Department of Physics and Astronomy
University of Heidelberg

Master Thesis

submitted by

Anna-Katharina Nitschke

Digital Twins of Patients In Urology - A Proposed Architecture

This master thesis has been carried out by Anna-Katharina Nitschke at the
Physics Institute of the University of Heidelberg
under the supervision of
Matthias Weidemüller

Digital Twins of Patients In Urology - A Proposed Architecture

The translation of the concept of a Digital Twin, a virtual representation of a physical asset, into a clinical setting would lead to a revolution in personalised medicine and is therefore the aim of this thesis. The complexity and heterogeneity of a human body and possible diseases raises new challenges for the design of such a patient-specific support system for assisting clinician's decision-making. In this thesis, a Digital Twin architecture is proposed that is potentially able to deal with the posed requirements and challenges. As this work is embedded in a larger project, challenges regarding collaborative algorithm development will be considered additionally. Conceptual advancements are exemplified for the treatment of prostate cancer and can be extended to other medical fields. An overarching architecture is presented that is encompassing the entire patient journey and can be split up into specific architectures, being specified to one decision task and combined though Bayesian inference. The proposed specific Digital Twin design uses ensemble learning, which combines the opinion of several voters. For this method, different ways of generating or selecting diverse and complementary individual voters, the voters' topological ordering and fusers combining their estimations are presented, implemented and evaluated. The best results could be obtained by a Stacking approach, for which a Feature Space Dimension Reduction was performed and a two-dimensional Neural Network was trained on the voters opinions. Second best performance could be achieved by a Redundancy-aware Feature Space Partitioning and trained Logarithmic Opinion Pool combination. It will be shown that the presented specific interface architecture is able to deal with distributed data sources and heterogeneous input types. It reduces the variance of the algorithm performance, assuring robust predictions, and introduces interpretability, potentially leading to an increase in acceptance by physicians.

Digitale Zwillinge von Patienten in der Urologie - Vorschlag einer Architektur

Die Übertragung des Konzepts eines Digitalen Zwillings in ein klinisches Umfeld, als digitale Abbildung eines spezifischen Patienten, würde zu einer Revolution in der personalisierten Medizin führen und ist daher das angestrebte Ziel dieser Arbeit. Die Komplexität und Heterogenität eines menschlichen Körpers und möglicher Krankheiten stellt neue Herausforderungen an die Gestaltung eines solchen patientenspezifischen Unterstützungssystems zur Entscheidungsfindung des Arztes. In dieser Arbeit wird eine mögliche Architektur eines solchen Digitalen Zwillings vorgeschlagen, welche in der Lage ist, die an sie gestellten Anforderungen und Herausforderungen zu bewältigen. Da diese Masterarbeit in ein größeres Projekt eingebettet ist, werden Herausforderungen bezüglich der kollaborativen Algorithmenentwicklung zusätzlich berücksichtigt. Konzeptionelle Weiterentwicklungen innerhalb dieser Arbeit werden beispielhaft für die Behandlung des Prostatakarzinoms aufgezeigt und können auf andere medizinische Bereiche übertragen werden. Es wird eine übergeordnete Architektur vorgestellt, welche den gesamten Patientenaufenthalt überspannt und aus Entscheidungsaufgaben-spezifizierten Architekturen, kombiniert durch bayessche Inferenz, besteht. Für das spezifischere Design wird die Verwendung von Ensemble-Learning vorgeschlagen, welches die Meinung mehrerer Wähler kombiniert. Für dieses Verfahren werden unterschiedliche Möglichkeiten zur Generierung bzw. Auswahl diverser und komplementärer individueller Wähler, die topologische Ordnung der Wähler und die Kombination ihrer Aussagen vorgestellt, implementiert und ausgewertet. Die besten Ergebnisse konnten durch einen Stacking-Ansatz erzielt werden, für den eine Reduktion der Parameteranzahl durchgeführt wurde und ein zweidimensionales neuronales Netzwerk anhand der Aussagen der Wähler trainiert wurde. Die zweitbeste Leistung konnte durch eine redundanzbewusste Aufteilung des Parameterraumes (Redundancy-aware Feature Space Partitioning) und einer logarithmische Meinungskombination mit erlernter Gewichtung der Wähler (Logarithmic Opinion Pool) erreicht werden. Es wird gezeigt, dass die vorgestellte spezifische Architektur in der Lage ist, mit verteilten Datenquellen und heterogenen Eingabetypen umzugehen. Sie reduziert die Varianz der Algorithmusleistung, gewährleistet robuste Vorhersagen und führt Interpretierbarkeit ein, was möglicherweise zu einer Erhöhung der Akzeptanz durch die Ärzte führt.

Contents

1 Motivation

> *"So we need two things: first, we need ways of predicting and detecting disease well before it becomes life threatening; and second, we need medicines that work for you and your unique body."*
>
> — Pieter Cullis [1]

The medical model dealing with this objective is called personalised medicine. Its implementation in form of a digital image of a person to record their state of health might have been treated as science fiction a few years ago, is now part of current research by the concept of patient-specific Digital Twins.

Originally, the concept of such a Digital Twin was first presented by Dr. Michael Grieves and John Vickers 2002 at University of Michigan as the "Ideal Concept for Product Lifecycle Management (PLM)". As they described it themselves [2]: "It is based on the idea that a digital informational construct about a physical system could be created as an entity on its own. This digital information would be a twin of the information that was embedded within the physical system itself and be linked with that physical system through the entire lifecycle of the system." Back then the success story of this concept was incalculable. A good twenty years later, this idea is widespread and used in many different settings. It has mainly evolved in the field of engineering and can be used for example to remotely control satellites, improve product development, develop city infrastructure, monitor wind farms or to map entire organisations [3][4].

As mentioned, the translation of the developed Digital Twin concept from the industry domain into healthcare is a current topic of research [5]. The three main fields of application of the Digital Twin concept are: hospital design, hospital management and patient care [6], of which this thesis will concentrate on the last point. In medicine, "one usually aims to provide therapies with greater effectiveness and fewer side effects through a better understanding of the physiological and pathological processes" [3]. Going into the development of digital twins for patients, we quickly leave the realms of analytical descriptiveness due to the current lack of understanding of the underlying disease. As engineering approaches do not apply to disease behaviour in humans, a switch from the mechanical simulation of organ systems or metabolic cycles to problem-specific machine learning predictions is needed. The Digital Twins should provide a frame of reference to analyse the evolution of the patient state [5] and support decision making. It is still an open question how this design should look like.

An article [7] published by Shaip illustrates how until now AI has shown the potential to power the next wave of healthcare innovation through processing massive data sets far beyond the scope of human ability. They could possibly help physicians to plan and provide better care, first of all by using advanced pattern-recognition capabilities for medical image analysis (highlighting image features, identifying early cancer predictors, ...). Furthermore through its ability to cross-reference, AI could help to discover new drugs. Moreover probable health concerns could be detected early through the analysis of patients electronic health record data. Two important areas of research regarding those tasks are Computer Aided Diagnosis (CAD) systems or Clinical Decision Support Systems (CDSS) [8].

An interesting research area for the application of machine learning is prostate cancer, as this is one of the few medical diseases for which sufficient data is available. This is because prostate cancer is the most common malignant cancer for men worldwide [9]. Its diagnosis and treatment improvement through the implementation of a Digital Twin in the clinical context therefore is a relevant and interesting example case. The prostate gland (prostate) is one of the internal sex organs in men and prostate cancer is a malignant tumour of the prostate. "In Germany, more than 58,000 prostate carcinomas are diagnosed each year and, at 25.4%, it is the most common malignant neoplasm in men. When it comes to cancer leading to death, prostate carcinoma is the third with 10.1% (approx. 12,000 men)" [10]. Obviously, the improvement of diagnostic and treatment procedures through personalised medicine are of great importance.

In this thesis, we propose a Digital Twin architecture of a prostate cancer patient in urology that would support the clinician's decision-making. In the Chapter 2, different types of challenges and requirements posed on these systems will be identified. Additionally, the presented concept will be discussed on a superior as well a specific level of the interface between patients, clinicians and the Digital Twin. Chapter 3 of this thesis therefore on the one hand presents an overview of options for implementing the proposed architecture to Digital Twin developers through a literature research and one the other hand shows a concrete implementation of this architecture for one specific decision task in urology, namely a biopsy decision. The results can provide support to the hypothesis stated in the second chapter. Nevertheless, the need for improvements and further developments is explained. In addition, the limits of the approach are evident and will be further discussed in the last chapter. The critical analysis and consideration of the results is crucially needed in the process of generating guidelines for projects with clinical collaborations.

2 A General Architecture of a Digital Twin in Urology

This thesis is embedded in a larger project called CLINIC5.1 (Comprehensive Life-sciences Neural Information Computing), initiated by the Department of Urology at the University Hospital Heidelberg, aiming to develop a Digital Twin (DT) of a prostate cancer patient. Therefore, the overall task was to develop a proof of concept for a Digital Twin that should be able to deal with the specific challenges and requirements of the urologists on site as well as being adaptable to other problems in medicine.

The current chapter is going to show a systematic comparison of clinical settings with currently available concepts and approaches in industry, and finding their translation into our project. Challenges and requirement, which a Digital Twin design needs to withstand are going to be identified and discussed. A possible design for the Digital Twin architecture will be discussed on the superior level, overarching the entire patient journey, as well as on the specific level, being specified on one decision task. The specific Digital Twin architecture is exemplified on the clinical decision task, whether or not a biopsy needs to be performed.

2.1 Overarching Interface Structure

2.1.1 Clinical Patient Journey

Journey: The process that a prostate cancer patient goes through is visualised in Figure 2.1. After a conspicuous finding by an urologist, the patient is forwarded to the urologists in the clinic. For a targeted diagnosis, the urologist primarily carries out a blood test, a palpation report called digital rectal examination (DRE) and a transrectal sonography (TRUS). Nowadays, magnetic resonance imaging (MRI) is often used for further diagnosis and allows a second estimation of the parameters taken by the urologist. If abnormalities occur in one of these diagnostic methods, a biopsy is carried out on the patient in the urology department. The biopsy is a targeted tissue sample removal from the prostate, which can be carried out both in a systematically and target-oriented way by superimposing ultrasound and MRI images. This step represents the first main decision made by the urologist based on the available parameters. The malignancy of the tumour is assessed on the basis of these tissue samples and represented through a Gleason Score (GS) assigned by the pathologist. If suspicion exists that the disease is not limited to the actual prostate,

further diagnostics must be carried out. For this assessment, the urologist classifies the cancer according to a defined frame, called TNM [11], which can be adjusted and updated by further measurements. Here, T describes the tumour size and whether it is organ-confined. N describes whether lymph nodes are involved and M the existence of metastasis. This choice represents the second main decision step of the urologist, named the staging decision. Based on the available measurement results, it is then assessed which form of therapy seems suitable for the patient. This third decision is finally determined in a consultation with the patient and incorporates a variety of possibilities such as chemotherapy, hormone therapy, prostatectomy or radiotherapy. After the therapy, the patient ideally continues to receive follow-up care and monitoring in order to be able to assess the success of the therapy and the patient's condition.

From this patient journey we can already see, that the parameters obtained by the performed measurements are not the same over the different decision-making processes. Therefore, it can be understood as complementary information acquisition in contrast to a process in which parameters are updated over time. In addition, one can recognise that the correctness of the decisions made by the urologist is always tested through the next measurement. For example, the results of the biopsy show whether the biopsy itself was necessary. This means that there is no obvious right or wrong decision available to the clinician through the currently used parametrisation of the patient and the correlation to his health state. The reason for this is a lack of understanding of the underlying processes and of the ability to model them in a precise way. Decisions are therefore made based on a combination of the physician's intuition and evidence-based knowledge determined from studies. Due to the biological nature and heterogeneity of tumours, a biological model is not applicable. Therefore, evidence can only give estimates of an individual patient's progress. We can call this soft or population evidence.

Evidence Generation: There are different ways of synthesising the evidence found in individual studies, such as systematic reviews, meta-analysis, or clinical practice guidelines. While systematic reviews and meta-analysis aim to identify all information and results relevant to a specific research question [12], the clinical practice guidelines aim to standardise care and improve quality for the individual patient[13]. Systematic reviews provide robust data for clinical decisions by evaluating whether the findings of individual studies are consistent across populations, settings, and treatment options. In addition, if the review is well-conducted, bias is minimised, although publication bias is usually not taken into account and the findings are generalised. Guidelines do not only consist of high-level evidence-based knowledge, but also contain lower-level of evidence and expert opinions, therefore including knowledge from different hospitals and groups of specialists.

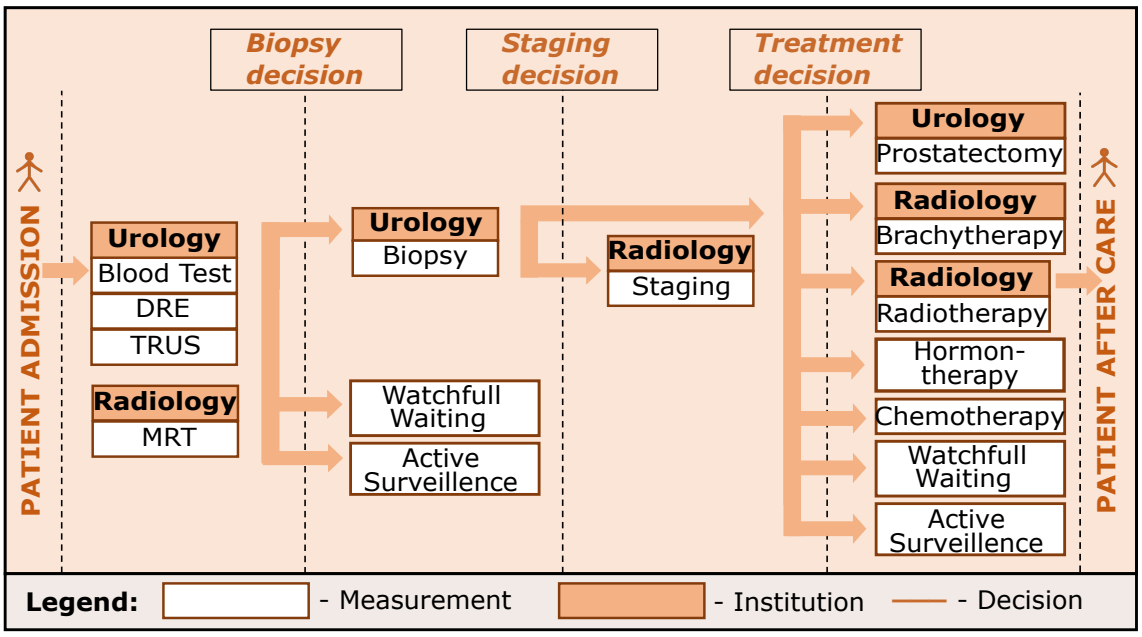

Figure 2.1: The visualisation of a prostate cancer patient's clinical journey. The journey of the real patient starts from the left. There are three main decisions that need to be carried out. For each decision (written in orange letters), an assignment of all necessary measurements (in boxes with a white background) and the responsible institution (in boxes with a dark orange background) is shown. The dark orange arrows indicate possible decision paths. After these three decisions, the patient receives his therapy and follow-up care. Abbreviations correspond to digital rectal examination (DRE), transrectal sonography (TRUS), magnetic resonance tomography (MRI).

One way of including this evidence generated over multiple trials is by including the Diagnostic Decision Support Systems (DDSS) many studies give as outcome. According to Shortliffe et al. DDSS methods can be devided into:"1) clinical algorithms, 2) clinical data banks that include analytic functions, 3) mathematical models of physical processes, 4) pattern recognition, 5) Bayesian statistics, 6) decision analyis, and 7) symbolic reasoning or artificial intelligence" [14]. The last mentioned methods are sometimes called expert systems [15]. In literature several overviews on expert systems for prostate cancer can be found [16][17][18]. Expert systems in the form of nomograms are widely used in the clinic. According to Shariat [19], "nomogram represents a graphical calculation instrument, that can be based on any type of function, such as logistic regression or Cox hazards ratio regression models". Every year new nomograms or updates to existing ones are published. This makes the identification of the relevant and most predictive ones a continuous challenge.

2.1.2 Digital Twin Concepts

Origin: Grieves and Vickers published a paper [20] presenting the origin of the Digital Twin concept and some definitions to rely on as following: "the Digital Twin is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level" [20]. One can distinguish between the Digital Twin Prototype (DTP), which is an informational set describing the virtual version and Digital Twin Instance (DTI), being an informational set describing the physical version. They are operated on in the Digital Twin Environment (DTE) in a predictive or interrogative way. The individual asset DTI's can be accessed through a computer construct called Digital Twin Aggregate (DTA). The concept of a Digital Twin presented by Dr. Michael Grieves was originally designed for a Product Lifecycle Management (PLM). Therefore, the interaction between the virtual and real systems was meant to allow the non-static connection during all lifecycle phases (creation, production, operation, disposal). The system evolves from being a DTP during the creation phase, giving a description of how to produce a physical version. As soon as the physical asset exists, information about the specific realisation is fed back to the virtual space in form of the DTI. From here on, information is transported between the systems in both ways to allow for adjustments and changes during the support/ sustain phase. At the final disposal/decommissioning phase the information gained from this individual system should be used for the next ones. In the context of medicine the DTP is of no interest, because the biological system itself already exists and we want to modulate it as good as possible. In the following DT will be used equivalently to DTI. A more explicit version of what a patient's DTE might look like, was presented by Croatti et al. [21] using the concept of mirror worlds. Not only the patient himself, but also the physician's team and the hospital (such as medical rooms and tools) should be transferred to the digital level, thus enabling a comprehensive view of the interaction with the patient.

Implementations: About twenty years later, this concept gained much attention and has been adapted and implemented in a wide range of use-cases. Many of these examples have implemented the concept as follows: Sensor data on the physical object are frequently collected in real-time and passed on to the virtual version. In turn this can provide predictions about the physical object based on physical simulations or models. Based on them, decisions can be made either manually by an operator or automatically by an algorithm, which in turn affect the object itself. Examples for this are the "Development of a digital twin for a flexible air separation unit using a pressure-driven simulation approach"[22], "A digital twin smart city for citizen feedback" [23] and "Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance" [24]. The options of application in the medical field are wide, ranging from hospital management, public health monitoring, drug development (such as SARS-CoV-2 [25]), therapy improvement (such as European citizens health record [26]) to personalised medicine. Boulos and

Zhang [27] have shown with their review of current Digital Twin concepts in the field of personalised medicine the variety of approaches, from modelling the full body, to one organ, to one cell, and so on down to the molecular level. "Given the intricacy and complex interconnections of the diverse types of systems within the human body, establishing an adequate, complete human Digital Twin may be far from reality" [27]. Therefore, instead of creating a holistic model, current studies mainly focus on modelling organ-specific functions and disease, such as such as Subramanian's presentation [28] of a virtual liver by means of including scientific knowledge about metabolism using ordinary differential equations.

Inclusion of Machine Learning: Due to the large amount of data provided by the physical asset some concept proposals have started to combine physics-based models with data-driven models in a so called hybrid analysis and modelling approach and can be used in many scientific and engineering applications ranging from wind power and weather forecasting to aircraft design [29]. Using this approach Corral-Acero et al. [30] have presented an interesting vision for precision cardiology. The idea has been to use a wide range of observational parameters to describe the patient state, such as genomics, lifestyle, environment and biological data. They have constructed a hybrid system, as statistical modelling allows for categorising patients according to the knowledge inferred from data and the mechanistic modelling the visualisation of the cardiovascular system. This provides more insights to support or reject the DT prediction, as it is based on a deduction from anatomic, mechanistic, and functional knowledge. This is possible as statistical models on the one hand can automatically extract known parameters as well as hidden patterns from data. On the other hand mechanistic models were able to provide interpretability to the clinicians and can be used to make predictions. More details can be taken from Figure C.2 in the Appendix.

Graphical Model: Kaptyn et al. [31] introduced in their work a Digital Twin concept developed for an self-aware aerial vehicle application. They aimed to develop a concept which is generalisable and therefore can be used to model any asset-twin system. The basis for their approach is the abstract formulation of the interface between the digital and real world, the asset and the twin. The physical state [S] of the asset is first represented as precisely and holistically as possible through the observational data [O] taken from it. The observational data is then forwarded to the Digital Twin that then defines the digital state [D] from it. The quantities of interest [Q] are the variables calculated from the physical state, allowing conclusions on whatever whatever one wants to know about the asset, such as its well-being. From this information a control input [U] is fed back to the asset and influences the asset. A reward [R] is assigned to the asset-twin system for this interaction. The interaction between these quantities is visualised by a graphical model in Figure 2.2. One can see that between time point zero t_0 and time point t_C there is an interaction with the physical space. This is called the calibration phase in which more and more

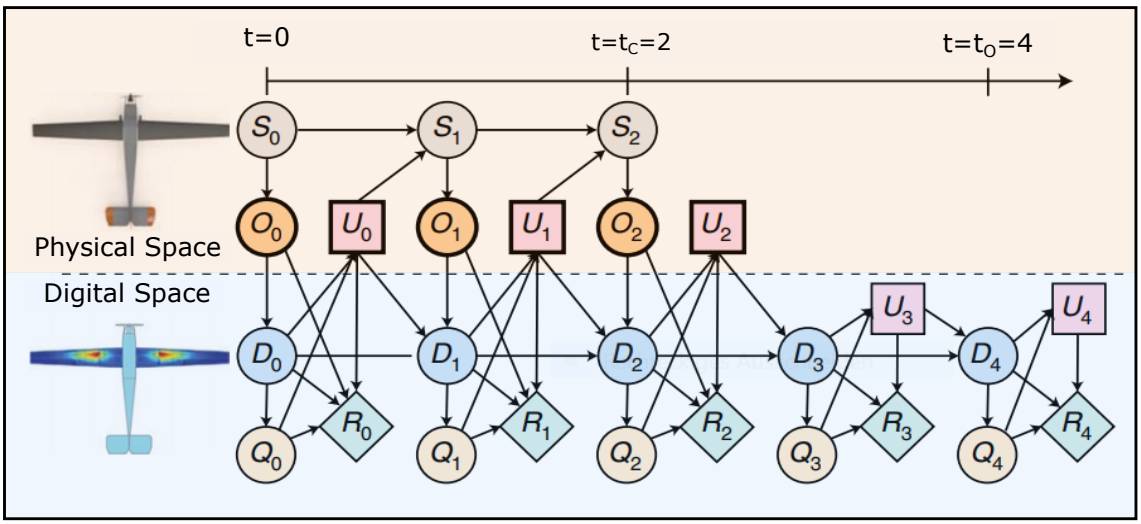

Figure 2.2: Visualisation of the interaction of the quantities defined to describe the Digital Twin concept developed by Kaptyn et al. [31]. The figure was added from their publication. To the asset (real world) belong the quantities: physical state [S], observational data [O] and control inputs [U]. To the Digital Twin (digital world) belong the quantities: digital state [D], quantities of interest [Q] and reward [R]. Besides the observational data being deterministic values, other quantities are estimated and therefore typically represented by a probability distribution. The graphical model shows the conditional dependence between the quantities through arrows.

information about the asset is gathered to adjust the Digital Twin to resemble the true asset more and more closely. After the calibration is finished, the operational phase starts in which the Digital Twin acts in a dynamic data-driven manner. For this phase it is necessarily that the quantity of interest "structural health" can be dynamically and reliably estimated through the calibrated simulation model. Therefore a physical understanding of the aerial vehicle state during its operational phase is needed.

2.1.3 Translating Overarching Concept

Based on the already existing general concepts for Digital Twins one could develop an interactive support system for clinicians that is able to assist during the whole cancer patient journey in a personalised way.

Context: Due to the fact that complex diseases such as prostate cancer are not understood in a physically describable way, mechanistic modelling can not be realised. But looking at the hybrid approach of Corral-Acero et al. [30], which enables one to process big data, one could try to find a new way of interpretability, for example through the use of model-agnostic methods or evidence-based knowledge.

The incomplete insight into the disease makes it necessary to work with machine learning models instead of physics-based simulations. The parallels between the challenges coming from modern AI and those from Digital Twins have been mentioned by Boulos and Zhang as being "data availability and quality issues; data integration and interoperability issues; data sharing issues, including concerns about intellectual property; data privacy and security across platforms and systems; and AI bias, (and poor) explainability and reproducibility issues" [27].

The general aim of the proposed Digital Twin concept is to improve currently used diagnosis and treatment strategies within the existing limits. Those limits correspond in the clinical context to evidence generated through current practice and research, summarised in the clinical guidelines. Staying within these bounds is given for supervised learning, since the data sets used, and in particular the labels, are often strongly biased by current procedures. This means that a good recommendation can be made form among the therapy options that are currently available to the patient, but no new form of therapy can be developed. If the case were otherwise, unsupervised learning could have been used to work out new clinical action guidelines. Basic definitions of words used in the context of machine learning are given in Table A.1 in the Appendix.

As the approach presented by Kaptyn et al. [31] has been the only concept that also allowed complementary parameters and not just time-evolving parameters, a closer look at this approach seems to be the most promising. In his doctoral thesis he indicated that his presented concept can also be applied to medical questions, such as hearth insufficiency (an example is given by the asset-twin system of a human heart in Figure C.1 in the Appendix).

Concept: While trying to translate the proposed interface structure of Kaptyn et al. [31], the following differences and similarities can be observed: Due to the complementary parameter acquisition, the clinical situation can be better represented by the calibration phase in which the digital version is updated and adjusted to more and more resemble the real patient. That means, that at each time-point,

representing a specific decision task in the clinical context, a interaction between the physical and the digital world is executed.

For the transfer to the clinical setting, a further layer is introduced due to the interaction of the physician with the real patient as well as DT. The previously mentioned "control input" corresponds in the clinical setting to the action taken on the patient. The overall performance of the asset-twin system can be measured by reviewing the decision correctness after performing the action (such as Gleason Score indicates whether biopsy was necessary) or after the health state of the patient has improved (PSA value evolution after treatment). The quantities used to describe the graphical model need to be adjusted as listed in Table 2.1.

Quantity	**Notation**	**Description**
Physical State	PS	real state of patient
Observational Data	O	parameters, allowing patient state estimation
Digital State	DS	model estimating patient state through observational data and knowledge inferred from training
Digital Decision	DD	decision suggestion forwarded to clinician
Clinician Decision	CD	decision of expert influencing the physical asset
Action	A	action taken on the patient
Reward	R	system feedback - reviewing prediction accuracy

Table 2.1: Summarising the quantities of a twin-patient system comprising the interacting components in the graphical model of Figure 2.3.

The superior interface structure is visualised in Figure 2.3. At the first time point, one has to decide whether a biopsy needs to be performed. The best possible representation of the patient's current health state is attempted based on the collected parameters that make up the observational data. The Digital Twin then uses this information as system input to generate a probabilistic decision as output. The output together with his own expertise on the given observational data will lead the clinician to a final decision, which in turn leads to an action on the patient. To be able to constantly improve the system's performance, the result of the biopsy and the change in the patient's state are fed back to the Digital Twin. From here on the procedure is repeated for the "Staging Decision" as well as "Treatment decision".

This observation makes clear, that for a specific realisation of a Digital Twin in this context, one can focus on finding a structure for the recurring specific interface at one decision task. It seems reasonable to use Bayesian inference to depict the information gain about the patient throughout the measurements and decisions made. This could be used to model the time evolution in the patient's journey, putting together individual decisions to form a bigger picture.

