLE 67: MOST COMMON STANDARD BIOMARKER COMBINATIONS FOR DIAGNOSING HAEMOLYSIS

Act_B12_R	CKD_epi	Fe	Ferritin	Foliumz_R	LD	Leuco	MCV	Reti_n	Transf	Trombo	ID
		X	X		X	X	X		X	X	**1**
	X	X	X		X	X	X	X	X	X	**2**
		X	X		X	X	X	X	X	X	**3**
					X	X	X	X	X	X	**4**
X	X			X	X	X	X	X		X	**5**
					X	X	X		X	X	**6**
					X	X	X	X		X	**7**

12.56 BIOMARKER FREQUENCIES FOR HAEMOLYSIS

TABLE 68: HAEMOLYSIS BIOMARKER FREQUENCIES (STANDARD BIOMARKER)

ID	Frequency	Relative Frequency
1	287	21.24%
2	253	18.73%
3	163	12.07%
4	101	7.48%
5	92	6.81%
6	73	5.4%
7	64	4.74%

12.57 BIOMARKER COMPARISON ANAEMIA OF CHRONIC DISEASE MICE DATASET

TABLE 69: MICE BIOMARKER FOR ANAEMIA OF CHRONIC DISEASE COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
99.65%	MCHC	TYBC	-5
99.61%	RDW	Transf	-3
99.52%	Kreat	Fe	-6
99.47%	Ery	Transf_verz	-3
97.11%	Transf	Ery	+1
97.07%	TYBC	Ret_He	-4
99.05%	Transf_verz	Ferritin	-2
89.09%	Ferritin	RDW	+6
87.8%	Fe	MCHC	+8
83.83%	Ret_He	Ureum	-1
17.76%	Ureum	Kreat	+8

12.58 BIOMARKER COMPARISON VIT B12/FOLIC ACID DEFICIENCY ANAEMIA MICE DATASET

TABLE 70: MICE BIOMARKER FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
100%	MCV	MCV	0
99.45%	Ery	Ery	0
99.4%	MCHC	MCH_n	-2
99.2%	RDW	RDW	0
75.7%	MCH_n	TYBC	-5
74.5%	Ret_He	Ferritin	-1
49.1%	Ferritin	Transf	-1
47.06%	Transf	MCHC	+5
46.91%	Transf_verz	Transf_verz	0
46.91%	TYBC	Foliumz_R	-1
36.53%	Foliumz_R	Ret_He	+5

12.59 BIOMARKER COMPARISON BONE MARROW DISEASE MICE DATASET

TABLE 71: MICE BIOMARKER FOR BONE MARROW DISEASE COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
100%	Leuco	Trombo	-2
100%	MCV	Leuco	+1
99.55%	Trombo	Ery	-2
99.55%	RDW	RDW	0
99.28%	Ery	Segment	-4
71.13%	Ret_He	Neutro	-2
64.34%	Transf	MCV	+5
50.68%	Neutro	Ureum	-2
15.11%	Segment	Ret_He	+3
12.85%	Ureum	Mg	-1
1.27%	Mg	Transf	+4

12.60 BIOMARKER COMPARISON HAEMOLYSIS MICE DATASET

TABLE 72: MICE BIOMARKER FOR HAEMOLYSIS COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
99.93%	MCV	Reti_n	-6
99.19%	Ery	LD	-2
99.04%	RDW	Ery	+1
97.56%	LD	RDW	+1
82.9%	Ret_He	Haptoglo	-3
61.95%	Reti_n	Ret_He	+1
16.14%	Ureum	MCV	+6
38.19%	Haptoglo	VitB1	-1
3.03%	VitB1	Mg	-2
1.48%	TE	Ureum	+3
0.67%	Mg	TE	+1

12.61 BIOMARKER COMPARISON ANAEMIA OF CHRONIC DISEASE MISSFOREST DATASET

TABLE 73: MISSFOREST BIOMARKER FOR ANAEMIA OF CHRONIC DISEASE COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
99.65%	MCHC	TYBC	-4
99.61%	RDW	Transf_verz	-4
99.47%	Ery	Fe	-5
97.11%	Transf	Ery	+1
97.07%	TYBC	Transf	+1
99.05%	Transf_verz	Ferritin	-1
89.09%	Ferritin	MCHC	+6
87.8%	Fe	Mg	-3
18.03%	NTproBNP	NTproBNP	0
17.76%	Ureum	RDW	+8
1.37%	Mg	Ureum	-3

12.62 BIOMARKER COMPARISON VIT B12/FOLIC ACID DEFICIENCY ANAEMIA MISSFOREST DATASET

TABLE 74: MISSFOREST BIOMARKER FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
100%	MCV	LD	-4
99.45%	Ery	RDW	-2
99.4%	MCHC	Ery	+1
99.2%	RDW	Reti_n	-3
89.57%	LD	TYBC	-5
74.5%	Ret_He	Ret_He	0
58.83%	Reti_n	Haptoglo	-4
47.06%	Transf	Transf_verz	-1
46.91%	Transf_verz	Transf	+1
46.91%	TYBC	MCHC	-+7
1.85%	Haptoglo	MCV	+10