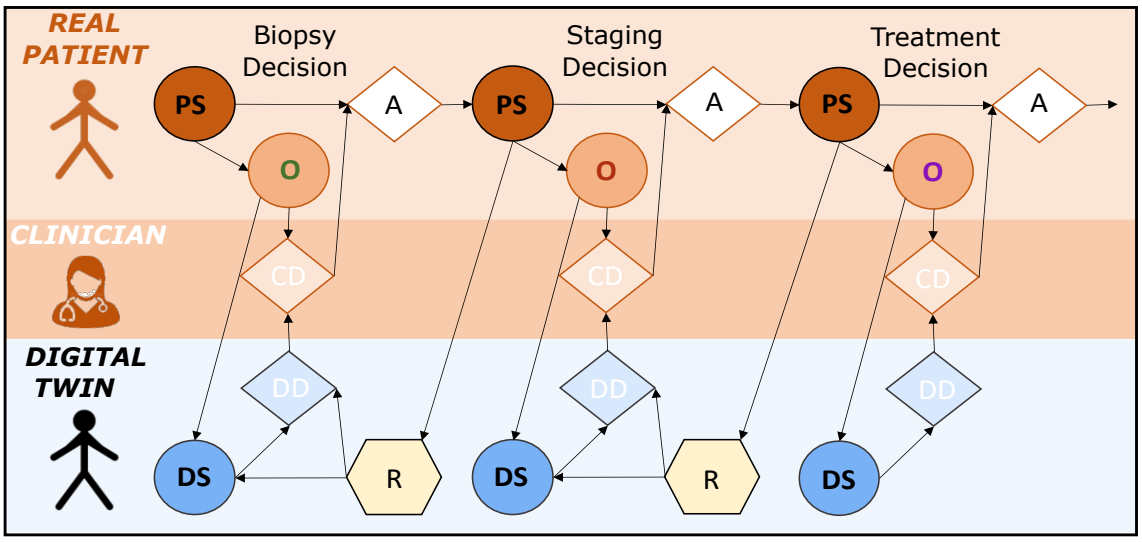

Figure 2.3: A Digital Twin concept developed for an clinical prostate cancer patient journey. The quantities used to model the asset-twin systems are described in Table 2.1. To the patient belong the quantities: physical state [PS], observational data [O] and actions [A]. To the Digital Twin belong the quantities: digital state [DS], digital decision [DD] and reward [R]. The third interface layer consists of the clinician with his quantity: clinician decision [CD]. The different colours used for the observational data are meant to imply that different information is taken from the physical state at each time-point. While the observational data are deterministic samples from probability distributions, other quantities are estimated and therefore typically non-deterministic probabilities. The visualised graphical model shows the conditional dependence between the quantities through arrows. A detailed description can be found in the text and Table 2.1.

2.2 Specific Interface Structure: Biopsy - yes or no?

2.2.1 Biopsy Decision

The prostate cancer patient journey visualised in Figure 2.1 can be broken down into three main decisions which the clinician has to perform: Biopsy, Staging, Therapy. For each decision, the clinician has parameters available that are intended to reflect the patient's state of health. Based on these parameters, decisions are made by relying on experience and supporting systems such as nomograms, but always within the boundaries of the clinical guidelines. For each decision a detailed look at the parameter acquisition, clinical guidelines, and decision support systems helps to understand the underlying process. A detailed look at the first decision was performed: Biopsy - yes or no?

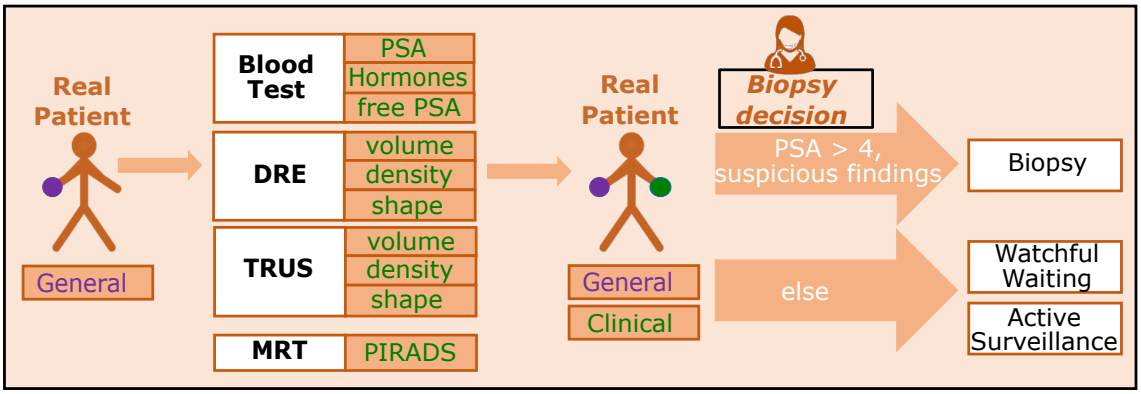

Figure 2.4: Visualisation of the clinical parameter acquisition and decision processes to determine whether the current patient is likely to have cancer and therefore has to undergo a biopsy. Parameters (in boxes with a dark orange background) are taken from the measurements (in boxes with a white background) - blood test, digital rectal examination (DRE), transrectal sonography (TRUS), magnetic resonance imaging (MRI). The colours in which the parameters are written are equivalent to the colours of the circles on the patient figure representing the assigned information. Further information on the parameters is included in the text.

Parameter Acquisition: The patient entering the clinic already holds a lot of important information for his primary cancer risk estimation. This is general information such as family history, age, BMI, race, pre-existing health condition, physical condition, and so forth. To decide whether a biopsy for this patient is going to be performed a few diagnostic measurements are conducted. From the blood test, factors about the performance of the prostate can be measured such as the PSA (prostate specific antigen) concentration. PSA is one of the serum markers for prostate cancer showing the highest prognostic strength when measured over 2 years [32]. Furthermore, one gains information about the hormone level and other

enzymes, which provide insight into tissue injuries (such as LDH- lactate dehydrogenase), bone metastasis development (such as AP - alkaline phosphatase) and affection of the liver (such as GPT - glutamat pyruvat transaminase). The digital rectal examination (DRE) allows the urologist to get a feeling for the size and shape of the prostate as well as the existence of any suspicious regions of rigidifications (palpable indurations). The transrectal sonography allows for a similar parameter collection, but through a visual estimation of the tissue density. Findings in the second visual inspection - multiparametric prostate magnetic resonance imaging (MRI) - are summed up in a so called PIRADS Score (Prostate Imaging Reporting and Data System) that attempts to assess the malignancy of the tumour in a structured reporting scheme [33]. The likelihood of clinically significant cancer is indicated through the assignments of a score ranging from 1 to 5 for each suspicious lesion. The significance of these levels was taken from definitions evolved by a representative group involving the American College of Radiology (ACR), European Society of Urogenital Radiology (ESUR), and AdMeTech Foundation [33]:

PI-RADS 1) very low (clinically significant cancer is highly unlikely)
PI-RADS 2) low (clinically significant cancer is unlikely)
PI-RADS 3) intermediate (clinically significant cancer is equivocal)
PI-RADS 4) high (clinically significant cancer is likely)
PI-RADS 5) very high (clinically significant cancer is very likely)

Clinical Guidelines: "Clinical practice guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" (Institute of Medicine, 1990, cited from [34]). In Germany, clinical guidelines can be extracted from the web page of the AWMF [35], the network of scientific medical societies in Germany. Given all the acquired information, the clinician then makes a decision whether or not a biopsy should be performed. The clinical guidelines give the threshold of a PSA value being higher than 4.0 ng/ml or any suspicious findings during the other measurements as being sufficient reason to justify the performance of a biopsy. If the comprehensive carcinoma diagnosis does not identify the need of an immediate active therapy, Active Surveillance (AS) can be conducted. This is a treatment strategy that involves active and close observation of the patient. Further treatments are only carried out if the patient's condition deteriorates or if there is a specific request for treatment. Due to factors such as patient age, the clinician can recommend a palliative instead of a curative therapy. This is called watchful waiting and focuses on the patient's quality of life and the management of complications of a disease.

Nomogram - PBCG: As decision support, clinicians can use nomograms such as the PBCG (Prostate Biopsy Collaborative Group) risk calculator [36] designed by Ankerst et al. (`https://riskcalc.org/PBCG/`). The risks of high-grade prostate cancer, defined through a Gleason score larger or equal to 7, were calculated through

the use of multinomial logistic regression. The relevant parameters for the cancer risk estimation [P] are age, prostate-specific antigen, digital rectal exam, African ancestry, first-degree family history, prior negative biopsy. The data was obtained from eight North American institutions and finally included 15,611 men undergoing 16,369 prostate biopsies (2006 - 2017). The researchers were able to show an area under receiver operating characteristic curve (auROC) of 75.5%, outperforming the widely used online Prostate Cancer Prevention Trial Risk Calculator (PCPTRC).

2.2.2 AI Based Decision Support Methods

As previously described, nomograms are the current state of support methods lack personalisation and accuracy. Therefore, alternative techniques are reviewed by searching for artificial intelligence (AI) methods developed to improve clinical decision making, currently focusing on the biopsy decision.

As the biopsy is part of the diagnostic procedure, the potential improvement of disease detection trough AI is of a certain interest. Due to the high costs and side-effects of a biopsy, cancer risk assessment tools should show a high performance in early stage detection, which AI promises to deliver through in-depth analysis of historical data. "For example, by looking at an entire patient population that closely matches the demographic of a specific individual in addition to the medical history of relatives, AI could conclude that a patient is very likely to develop a malady [...] years before a doctor could ever accurately make a diagnosis" [7].

Systematic reviews of the currently developed machine learning applications in urology are helpful to get an overview about possible decision support systems. Salem et al. [16] systematically reviewed in 2021 the urological health care support systems. They found Artificial Neural Networks (ANN) to be the mainly used model and prostate cancer to be a domain to which they have commonly been applied to. The use of ANN makes support system development easier because they bypass the need for quantifiable expert knowledge. "However, their analytical hidden layer of nodes black box phenomenon has been a subject for wide criticism and rejection from clinicians due to lack of transparency and understanding of its function" [16]. Also interesting is that the review of Shariat et al. in 2009 [37] about support systems developed for the prediction of prostate cancer in the initial biopsy showed that nearly all of them have been using an ANN as model. A few of these DSS also presented comparisons with different machine learning models in their publication, such as Logistic Regression (LR) [38] or Support Vector Machines (SVM) and Random Forest (RF) [39].

Interactions of the model with the clinician have been summarised as guiding the expert, supporting the none-expert or replacing the clinician completely [16]. The way of interaction of the Digital Twin has to be considered from the start. The goal is to guide experts to be able to make personalised decisions. Moreover, we want to support less experienced clinicians, so that for example interobserver reproducibility of the PIRADS Score assignment is increased, as this can vary depending on the level of experience as studies have shown [40].

Furthermore, focus was also placed on the translation from application development to actual clinical implementation and successful usage by Kawamoto et al. [41]. Although the clinical decision support systems show high potential to improve patient care, often a gap between development, implementation, and usage can be

observed, even though the reasons for this are not always clear. With the use of multiple logistic regression analysis, Kawamoto et al. [41] "identified four features as independent predictors of improved clinical practice: automatic provision of decision support as part of clinician workflow ($P < 0.00001$), provision of recommendations rather than just assessments ($P = 0.0187$), provision of decision support at the time and location of decision making ($P = 0.0263$), and computer based decision support ($P = 0.0294$)". These results make it clear that the effort reduction for clinicians to interact with the recommendation system needs to be taken into account from the outset.

2.2.3 Translating Specific Concept

In a clinical urology setting, despite the consideration of individual patient factors (such as age, comorbidities, etc.), the decision-making scope of the physicians is heavily regulated and specified by the guidelines. In general, for safety reasons the Department of Urology at the University Hospital Heidelberg recommends their patients a biopsy as soon as any abnormalities in the diagnostic procedure have been observed. PSA based diagnosis significantly reduced the mortality rate for prostate cancer (PCa). On the other hand in many biopsies no clinically significant PCa is detected [42]. Preferably only men with a clinically significant PCa should be identified and diagnosed. Therefore a decision support system fine-tuning the decision options within the frame of the clinical guidelines is sought.

Context: The general decision process for which we want to construct a clinical support system concept is visualised in Figure 2.5. Requirements for the concept of the Digital Twin come from different perspectives. Regarding ethical implications, Boulos and Zhang have stated the importance of governance mechanisms and policies to safeguard the data privacy and control transparent usage [27]. Design criteria have been identified by Schwartz et al.[43], namely, "clear data visualisation", "prioritisation of interventions", "ease of adding and removing data sources", and "integration into clinical workflow". Further criteria used for this proposal of a DT concept were found through the inspection of the clinical setting in which the Digital Twin will be placed, such as:

1. Robust predictions (possibility to generalise to other institutions; still has to work even if some data or algorithm is missing)
2. Inclusion of evidence-based knowledge
3. Personalised and interpretable decision support (intuitive for physicians)

Since the master's thesis was carried out as part of a cooperation project, the desired concept of a Digital Twin should take into account the situations that can arise in the context of this project. The possible occurring situations can be as following: a) One algorithm is using all available parameters for its prediction. b) One algorithm is using just a few of the available parameters for its prediction. c-d) Several algorithms for which each uses a non-/ overlapping set of available parameters, together using the entire parameter space. e-f) Several algorithms for which each uses a non-/ overlapping set of available parameters, together using only a part of the entire parameter space.

As there might occur situations in which several algorithms are trying to answer the same questions by looking at different parameters, a solution needs to be found how these different predictions can be combined and presented to a physician. This is the situation in the current project and raises the question of how such a combi-

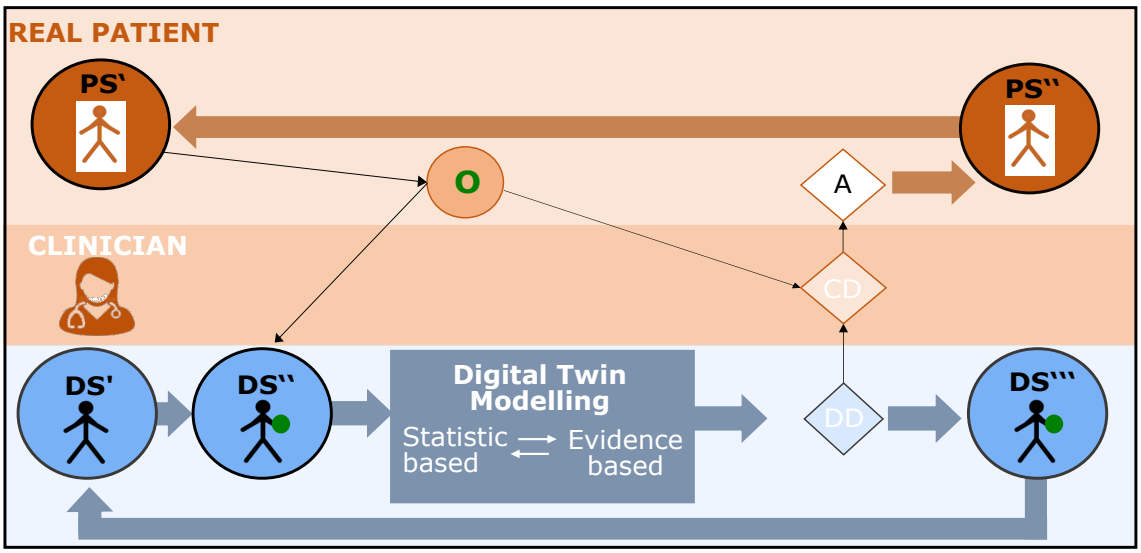

Figure 2.5: Visualisation through a graphical model for one of the clinical decision processes and the three-layer specific interface architecture between real patient, clinician and Digital Twin. The bold arrows represent the recurrence of the interface structure during the state evolution of the real (coloured with orange) and digital (coloured with blue) patient. The physical state [PS'] is changed through the action on the patient to [PS"]. From [PS'], observations [O] are taken and given to the clinician and the Digital Twin, leading to the evolution of the digital state from [DS'] to [DS"], represented by adding the blue coloured ball equivalent to the observation colour. The digital state is then analysed through a Digital Twin model, being a combination of statistic based and evidence based methods. The model returns a digital decision [DD] which evolves the digital state to [DS"'] and is given to the clinician. The final clinical decision [CD] is then performed by the clinician on the basis of O and DD.

nation can best be carried out. Further information about the different cooperation partners' individual algorithms and their tasks can be viewed in Figure C.3 in the Appendix. Moreover, there might be situations in which not all available parameters are used by the developed algorithms, but could contain so far unknown meaningful information, which is why their inclusion could be advantageous. As the precise record of the current physical condition of the patient is sought, all available information should be used in the best possible way. This in turn raises the question of how data can best be used to train models under these specific conditions.

Another challenge arising from collaborative projects and usage of clinical data is the need to learn over distributed data chunks stored in different databases. Legal or commercial constraints might come up and hamper the usage of distributed data because they do not allow sharing raw data sets and merging them into a common repository.

The resulting requirements that the concept must meet while answering the two previously mentioned questions are the following:

- Dealing with different data types (images, scores, continuous parameters)
- Dealing with data being distributed between different institutions (urology, radiology)
- Need to generate best possible performance by having only small data sets, but potentially many features available
- Need to work reliably, even if measurements have been left out or executed in an atypical way
- Need to allow evolution and improvement of existing algorithms, as well as new ones to be added to the system

Concept: In this regard, the ways of dealing with multi-view data (heterogeneous data providing complementary information) are reviewed by Li et al. [44] as follows: One can distinguish four types of multi-view data, in which the collaborative projects face the second data set type, which deals with the inclusion of distinct features for the same patients (the complete overview can be viewed in Table A.2 in the Appendix). They presented a general overview of how data fusion can be included through machine learning techniques categorised into early, intermediate or late integration methods. In early integration methods, features from different data are concatenated into a single feature vector before fitting an unsupervised or supervised model. The presented early integration method called concatenation involves the combination of all features to one single input vector before passing forward to a supervised model. Even if this approach seems straightforward and intuitive, constructing a model that is able to deal with this kind of input vector is not that easy. It might require further feature preprocessing. Interestingly, the previously mentioned late integration method for which separate models (base models) are first trained on the individual data subsets involves a combination of those individual outputs to a final response. Therefore this approach, which is called ensemble learning, is in a way a combination of the two questions raised previously. "Reviewing the literature of ensemble learning, one can find several theories that tend to explain the success of ensemble learning algorithms, such as the concept of diversity, the concept of margin, the ambiguity decomposition [45], the bias-variance decomposition and the bias-variance-covariance decomposition [46]" [47].

Ensemble learning approaches have been categorised into: (1) algorithms that use heterogeneous predictive models on the full dataset such as stacking; (2) algorithms that manipulate the instances of the datasets such as bagging, boosting, random forests, and bagging with subspaces; (3) algorithms that manipulate the learning algorithm such as random forests, neural networks ensemble, and extra-trees ensemble; (4) algorithms that manipulate the features of the datasets such as random forests, random subspaces, and bagging with subspaces [48].

We will focus on a combination of the algorithms in the first and forth category, allowing heterogeneous algorithms trained on non-overlapping feature spaces to be combined. The combination can either be a simple voting method or a more elaborated method using an algorithm (meta model) to learn the correct combination of the base models predictions. The different methods for feature space selection as well as the combination methods will be presented in Section 3.4 and 3.5 respectively. We have decided to not choose method (2) dealing with instance selection, because in most medical application the lack of data is already a significant problem and a separation into different sub-data sets along the instances would amplify this issue. Method (3), i.e. manipulating the algorithm itself, is disregarded, because we aim to construct a general setting in which collaboration partners and other institutions can apply their externally developed algorithm to a bigger picture.

3 Protocol for a Specific Interface: Biopsy - yes or no?

In this chapter a general protocol for the implementation of the proposed specific Digital Twin interface structure is given. This is done through a systematic research analysis of suitable methods. The paragraphs titled "Context" present an overview of each sections findings. The titles of the section's represent the individual steps a Digital win developer has to go through to be able to construct a suitable algorithm for the problem at hand. Additionally, this thesis exemplifies this process by implementing a selection of the introduced methods for the Biopsy Decision. Finally, the problem-specific considerations and implementations are described and explained in the paragraphs titled "Application".

3.1 Data Cleaning and Preprocessing

For the clinical implementation of decision support systems, the ability to generalise the knowledge that is inferred by the algorithm is mandatory [30]. Hence, the data set quality needs to be ensured to avoid biases influencing the model's predictions [49]. First of all, clinical expert knowledge is needed to perform meaningful data preprocessing. Furthermore, the data needs to be specifically refined and cleaned to enable machine learning algorithms to make reliable data-driven decisions [50].

Context: In a medical context, huge challenges for machine learning often arise from the sets of data available for the learning procedure. In contrast to other areas of application for machine learning, such as industry, medical data sets tend to be very small, including only a few hundred instances. Moreover, data is often collected from only one institution, for which reason it might be exposed to biases coming from the respective patient demographics. This phenomenon is called selection bias and leads to a lack of generalisability [51]. To address this issue, efforts for the advancement of open access data-sharing platforms between institutions should be established [49]. Additionally, synthetic cases of a representative wider population could be generated and used for model improvement [30].

The development of new forms of therapy or recommendations for actions is difficult because data is often used in a retrospective way. This means that behaviour that has not been observed so far can also not be learned by the model. The guidelines are therefore contained implicitly in the data set itself and are thus also passed on to the algorithm. Prospective studies, on the other hand, are carried out in a very

limited quantitative framework so that they are not suitable for training a model, but rather for statistical analysis.

A problem associated with supervised learning is that patients which have been classified incorrectly by the doctor will have an impact on the learning process of the algorithm, which means that it learns to mimic the behaviour of the doctor. Such systematic errors could occur in areas in which medically sound knowledge is not available, such as in the transition from benign to malignant tumours.

Those limitations need to be considered when trying to find a suitable match between research questions that need to be solved and available data. When a data set is collected, **Data Cleaning and Preprocessing** needs to be performed before any kind of analysis can be executed. This procedure can basically be summarised through the following steps given by Nielsen [52]:

- *Check for Missing Values* - In general there are two different ways of dealing with missing values, either by dropping or by imputing. Missing values can be non-existing variables or abnormal numbers filtered out through domain knowledge. They can either be missing at random, missing completely at random or missing not at random. Columns with large amounts of missing rates should be dropped. Otherwise imputation with mean, median, random or just a fixed number can be used.
- *Outlier Detection* - Outliers can have huge impact on the generated model, depending on the used score function and model type. Therefore, it is important to understand where the outliers come from and whether they represent real scenarios. Depending on these insights they can be kept or should be removed.
- *Data Scaling* - Data Scaling is needed for facilitating the algorithm learning process. The two main methods are standardisation for Gaussian distributed data or normalisation. Max-min normalisation methods scale between zero to one by calculating the difference between any value and the minimum value, divided by the difference of the maximum and minimum values. Therefore, these normalisation techniques are largely affected by outliers.
- *Data Imbalance Correction* - Balanced data sets between the classes are needed so that the algorithm does not learn any bias from it.
- *Feature Engineering* - This is a wide range of research and often involves a lot of expert knowledge and tedious work. Depending on what kind of model is used and its ability to deal with more complex classification tasks, these steps can be transferred to the algorithm itself.

Application: For the exemplification of the specific DT interface concept on the Biopsy Decision, a publicly available data set of real patients has been chosen. The problem was that a suitable data package was not easy to find, since most of them contained either very few patients (between 100-500) and/or very few parameters.

Finally, we requested the **PLCO (Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial) data set** [53] from the Cancer Data Access System ("CDAS") and obtained a data transfer agreement. The data is the result of a large-scale, randomised study aiming to determine the impact of screening tests. The trial enrolled 76,693 men at 10 U.S. study centres who were randomly assigned to receive either annual screening (n = 38,343) or usual care (n = 38,350).

Details about the merging procedure of the data set from the different files provided by the CDAS can be found in the Appendix B.1. From the data set only patients with an assigned *Gleason Score (GS)* have been extracted, as we want to use it as a label for supervised learning. The proportion of patients in each class is visualised in Table 3.1.

GS	Frequency Count	Proportion of Patients [%]	Malignant	Label
2.0	32	0.4	no	0
3.0	65	0.8	no	0
4.0	266	3.1	no	0
5.0	639	7.4	no	0
6.0	4199	48.9	no	0
7.0	2405	28.0	yes	1
8.0	605	7.0	yes	1
9.0	322	3.7	yes	1
10.0	56	0.7	yes	1

Table 3.1: This table is representing the patient distribution for the *Gleason Score*.

Finally, the target label is chosen in binary form: non-malignant, i.e. *GS* between 2 and 6 and malignant, i.e. *GS* higher or equal 7. In principle, mapping to the *Gleason Score* itself would be possible and meaningful, since it could provide even more information for the doctors. In our case, however, binarisation was necessary to create a balanced data set between the classes,since there are few values with low and high *GS* such as 2, 3 and 10. In the end, we have an approximate proportion of 40% to 60%.

3.2 Data Analysis

In order to perform data analysis, the data has been preprocessed, so that trends and relations can be identified. In the following the word *feature* is referring to the observational parameters obtained from the patient (for further definitions see Table A.1). Several aspects of interest and importance are the:

- feature distributions and data types
- redundancies between features
- correlation strengths of features with the label

Context: Exploratory Data Analysis is an essential step in any research analysis [54]. It deals with the examination and assessment of data for which little is known about its relationships. This approach for data analysis includes graphical (such as histogram, box-plot, scatter-plot) and statistical techniques (such as variance, skewness, correlation) to identify trends and patterns or to verify assumptions [54].

Histograms can be used to observe the data distribution. Even though, finding out the actual feature distribution is a complex task and a field of research on its own, it is sufficient to get sense of whether methods that require Gaussian-distributed data can perform well on this data. In addition, a check for feature relevance can be performed, meaning that variables with zero variance over the whole data set are going to be excluded.

A feature is called redundant if it can be derived from another feature [55]. This means that by measuring the correlation between different features, their degree of redundancy can be estimated. Therefore, methods evaluating the redundancy between features as well as the correlation between the features and the label can be reviewed together. The population correlation coefficient is approximated as closely as possible in inductive statistics [56]. The different methods for estimating this metric depend on the types of data present. Common input parameters can be assigned to the following data types: numerical variables (integer variables - number of positive biopsy cores, floating point variables - blood pressure) and categorical variables (Boolean variables - preexisting conditions, ordinal variables - Gleason score, nominal variables - blood types) [57]. In addition, Yu and Liu [58] have highlighted that the approaches can also be distinguished between dealing with linear or non-linear feature correlations. “Two variables may be related by a nonlinear relationship, such that the relationship is stronger or weaker across the distribution of the variables” [59].

Brownlee [57] summarised which methods can be used in which scenarios. The **Pearson Correlation Coefficient** can be used if input and output features are numerical, follow a normal distribution and show a linear correlation [60]. In the case of non-Gaussian numerical features with a non-linear correlation Spearman’s Correlation Coefficient can be deployed [56]. The variables must be at least ordinal

and monotonically related [56]. Brownlee explained that in the case of one feature being categorical and the other one being numerical, Kendall's Rank Coefficient can be applied to non-linear correlations and ANOVA (ANalysis Of VAriance) with linear correlations. Chi-squared coefficients can be deployed on categorical values and compare the obtained to the expected frequency of a class for a given value [61]. Other powerful methods use the approach of information theory. Entropy measures can be applied to nominal features and continuous features if they have been discretised in advance [62]. The information-theoretical concept of entropy $H(x, y)$ can for example be used to calculate the **Mutual Information** $MI(x, y)$ or **Symmetrical Uncertainty Coefficient** $SU(x, y)$ between feature x and label y [63]. Symmetrical Uncertainty "compensates for information gain′s bias toward features with more values and normalises its values to the range [0, 1] with the value 1 indicating that knowledge of the value of either one completely predicts the value of the other and the value 0 indicating that X and Y are independent" [64].

$$\begin{aligned} H(x) &= -\sum_i p(x_i) \log_2 p(x_i) \\ H(x, y) &= -\sum_{i,j} p(x_i, y_j) \log_2 p(x_i, y_j) \\ MI(x, y) &= H(x) + H(y) - H(x, y) \\ SU(x, y) &= 2\frac{MI(x, y)}{H(x) + H(y)} \end{aligned} \tag{3.1}$$

Application: Originally a data set of 8,589 patients and 56 features was extracted from the PLCO data base. A Table A.3 with an overview of the features, their original name assigned from the PLCO study and a short definition obtained from the provided documentation files, is displayed in the Appendix. Furthermore, a Figure C.4 with the representation of the features through the use of histograms is presented in the Appendix. From this, the feature types and their domains of definition were able to be extracted, allowing for the detection and exclusion of outliers. General information about the feature mean value and standard deviation depending on the label, as well as their rate of missing values is shown in Table A.4 in the Appendix. One can already deduct that some values show different means between the two classes. This might indicate a correlation, but is an insufficient observation on its own.

More sophisticated measures as previously introduced could be used for further analysis. Regarding the overview given by Table A.4 one can see that the data includes a wide range of different data types. As the statistical correlation measures can mostly deal with just a specific combination of data types, they can not be applied to the whole data set straight away. One idea could be to use a categorisation of the numerical features. This approach is often used, but leads to some information loss about the variables. A second idea could include the mapping of pairwise

selected features to their correct correlation measure and an overall representation of the obtained results in one heat map. This might lead to strong misinterpretations as the different measures do not scale the same way. Hence, Mutual Information was used. The probability estimation for the calculation of the feature entropy was specified for categorical values as $p_{i,categorical} = \frac{countsvalue_i}{\sum_k countsvalue_k}$. For numerical values a probability function was fitted to the data and read out for the specific feature values $p_{i,numerical} = pdf(value_i)$. For this purpose, the best-fitting function from a range of standard functions such as Cauchy, exponential, logarithmic, Gamma, Rayleigh and polynomial distribution was identified (visualised in Figure C.5 in the Appendix).

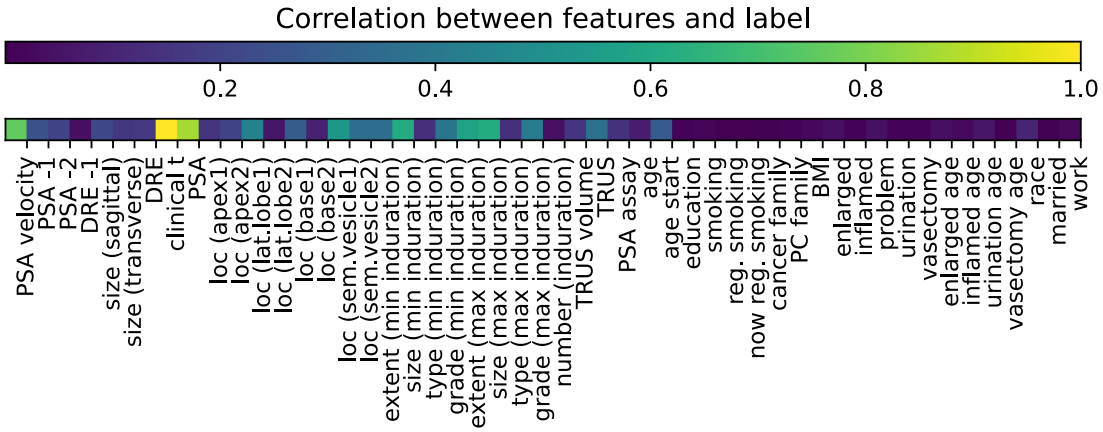

Figure 3.1: The symmetric uncertainty between features and label, normalised through the division by the maximal correlation.

As expected the Symmetric Uncertainty shows a symmetric correlation strength between the features and high values for the self-correlation (see Figure C.6 in the Appendix). Examining the correlation with the label (see Figure 3.1), one can see that parameters such as *PSA* and *clinical t* show strong correlation, as one would expect from their clinical significance. *DRE* shows a surprisingly low correlation strength, especially as other parameters obtained during the DRE procedure such as *size (min and max induration)* represent higher correlation values. Generally, one has to be careful with interpreting the correlation strength, as values that are nearly constant for all patients such as *extent (min and max induration)* also seem to show a strong correlation. It might be advisable to check the value distribution (see Figure C.4 in the Appendix) to avoid an over-interpretation of the importance of such features.

3.3 Machine Learning Models

The late integration method for multi-view data presented in this thesis requires base models to be trained individually before combining their predictions. Therefore, a general overview of the existing machine learning categories is necessary to identify suitable types of algorithms. Furthermore, the performance of an early integration method, using feature vector concatenation and training of single models, is evaluated and used as a baseline comparison for later model performance.

Context: Machine Learning (ML) is a wide field of research with many different approaches and settings aiming to mine previously unknown knowledge from existing data [65]. As the task at hand is a supervised classification, only machine learning algorithms of this type are reviewed. A general way for the categorisation of ML models is provided by the **Bayes' Rule** that comes from the field of probability theory. It describes a way for calculating a conditional probability and can depict how knowledge is inferred by a model M by training on a data set D [66].

$$P(M|D) = \frac{P(D|M) \cdot P(M)}{P(D)} \tag{3.2}$$

$P(M|D)$ is the posterior probability of the model given the data and is usually of interest. The knowledge about the probability of the model is given by $P(M)$. $P(D)$ is called the evidence of the data and serves as a normalisation factor. The last factor $P(D|M)$ is called inverse probability or likelihood and measures how likely the data would be for the model with the current parameters. Generative models (such as Naive Bayes, Bayesian Networks and Density Tree) learn the likelihood and discriminative models (such as Logistic regression, Support Vector Machines, Artificial Neural Networks, k-Nearest Neighbour and Decision Tree) estimate the posterior probability directly. Observations show that the "discriminative performances of state-of-the-art generative models are still far behind discriminative ones" [67].

It is possible to distinguish between stable (such as Naive Bayes, k-Nearest Neighbour and Support Vector Machines) and unstable learners (such as Decision Trees). Ting et al. [68] explain "[..] [U]nstable learners will generate substantially different models when there is a small perturbation on the training data; whereas stable learners generate models that differ little in the context of small perturbations."

Furthermore, one can distinguish between parametric (such as Logistic Regression, Linear Discriminant Analysis, Perceptron and Naive Bayes) and non-parametric (such as k-Nearest Neighbours, Decision Trees and Support Vector Machines) ML algorithms. Algorithms that simplify the data to a set of parameters, assuming the data can be described by a known form, are called parametric machine learning algorithms [69]. Individual models will not be explained in detail, but can be investigated for example in the book by M. Kubat [70].

Application: Ensembles that use individual classifiers based on different classification models make use of the bias which each of this classifier types induces on the inferred knowledge [8]. In order to ensure that this effect can be used, a variety of base models is included. Methods from the following machine learning families are considered: **Logistic Regression (LR), Support Vector Machine (SVM), Artificial Neural Network (NN)/ Multi-Layer Perceptron (MLP), k-Nearest Neighbour (kNN)**.

As mentioned in Section 4.1, the correct treatment of missing values is essential for a successful implementation of machine learning models. There are many model-specific approaches possible (mentioned in the Appendix B.2). Due to the demand for a general approach, model-specific solutions have been neglected. Imputation methods have been excluded, as they might have a huge impact on the prediction accuracy for an individual patient. To deal with this aspect, a **complete data set** ($D_{complete}$) with no missing values was generated from the original data set ($D_{original}$) for the testing of individual models on a concatenated feature vector. In this case, feature concatenation of course was not necessary, but has been simulated as the data contains heterogeneous feature sets (features originate from different institutions).

The aim was to keep as many features and patients as possible at the same time. This is a trade-off, as keeping more features makes it more likely that patients will not have a full set of parameters. The threshold for dropping features with a missing value rate larger than this percentage was set as low as possible by still including many features. By comparing the distribution of the missing value rates it could be observed that just a few of these parameters fall in the range between 10% and 70% and therefore most lie outside these bounds (can be viewed in Figure C.7 in the Appendix). This means that if we were to include all features with a missing values rate of up to a maximum of 20% and were to then add just one more feature, a dramatic decrease in the number of patients would be observed. Nevertheless, to ensure the predictive strength of the feature set, features that correlate strongly with the label should be defined beforehand and then get included. In this regard, we decided to include *PSA*, *clinical t* and *DRE*. After dropping all patients with missing values for these three features, the threshold was set to 20%, as a trade-off between patient and feature number (missing value rate distribution and trade-off between patient and feature number can be seen in Figure C.7 in the Appendix). After fixing the threshold, all patients with incomplete features were dropped as well. $D_{complete}$ includes 539 patients and 24 features: *PSA velocity*, *PSA -1*, *PSA -2*, *DRE -1*, *size (sagittal)*, *size (transverse)*, *DRE*, *clinical t*, *PSA*, *age*, *education*, *smoking*, *PC family*, *BMI*, *urination*, *reg. smoking*, *cancer family*, *enlarged*, *inflamed*, *problem*, *vasectomy*, *race*, *married* and *work*.

The model parameters are internal configuration variables that are estimated from data during the training procedure. Search strategies for fitting parameters have

been implemented from python libraries. As this process depends on the data presented, the parameters will vary and impact the generalisability of the model. A method to find the parameters that yield the best performance on the test set is called **10-fold cross-validation**. Therefore, the data set is divided into 10 units, which are called "folds". In each of the 10 runs, one fold is selected as being the test set and the model is trained on the other nine folds.

Model hyperparameters, as being external configuration variables, cannot be estimated from data and need to be searched for. Hence, a 10-fold cross-validated **Grid Search** over a parameter grid was implemented (analogous to [39]). Due to computational complexity the grid was first chosen with a wide range and sparsity. After identifying a parameter range of interest, the grid was fine-tuned. Grid Search is a brute force approach, prone to miss the best hyperparameter combination through the manually defined hyperparameter search space. Hence, one could consider using more sophisticated methods such as Random Search, Coarse to Fine Search, Bayesian Search, Genetic Algorithm (can be studied in [71]). But as the fine-tuned optimisation of the individual algorithm was not the main aspect of this thesis, Grid Search was sufficient. The following hyperparameter could be determined for the base models over $D_{complete}$:

SVM) regularisation parameter C = 13.9, kernel = radial basis function , kernel coefficient gamma = $1.9{\cdot}10^{-6}$

LR) regularisation parameter C = 0.05, optimiser = Newton-CG, penalty = l2

kNN) number of neighbours = 18, power parameter p for Minkowski metric = 4, weights = distance

MLP) number of hidden layers = 1, number of neurons per hidden layer = 100, learning rate = 0.001, momentum = 0.7 , optimiser = Stochastic Gradient Descent

A collection of available performance measurement tools has been used in combination by Bibault et al. [72] and shown to be able to give a broader insight into how and in which way the performance of presented classifiers differ from each other. The choice for measures of interest also depends on the posed question. For example, the True Positive Rate (TPR) of the algorithm would indicate the diagnostic yield of the model, the False Positive Rate (FPR) on the other hand would indicate how likely it is that an unnecessary biopsy might be performed [73]. Used measures that were selected were analogous to the ones used by Bibault et al. [72]: **Accuracy** = $\frac{\text{true positives}}{\text{total number samples}}$, **Precision** = $\frac{\text{true positives}}{\text{true positive + false positive}}$, **Recall** = $\frac{\text{true positives}}{\text{true positive + false negative}}$, **F1-Score** = harmonic mean of precision and recall, **auROC** = area under curve of true and false positive rate at different thresholds, **prAUC** = area under curve of precision and recall for various thresholds (for further detail see [74]).

For the different selected models the cross-validated auROC performance can be seen in Figure 3.2. The mean and standard deviation of the auROC are reported in the legend of the figure and are visualised through using a Shader Graph for mapping the **Receiver Operating Characteristic (ROC) Curve**. On this measure the SVM performs best, followed by LR and MLP. kNN seems to not be able to learn the correct correlation. In contrast, it assigns the labels in reverse. One option to address this situation would be to reverse its prediction and therefore improve its performance. As this approach is not well elaborated, it was not used.

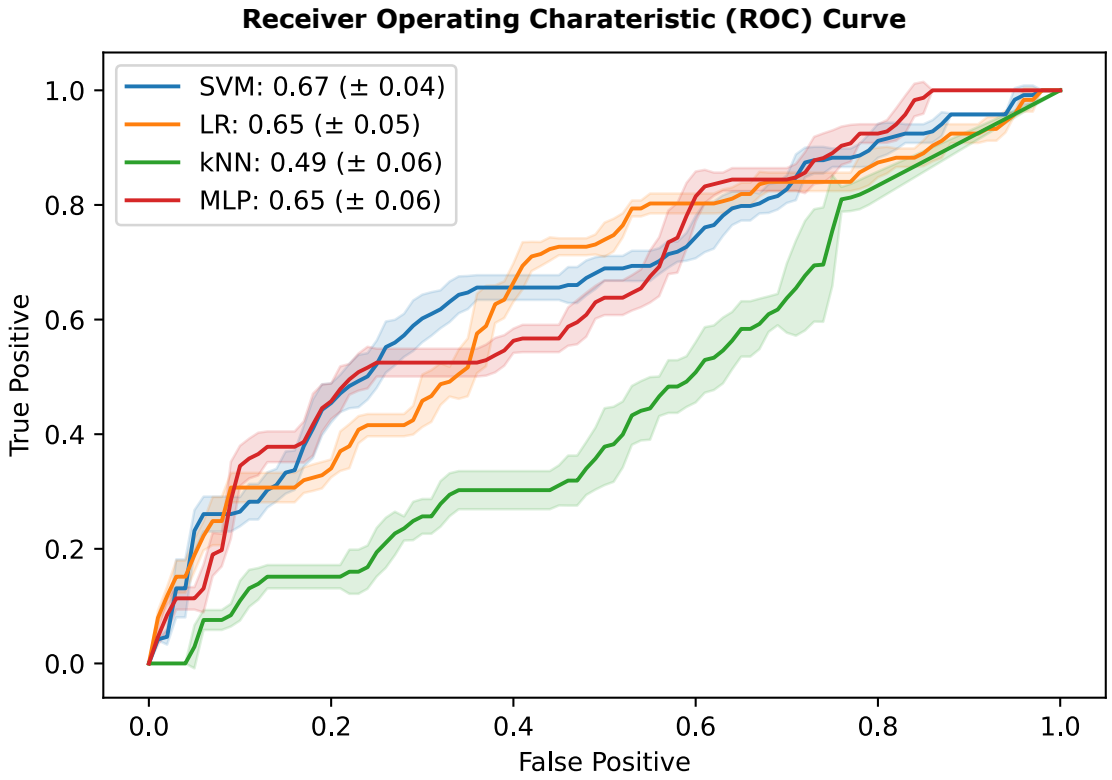

Figure 3.2: A performance analysis of the different machine learning models by observing their cross-validated ROC curves in a Shader Graph. The legend includes the area under Receiver Operating Characteristic Curve (auROC) mean and standard deviation.

Shariat et al. have argued in their review of PC predictive tools: "When judging a new tool, one should compare it with established prediction tools, in order to determine whether the new prediction tool offers advantages over the available alternatives" [37]. Hence, we chose a widely used nomogram, the PBCG Risk Calculator for comparison, as it performs on the same task (biopsy malignant/ non-malignant). A test data set of 50 patients with information about age, prostate-specific antigen, digital rectal exam, African ancestry, first-degree family history and prior negative

biopsy was selected. The corresponding probabilities have been calculated using the online tool. Surprisingly, the auROC did not yield similar results as stated in the publication. Instead of 75.5%, an auROC of 48% was achieved. Other measures returned for a classification threshold of 60%: Accuracy = 0.66, Precision = 0.36, Recall = 0.29, F1-score = 0.32, prAUC = 0.31. Comparing the auROC, the base models outperform the nomogram for this trial. In comparison to the published result, it would be the other way around. To be able to examine the other performance measures, the threshold at which the model responses are binarised [operating point] need to be selected.

Choosing the **Operating Threshold** for a classifier strongly influences its performance. One widely-selected method is to choose the point along the ROC curve that is furthest away from a straight line with slope 1. That point would represent the case in which the ratio between true positive and false positive counts is maximised. Decision goals can be defined in different ways, seeking the maximisation of different quantities. This might be due to the unequal weighing of the benefit of true positive $[B(TP)]$/ negative $[B(TN)]$ diagnosis as well as the costs for false positive $[C(FP)]$/ negative $[C(FN)]$ diagnosis, or the prior probability for a positive $[p_P]$ and negative $[p_N]$ diagnosis might be different. Then the slope S for optimal operation can be selected as $S = \frac{p_N}{p_P} \cdot \frac{B(TN)+C(FP)}{B(TP)+C(FN)}$ [75]. If the decision goal is to maximise the quote for catching positive diagnoses, one can try to obtain an increase in Precision without overly decrease in Recall. As the F1-Score is the harmonic mean of the two quantities, one can try to maximise it. If on the other hand one tries to maximise the overall percentage of correct diagnostic decisions, one can try to optimise with regard to the algorithm accuracy.

As one can infer from Figure 3.2 there might not be a clear and unique option for selecting the optimal operating point. Following measures are of certain importance for the biopsy decision task: the False Negative rate - miss of high-risk patients and the False Positive rate - execution of unnecessary biopsies. Hence, accuracy was considered most interesting. Setting a classification threshold for the single classifier was done by observing the performance scores over a range of thresholds between zero and one. A 10-fold cross validation was performed to be able to estimate the mean scores and their standard deviation, while trying to get a sense of performance stability. It could be observed that the performance score dependence on the threshold varied a lot over different data set configurations (two examples can be viewed in Figure C.8 in the Appendix). The final result therefore depends very much on which decision criteria was used and which score was considered decisive to set the threshold. The threshold was finally set to 0.35. The mean and the standard deviation of the performance scores for each type of base model is summarised within Table 3.2. One can see that Precision, Recall and therefore F1-Score vary a lot during the cross validation and do not show a stable performance for this threshold. On the other hand, the accuracy and auROC are more stable, with the

SVM being the most accurate model, followed by the MLP. The kNN shows overall a bad performance, indicating that the classification task might be too complicated and the decision boundary not suited for this type of machine learning algorithm. The overall observations suggest that the nomogram is not appropriate for this data set. Therefore, the best performing base model was then used as a benchmark for performance comparison.

Scores	SVM (Mean ± Std)	LR (Mean ± Std)	kNN (Mean ± Std)	MLP (Mean ± Std)
Accuracy	0.76 (±0.03)	0.76 (±0.04)	0.60 (±0.03)	0.74 (±0.04)
Precision	0.57 (±0.08)	0.54 (±0.14)	0.26 (±0.06)	0.46 (±0.15)
Recall	0.24 (±0.06)	0.28 (±0.08)	0.34 (±0.10)	0.24 (±0.08)
F1-Score	0.33 (±0.07)	0.37 (±0.10)	0.29 (±0.07)	0.32 (±0.10)
auROC	0.66 (±0.04)	0.65 (±0.05)	0.49 (±0.06)	0.67 (±0.06)
prAUC	0.46 (±0.08)	0.48 (±0.08)	0.26 (±0.04)	0.44 (±0.09)

Table 3.2: Performance scores of individual machine learning models applied to the complete data set.

3.4 Feature Subset Selection for Base Models

As medical data is low in patient number but high in feature dimensions, single classifiers trained with it may tend to include strong biases and exhibit large variance. Skurichina and Duin mentioned the approaches on stabilising decisions through regularisation, noise injection, or combined decision [76]. Likewise, the previously discussed late integration approach can be simulated through an ensemble learning strategy using feature subset selection and a combined decision of parallel base models. For this reason, ensemble learning that uses a combination of the opinion of several voters is proposed as interesting approach for a specific DT interface architecture. The feature spaces over which the voters are trained to perform a certain task can either be selected freely, for which a variety of selection methods exist, or can be externally fixed, for example through data set distribution between institutions. An investigation of different scenarios and methods is carried out following the question whether this procedure might be a generalisable proposal on how to deal with medical data of the mentioned kind.

Context: In general, the following two points can be summarised under Feature Subset Selection (FSS):

- *Feature Selection* for which one specific feature subset with lower dimension is selected depending on some feature weighting method.
- *Feature Space Partitioning* subdivides the original feature space in several subsets.

There are two main types of **Feature Selection (FS)** techniques: supervised and unsupervised [77]. Supervised methods can be further divided into model specific wrapper and model independent filter methods, where the latter is of interest for a general approach [78]. The advantage in efficiency for filter methods, through the separation of the biases coming from the feature selection algorithm and the trained classifiers, has been described by Tang et al. [77]. As the filter approach is based on selecting high scoring or ranking features, it first needs to be discussed how the "goodness" of a feature can be evaluated. As explained by Yu et al., "a feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features" [64]. As summarised by Gheyas and Smith, "frequently used filter methods include t-test [79], chi-square test [61], Wilcoxon MannWhitney test [80], mutual information [81], Pearson correlation coefficients [63] and principal component analysis [82]" [78]. The correlation measures introduced in Section 4.2 can be used to measure the "goodness" of a feature. This approach is generally known as correlation-based feature selection.

Feature Space Partitioning (FSP) is a general base model diversification strategy of ensemble learning [8]. Here, the single subsets can be viewed as a different projection of the training set[83]. Theoretical and empirical results previously showed that independence between the base models is achieved using this method

[84]. Furthermore, ensemble methods combining "multiple models with different features (different explanations) usually perform well because averaging over those stories makes the predictions more robust and accurate" [85]. Turner et al. [86] explicitly mentioned advantages such as the reduction in base model over-fitting, computational complexity, time and correlation between base classifiers. Its usefulness for simple machine learning methods that often struggle with high dimensional data was explained by Skurichina and Duin [87] via the generated relative increase of training instances per parameter.

Type	Feature Weighting	Selection Method
manual	no	Domain knowledge Feature type Feature origin Feature granularity
algorithmic	yes	Redundancy Best scoring Probability

Table 3.3: This table is giving an overview of the different feature subset selection methods discussed for this proposal.

As visualised by Table 3.3 two types of Feature Subset Selections Methods are distinguished: manual and algorithmic selection. **Manual Selection** can be used to generate subsets with regard to impact factors of the Digital Twin environment. Grouping can be done according to:

- *Domain Knowledge:* Analogous to the concept of Mixture of Expert (explained in detail by Yuksel et al. [88]) domain knowledge can be used to find a division of the predictive modelling problem into subtasks. Therefore, a complex task can be broken down by using the divide-and-conquer strategy [89].
- *Feature Type:* This shows a possible way for dealing with heterogeneous data as discussed previously [90].
- *Feature Origin:* This approach can be used if features, coming from different institutions and/or measurements performed on the patient, deliver complementary information. Moreover, it represents a way of treating multi-view data and simulate the situation currently present in many collaborative projects between institutions, where data sharing is not easily possible.
- *Feature Granularity:* Granularity describes the information content of a variable, such as the *PIRADS score* is condensed and the *PSA value* granular [91]. Using these categories hierarchical levels could be set up and individual algorithms could be trained at each level. This procedure attempts to isolate redundant information.

The **Algorithmic FSP** methods use **Feature Weighting** and can select the subsets by using:

- *Redundancy Consideration:* Just the correlation-weighted selection of features that are less correlated compared to the previously selected ones than to the label is allowed. If there are no non-redundant features left, a new subset is started by selecting the feature with the strongest label correlation. Explicit description of the implementation can be found in the publication of Piao et al. [84]. This method will be used in the further analysis.
- *Best Scores:* Feature weights are used as rank indicator. A set of best-performing features are selected, for which the subset size is defined by a threshold.
- *Probability Guidance:* The selection of features is done by using a multinomial distribution with weights as probability measure. For this approach the hyperparameter *subset size* needs to be manually defined. This procedure was presented by Elshrif and Fokoue [48].

For many methods, the subset size has to be provided and needs to be chosen carefully as it has a significant of impact on the ensemble performance. Breiman [92] recommends the use of $d = \sqrt{p}$ for classification tasks, with d being the dimension of the subsets and p the number of features available.

When comparing the quality of the different proposed methods usually the performance of the full ensemble is examined (such as in the paper of Piao et al. [84]). For a generalisable guidance, a meta-model independent measure is sought. Indicators for a good ensemble accuracy are well-performing, as well as diverse and independent base classifiers [84]. This positive correlation has been shown empirically by Dietterich et al. [93]. The statistical independence of the base models needs to be assured with a measurement on the data set and might not be so easy to handle, as the decision correlation may vary a lot with the classification difficulty of the present patient [94].

Diversity measures can be either base model-independent, or base model-dependent. The base-model-independent method, **Feature Subset Relevance (FSR)** will be introduced in more detail, as it will be used within this section, as analogously presented by by Biesiada and Duch [63] to evaluate the goodness of selected feature subsets. They calculated the relevance [FSR] of the feature subset $[X_k]$ (with k features) from the average correlation coefficient within the feature subset $[r_{kk}]$ and the average correlation with the class $[r_{kc}]$ as follows:

$$\mathrm{FSR}(X_k, C) = \frac{k \cdot r_{kc}}{\sqrt{k + (k-1) \cdot r_{kk}}} \tag{3.3}$$

Base model-dependent methods are reviewed by Woźniak et al. [8] and divided up into pairwise (between two base models) or non-pairwise (between base model and entire ensemble) methods (for details see [47]). For the following analysis of base model error diversity measures **Generalisation Diversity (GD)** and **Compound Diversity (CD)** are further considered, analogous to Roli et al. [95]. GD was originally developed by Partidge and Krzanowski, who "argued that maximum diversity is achieved when failure of one classifier is accompanied by correct classification by the other classifiers and minimum diversity occurs when two classifiers fail together" [47]. Therefore, the quantity is calculated using the probability that two randomly selected classifier fail p(both fail) and the probability that one randomly selected classifier fails p(one fails), as follows [96]: $GD = 1 - \frac{p(\text{both fail})}{p(\text{one fails})}$. Giacinto and Roli introduced CD as a pairwise measure based on the compound error probability between two models c_i and c_j [97]: $CD = 1 - p(c_{\text{i,fails}}, c_{\text{j,fails}})$.

Tang, Suganthan and Yao stated „three possible applications of the diversity measures in an ensemble learning algorithm: generating individual classifiers, visualising relevant properties of the ensemble and selecting base classifiers" [47]. By means of their theoretical and experimental analysis they claim that the investigated diversity measure are only suited for the first two purposes. For this reason, they were used as descriptive tools for the selected feature subsets. For the ensemble characterisation in this selection, the mean of CD, as well as the mean of the Pearson Correlation (PC) Coefficient, the GD and FSR were calculated.