12.63 BIOMARKER COMPARISON BONE MARROW DISEASE MISSFOREST DATASET

TABLE 75: MISSFOREST BIOMARKER FOR BONE MARROW DISEASE COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
100%	Leuco	Ery	-2
99.55%	Trombo	Trombo	0
99.55%	RDW	RDW	-4
99.28%	Ery	Mg	-4
50.68%	Neutro	NTproBNP	-3
12.85%	Ureum	Leuco	+5
1.27%	Mg	TN_T_HS	-2
14.39%	NTproBNP	Trombo_cit	-2
0.9%	TN_T_HS	Ureum	+3
0.18%	Trombo_cit	Neutro	+5
0.0%	Testost_lum	Testost_lum	0

12.64 BIOMARKER COMPARISON HAEMOLYSIS MISSFOREST DATASET

TABLE 76: MISSFOREST BIOMARKER FOR HAEMOLYSIS COMPARED TO THE ACTUAL PROCESS

Frequency	Biomarker	Importance Order	Position Change for importance order
99.26%	MCHC	Ery	-1
99.19%	Ery	RDW	-1
99.04%	RDW	Mg	-8
97.56%	LD	Reti_n	-1
61.95%	Reti_n	NTproBNP	-1
19.39%	NTproBNP	LD	+2
3.03%	VitB1	Myelo	-1
1.7%	Myelo	TN_T_HS	-2
1.48%	Meta	MCHC	+8
1.18%	TN_T_HS	VitB1	+3
0.67%	Mg	Meta	+2

12.66 ASSESSMENT IRON DEFICIENCY ANAEMIA PREDICTIONS

TABLE 77:EXPERT ASSESSMENT FOR IRON DEFICIENCY ANAEMIA BIOMARKERS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
Transf_verz	X	Ferritin	X
Ferritin	X	Transf_verz	X
Transf	X	MCH_n	X
TYBC	X	MCHC	X
MCH_n	X	Fe	X
Fe	X	RDW	X
Ret_He	X	Transf	X
MCHC	X	TYBC	X
MCV	X	MCV	X
aTTG_n	X	Ery	X
RDW	X	Testost_lum	O

12.67 ASSESSMENT ANAEMIA OF CHRONIC DISEASE PREDICTIONS

TABLE 78: EXPERT ASSESSMENT FOR ANAEMIA OF CHRONIC DISEASE PREDICTIONS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
Transf_verz	X	TYBC	X
Transf	X	Transf	X
TYBC	X	Fe	X
Fe	X	Transf_verz	X
Ret_He	X	Ery	X
Haptoglo	O	Ret_He	X
NTproBNP	X	Ferritin	X
Ferritin	X	RDW	X
TN_T_HS	O	MCHC	X
Ery	X	Ureum	O
Myelo	O	Kreat	X

12.68 ASSESSMENT VIT B12/FOLIC ACID DEFICIENCY ANAEMIA PREDICTIONS

TABLE 79: EXPERT ASSESSMENT FOR VIT B12/FOLIC ACID DEFICIENCY ANAEMIA PREDICTIONS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
MCV	X	MCV	X
Ery	X	Ery	X
Transf	O	MCH_n	X
TYBC	O	RDW	X
Transf_verz	O	TYBC	O
NTproBNP	O	Ferritin	O
MCH_n	X	Transf	O
aTTG_n	X	MCHC	X
Ferritin	O	Transf_verz	O
LD	X	Foliumz_R	X
TN_T_HS	O	Ret_He	O

12.69 ASSESSMENT BONE MARROW DISEASE PREDICTIONS

TABLE 80: EXPERT ASSESSMENT FOR BONE MARROW DISEASE PREDICTIONS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
Trombo	X	Trombo	X
Leuco	X	Leuco	X
aTTG_n	O	Ery	X
Ery	X	RDW	O
Trombo_cit	X (=trombo)	Segment	X
NTproBNP	O	Neutro	X
Mg	O	MCV	X
TYBC	O	Ureum	O
Segment	X	Ret_He	O
TN_T_HS	O	Mg	O
Transf	O	Transf	O

12.70 ASSESSMENT HAEMOLYSIS PREDICTIONS

TABLE 81: EXPERT ASSESSMENT FOR HAEMOLYSIS PREDICTIONS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
Ery	X	Reti_n	X
RDW	X	LD	X
Mg	O	Ery	X
Reti_n	X	RDW	X
NTproBNP	O	Haptoglo	X
LD	X	Ret_He	O
Myelo	O	MCV	X
TN_T_HS	O	VitB1	O
MCHC	X	Mg	O
VitB1	O	Ureum	O
Meta	O	TE	O

12.71 ASSESSMENT ANAEMIA SEVERITY PREDICTIONS

TABLE 82: EXPERT ASSESSMENT FOR ANAEMIA SEVERITY PREDICTIONS

Ranking 1	Assessment Ranking 1	Ranking 2	Assessment Ranking 2
Ery	X	Ery	X
VitB1	O	MCHC	X
Transf_verz	O	MCH_n	X
MCH_n	X	RDW	X
MCHC	X	Fe	O
Fe	O	Transf_verz	O
aTTG_n	O	MCV	O
PSA	O	Testost_lum	O
Transf	O	Mg	O
TYBC	O	VitB1	O
NTproBNP	O	Ureum	O