Application: Generally, a wide comparison of the different abovementioned methods could be executed. For the sake of manageability, just a selection of them had been used. The assignment of the base models to a data subsets was done using a heuristic optimisation approach. The base model performance was used as indicator. As the overall goal is the improvement of the classification performance over the PLCO data set with the features fixed to the ones included in $D_{complete}$, various approaches using FS are contemplated.

The **Reduction in Feature Space Dimension** using FS might be able to improve the model performance if there are not enough instances available for the training procedure to deal with the noise level. This is a trade-off between gaining information through including a new feature and at the same time increasing the classification complexity. For the structured assessment of the impact of the reduction in the amount of input variables on the model performance, the feature ranks using Symmetric Uncertainty were determined (can be viewed in Table A.5 in the Appendix). Hence, the features were ordered according to their correlation strength represented in Figure 3.1. Even if they were calculated on a smaller data set, no major differences in their correlation strength ranking could be observed. According to this feature ordering, the contributing features for each subset size were chosen. The ML model performance scores were calculated at an Operating Threshold of

0.35.

Maximal accuracy of the models can be observed for different sizes (full analysis is summarised by Table A.6 in the Appendix). The SVM performed best with 76.7 (± 3.1)% for 11 features. Logistic Regression received an accuracy of 75.8 (± 2.7)% for 6 features. k-NN with 67.5 (± 4.7)% and MLP with 74.9 (± 2.9)% showed best results for a set size of 3. Those maximal values can be found surprisingly early, for very few features. Comparing further scores for the SVM model in Figure 3.3 also indicates that after a feature size of four the model does not improve much. This indicates that the first four features *PSA-2*, *clinical t*, *PSA velocity* and *PSA* are the most informative ones. Furthermore, this analysis points out that feature subsets should not be chosen smaller than four. For a stacking approach (further explained in the next section) on this data set, the feature dimension is set to 8. The results will depend on the used combination method and are discussed in the next section. With these results a small improvement in accuracy could already be achieved compared to the performance on the complete data set ($\Delta\text{Accuracy}_{\text{SVM}} = 0.1$, $\Delta\text{Accuracy}_{\text{LR}} = 0.0$, $\Delta\text{Accuracy}_{\text{kNN}} = 0.8$, $\Delta\text{Accuracy}_{\text{MLP}} = 0.1$).

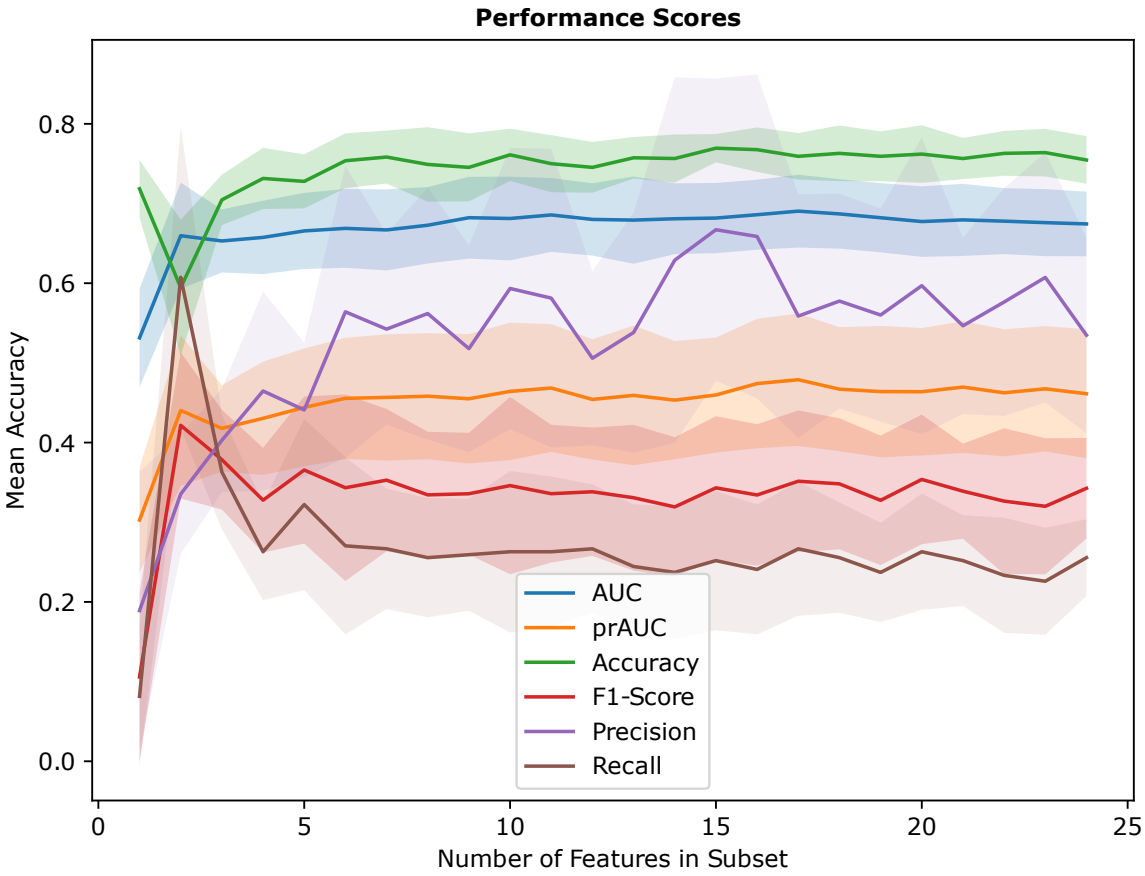

Figure 3.3: A performance analysis of the Support Vector Machine (SVM) over different data sets varying in the number of included features.

In an alternative approach to strive for a better classification performance, an **Increase in Instance Dimension** is pursued. By using FSS, the relative amount of patients per feature is already increased due to reduced feature set size per base model. Additionally, the reduction in feature set size through FSP can be used to increase the absolute patient number, as potentially more patients hold complete data of these fewer values. These patients will be selected for training of the individual base models. This means that each algorithm can identify partially different patients.

Following this approach, both a manual and an algorithmic FSP method are implemented. For the manual selection of the feature subsets, **Missing Value Clustering** attempts to increase the number of patients as much as possible. It can be observed that the data clusters appear to represent the different measurements performed on the patient and do not cluster randomly. For this reason, three informative subsets using *Domain Knowledge* have been selected (can be seen in Figure C.9 in the Appendix), to which the base line has been added, respectively, as one can expect the baseline to be just informative in combination with other features because they show low correlation with the label themselves (see Figure 3.1). The three data sets generated through a Missing Value Clustering are:

Set1) *PSA velocity, PSA -1, PSA -2, DRE -1*
Set2) *DRE, size (sagital), size (transverse)*
Set3) *PSA, age, clinical t*

The baseline features identified from the data set, which are added to the different subsets are: *education, race, married, reg. smoking, cancer family, work, inflamed, vasectomy, enlarged, problem, smoking, PC family, BMI and urination.* As an exception, *inflamed* was dropped from Set 2, as it was lowering the amount of available training instances dramatically.

For each set the number of patients [# Patients] with a complete feature vector is determined and used for the training of the base models. Table 3.4 depicts how much the number of instances could be increased for Set 3. On the other hand, Set 1 and Set 2 have an even smaller number of training objects, as a separate set of instances needs to be kept for the training of the combination method. To exclude class biases, the percentage of non-malignant and malignant [% Malignant] patients is calculated again. A shift towards non-malignant patients can be observed especially for Set 1 and 2, but should be still in a reasonable range. Surprisingly, the FSR for Set 1 is maximal. Even though *PSA velocity* is significantly correlated with *PSA -1* and *PSA -2*, their strong correlation to the label balance out this factor. For the first and second subset SVM has been selected, as it was showing the highest accuracy. For the last set, LR has been selected according to its good performance and additionally, the selection of a different classifier family might be advantageous for the ensemble performance. A complete comparison of the models accuracy and

auROC can be found in Figure C.10 in the Appendix. The error Diversity Measures returned for these three subsets: GD = 0.24, mean(CD) = 0.83, mean(PC) = 0.69.

	# Patients	% Malignant	FSR	Model	Accuracy
Set1	365	28	0.07	SVM	0.72 (±0.04)
Set2	266	22	0.02	SVM	0.78 (±0.04)
Set3	3625	37	0.05	LR	0.62 (±0.02)

Table 3.4: Summary of important quantities for each subset generated through Missing Value Clustering. The quantities include the number of patients [# Patients], percentage of malignant patients [% Malignant], Feature Subset Relevance [FSR], best performing model and its accuracy per selected subset.

The algorithmic FSP method, on the other hand, should form higher-quality subsets. **Redundancy aware FSP** was chosen as a promising approach because of the strong impact the feature redundancy can have on the performance of ensembles generated by FSP as emphasised by Marina Skurichina et al.[94]. For this algorithmic FSP method an end criterion must be chosen. Otherwise, feature subsets might become too small in order to allow a base model to be trained with it, as discussed for Feature Space Dimension Reduction. For example, if there are not many informative features left, the algorithm would tend to form subsets of size one. One option is to define that all the following features are put into one subset as soon as this case occurs. Alternatively, one can chose a minimum subset size. The subsets generated using the second approach with the previously determined minimum subset size of 4, are:

Set1) *PSA -2, inflamed, race, reg. smoking, work*
Set2) *clinical t, education, vasectomy, married*
Set3) *PSA velocity, DRE -1, PSA, PC family*
Set4) *PSA -1, smoking, BMI, cancer family*
Set5) *age, DRE, size (transverse),size (sagittal), urination, problem, enlarged*

All informative quantities regarding these subset are given in Table 3.5. Except for the fifth subset, sets with a higher instance number were able to be generated by using this Redundancy-aware FSP. The percentage of malignant patients is lowered similarly to the manual selection. The choice of the model for each set was performed analogously to the previous procedure. A general overview of the accuracy performance over all models and sets are shown in the box-plot in the Appendix C.11. The achieved Accuracies of the base models are similar to the ones from the manual selection in a range between 0.6 and 0.8. This is surprising, as one would have expected that the subsets generated by using the algorithmic FSP would be better suited for simple machine learning models, as they account for redundancy

within the features. It is all the more interesting to consider the influence of the FSP methods on the ensemble performance using different combination methods. Also the error diversity measures returned similar values as before: GD = 0.34, mean(CD) = 0.82, mean(PC) = 0.61.

	# Patients	% Malignant	FSR	Model	Accuracy
Set1	623	28	0.05	SVM	0.71 (± 0.03)
Set2	7857	40	0.03	SVM	0.60 (± 0.02)
Set3	554	28	0.06	LR	0.72 (± 0.03)
Set4	798	29	0.04	SVM	0.71 (± 0.05)
Set5	291	23	0.02	LR	0.75 (± 0.03)

Table 3.5: Important quantities are summarised for each subset generated through using Feature Weighting and Redundancy aware Selection. The quantities include the number of patients [# Patients], percentage of malignant patients [% Malignant], Feature Subset Relevance [FSR], best performing model and its accuracy per selected subset.

3.5 Combination Methods

> *Diversity and independence are important because the best collective decisions are the product of disagreement and contest, not consensus or compromise.*
>
> — James Surowiecki [98]

Throughout this section, the impact of the inclusion of this "wisdom of crowds", indicated by James Surowiecki via the different feature selection and combination methods onto the ensemble performance, will be analysed. Therefore different ways of achieving one final output through the combination of the opinions of several voters, need to be introduced and further observed. As a generalisable approach is sought, only base model independent solutions are investigated.

Context: Using the terminology introduced by Rokach [90], two main methods for combining base-classifiers outputs can be distinguished: Weighting Methods, useful if the base-classifiers perform the same task and have comparable success; and Meta-Learning Methods, suited to cases in which certain classifiers consistently in-/ correctly classify instances [90]. They can be distinguished from each other by the need for a training procedure at the combination level. In particular, Meta-Learning Methods require training and weighting methods do not. For this reason, Meta-Learning Methods can only be applied if access to the complete data set is available to the algorithm developer. Otherwise, Weighting Methods need to be considered.

Weighting Methods can be further split up into the following categories: label type [soft, hard or ranked]; use of parameter estimates (PE) from base model training [yes or no] and use of weights [yes or no].

Voting Methods are the simplest form for combining of several model outputs, not using any base model training parameters or weights. They have been summrized by the overview paper from Merijn van Erp et al. [99] and are represented in Table 3.6. The most simple methods are **Plurality Voting** for which the class with the most votes wins and Majority Voting for which the class with more than half of the votes wins. Borda Count is a ranked Voting Method. In the Pandemonium Method all voters give their confidence to the estimated most likely class and the class with the highest total confidence wins. For the **Probability Sum** and Product Rule, the voters give their confidence for every class, then these confidences are summed up/ multiplied for each class and the class with the highest value wins. The method comparison experiments of Merijn van Erp et al. [99] yielded best results for Borda Count, Product Rules and Probability Sum already for small ensemble sizes. Rokach [90] mentioned an alternative method called **Entropy Weighting**, that is assigning each classifier a weight that is inversely proportional to the entropy of its classification vector.

Method Type	Label	PE	Weighting	Method
Voting	hard	no	no	Plurality Voting [99] Majority Voting [99]
	ranked	no	no	Borda Count [99]
	soft	no	no	Pandemonium [99] Sum Rule [99] Product Rule [99] Entropy Weighting [90]
Classifier Weighting	soft/hard	no	yes	Performance Weighting [90] DEA Weighting Method [100] Logarithmic Opinion Pool [101]
Training Parameter	hard	yes	no	Naive Bayes [90]
	soft	yes	no	Decision Templates [90] Dempster-Shafer [100] Logarithmic Opinion Pool [101]

Table 3.6: This table gives an overview of the different Weighting Methods discussed for the combination of several base models. The abbreviation PE is used for parameter estimation.

There are methods that take into account scenarios such as the unequal performance of the base models, the non-uniform representation of the label classes and the unequal likelihood of instances. They do so by assigning weights respectively to balance out the just mentioned factors. One simple extension of the Voting Methods can be done by using class weights, multiplied with the base model predictions during the voting procedure to cancel out unequal probabilities in the original data [102]. Finding weights that result in an overall increase in performance compared to the individual base models or simple Voting Methods is a challenge. The weights can for example represent the performance of each base model (single measures such as accuracy or several measures from Data Envelopment Analysis). Alternatively, the **Logarithmic Opinion Pool** calculates the class that maximises the exponential function from the sum of logarithmic probabilities assigned by each classifier and weighted by a classifier dependent factor [101].

For combination methods that use information generated through the training process of the base models for a hard classification problem, one can use Naive Bayes and Behaviour-Knowledge Space. Naive Bayes utilises the confusion matrices of the individual base classifiers [103]. Behaviour-Knowledge Space generates lookup-tables by going through a data set and examining the prediction patterns of the base models. As it is important to know how the classifiers vote for one specific instance, feature subspaces would disturb this method and can therefore not be used for the proposed ensemble setup. For soft classification **Decision Template**,

using similarity of current decision profiles to decision templates generated from the classifier responses to the training set, or **Dempster-Shafer**, which maximises the belief function instead of just using similarity measures such as decision templates, are suited [103].

"**Meta-Learning** [104] is loosely defined as the learning of meta-knowledge about learned knowledge" [105]. In this context it describes how a meta model can learn information - i.e. finding the correct final system response from the predictions of the base models. Those can generally be divided into global (defined over full feature space) or local (defined of feature subset) base models. The different concepts reviewed in the literature can be split up into the following cases:

- Combining global models of same type - Bagging, Boosting, Random Forest and Random Subspace (regarding instance selection)
- Combining global models of different types - Stacking
- Combining a selection of global models - Grading [90]
- Combining local models of different types - Mixture of Experts [88]

Stacking is a global ensemble learning method, training a meta algorithm on the base models' outputs and typically not providing access to their input features [106]. The idea is to use the individual base model biases to generate an ensemble with uncorrelated errors [102]. However, Džeroski and Ženko [107] showed with their empirical analysis that several state-of-the-art stacking methods with heterogeneous classifiers showed best comparable performance as the best base classifier. Still Stacking off models trained on the reduced feature set size is considered in this thesis to be a promising method for the performance improvement compared to the individual models trained on the full feature dimension.

As obvious from previous discussions, the usage of local models is more suitable for the developed Digital Twin architecture proposal. Coming from the concept of Weighting Methods, the crucial distinguishing feature of **Mixture of Experts** is that the weights assigned to each experts are assigned depending on the model input parameters by using a gating function [90]. The gate is trained on mapping the base model input features onto the base model outputs. Neural Networks are often used for this task.

Meta models can be trained either only on the predictions made by the base models or combined with the input parameters. Alternatively, Džeroski and Ženko [107] achieved a performance improvement using a dimension extension of the meta-level features to $N \cdot (2m+1)$ with N classifiers and m classes. To the probability response of each classifier for each class they added the probability responses multiplied by the maximum probability returned by the classifiers, as well as the entropy of the probabilities returned by the classifiers over the different classes. Moreover, the meta models can be trained either directly together with the base model or after completing the base model training procedure on a separate holdout training set.

Application: In order to ensure that the comparison drawn between the different combination methods covers a wide range of real life scenarios and applications, methods from each type have been selected. However, for simplicity only methods using soft labels have been used. Of the Voting Methods the Entropy Weighting and the Summation Method are used. Accuracy Weighted Plurality Voting as well as Logarithmic Opinion Pool (LOP) are selected from the Classifier Weighting Methods. Decision Templates and Dempster-Shafer are used as representatives of the more elaborated methods that use information from the base model training procedure. Logarithmic Opinion Pool was modified in such a way, that it was not using fixed weights corresponding to the base model accuracy, but was able to learn them (LOP trained). Using simple meta-models such as Logistic Regression (LR) might allow an interpretation of the final decision-making process. In this case, the meta model input was once chosen to be only the base model outputs (LR 1) and once additionally the base model inputs (LR 2). However, more complex base models such as Neural Networks could also have been taken into consideration. The implementation of a one-layer Neural Network analogously to the one presented by Wozniak and Zmyslony [108] was performed. Additionally, a Mixture of Experts using a Neural Network with one hidden layer (activation function = Relu) as gating function was considered.

FSS Method	Performance Measure	Weighting Method		Meta-Learning Method	
		Method	Value	Model	Value
Stacking	auROC	LOP	0.73	Linear Fuzer	0.71
	Acc	Weighted Average	0.75	2D NN	0.77
Missing Value Clustering	auROC	Voting	0.70	LOP trained	0.69
	Acc	Prob. Sum	0.73	2D NN	0.71
	Acc	Weighted Average	0.73		
	Acc	Entropy Weighting	0.73		
Redundancy aware FSP	auROC	Prob. Sum	0.71	LOP trained	0.73
	auROC	Weighted Average	0.71	LOP trained	0.75
	Acc	Weighted Average	0.72		

Table 3.7: This table is summarises the best performing combinations achieved by varying Feature Subset Selection Methods (FSS) or the way of combining the base model outputs through either Weighting Methods or Meta-Learning Methods. The performance measures accuracy (Acc) and area under Receiver Operating Curve (auROC) are specified. The standard deviation is not shown, as it is the same for all values, namely ± 0.02. The overview of all the other measures and methods can be found in Table A.7 and Table A.8 in the Appendix.

For each abovementioned method, the performance scores over 10-fold cross validation were measured and recorded (can be viewed in Table A.7 and Table A.8

in the Appendix). The most important information regarding the best scores in auROC and accuracy is summarised in Table 3.7.

The performance of Stacking while using the Weighting Methods and Meta-Learning Methods can be seen in Figure 3.4. As Stacking is usually performed using meta models, it is surprising that the overall best auROC of 0.73 (±0.02) was achieved by LOP. On the other hand for accuracy, the 2D-NN was best by 0.77 (±0.02) (see Table 3.7). It needs to be mentioned that Decision Templates and Dempster-Shafer have not been able to show good results for any FSS method, as one can see from Figure 3.4. With regard to the statement of Džeroski and Ženko [107] the combination methods are further compared to the best base model performance. Here, the SVM achieved an accuracy of 0.75 on a feature set size of 8, which we used for stacking. For the weighting methods, Weighted Average scored the same value. Stacking by using Mixture of Experts with a gating function of a 2D-NN was able to receive an accuracy of 0.77. To be fair, one should actually compare Stacking to the maximal possible accuracy through feature dimension reduction from the complete data set, which would be 0.77 (±0.03) for SVM. Hence, no improvement could be generated.

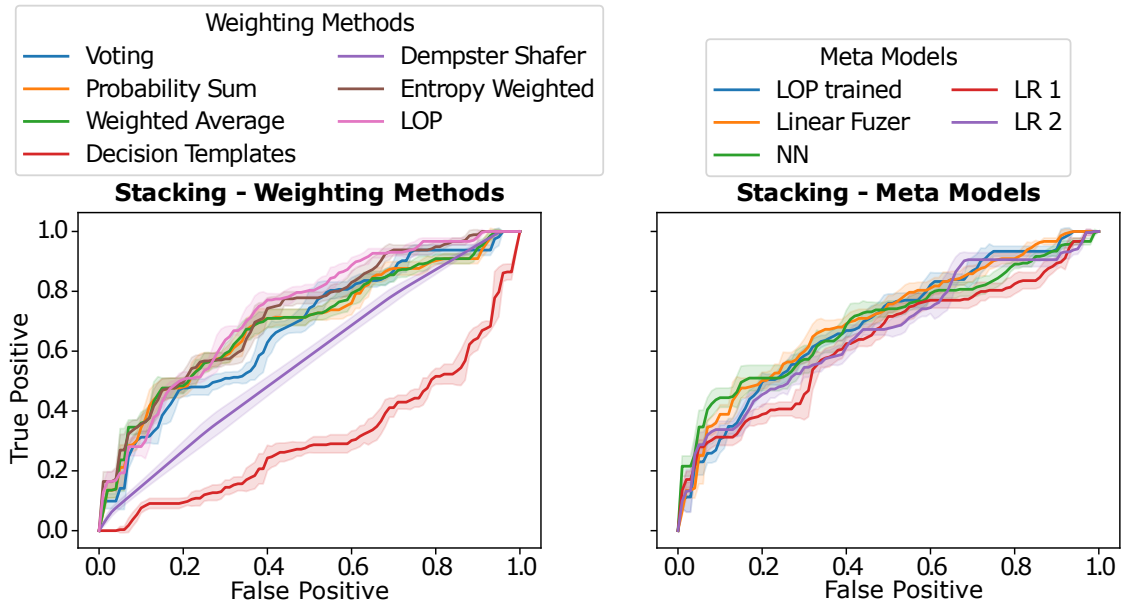

Figure 3.4: Plotting receiver operating curve for performance visualisation for weighting methods and meta models for stacking.The methods and models are further explained in the text. Abbreviations are used for the Logarithmic Opinion Pool (LOP), with weights assigned to the base model accuracy as well as its trained version (LOP trained). Logistic Regression is either trained only on the base model output (LR 1) or additionally on the base model input (LR 2).

For the manually selected feature subsets using Missing Value Clustering, the Weighting Methods even slightly outperformed the Meta-Learning Methods with Probability Sum, Weighted Average and Entropy Weighting showing a maximal accuracy of 0.73 (±0.02) and Voting presenting a auROC of 0.70 (±0.02) (see Figure 3.5). These might be the reason as one base model is outperforming the others, with SVM having an accuracy of 0.78 (±0.04). Therefore, mainly following the estimation of this voter would already show a good performance. Nevertheless, the combination methods try to find a compromise between the base model statements, as expected, but in this case with the result of losing some accuracy.

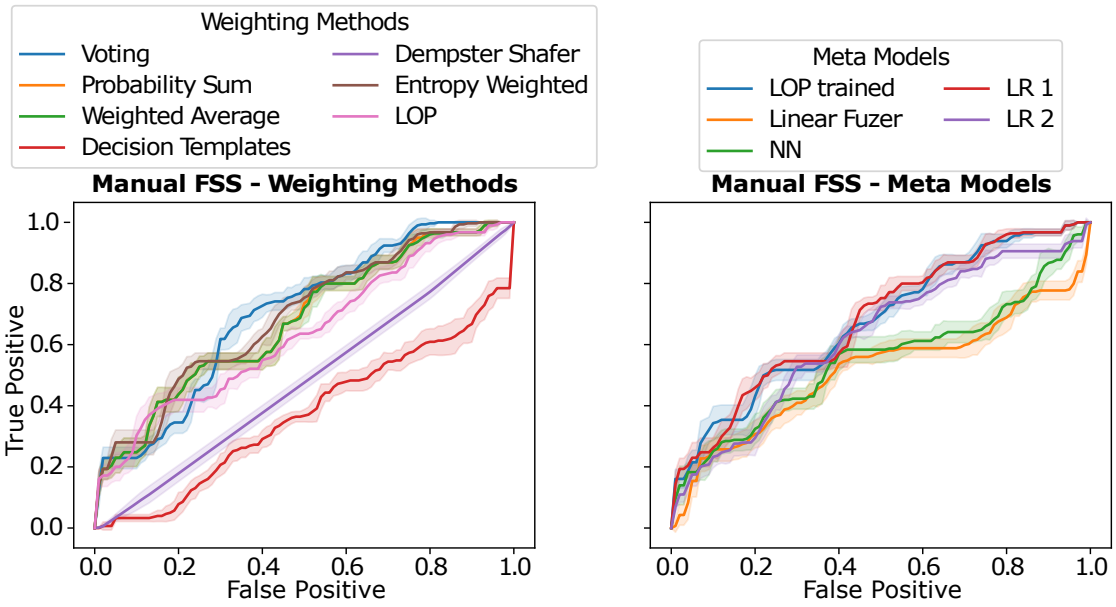

Figure 3.5: Plotting receiver operating curve for performance visualisation for weighting methods and meta models on manually selected feature sets. The methods and models are further explained in the text. Abbreviations are used for the Logarithmic Opinion Pool (LOP), with weights assigned to the base model accuracy as well as its trained version (LOP trained). Logistic Regression is either trained only on the base model output (LR 1) or additionally on the base model input (LR 2).

Finally for the Redundancy-aware FSP, the Meta-Learning Methods outperformed the Weighting Methods with LOP Trained scoring an accuracy of 0.75 (±0.02) and auROC of 0.73 (±0.02) (see Figure 3.6). This is at least performing as well as the best base model on Set 5, with LR achieving an accuracy of 0.75 (±0.03).

We aimed to answer the question of whether the performance of ensemble methods outperforms single classifiers on small data sets with a large feature space dimension.

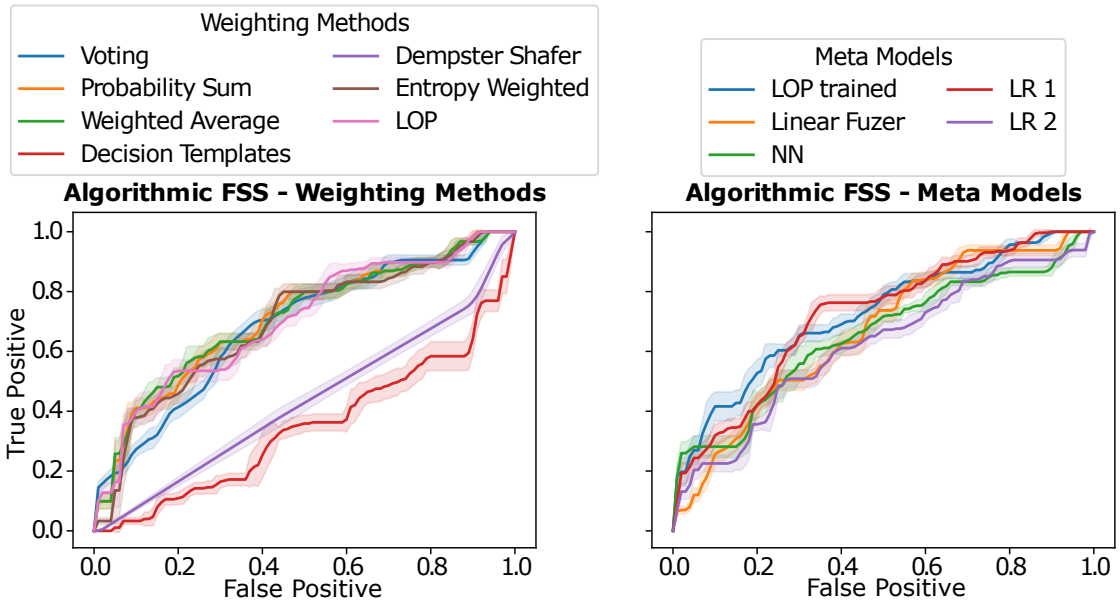

Figure 3.6: Plotting receiver operating curve for performance visualisation for weighting methods and meta models on algorithmic selected feature sets. The methods and models are further explained in the text. Abbreviations are used for the Logarithmic Opinion Pool (LOP), with weights assigned to the base model accuracy as well as its trained version (LOP trained). Logistic Regression is either trained only on the base model output (LR 1) or additionally on the base model input (LR 2).

The following trends could be observed:

- A performance improvement through Reduction in Feature Space Dimension for individual classifiers was shown.
- Reduction in Feature Space Dimension for individual classifiers could hold the same results as Stacking of individual classifiers for a fixed feature number.
- An improvement of the performance of the combination methods could be achieved through selecting meaningful subsets (Redundancy-aware FSP), even if Missing Value Clustering was generating the best performing base model.
- A clear reduction in the variance of the performance over 10-fold cross validation could be observed for the combination of several model outputs, as intended.
- There were no improvements from Stacking to approaches to Increase in Instance Dimension (Missing Value Clustering and Redundancy-aware FSP) observable. This might be due to the lack of suitable data to show the advantages of the FSP methods.
- FSP was shown to be a good way of dealing with data being presented in a distributed fashion. It accomplished nearly the same results in performance as would simple ML algorithms presented to the full data set.

As quite a few results obtained from this study contradict claims about the strengths and usefulness of FSS, experimental methodologies might need to be further adjusted. Nonetheless, the proposed concept for a specific Digital Twin architecture is now set up and ready to be studied in more detail. Impact factors, which should be investigated in more detail are the distributions of parameters, the existence of different forms of bias, the correlation strengths between features and the percentage of instance-feature ratio. By generating synthetic data that represents these properties accordingly, a systematic study could be executed. This would enable a good understanding of the limits and suitable areas of application of the different combination methods. Additionally, more effort could be spent on the optimisation of the algorithms, as until now mainly simple methods have been used for fixing hyperparameters. Different base models and combination methods could be tested as well. For now, the scope of reviewed algorithms was limited due to time constraints.

3.6 Inclusion of Evidence

The possible scope of including prior knowledge was described by Bundage to "overcome problems of limited data and help to meet certain restrictions, such as biological processes or laws and guidelines, which is important for trustworthy AI" [109]. For this purpose, a distillation of available knowledge in the clinical context must first be carried out. Secondly, possibilities of including knowledge into the ensemble-based Digital Twin need to be reviewed, so that usable representation formats for our DT can be identified.

Context: Sources of knowledge have already been described in Section 2.1 as being systematic reviews, meta-analysis or clinical practice guidelines. The current way of including knowledge in form of nomograms has also been explained. It has been shown that those methods struggle with enabling personalised health care and good performance. To be able to think about the inclusion of prior knowledge into individual machine learning models, possible sources of knowledge and their representations need to be discussed. Laura von Rueden et al. [110] presented a review on this so-called **Informed Machine Learning**. Knowledge, defined as being validated and useful information, can have different sources and was subdivided into scientific knowledge (natural science, engineering), world knowledge (Vision, Linguistics, Semantics) and Expert Knowledge (Intuition, less formal) (division taken from [110] Figure 2). Depending on the source type the knowledge can be formalised by using different representation methods, these being algebraic equations, differential equations, simulation results, spatial invariances, logic rules, knowledge graphs, probabilistic relations or human feedback. Finally, the implementation into a machine learning pipeline can be either achieved through the generation of training data, or the construction of a hypothesis set, the adaption of the learning algorithm, or the validation of the final hypothesis. Regarding the clinical context, we have already discussed the sources of knowledge and can now try to find a representation of them. First of all, knowledge coming from scientific research can be represented trough nomograms including algebraic equations or guidelines giving logic rules to the clinicians. Obviously, another important source is expert knowledge, which can be represented as human feedback to the system. Knowledge graphs about the relations between the used parameters can come from any of the two sources.

Depending on the type of knowledge available for the task at hand, approaches for their implementation can "range from methods that strictly enforce physical consistency in data science models (e.g., while designing model architecture or specifying theory-based constraints) to methods that allow a relaxed usage of scientific knowledge where our scientific understanding is weak (e.g., as priors or regularisation terms)" [111]. The publication presented by Karpatne et al. described two different ways of including scientific knowledge in the design of data models [111]. One way is through specifying the model response in a theory-guided way. For example in linear models, the Gumbel distribution can be selected as a response function

to account for very rare events. This is a process which can be integrated into the individual base models. The second approach is the choice of an appropriate model architecture, in which the researcher has a high degree of freedom. These design considerations might be informed by physical knowledge. They presented two promising directions using scientific knowledge while constructing ANN models: 1) by using a modular design, 2) by specifying the connections among the nodes in a physically consistent manner. For the modular design, domain knowledge can be used to divide the problem into sub-problems, from which every sub-problem can be learned on a different ANN, whose inputs and outputs are connected with each other in accordance with the process of the sub-problems (serial connection). Even though we are using a parallel modular design, the previous chain of arguments can nevertheless be transferred to our method.

Concept: "[I]ntegrating knowledge into machine learning is common, e.g. through labelling or feature engineering" [110]. Throughout this Chapter, we have shown how knowledge can be used to construct meaningful feature subsets. Furthermore, we showed how crucial it is for the data preprocessing step to be able to extract high quality data and to avoid including any bias.

Given the classification of prior knowledge in machine learning by von Rueden et al. [110] guidelines be described as logic rules. Hence, an important part of the Digital Twin in the clinical context is the inclusion of clinical guidelines as a framework. This helps to prevent the algorithm from suggesting recommendations for actions that lie outside the doctor's permitted scope of action. As already mentioned, the Digital Twin should not create new guidelines or even replace the existing ones, but should make recommendations for decisions in areas in which there is still room for manoeuvre or in which guidelines do not provide any concrete instructions for action. Since the guidelines are not quantified, their implementation is another area of research. Not all recommendations for action have to be understood as a hard limit. Basically, a distinction can be made between different recommendation grades [A - must, B - should, 0 - can], evidence grades [1++, 1+, 1-, 2++, 2+, 2-, 3, 4] and consistency levels [strong consensus $>$ 95%, consensus $>$ 75%95%, majority approval 50%75%, dissent $<$ 50%] (for further detail see [112]).

A possible proposal for Informed Machine Learning through the validation of the final hypothesis stated by the algorithm could be the visualisation of the Digital Twin's recommendations for action together with the presentation of the existing guidelines. The strength of the recommendation of the algorithm could be seen analogously to the degree of recommendation of the guidelines and may be presented in the form of a traffic light system [must - green, should - yellow, can - orange]. The certainty behind this recommendation, analogous to the level of evidence of the guidelines, should be visible for every recommendation for action. The next section will deal with the challenge of estimating the certainty of the DT recommendation.

The inclusion of previously existing Decision Support Systems is questionable. At first sight, implementation as a base module would be conceivable. In this way one could, for example, try to increase the doctors' acceptance of the new concept. Through an interactive examination of the Digital Twin, the physician could investigate how the addition of further basic modules and parameters influences the decision recommendation. However, the question is: Is it smart to use an algorithm as a base module for which one knows that it does not achieve very good performance because it is based, e.g., on a very simple regression model? Alternatively, it would be possible to make it available to the doctor as a comparison, analogous to the guidelines. Within this thesis, it was shown that the Digital Twin could achieve better performance than the nomogram.

If the statements for an individual patient from nomogram and DT contradict each other, it is unclear which algorithm is correct. It is therefore necessary to clarify how the doctor should deal with such situations. In order to be able to make a decision, the clinician would need to have a very detailed decision analysis, for example using methods of interpretation, which will be discussed in the next Section 4.7. On their basis, the doctor must be able to understand why the Digital Twin has decided differently for the current patient and a deviation from the standard procedure inferred by the currently used support system should be favoured. If, in return, the statements match and the doctor is not in the dilemma of deciding which algorithm to trust, the main improvement generated by the DT would come from the increase in interpretability and reliability of made decisions.

3.7 Interpretability and Personalisation

The goal of science is to gain knowledge. With the methods used in machine learning the general knowledge extraction processes undergo changes, where the „model itself becomes the source of knowledge instead of the data“ [60]. To be able to extract this knowledge, interpretability methods can be applied to otherwise black box machine learning models that capture this information. Likewise, Corral-Acero et al. [30] mentioned a lack of translation from research into the clinic due to a lack of clinical interpretability for the Digital Twin technologies in cardiology. This is due to a deficit of external validation and potentially obscure model failures. In general one can say, that a model providing explanations would be more useful than one just providing predictions.

Context: Machine learning models can be distinguished according to how interpretable they are. One can range them from interpretable models (such as Linear Regression, Logistic Regression, GLM, GAM, Decision Tree, Decision Rules, RuleFit, Naive Bayes Classifier, K-Nearest Neighbours), which are understandable in human terms, to explainable models (such as Random Forest, Neural Network, Convolutional Neural Network), that are too complicated to be understood without external methods. That means, “the less interpretable a model the harder it is to learn from it” [113]. On the other hand, more complex systems that can deal with less-guided learning tasks also have a higher chance of teaching us something new. This can be seen as a trade-off between intrinsic interpretability and accuracy. Understanding intrinsic interpretability requires model-specific knowledge about the learning procedure. For explainable models on the other hand model-agnostic methods can be used for the explanation of its predictions. Because for the Digital Twin concept a general approach is sought, only model-agnostic methods will be further reviewed. They can be split up into global model-agnostic methods (methods describing how features affect the prediction on average) and local model-agnostic methods (methods aiming to explain individual predictions).

In his book, Christoph Molnar [114] gives an overview of the currently-used methods and some short explanations. **Global model-agnostic methods**, which are useful to observe general mechanisms, such as partial dependence plots, accumulated local effect plots and global surrogate models get explained. **Local model-agnostic methods** are good for explaining individual predictions. Regarding these methods, he lists the following selection:

- Individual Conditional Expectation Curve (ICE), which describes the impact of a single feature on the model prediction.
- Local Surrogate Model (LIME), which replaces the complex model by a simple, intrinsically interpretable surrogate model.
- Scoped rules (anchors), which list the feature values that keep the model prediction at a certain statement.

- Counterfactual Explanations, which list the feature values that need to be changed to receive a desired model estimation.
- Shapley Values and SHAP Force plots, that represent the impact strength of each feature to the obtained prediction.

Concept: Global model-agnostic methods seem to be useful for an understanding of the overall trends and correlations between the features and label. It helps to answer questions such as: How was the model prediction affected by feature joint effect? - This can be measured with H-statistic or SHAP-interaction value and visualised with 2D-PDP. How robust is the model performance to uncertainties in the features? - This can be measured with permutation feature importance. How robust is the model prediction to uncertainties in the features? - This can be measured with PDP-based feature importance measure or SHAP-feature importance and visualised with ALE. One could think of different ways to use this information about the base models as a further input to the meta model.

But as the personalised conception of the Digital Twin concept is one of its greatest strengths, local model-agnostic methods are further reviewed. A **Local Accuracy** and **Consistency** measure was sought, answering the question: How was the performance for similar patients? For this, similar patients could be selected for example by comparing Shapley or feature values. A method using k-NN regarding the patient features were selected. A local neighbourhood around the patient was defined by using the maximal and minimal values for each feature represented within the k nearest neighbours. Within this region feature combinations was sampled uniformly and randomly. Over those synthetic patients, the algorithm Consistency was calculated. This was done by measuring how often these candidates were assigned to the same class. The base and meta model accuracy was moreover calculated over the k nearest neighbours.

Another possible factor influencing the algorithm performance might be the training data density around the patient. To estimate how well a single data point can be represented by the general data distribution, the k-NN approach was further developed. First of all, the mean distance and the distance quantiles have been extracted from a histogram showing the distances between the data points. Finally, the density is approximated by the distance to the nearest neighbour scaled according to the distance distribution in training data. From the two just described measures one could think of constructing a confidence score proportional to the Local Accuracy and Consistency, scaled by the density estimate. Certainly, the approaches just-described can be further expanded and improved. However, this type of implementation was initially intended to provide a general insight into possible perspectives on the realisation of algorithm interpretability.

Especially **SHAP Force Plots** could be very useful in healthcare by visualising the impact strengths of the different input parameters on the final outcome. The most promising expectations from the Digital Twin concept can be found in its ability of testing decisions in a simulated environment [115]. Therefore, an evaluation of the decision can be made without even having to perform the action in the real world [6]. One possible approach could be the usage of Counterfactual Explanations to tell the doctor how initial conditions would need to be changed to receive a better outcome.

For the visualisation of how these interpretability measures can be used, a possible implementation is shown in Figure 3.7. An example case, the clinical patient journey of Max Müller is reviewed. The observations one has about this patient are enumerated. The right box is including features, that are considered especially important by the guidelines and along which the further treatment of patients is organised. As Max has a PSA value higher then the threshold of 4 ng/ml, the guidelines would suggest a biopsy. Additionally, the complete selection of information could now be passed to the Digital Twin, which was initialised with an algorithmic feature subset selection method (Algorithmic FSS, explicitly Redundancy-aware FSP) considering feature redundancy and a Logarithmic Opinion Pool (LOP) for base model combination. The ensemble returns a probability of 0.27 of Max having a malignant tumour. As the threshold is placed at 0.35, the DT would not suggest taking a biopsy. Because the clinician is in the delicate position of having to decide what action should be taken, he needs to understand the reasons for the DT suggestion and estimate how trustworthy this recommendation is.

The Local Accuracy and Consistency of the ensemble as well as of the base models are presented by the DT in Figure 3.7. One can see that the ensemble learned to not weight the opinion of base model two very heavily, as its predictions differ a lot from the final ensemble prediction. This is beneficial, as we can see that the Local Accuracy of this model is very low. Except for base model three the Consistency in the patient neighbourhood is very high. In the SHAP Force Plot one can now see in more detail how strongly the individual parameters were voting for (positive - red) or against (negative - blue) Max exhibiting a malignant prostate tumour. This plot explains why the Digital Twin was shifting its prediction from the overall mean probability for this class to occur. As one can infer, evidence-based meaningful parameters such as the patient *age*, *clinical t* and *PC family* have been considered by the algorithm as well. Still, the physician decides to perform a biopsy and a GS of 6 is determined for Max. This means that for this example an unnecessary biopsy could have been prevented. In any case, the difficulty of the decisions made by the clinician should not be underestimated because they carry a lot of responsibility. Therefore, this situation is very hard to handle, even if the Digital Twin performance is very high. A systematic inclusion into or combination with the guidelines would be needed.

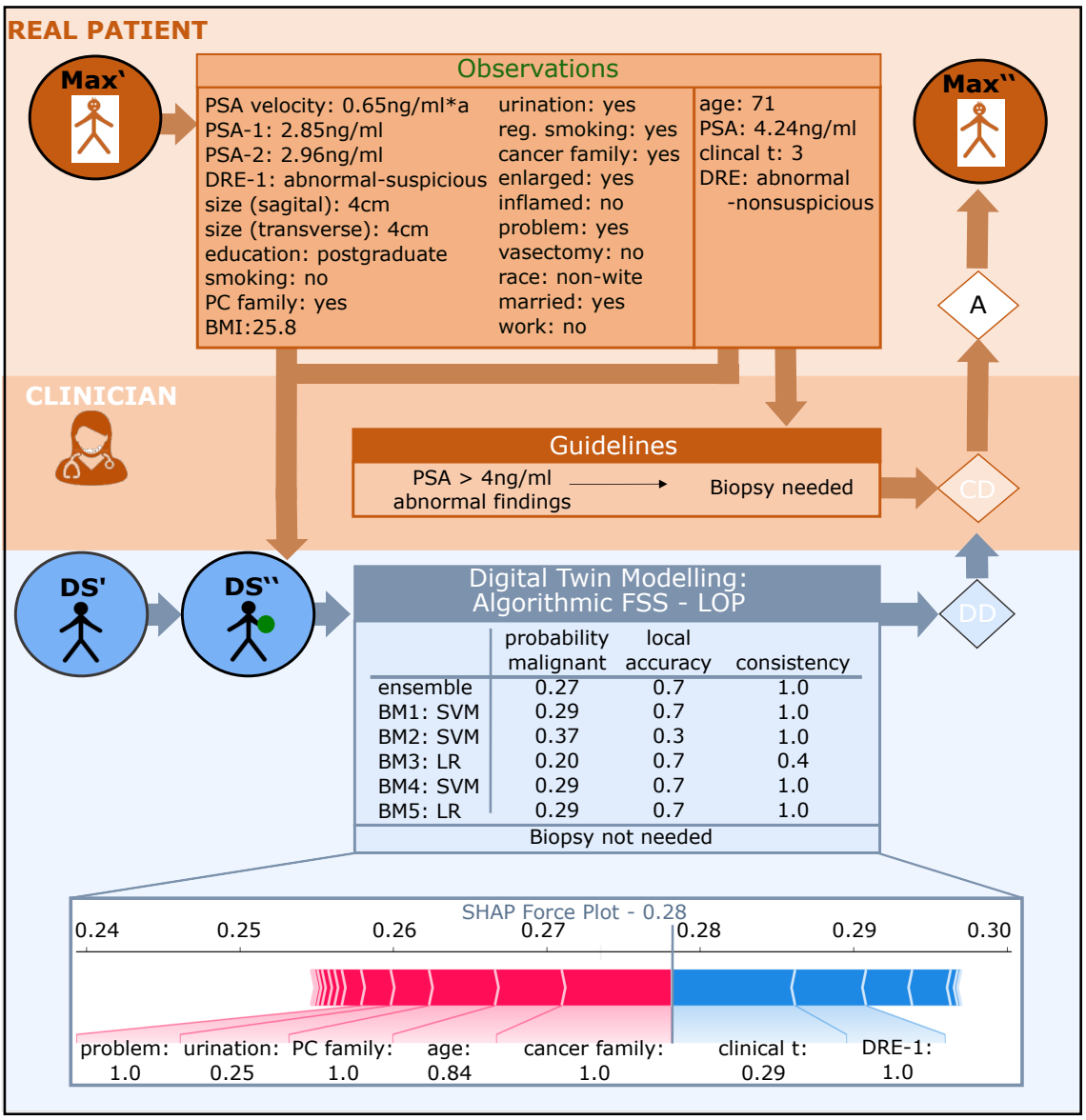

Figure 3.7: The interaction with a Digital Twin during clinical decision-making process is exemplified on the patient Max Müller. A possible integration of interpretability methods is presented. The Digital Twin was initialised for this decision task with an algorithmic feature subset selection method (Algorithmic FSS) and a Logarithmic Opinion Pool (LOP) for base model combination.

In the context of applied machine leaning, there is a second type of interpretability which should be mentioned here as well. This can be understood as the ability of the algorithm to enable the doctor to develop his own intuition for how the algorithm works. The proposed approach, using a modular structure, introduces exactly this kind of ability to construct an algorithm capable of allowing the physician to develop an intuition about its decision-making process. This could be realised by enabling

the doctor to freely select and deselect base models that contribute to the decision. Accordingly, the combination method would have to be trained and implemented for all different combinations of existing base modules. As a result, the doctor would on the one hand be able to develop a feeling for the impact of the base modules and their input parameters on the decision-making process. In addition, the physician would be able to express his distrust in a parameter or base module.

4 Achievements and Potentials of the Proposal

Did the proposed architecture tackle current challenges and achieve posed requirements?

As we have seen in the previous chapter the approach of using an ensemble-based algorithm as a specific interface structure of the Digital Twin entails many advantages and can overcome many of the challenges placed on a Digital Twin in healthcare. Many of them are related to the circumstances under which the data may be collected and accessed, such as:

- *Big data and distributed data sources:* As more and more data is generated, the computational power for processing and storing this data needs to be increased. One solution for the storage of big data is to diversify the storage location. Accordingly algorithms need to get access to these different locations or need to be trained on-site. This is enabled through the modular approach, allowing a local training of base models and a final combination of their output through a meta model.
- *Large feature space dimension:* As introduced in this thesis, the same approach can be helpful to deal with settings in which support system developers are dealing with small data sets that span a large feature space. Breaking down the dimensions or splitting up the dimensions increases the relative amount of patients and therefore can be advantageous for the training of machine learning algorithms.
- *Heterogeneous Features:* Medical Decisions are made on the basis of a variety of data, including different data formats. Therefore, the combination of this data for medical decision support systems is not straightforward, as algorithms usually process only one class of data (text, image, numbers, etc.). By using an ensemble learning approach to combine the predictions made on different parts of the data, each algorithm in the ensemble can process its corresponding data type.
- *Adaptability/ Upgrading:* The inclusion of new decision support tools and practices is a tedious and time-consuming process. The modular approach makes the implementation of newly developed algorithms easier for support system developers because they can use the existing infrastructure. This also facilitates the usage by the doctors, because they do not always have to adapt to new software, as the interface does not change much. Moreover, this enables the improvement of the system by upgrading its building block once in a while.

- *Privacy and IP issues:* Analogously this setting is needed if collaborative projects struggle with privacy issues or IP security and want to keep the collected data and generated algorithms in place [105]. The distributed data paradigm is strongly connected with the big data analysis problem [8]. This situation should not be confused with Federated Learning in which one algorithm is chaired and trained with multiple decentralised data sets.

Additionally, the requirements placed on a medical Digital Twin had to be considered for the design of the proposed architecture (see Section 2.2). An important requirement for support systems is its reliable and robust prediction. As ensemble methods take into account several opinions and votes of different classifiers, their variance decreases through averaging over these stories. Furthermore, this system can be prepared to deal with missing values or even missing measurements, or measurements being executed in a different ordering, as the structure of the algorithm can be adjusted very easily and meta-models can be re-trained.

In the medical context, interpretability of algorithms is mandatory, as decisions made upon the basis of them can influence the patients' outcome strongly. By using machine learning models with intrinsic interpretability as well as through the usage of model-agnostic interpretability methods this requirement is fulfilled.

Another goal is to increase the acceptance of the algorithm by physicians. On the one hand, this can be done by increasing the understanding of the decision-making process of the algorithm by incorporating known knowledge and by respecting the doctor's competence. As the modular design allows the software structure to be accessed in an interactive way by the doctor, this increases its transparency and understandability. Moreover, acceptance can be increased through an understandable and intuitive visualisation of the algorithm and interface with the doctor. As the Digital Twin is meant to be used as a support system, it is supposed to assists and not replace the doctor. Therefore, a third interaction layer in the graphical visualisation of the architecture has been introduced.

The usage of informed learning introduces the ability to adapt to the evolving state of knowledge recorded within the clinical guidelines. Other decision support tools usually need to be updated regularly to include new characteristics found or changes in the population. By using online-learning, the algorithms used for the Digital Twin can improve over time. This is done by regularly including new patients into the data base, on which the model can be retrained.

On the one hand, the Digital Twin should represent a decision-making aid for the current challenge and, on the other hand, it should be able to make a prediction about the future development of the patient if a certain option is chosen. This is implicitly fulfilled in that the statements made by the algorithms are an estimate of the biopsy result and thus correspond to a prediction. At the same time, of course,

a recommendation for the behaviour of the doctor is also made.

Is the proposed design limited to the specific example of the biopsy decision?

No, it is not. On the contrary, the design can be adjusted to fit other decision tasks mentioned in section 2.2.1 such as the staging decision. Therefore, base models that are able to deal with pathological images, base line information and other parameters collected during the biopsy need to be consulted. The specific architecture applied to different questions during the patient journey is scaled to a big picture through the usage of Bayesian inference realised through the update in the global patient state after each clinical decision. In any case, the explicit formulation of the overarching Digital Twin might not be straightforward, as Li et al. discussed in their review [44] the difficulties of including knowledge through selecting priors for a Bayesian method dealing with multiview data. First of all, this is because finding useful information as prior features might not be so easy and secondly because proper class conditional distributions are difficult to assume.

One can further advance the algorithm to handle new populations from different countries, ethnicities, or even just different institutions, which would entail the need for recalibration. This would be generally necessary if the solution was implemented for a broader patient cohort, as otherwise there might be the problem of sampling bias due to intrinsic and extrinsic demographic heterogeneity of the training data [30]. This fact was taken into account by selecting a database source covering several institutions.

Going one step further, the proposed architecture can be seen as a general approach on clinical patient care, independent of an explicit clinical setting. The approach is first specified for a specific application by choosing suitable base models.

What are current limitations and how can they be tackled?

"Any model is a simplified representation of the reality, with a limited scope and dependence on assumptions made" [30]. Therefore, it is crucially important to communicate these limitations, to avoid wrong usage of the support system as well as to maintain trust among physician, researchers and society.

The abilities of the developed approach for a Digital Twin concept are strongly correlated with the current clinical practice and the available data. Therefore, it is for example not possible to develop new treatments or practices. Using supervised machine learning allows for the identification of structures within the data, but nothing beyond that. Therefore, more data needs to be stored to capture more care events or alternatively prospective studies need to be performed to be able to evaluate whether there are better treatment options.

The performance of algorithms strongly depends on the data quality. Because in the medical context model developers often lack a sufficient amount of data to be able to perform well, feature engineering or feature selection is often used. The proposal of a modular Digital Twin presents a new way of dealing with this issue. Still it can be seen as a trade-off trying to minimise correlation between increasing the relative amount of training data (input feature) and losing information between features by splitting them up. In the case that all data is available together and in sufficient quantity, then approaches with better performance could certainly be developed.

Medical Digital Twins in Urology only would have sparse temporal data. If additional measurements or practices would be introduced to record parameters that are taken at more or less regular time intervals, alternative Digital Twin designs could be taken into consideration. In particular, approaches would need the capability to gain knowledge from the temporal evolution of these parameters. These models could be of interest for active patient monitoring and aftercare in urology. In these cases, an assessment of the potential course of the patient's condition and the need for clinical intervention is required.

What is problematic when applying supervised machine learning in a straightforward manner to clinical treatment data is that one must always include currently performed practices. Therefore, generalisation to other institutions with slightly different intervention patterns will be challenging from the very first stage. Hence, the usage of alternative forms of supervision might be advantageous. One option is to use a Joint Modelling [116] of the action and patient state by using causal inference for counterfactual reasoning. With this approach one should always monitor the patient state without any kind of treatment as the true label.

For medical data there are no natural fixed time points for measurements across patients. Furthermore, treatments are also not timed regularly and can have arbitrary forms i.e. doses. One way of treating this data was introduced by Soleimani, Subbaswamy and Saria during a conference on uncertainty in artificial intelligence in 2017 [117]. They explained the combination between Marked Point Process (MMP) for "when" to observe something and Gaussian Processes (GP) for "what" kind of relationships to observe. To be able to estimate the patient state without any intervention they referred to using Continuous Gaussian Processes (CGP). Applied to urology, this approach could be integrated as follows: Of general interest are the patient's quality-corrected life years, which weight the quantity and quality of the patient's remaining life. This can be understood as the label of interest whose course one would like to observe. It should be shown how the value would develop with and without intervention. Additionally, the expected outcomes of the diagnostic method - Biopsy can be viewed for a further assessment of the need of intervention.

Because for medical Digital Twins in Urology no good mechanical or systemic understanding of the disease is currently available, simulation can just be performed indirectly via ML estimates, but not explicitly through using mathematical description to drive the system forward. A possible approach for a simulation-based medical Digital Twin has been presented by Masison et al. [118]. Their central principle of the DT architecture also has been a "separation of computational algorithms for the different dynamic processes, eliminating dependencies that make model modifications and extensions cumbersome or impossible in complex models" [118]. They have argued as well that a decentralised modular software platform would allow the system to be scalable and adapt to the current state of knowledge and data and allow the separate analysis of different data types. They call their proposal a hub- and-spoke modular design, as different modules access their input data from a central data structure representing the global model state. The difference to the scenario faced in urology and hence to the concept proposed by this thesis is that they have access to parameters that are updated over time. Moreover, they can model the dynamic evolution of the system in discrete time steps through agent-based models. Therefore, the same submodels are applied to the same set of parameters at different time points. For prostate cancer patients, complementary parameters are collected, for each of which suitable models are developed. Another fundamental difference is that the models do not pursue the same task/question and are sequentially applied to the system in a fixed order. The presented proposal, on the other hand, strives for the combination of several statements on one topic in order to enable a comprehensive consideration of several factors influencing this decision. Nevertheless, this publication shows further advantages and possibilities of how the modular approach can be used to tackle further problems. Masison et al. includes the following challenging points [118]: "1) lack of transparency in the implementation of computational models, 2) intertwined component models and simulation processes dependent on each other, 3) use of incompatible data structures and computer languages, 4) brittle architectures that do not easily accommodate extensions of a model, and 5) software environments that do not easily support distributed collaboration".

A further improvement could be achieved if the information compression for decision making currently performed in the clinic (e.g. in the PIRAD score) were replaced by an alternative feature extraction method that could better preserve this information. For example, an AI could map MRI images to a multidimensional latent space, allowing image features to be represented better than through the PIRAD score.

What are the ethical questions or worries?

The statement of Corral-Acero et al. [30] about the need of transparency and honest communication is supported by other researchers as well. For example, Frederike Kaltheuner [119] has noted, that there is much of misguided AI in the medical domain and mentioned the importance for the scientific understanding of the limits of accuracy. Furthermore, she referred to papers that show massive gaps in accuracy for retrospectively-developed tools applied on prospective clinical setting. This shows the need for guidelines and benchmark tests. Until these are uniformly decided on and used, every developer should consider the advantages and risks of their algorithm with great care. There are already a few approaches that aim to ensure robust validation, for example Kim et al. [120] state three design criteria: diagnostic cohort design, the inclusion of multiple institutions, and prospective data collection.

Another important point to be aware of is how much influence the selected training data set has on the algorithm's decisions. If racist or discriminatory behavioural structures are contained in the data, the algorithm will learn to continue to carry them out. Therefore, a conscious questioning of the current practices with which the data set was created is absolutely necessary. Potential sources of bias in care have been explicitly listed by Schwartz et al. [43] as follows:

- Historical inequities
- Exclusion of women, non-white people, or members of other ethnicities from studies
- Values shaping access to resources, technology, and information

As in many situations it is difficult to exclude biases completely, it is still important to refer to them and to make sure that these algorithms are just applied to relevant demographic groups.

This chapter summarised again the huge potential of improving current clinical practice by using Patient Digital Twin as a Support System. The proposed approach within this thesis is already tackling many of the currently identified challenges. As discussed, there are still many open risks and questions, that need to be studied in more detail. Hence, the improvement, scaling, critical analysis and risk assessments of the current analysis is crucially important on the way towards generating guidelines for the Digital Twin development in a clinical environment.

Part I

Appendix

A Tables

Word	Description
machine learning (ML)	use mathematical models for inference on data & building mathematical models using data
supervised learning	trained on labelled data - data including input and desired output values
feature	input variable to model; quantitative description of items; clinical parameters
instance	item; in clinical context - patient
bias	difference between average and true model
variance	describes difference between models estimated from different data sets
input data	full data set, divided into training and test data
training data	model is built/learned with this data
test data	model accuracy is tested on this data
instance	in clinical context - patient
classification	supervised learning in which data instances are separated into categories according to their features
Output variables	predictions from model; response variable
soft prediction	probabilities for the instance being in a certain class is returned from the model
hard prediction	only the most probable class according to the model is transmitted
ranked prediction	the classes are returned in an ordered listing

Table A.1: Listing of basics of Machine Learning partly taken from the page [121].

Integrative method	Multi class data (type-1)	Multi feature set data (type-2)	Tensor data (type 3)	Multi relational data (type-4)
Feature concatenation		Classification Regression Feature Selection		
Bayesian models or networks	Classification Feature Selection	Classification Regression Feature Selection Pathway analysis		
Ensemble learning	Classification Feature Selection	Classification Regression Feature Selection		
Multiple Kernel Learning	Classification	classification Regression Clustering		Association study
Network-based methods				Association study
Multi-view matrix or tensor factorisation	Classification Feature Selection	Classification Feature Selection Pathway analysis Clustering	Classification Clustering	Association Study
Multi-modal learning		Classification Clustering Association study		

Table A.2: The Table taken from Li et al. [44] is showing machine learning methods handling four types of multi-view data.

Feature Name Original	Feature Name	Description
pros_gleason	GS	Gleason Score obtained from the Biopsy
psa_assay	PSA assay	Whether or not Hyvritech Tandem PSA Assay was used
psa_value	PSA 0	Currently measured PSA value
psa_encounter-1	PSA -1	PSA value from the previous encounter
psa_encounter-2	PSA -2	PSA value from the encounter prior to the psa_encounter-1
vpsa	PSA velocity	Calculated from three previously measured PSA values
dre_result	DRE	Digital Rectal Examination screening result
dre_encounter-1	DRE -1	Result of the DRE obtained from the previous encounter
sizesag	size (sagittal)	Size of gland in sagittal direction
sizetran	size (transverse)	Size of gland in transverse direction
loc_1	loc (apex1)	Location of induration - left apex
loc_2	loc (apex2)	Location of induration - right apex
loc_3	loc (lat.lobe1)	Location of induration - left lateral lobe
loc_4	loc (lat.lobe2)	Location of induration - right lateral lobe
loc_5	loc (base1)	Location of induration - left base
loc_6	loc (base2)	Location of induration - right base
loc_7	loc (sem.vesicle1)	Location of induration - left seminalvesicle
loc_8	loc (sem.vesicle2)	Location of induration - right seminalvesicle
extent_min	extent (min induration)	Extent of this area of smallest induration
sizein_min	size (min induration)	Approximate size of this area of smallest induration
typein_min	type (min induration)	Type of this area of smallest induration (non-nodular/ diffuse/ nodular)
grade_min	grade (min induration)	Grade of this area of smallest induration
extent_max	extent (max induration)	Extent of this area of biggest induration
sizein_max	size (max induration)	Approximate size of this area of biggest induration
typein_max	type (max induration)	Type of this area of biggest induration (non-nodular/ diffuse/ nodular)

grade_max	grade (max induration)	Grade of this area of biggest induration
pros_clinstage_t	clinical t	Prospective clinical t value
max_sbcd	number (induration)	Number of indurations found during one examination
trus_volume	TRUS volume	The volume of the TRUS prostate dimensions
trus_result	TRUS	Result of the TRUS procedure (suspicious)
age	age	Patient age
ph_first_cancer_age	age start	Age at First Personal History of Cancer
educat	education	Grade of education
cigar	smoking	Whether the patient has ever been/ is smoking
smoked_f	reg. smoking	Whether the patient has ever been smoking regularly
rsmoker_f	now reg. smoking	Whether the patient is smoking regularly
fh_cancer	cancer family	Family history of any cancer
pros_fh	PC family	Family history of prostate cancer
bmi_curr	BMI	Body-mass-index
enlpros_f	enlarged	Existence of enlarged prostate stated before
enlprosa	enlarged age	Age at which existence of enlarged prostate stated before
infpros_f	inflamed	Existence of inflamed prostate stated before
infprosa	inflamed age	Age at which existence of inflamed prostate stated before
prosprob_f	problem	Existence of problems with prostate stated before
urinate_f	urination	Frequency of nocturnal urination
urinatea	urination age	Age at which nocturnal urination started
vasect_f	vasectomy	Whether a vasectomy has been performed
vasecta	vasectomy age	Age at which a vasectomy has been performed
race: white	race	Binary categorisation for white race - Weird category!
Married Or Living As Married	marriage	Marriage
current_working	work	Work

Table A.3: An overview of the features coming from the original PLCO data set, their original name and a short description. Further details can be found in the PLCO documentation file [53].

Feature Name	Full cohort Mean	Full cohort Std	Malignant Mean	Malignant Std	Non-Malignant Mean	Non-Malignant Std	Datatype	Missing Values [%]
PSA velocity	1.10	2.45	1.86	4.22	0.80	1.05	numeric	86
PSA -1	4.73	4.77	5.63	7.68	4.37	2.88	numeric	86
PSA -2	4.14	3.90	4.55	5.75	3.99	2.86	numeric	86
DRE -1	2.21	0.89	2.15	0.90	2.23	0.89	category	89
size (sagittal)	3.82	0.75	3.74	0.70	3.84	0.76	numeric	92
size (transverse)	4.07	0.76	4.00	0.69	4.10	0.78	numeric	92
DRE	2.18	0.75	2.20	0.79	2.18	0.73	category	92
clinical t	3.58	1.06	3.89	1.24	3.39	0.87	category	1
PSA	9.44	12.88	13.20	18.92	7.26	6.42	numeric	37
loc (apex1)	0.35	0.51	0.41	0.53	0.33	0.49	numeric	93
loc (apex2)	0.29	0.49	0.28	0.46	0.29	0.51	numeric	93
loc (lat.lobe1)	0.37	0.52	0.47	0.54	0.32	0.50	numeric	93
loc (lat.lobe2)	0.36	0.51	0.37	0.50	0.36	0.52	numeric	93
loc (base1)	0.21	0.43	0.25	0.49	0.20	0.40	numeric	93
loc (base2)	0.33	0.50	0.29	0.47	0.35	0.52	numeric	93
loc (sem.vesicle1)	0.00	0.07	0.01	0.12	0.00	0.00	numeric	93
loc (sem.vesicle2)	0.00	0.06	0.01	0.10	0.00	0.00	numeric	93
extent (min induration)	1.01	0.14	1.01	0.21	1.01	0.09	category	93
size (min induration)	1.32	0.66	1.44	0.73	1.26	0.61	category	93
type (min induration)	1.97	0.91	2.06	0.90	1.92	0.92	category	93
grade (min induration)	1.75	0.80	1.90	0.84	1.68	0.76	category	93
extent (max induration)	1.01	0.17	1.02	0.25	1.01	0.10	category	93
size (max induration)	1.36	0.69	1.49	0.75	1.30	0.65	category	93
type (max induration)	2.04	0.92	2.15	0.90	1.99	0.92	category	93
grade (max induration)	1.83	0.81	1.98	0.85	1.75	0.78	category	93
number (induration)	1.15	0.40	1.13	0.37	1.16	0.42	category	93

TRUS volume	43.36	22.37	41.25	20.89	44.77	23.22	numeric	82
TRUS	3.12	1.32	3.30	1.24	3.05	1.35	category	93
PSA assay	0.51	0.50	0.45	0.50	0.53	0.50	binary	77
age	63.48	5.07	63.92	5.11	63.20	5.02	numeric	0
age start	55.78	11.09	56.11	9.53	55.53	12.22	numeric	98
education	4.94	1.65	4.92	1.68	4.95	1.64	category	3
smoking	0.46	0.82	0.47	0.83	0.45	0.82	category	3
reg. smoking	0.59	0.49	0.59	0.49	0.59	0.49	binary	3
now reg. smoking	0.15	0.36	0.15	0.36	0.15	0.36	binary	4
cancer family	0.55	0.50	0.54	0.50	0.55	0.50	binary	3
PC family	0.12	0.31	0.11	0.31	0.12	0.32	category	3
BMI	27.23	3.89	27.34	4.09	27.15	3.76	numeric	4
enlarged	0.26	0.44	0.24	0.43	0.27	0.45	binary	3
inflamed	0.10	0.30	0.10	0.29	0.11	0.31	binary	2
problem	0.30	0.46	0.27	0.45	0.31	0.46	binary	3
urination	1.30	0.91	1.30	0.91	1.30	0.91	category	3
vasectomy	0.27	0.44	0.26	0.44	0.28	0.45	binary	3
enlarged age	4.31	0.91	4.32	0.93	4.31	0.90	category	75
inflamed age	3.60	1.30	3.60	1.34	3.61	1.28	category	92
urination age	4.14	1.04	4.21	1.03	4.10	1.04	category	65
vasectomy age	2.87	0.69	2.86	0.69	2.87	0.69	category	74
race	0.89	0.32	0.87	0.34	0.90	0.30	binary	3
married	0.86	0.35	0.85	0.36	0.86	0.34	binary	3
work	0.42	0.49	0.40	0.49	0.43	0.49	binary	3

Table A.4: The table is giving an overview of all the parameters extracted from the original PLCO data set with representation of their mean and standard deviation for each class, as well as their number of missing values and data type.

feature	PSA -2	clinical t	PSA velocity	PSA	PSA -1
SU	0.037249	0.037119	0.034457	0.03167	0.029544

feature	BMI	age	inflamed	size (transverse)	education
SU	0.010909	0.010855	0.010614	0.009194	0.007557

feature	race	DRE	size (sagittal)	vasectomy	reg. smoking
SU	0.007345	0.005601	0.005011	0.004936	0.004393

feature	urination	DRE -1	married	PC family	smoking
SU	0.004073	0.00142	0.000651	0.000355	0.000353

feature	work	problem	enlarged	cancer family
SU	0.00017	0.000123	0.000042	0.000003

Table A.5: Features included in the data set $D_{complete}$, ranked according to their Symmetric Uncertainty score. The table is starting with the highest scoring feature *PSA -2*.

Number of Features	Accuracy SVM [%]	LR [%]	kNN [%]	MLP [%]
25	76.3 (± 2.1)	74.0 (± 2.8)	58.5 (± 3.0)	73.1 (± 4.5)
24	76.2 (± 2.7)	74.1 (± 2.9)	57.5 (± 4.0)	73.4 (± 3.7)
23	75.0 (± 2.3)	74.1 (± 2.6)	57.2 (± 4.4)	72.1 (± 2.0)
22	75.6 (± 1.6)	74.6 (± 2.9)	58.3 (± 2.9)	72.8 (± 2.3)
21	76.1 (± 2.7)	74.9 (± 2.6)	57.6 (± 4.0)	72.6 (± 2.2)
20	76.2 (± 2.9)	75.1 (± 2.8)	57.7 (± 2.5)	70.7 (± 3.6)
19	76.5 (± 3.2)	74.8 (± 2.4)	58.9 (± 2.6)	72.9 (± 2.9)
18	75.6 (± 2.8)	74.7 (± 2.7)	59.4 (± 3.4)	72.2 (± 2.7)
17	75.7 (± 3.0)	74.3 (± 3.0)	59.9 (± 3.2)	73.8 (± 2.4)
16	75.4 (± 3.3)	75.0 (± 2.6)	59.6 (± 3.8)	74.5 (± 3.8)
15	75.2 (± 3.2)	74.9 (± 2.1)	57.8 (± 2.9)	73.5 (± 2.4)
14	76.1 (± 2.8)	73.8 (± 2.8)	57.5 (± 2.7)	72.6 (± 3.0)
13	74.6 (± 3.2)	74.0 (± 2.6)	56.7 (± 2.6)	71.5 (± 3.3)
12	75.6 (± 3.1)	73.8 (± 2.7)	60.4 (± 3.8)	73.6 (± 3.2)
11	76.7 (± 3.1)	74.1 (± 2.7)	60.7 (± 3.1)	74.5 (± 3.2)
10	75.0 (± 3.0)	74.0 (± 2.5)	64.6 (± 2.8)	74.4 (± 4.0)
9	76.5 (± 3.3)	73.9 (± 3.2)	64.3 (± 3.1)	73.9 (± 4.4)
8	75.3 (± 4.3)	75.1 (± 3.2)	61.8 (± 2.8)	74.3 (± 3.9)
7	74.7 (± 4.0)	75.2 (± 3.3)	63.2 (± 3.2)	73.8 (± 1.4)
6	74.8 (± 4.0)	75.8 (± 2.7)	64.0 (± 3.7)	74.6 (± 3.1)
5	73.5 (± 3.9)	75.2 (± 2.9)	63.6 (± 3.8)	74.8 (± 2.8)
4	72.6 (± 3.6)	75.7 (± 3.1)	63.9 (± 3.5)	73.9 (± 2.8)
3	70.7 (± 3.5)	75.7 (± 3.1)	67.5 (± 4.7)	74.9 (± 2.9)
2	62.7 (± 6.8)	73.8 (± 3.1)	63.4 (± 2.7)	71.9 (± 3.1)
1	75.1 (± 0.6)	75.4 (± 0.9)	62.6 (± 5.1)	73.0 (± 3.1)

Table A.6: Mean and standard deviation received for the Accuracy [%] of the machine learning models from 10-fold cross validation, for varying feature set size. The features contributing to the subset were chosen according to their rank, listed in Table A.5. Maximal Accuracy can be observed for a set size of: 11 - SVM, 5 - LR, 3 - kNN and 3 -MLP. The threshold for the Accuracy calculation was kept at 0.35.

		Voting	Prob. Sum	Weighted Average	Decision Templates	Dempster Shafer	Entropy Weighting	LOP
Stacking	Accuracy	0.70(±0.02)	0.74(±0.02)	0.75(±0.02)	0.69(±0.02)	0.31(±0.02)	0.73(±0.02)	0.72(±0.01)
	Precision	0.50(±0.04)	0.59(±0.03)	0.62(±0.03)	0.00(±0.00)	0.31(±0.02)	0.57(±0.03)	0.55(±0.03)
	Recall	0.31(±0.02)	0.48(±0.03)	0.48(±0.03)	0.00(±0.00)	1.00(±0.00)	0.48(±0.03)	0.44(±0.02)
	F1-Score	0.39(±0.02)	0.53(±0.03)	0.54(±0.03)	0.00(±0.00)	0.47(±0.03)	0.52(±0.03)	0.49(±0.02)
	auROC	0.67(±0.02)	0.69(±0.02)	0.69(±0.02)	0.31(±0.02)	0.57(±0.02)	0.72(±0.02)	0.73(±0.02)
	prAUC	0.51(±0.03)	0.53(±0.03)	0.54(±0.03)	0.22(±0.02)	0.39(±0.03)	0.58(±0.03)	0.57(±0.03)
Manual FSS	Accuracy	0.61(±0.02)	0.73(±0.02)	0.73(±0.02)	0.69(±0.02)	0.31(±0.02)	0.73(±0.02)	0.72(±0.02)
	Precision	0.42(±0.03)	0.85(±0.08)	0.85(±0.08)	0.00(±0.00)	0.31(±0.02)	0.75(±0.05)	1.00(±0.00)
	Recall	0.77(±0.02)	0.16(±0.03)	0.16(±0.03)	0.00(±0.00)	1.00(±0.00)	0.19(±0.03)	0.1(±0.03)
	F1-Score	0.54(±0.03)	0.27(±0.05)	0.27(±0.05)	0.00(±0.00)	0.47(±0.03)	0.31(±0.05)	0.17(±0.04)
	auROC	0.70(±0.02)	0.68(±0.02)	0.68(±0.02)	0.36(±0.02)	0.44(±0.01)	0.69(±0.02)	0.64(±0.02)
	prAUC	0.54(±0.03)	0.54(±0.03)	0.54(±0.03)	0.24(±0.02)	0.22(±0.03)	0.55(±0.03)	0.51(±0.03)
Alg. FSS	Accuracy	0.36(±0.02)	0.69(±0.02)	0.72(±0.02)	0.69(±0.02)	0.31(±0.02)	0.69(±0.02)	0.7(±0.02)
	Precision	0.32(±0.02)	0.46(±0.08)	1.00(±0.00)	0.00(±0.00)	0.31(±0.02)	0.48(±0.07)	0.9(±0.3)
	Recall	1.00(±0.00)	0.10(±0.03)	0.10(±0.03)	0.00(±0.00)	1.00(±0.00)	0.15(±0.04)	0.03(±0.01)
	F1-Score	0.49(±0.02)	0.16(±0.04)	0.18(±0.04)	0.00(±0.00)	0.47(±0.03)	0.23(±0.05)	0.06(±0.02)
	auROC	0.68(±0.02)	0.71(±0.02)	0.71(±0.02)	0.35(±0.02)	0.43(±0.02)	0.69(±0.02)	0.70(±0.02)
	prAUC	0.54(±0.03)	0.55(±0.03)	0.55(±0.03)	0.23(±0.02)	0.25(±0.06)	0.49(±0.03)	0.55(±0.03)

Table A.7: Performance overview of the Weighting Methods for Stacking, the Manual Feature Subset Selection (Manual FSS - Missing Value Clustering) and the Algorithmic Feature Subset Selection (Alg. FSS - Redundancy aware FSP). Selected Weighting Methods are Voting, Probability Sum (Prob. Sum), Weighted Average, Decision Templates, Dempster Shafer, Entropy Weighting and Logarithmic Opinion Pool (LOP).

		LOP trained	Linear Fuser	2D NN	LR 1	LR 2
Stacking	Accuracy	0.72 (±0.02)	0.7 (±0.02)	0.77 (±0.02)	0.69 (±0.01)	0.74 (±0.02)
	Precision	0.55 (±0.03)	0.63 (±0.07)	0.70 (±0.03)	0.50 (±0.03)	0.64 (±0.04)
	Recall	0.41 (±0.02)	0.07 (±0.02)	0.44 (±0.03)	0.34 (±0.02)	0.31 (±0.04)
	F1-Score	0.47 (±0.02)	0.12 (±0.03)	0.54 (±0.03)	0.40 (±0.03)	0.42 (±0.04)
	auROC	0.69 (±0.02)	0.71 (±0.02)	0.70 (±0.02)	0.63 (±0.02)	0.67 (±0.02)
	prAUC	0.54 (±0.03)	0.55 (±0.03)	0.61 (±0.03)	0.52 (±0.03)	0.53 (±0.03)
Manual FSS	Accuracy	0.69 (±0.02)	0.69 (±0.02)	0.71 (±0.02)	0.69 (±0.02)	0.64 (±0.01)
	Precision	0.48 (±0.03)	0.00 (±0.00)	0.60 (±0.07)	0.49 (±0.05)	0.42 (±0.03)
	Recall	0.42 (±0.04)	0.00 (±0.00)	0.18 (±0.03)	0.25 (±0.05)	0.41 (±0.03)
	F1-Score	0.45 (±0.03)	0.00 (±0.00)	0.28 (±0.04)	0.33 (±0.05)	0.42 (±0.03)
	auROC	0.69 (±0.02)	0.51 (±0.03)	0.55 (±0.03)	0.68 (±0.02)	0.63 (±0.02)
	prAUC	0.53 (±0.03)	0.36 (±0.05)	0.45 (±0.04)	0.54 (±0.03)	0.46 (±0.03)
Alg. FSS	Accuracy	0.75 (±0.02)	0.69 (±0.02)	0.72 (±0.02)	0.71 (±0.02)	0.64 (±0.01)
	Precision	0.70 (±0.04)	0.00 (±0.00)	1.00 (±0.00)	0.54 (±0.04)	0.41 (±0.03)
	Recall	0.31 (±0.04)	0.00 (±0.00)	0.09 (±0.02)	0.31 (±0.03)	0.36 (±0.02)
	F1-Score	0.43 (±0.05)	0.00 (±0.00)	0.17 (±0.03)	0.39 (±0.03)	0.38 (±0.02)
	auROC	0.73 (±0.02)	0.67 (±0.02)	0.65 (±0.02)	0.72 (±0.02)	0.62 (±0.02)
	prAUC	0.60 (±0.03)	0.47 (±0.03)	0.55 (±0.03)	0.56 (±0.03)	0.47 (±0.03)

Table A.8: Performance overview of the Meta-learning Methods for Stacking, the Manual Feature Subset Selection (Manual FSS - Missing Value Clustering) and the Algorithmic Feature Subset Selection (Alg. FSS - Redundancy aware FSP). Selected Meta-learning Methods are Logarithmic Opinion Pool with trained weights (LOP trained), one dimensional Artificial Neural Network (Linear Fuser), Artificial Neural Network with one hidden layer (2D NN), Logistic Regression on base model outputs (LR 1) and inputs (LR2).

B Further Information

B.1 Data Preprocessing PLCO

A merging process between several files had to be performed between the main file including all general information and main findings, the diagnostic procedures file including detailed information about procedures like PSA measurement, TRUS and Biopsy, the screening file including information about regular PSA and DRE measurements and the sub-screening file with even more details about the findings. During the merging process, filter categories were introduced and some variables were renamed. The merge was done based on the patient id and study year (*study_yr*) and randomised biopsy day. The procedure file was filtered to extract the following:

1. Biopsy related data, this will be used later to filter our data for the biopsy cases. Here, only the biopsy information (*has_biopsy*: yes = 1), result (now called: *biopsy_result*), staging/diagnosis (*bx_diag_staging*) and merge keys + *proc_days* (now: *biopsy_days*) were extracted.
2. PSA data gets extracted, same as in 1. only *psa_res* and *psa_assay* and *proc_days* get extracted. *proc_days* is renamed to *PSA_proc_days*
3. TRUS procedure data, the results is renamed to *trus_result* and only the TRUS dimensions 1-3, TRUS volume and result are relevant.
4. Combining TRUS results with biopsy and PSA results: As merge keys are not unique, to each biopsy the most recent TRUS scan is being assigned by using the *trus_days* and *biopsy_days* information. Similar for PSA measurement.

From the sub-screening file min and max values of the descriptive variables (extent, size, type, grade) for visits with multiple findings of locations have been added. The screening file introduce new features like PSA velocity, older PSA (up to 2 years) and DRE (only one year) results. The PSA velocity is calculated from three following PSA measurements (PSA_{1-3}) as: $\mathrm{PSA}_{\mathrm{Velocity}}\ \left[\frac{\mathrm{ng}}{\mathrm{ml{\cdot}a}}\right] = \frac{1}{2} \cdot \left(\frac{\mathrm{PSA}_2 - \mathrm{PSA}_1}{\mathrm{t}_1\mathrm{[a]}} + \frac{\mathrm{PSA}_3 - \mathrm{PSA}_2}{\mathrm{t}_2\mathrm{[a]}}\right)$. Additionally, the PSA value is chosen to be the most recent either from the procedures or the regular screening. If there is closest PSA value recorded is older than 360 days, it is omitted, thus resulting in a missing value.

Changes to original variable encoding were needed:

- *psa_assay* is binary encoded for Hybritech, originally being categorised into [1 = Hybritech, 2 = Abbott, 4 = Diagnostic Products, 5= Bayer, 8 = Other].
- *indu_loc_side* is rescaled. *No marked location* = 9 is relabelled to 0. Left and right are both labelled with 1 and *both sides* is labelled with 2.
- *dre_result* is changed from [1 = negative, 2 = abnormal + suspicious, 3= abnormal + not-suspicious] to [1 = negative, 2 = abnormal + not-suspicious,

3= abnormal + suspicious], similar for TRUS results.
- *pros_fh* (Family history) with unclear/relative unknown is set to 0.5.
- Binary encoding for working question is chosen, for which originally 7 categories existed [1 = Homemaker, 2 = Working, 3 = Unemployed, 4 = Retired, 5 = Extended Sick Leave, 6 = Disabled, 7 = Other].
- *proc_psares* is capped at a level of 80, this does not influence the PSA-velocity.
- *maritial* is binary encoded to married or not, with originally 5 categories [1 = Married Or Living As Married, 2 = Widowed, 3 = Divorced, 4 = Separated, 5 = Never Married].
- Race is binary encoded regarding the white race [yes or no].

B.2 Dealing with Missing Values

Ideas collected on how one can deal with missing values are the following:

- *Impute Constant Value:* Interpretation of the missing digits as a separate category, which is represented by a value. Certain models like trees can easily deal with this. Other simple models like kNN might be confused, as a proper distance metric is hard to define.
- *Impute Mean:* Missing values can be estimated from the rest of the data, for example using the mean. Even if this approach guarantees output from all models, it might lead to severe impact on the model prediction and therefore could have serious consequences.
- *Feature Space Augmentation:* A doubling of the dimension (dim=2*N) is done by appending a binary vector (dim=N) which represents whether the parameters are present (0 = yes, 1 = no). Like in the first approach, this method might depend strongly on the model in question and can not be implemented generally.
- *Exhaustive Ensemble Training:* Feature subsets can be created in such a way that it is taken into account which feature is missing with which probability, so that there is a high probability that at least one model always has complete input values and is therefore meaningful. The ensemble potential failure might be lowered by finding the subspace with the highest variance in absence probability and replace the worst feature with best feature that is not yet part of the subspace. If many values are missing so that no model can get through, it might also be a statement to the doctor that the currently available information about the patient is not sufficient for a decision-making process. For this approach, however, it would be necessary to carry out a computationally expensive training process that runs through all possible scenarios for the combination method of the algorithms.
- *Use Input Uncertainty:* Another model-specific solution could be the usage of algorithms that can deal with input uncertainty. For missing values one could use the centre point of the parameter distribution (multivariate Gauss) and their maximum deviation.

C Additional Figures

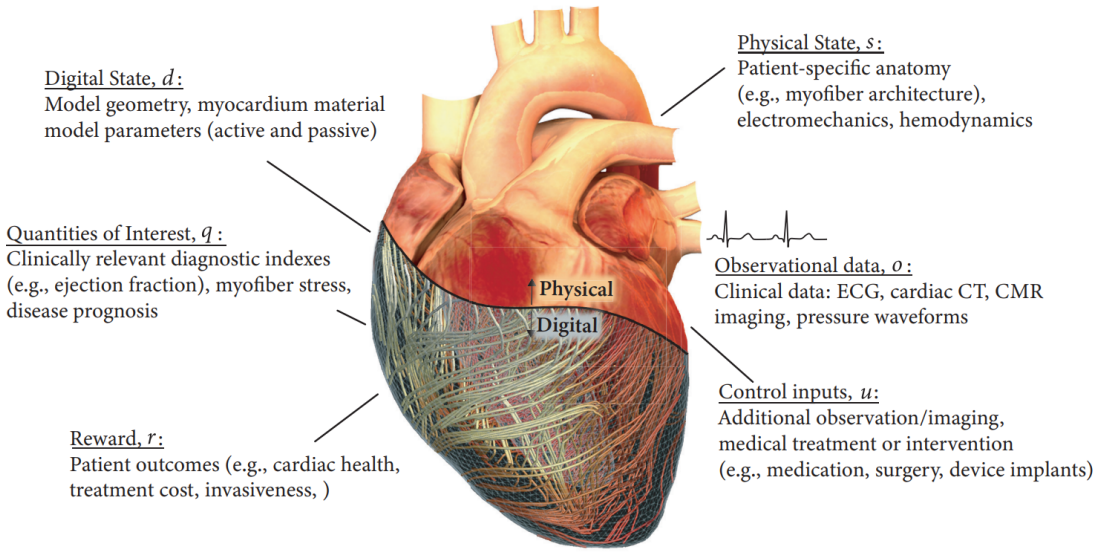

Figure C.1: Illustration of the quantities defined for the probabilistic graphical model concept of Kaptyn et al. [31] applied the asset-twin system of a human heart. The Figure was taken from the PhD thesis of Kaptyn [122] with permission.

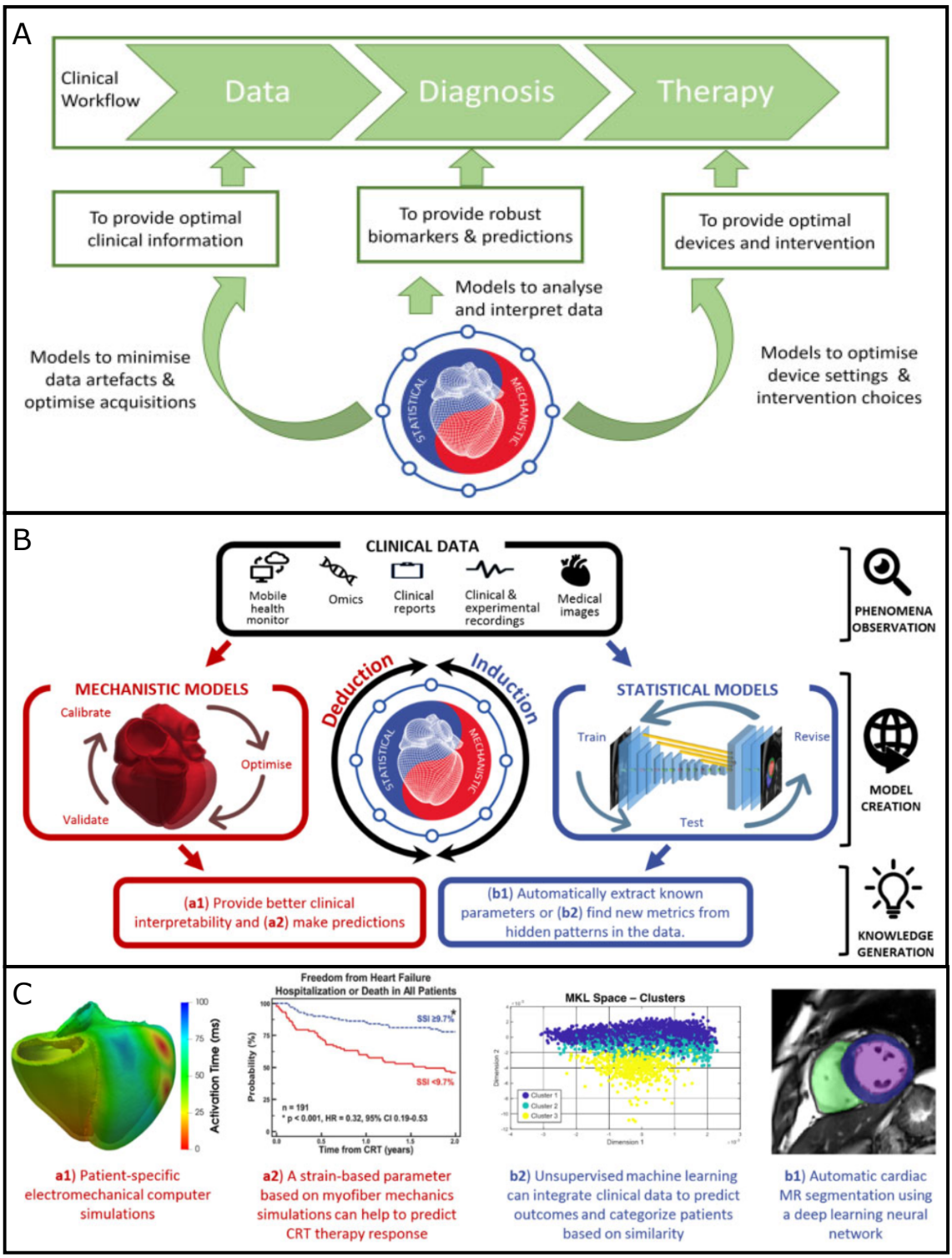

Figure C.2: The cardiovascular Digital Twin presented by J. Corral-Acero et al. [30] is a hybrid system combining mechanistic and statistical models. Figures were taken from this publication with permission. They show: (A) the DT as support system for clinical workflow: information extraction from patient data, knowledge inference for diagnosis and risk stratification, personalised therapy decision; (B) the complementary interplay between statistical and mechanistic model at the level of knowledge generation; (C) four examples of knowledge generation implementations.

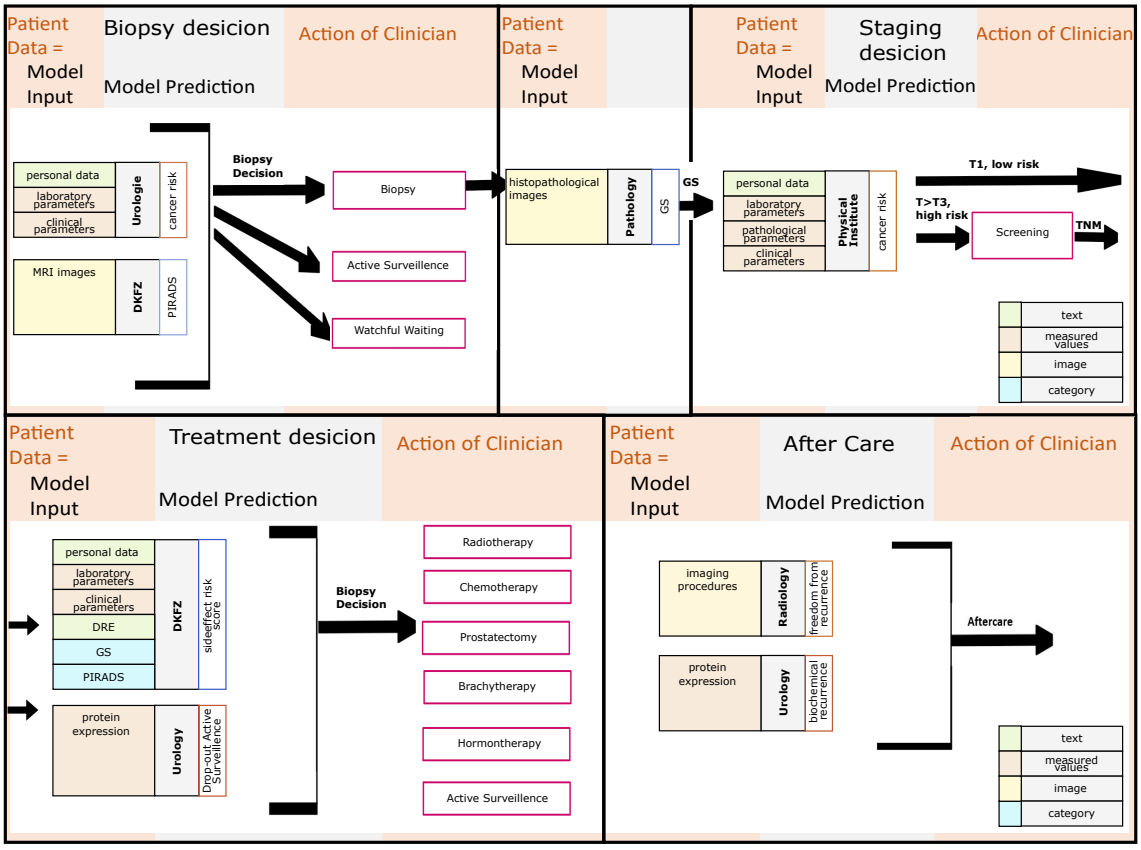

Figure C.3: Overview of the individual algorithms and their task within the cooperation project "Clinic5.1". Each of the four boxes represent one decision during the cancer patient journey. Starting at the top left- biopsy decision, going to top right - staging decision, going to bottom left - treatment decision and finally the bottom right - aftercare. Each box consists of the model input phase, the model prediction and the action of the clinician backed with colours (orange - real world, grey - digital world). The only exception no really fitting into the schemata is the Gleason Score extraction in the upper centre. Otherwise, in the tripartite boxes on can see the input parameters with colour codes explained by the legend at the bottom right, the name of the cooperation partner and the model output (probabilities in orange, scores in blue). The arrows show that an overall decision needs to be made to one of the possible actions (pink boxes).

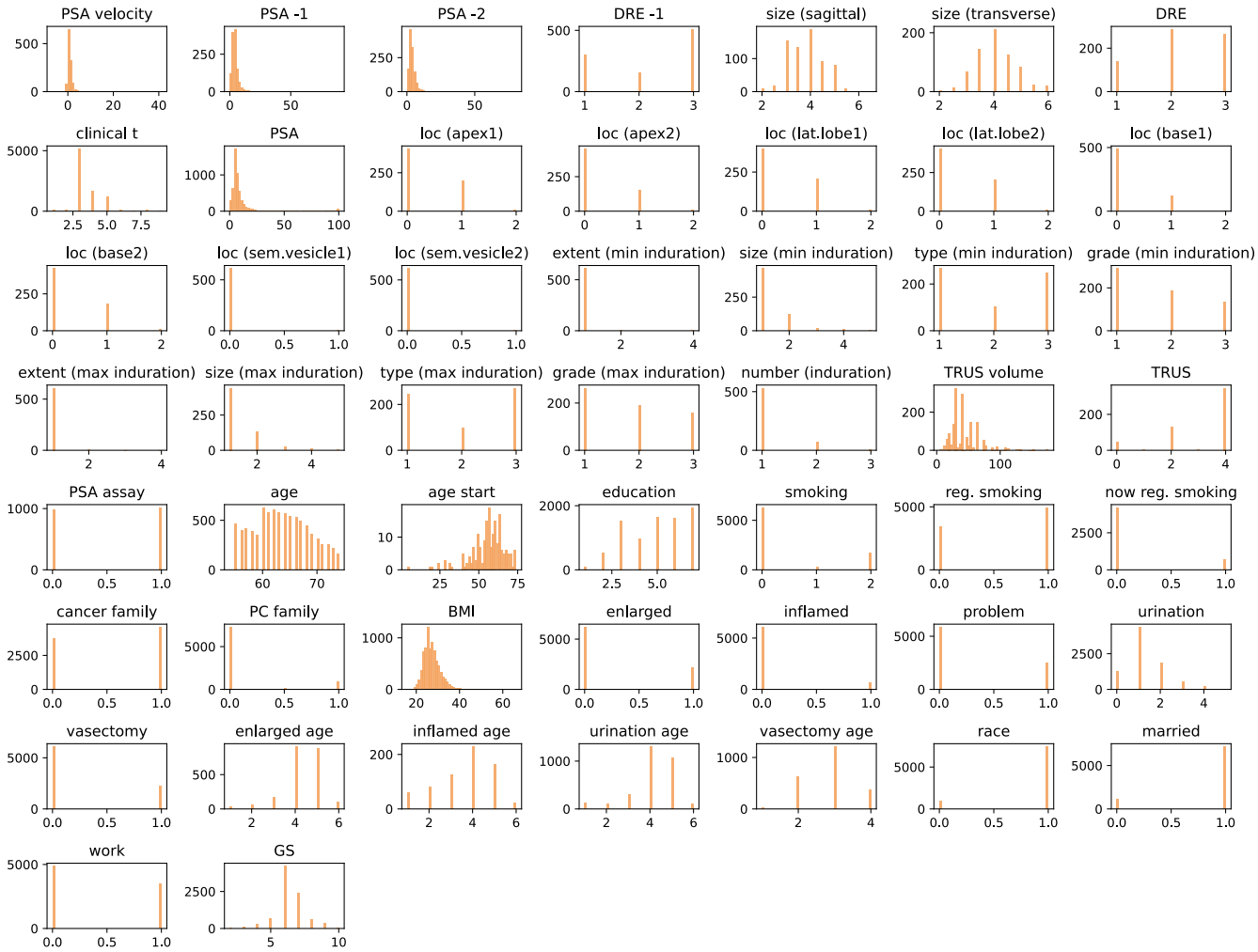

Figure C.4: This Figure is representing the distribution of the features extracted from the original PLCO data set. On the x-axes the existing values for the quantity are written down. On the y-axes the the amount of encountered patients with this feature value are given.

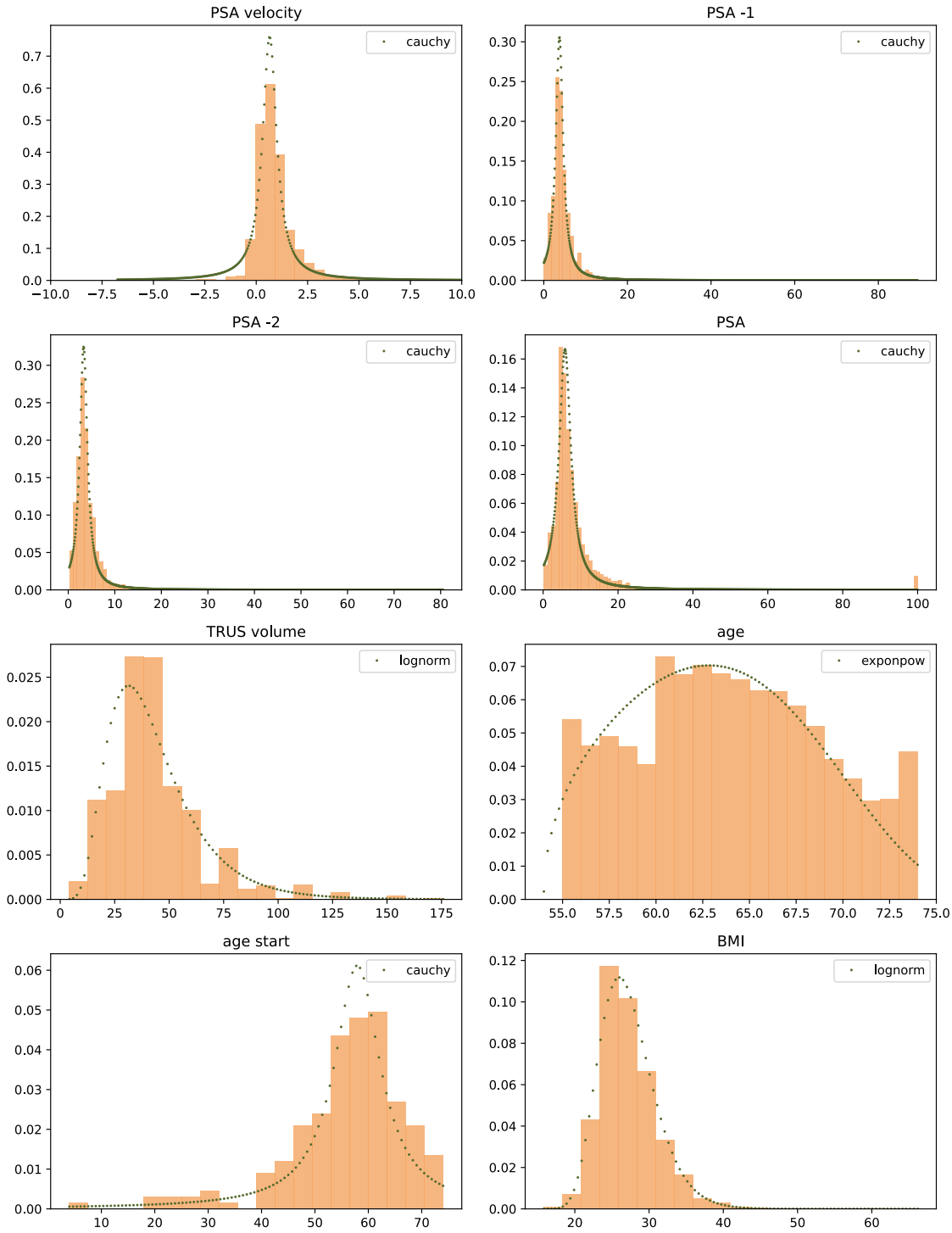

Figure C.5: Fit of the probability distribution for the numerical features. The label is showing which function family has been observed to fit the data best. On the x-axes the existing values for the quantity are written down. On the y-axes the the amount of encountered patients with this feature value are given.

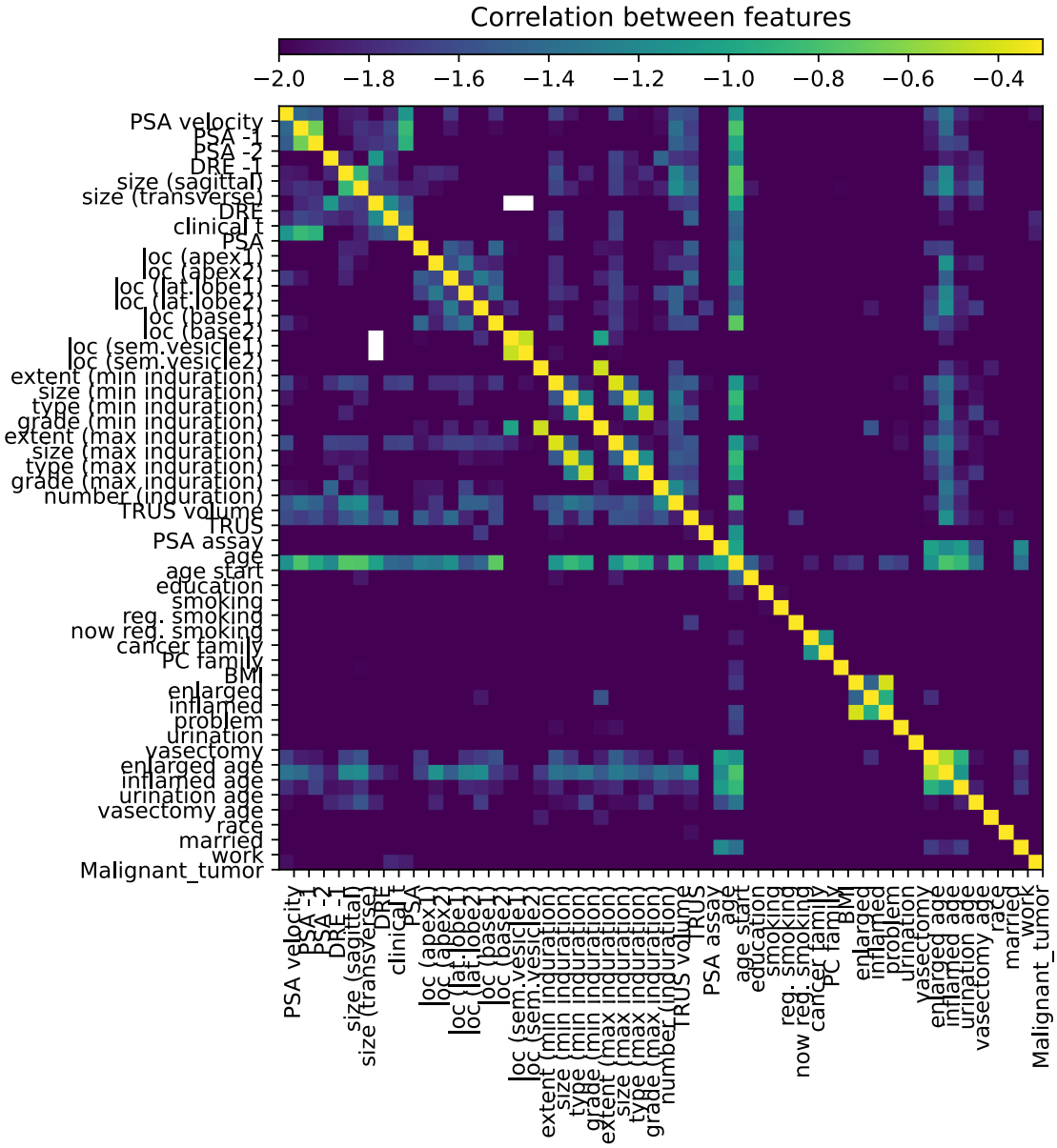

Figure C.6: Symmetric Uncertainty as a measure for the correlation strength between the features extracted from the PLCO data set, scaled by the logarithm to base 10. In general, the score is showing a symmetric correlation strength between two features, as expected. Some features, like TRUS, TRUS volume, clinical t, PSA and age, show stronger correlations to other features.

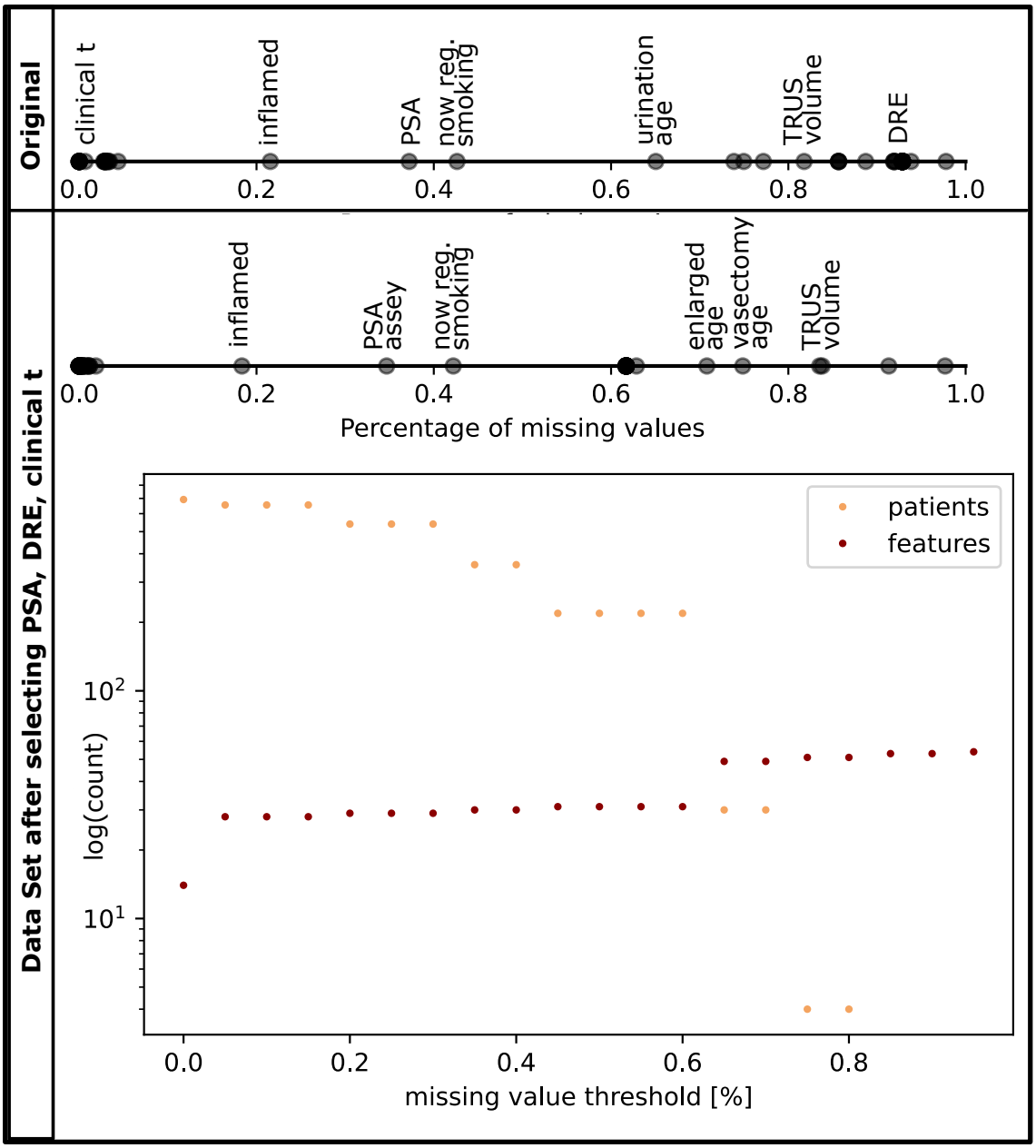

Figure C.7: Overview missing value rate of the features in the original data set, as well as after making sure that "PSA", "DRE" and "clinical t" are included as informative features. The lower plot is showing how the number of features and patients scale with selecting a threshold t for the allowed percentage of missing values. One can see how the two quantities contradict each other and an trade-off needs to be found. As the patient number is decreasing after the feature count is stagnating between 10% and 60%, the threshold is set to 20%.

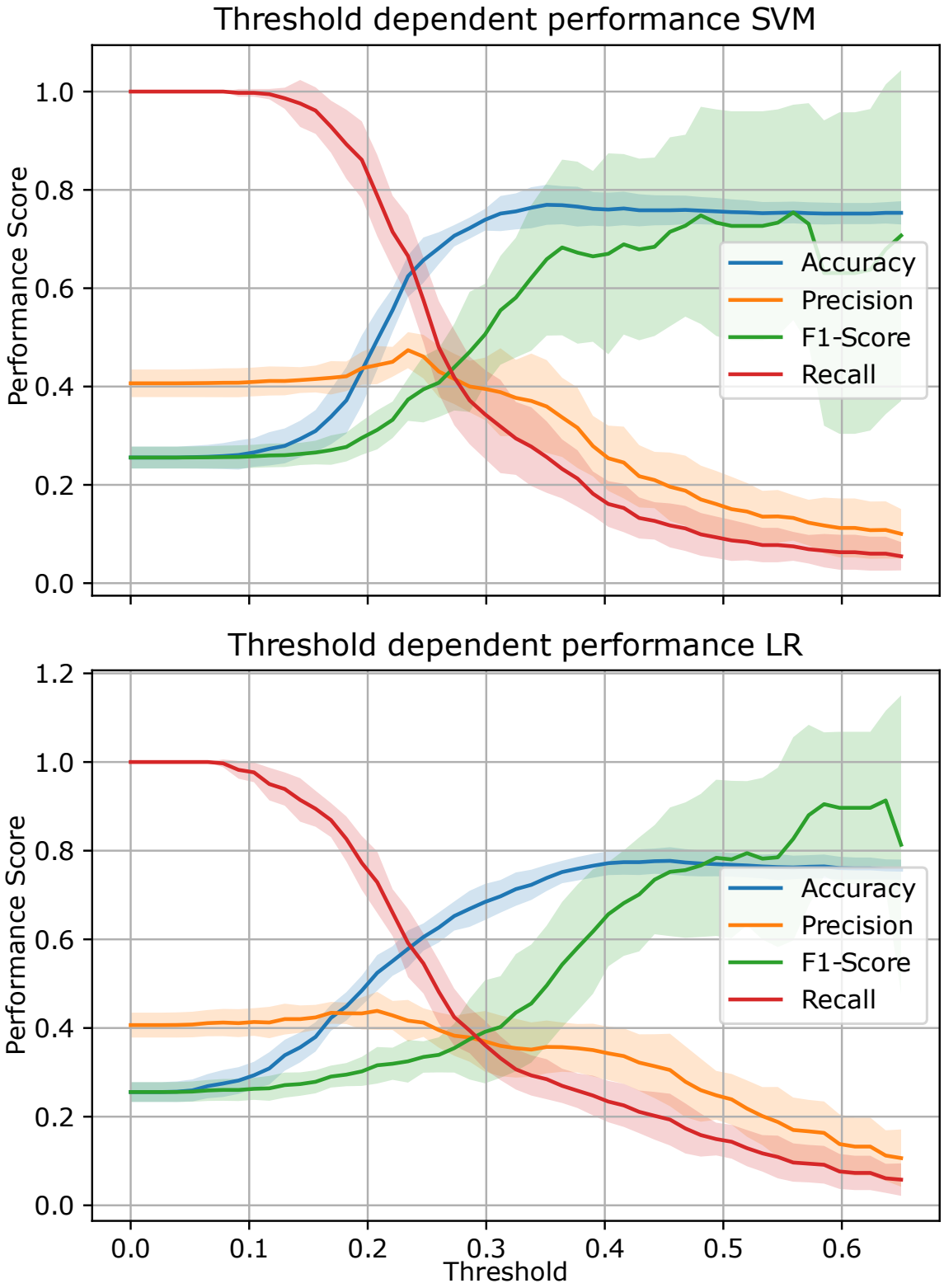

Figure C.8: The two plots represent the variance of the performance scores over the classification threshold for SVM and LR, observed during 10-fold cross validation. They illustrate how strong the chosen threshold is influencing the model performance and that its choice is depending a lot on the training data. The large variance of the F1-Score makes this measure not suited for threshold selection. The accuracy for both models is maximal at a value around 0.35.

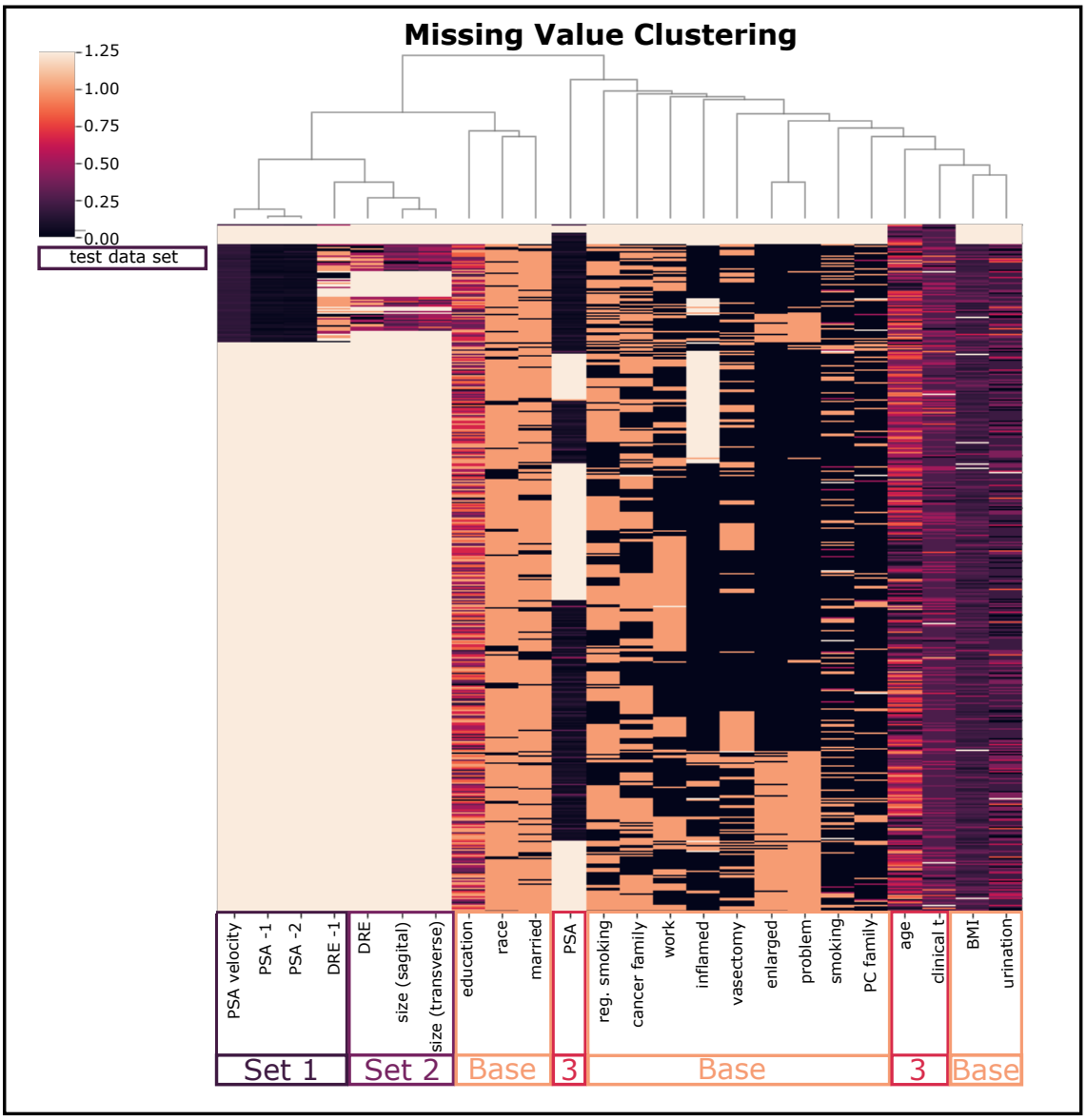

Figure C.9: The clustering plot is showing the features selected for $D_{complete}$. There values are min-max-scaled and missing values are assigned 1.25. One can see how 3 subsets can be distinguished through clustering on different patient cohorts. They have been selected, with the aim of increasing the amount of patients showing a full input vector for these subsets. To each of these informative clusters the baseline is assigned.

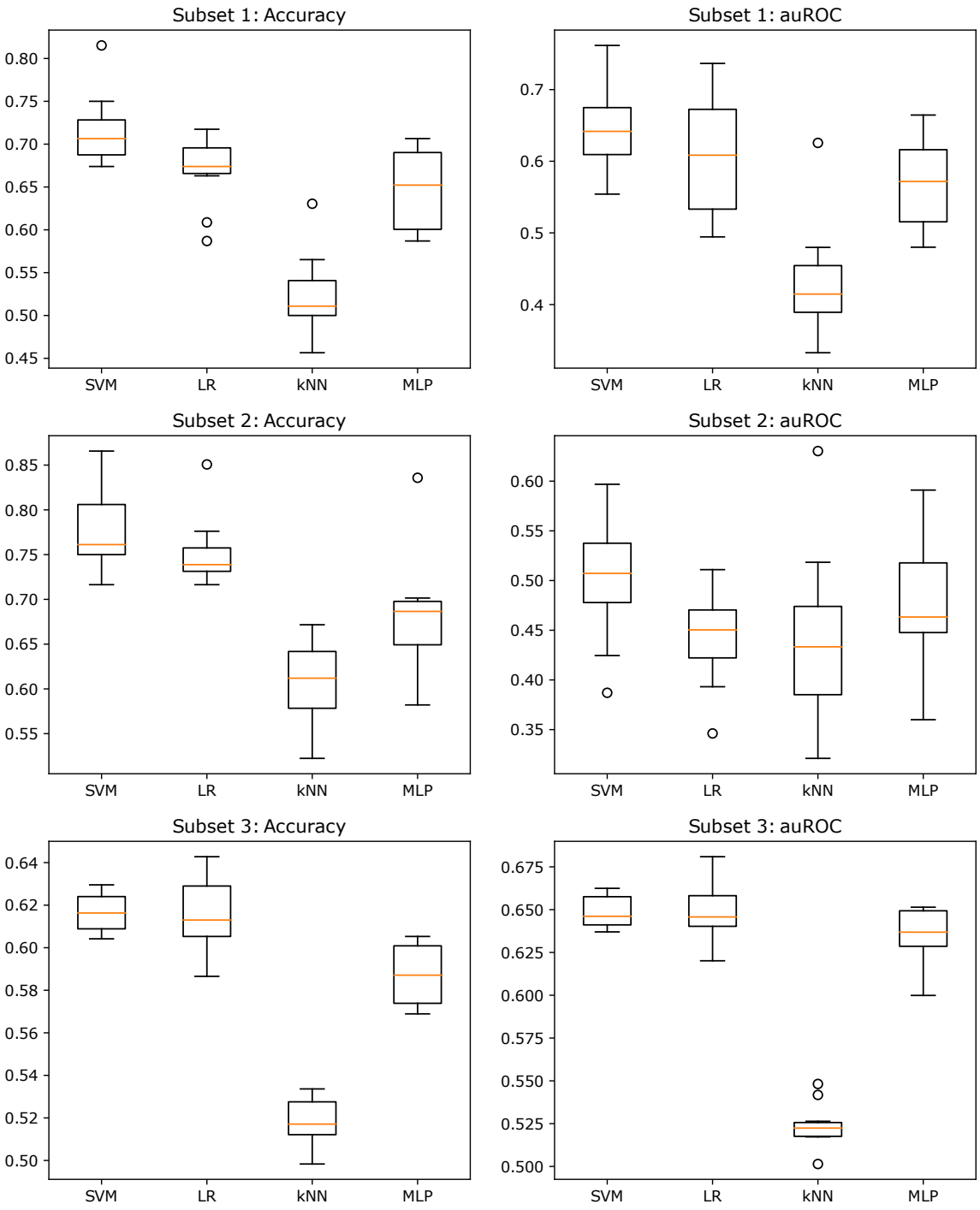

Figure C.10: The Figure is representing Boxplots of the performance scores auROC and Accuracy for the different machine learning models on the three manually selected feature subsets. In the Boxplots the medians and the interquartile ranges of the cross-validation are visualised.

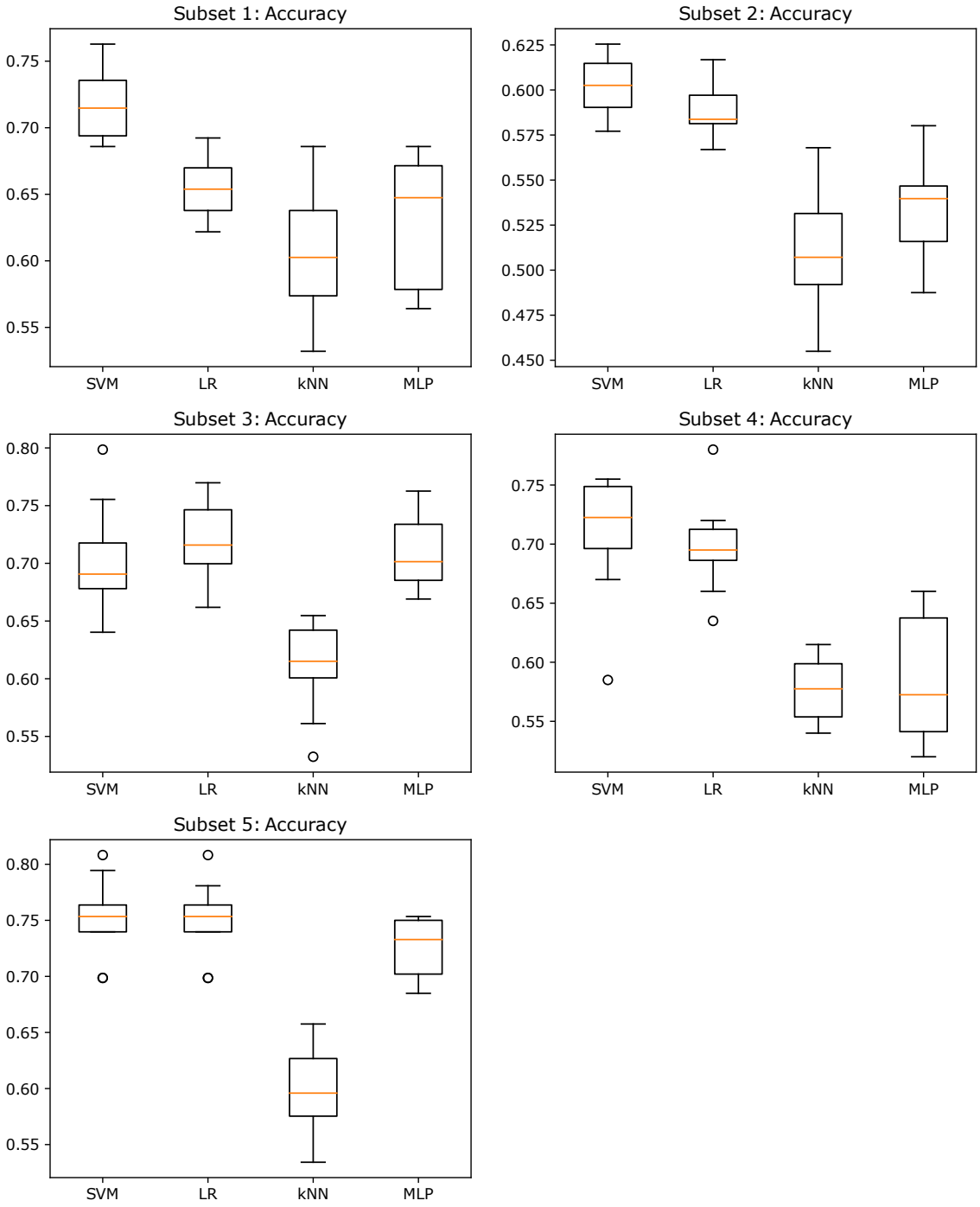

Figure C.11: The Figure is representing Boxplots of the accuracy for the different machine learning models on the five feature subsets, selected by algorithmic feature rank and redundancy comparison. In the Boxplots the medians and the interquartile ranges of the cross-validation are visualised.

D Bibliography

[1] Pieter Cullis. *The Personalized Medicine Revolution: How Diagnosing and Treating Disease Are About to Change Forever*. Greystone Books Ltd, 2015. Google-Books-ID: oOyWBQAAQBAJ.

[2] Michael Grieves. Origins of the Digital Twin Concept. In *Origins of the Digital Twin Concept*, 2016.

[3] Digital Twin in der Medizin: Mit KI und Confidential Computing zu besseren Behandlungsmethoden. https://business-services.heise.de/specials/moderne-it-infrastruktur/home/beitrag/digital-twin-in-der-medizin-mit-ki-und-confidential-computing-zu-besseren-behandlungsmethoden-4184.

[4] 3 Great Examples of Digital Twin Technology In Action. https://www.challenge.org/insights/digital-twin-examples/.

[5] Ignacio Rojas, Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, and Francisco Ortuño, editors. *Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 68, 2020, Proceedings*, volume 12108 of *Lecture Notes in Computer Science*. Springer International Publishing, 2020.

[6] Cecilio Angulo, Luis Gonzalez-Abril, Cristóbal Raya, and Juan Antonio Ortega. A Proposal to Evolving Towards Digital Twins in Healthcare. In Ignacio Rojas, Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, and Francisco Ortuño, editors, *Bioinformatics and Biomedical Engineering*, Lecture Notes in Computer Science, pages 418–426. Springer International Publishing, 2020.

[7] Shaip. How ai will power the next wave of healthcare innovation. https://becominghuman.ai/how-ai-will-power-the-next-wave-of-healthcare-innovation-695a2196aae8, 2021.

[8] Micha Woniak, Manuel Graña, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014.

[9] Stephen W. Leslie, Taylor L. Soon-Sutton, Hussain Sajjad, and Larry E. Siref. Prostate cancer. In *StatPearls*. StatPearls Publishing, 2022.

[10] C. Börgermann, H. Loertzer, H.-J. Luboldt, P. Hammerer, P. Fornara, M. Graefen, and H. Rübben. Psa quo vadis? *Der Urologe*, 48(9):1008–1017, 2009.

[11] James D. Brierley, Mary K. Gospodarowicz, and Christian Wittekind. *TNM Classification of Malignant Tumours*. Wiley, 8 edition, 2017.

[12] Aminu Bello, Ben Vandermeer, Natasha Wiebe, Amit X. Garg, and Marcello Tonelli. Evidence-Based Decision-Making 2: Systematic Reviews and Meta-Analysis. In Patrick S. Parfrey and Brendan J. Barrett, editors, *Clinical Epidemiology*, volume 2249 of *Methods in Molecular Biology*, pages 405–428. Springer US, 2021.

[13] Tae Won Yi, Sine Donnellan, and Adeera Levin. Evidence-Based Decision Making 4: Clinical Practice Guidelines. In Patrick S. Parfrey and Brendan J. Barrett, editors, *Clinical Epidemiology*, volume 2249 of *Methods in Molecular Biology*, pages 455–466. Springer US, 2021.

[14] E.H. Shortliffe, B.G. Buchanan, and E.A. Feigenbaum. Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. *Proceedings of the IEEE*, 67(9):1207–1224, 1979.

[15] Randolph A. Miller and Antoine Geissbuhler. Diagnostic Decision Support Systems. In Eta S. Berner, editor, *Clinical Decision Support Systems: Theory and Practice*, Health Informatics, pages 99–125. Springer, 2007.

[16] Hesham Salem, Daniele Soria, Jonathan N. Lund, and Amir Awwad. A systematic review of the applications of Expert Systems (ES) and machine learning (ML) in clinical urology. *BMC Medical Informatics and Decision Making*, 21:223, 2021.

[17] Jared M. Campbell, Elspeth Raymond, Michael E. O'Callaghan, Andrew D. Vincent, Kerri R. Beckmann, David Roder, Sue Evans, John McNeil, Jeremy Millar, John Zalcberg, Martin Borg, and Kim L. Moretti. Optimum Tools for Predicting Clinical Outcomes in Prostate Cancer Patients Undergoing Radical Prostatectomy: A Systematic Review of Prognostic Accuracy and Validity. *Clinical Genitourinary Cancer*, 15(5):e827–e834, 2017.

[18] Katharina Boehm, Alessandro Larcher, Burkhard Beyer, Zhe Tian, Derya Tilki, Thomas Steuber, Pierre I. Karakiewicz, Hans Heinzer, Markus Graefen, and Lars Budäus. Identifying the Most Informative Prediction Tool for Cancer-specific Mortality After Radical Prostatectomy: Comparative Analysis of Three Commonly Used Preoperative Prediction Models. *European Urology*, 69(6):1038–1043, 2016.

[19] Shahrokh F. Shariat, Pierre I. Karakiewicz, Guilherme Godoy, and Seth P. Lerner. Use of nomograms for predictions of outcome in patients with advanced bladder cancer. *Therapeutic Advances in Urology*, 1(1):13–26, 2009.

[20] Michael Grieves and John Vickers. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In Franz-Josef Kahlen,

Shannon Flumerfelt, and Anabela Alves, editors, *Transdisciplinary Perspectives on Complex Systems*, pages 85–113. Springer International Publishing, 2017.

[21] Angelo Croatti, Matteo Gabellini, Sara Montagna, and Alessandro Ricci. On the Integration of Agents and Digital Twins in Healthcare. *Journal of Medical Systems*, 44(9):161, 2020.

[22] Robert Kender, Florian Kaufmann, Felix RöSSler, Bernd Wunderlich, Dimitri Golubev, Ingo Thomas, Anna-Maria Ecker, Sebastian Rehfeldt, and Harald Klein. Development of a digital twin for a flexible air separation unit using a pressure-driven simulation approach. *Computers & Chemical Engineering*, 151:107349, 2021.

[23] Gary White, Anna Zink, Lara Codecá, and Siobhán Clarke. A digital twin smart city for citizen feedback. *Cities*, 110:103064, 2021.

[24] Qiuchen Lu, Xiang Xie, Ajith Kumar Parlikad, and Jennifer Mary Schooling. Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. *Automation in Construction*, 118:103277, 2020.

[25] Sakshi Piplani, P. Singh, D. Winkler, and N. Petrovsky. In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. *Scientific reports*, 2021.

[26] H. Lehrach and A. Ionescu. The Future of Health Care: Deep data, smart sensors, virtual patients and the Internet-of-Humans. https://www.semanticscholar.org/paper/The-Future-of-Health-Care2016.

[27] Maged N. Kamel Boulos and Peng Zhang. Digital Twins: From Personalised Medicine to Precision Public Health. *Journal of Personalized Medicine*, 11(8):745, 2021.

[28] Kalyanasundaram Subramanian. Digital Twin for Drug Discovery and DevelopmentThe Virtual Liver. *Journal of the Indian Institute of Science*, 100(4):653–662, 2020.

[29] Suraj Pawar, Shady E. Ahmed, Omer San, and Adil Rasheed. Hybrid analysis and modeling for next generation of digital twins. *Journal of Physics: Conference Series*, 2018(1):012031, 2021.

[30] Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Rodero, Filip Loncaric, Yingjing Feng, Andrew Gilbert, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, Manuel Villegas Martinez, Mariana Sousa Santos, Mehrdad Shamohammdi, Hongxing Luo, Philip Westphal, Paul Leeson,

Paolo DiAchille, Viatcheslav Gurev, Manuel Mayr, Liesbet Geris, Pras Pathmanathan, Tina Morrison, Richard Cornelussen, Frits Prinzen, Tammo Delhaas, Ada Doltra, Marta Sitges, Edward J Vigmond, Ernesto Zacur, Vicente Grau, Blanca Rodriguez, Espen W Remme, Steven Niederer, Peter Mortier, Kristin McLeod, Mark Potse, Esther Pueyo, Alfonso Bueno-Orovio, and Pablo Lamata. The Digital Twin to enable the vision of precision cardiology. *European Heart Journal*, 41(48):4556–4564, 2020.

[31] Michael G. Kapteyn, Jacob V. R. Pretorius, and Karen E. Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1(5):337–347, 2021.

[32] Ulf-Håkan Stenman, Per-Anders Abrahamsson, Gunnar Aus, Hans Lilja, Chris Bangma, Freddie C. Hamdy, Laurent Boccon-Gibod, and Peter Ekman. Prognostic value of serum markers for prostate cancer. *Scandinavian Journal of Urology and Nephrology*, 2009.

[33] ACR ESUR AdMeTech. PI-RADS. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/PI-RADS, 2019.

[34] NIH. Clinical practice guidelines. https://www.nccih.nih.gov/health/providers/clinicalpractice, 2022.

[35] AWMF. S3-leitlinie prostatakarzinom. https://www.awmf.org/leitlinien/detail/ll/043-022OL.html, 2021.

[36] Donna P. Ankerst, Johanna Straubinger, Katharina Selig, Lourdes Guerrios, Amanda De Hoedt, Javier Hernandez, Michael A. Liss, Robin J. Leach, Stephen J. Freedland, Michael W. Kattan, Robert Nam, Alexander Haese, Francesco Montorsi, Stephen A. Boorjian, Matthew R. Cooperberg, Cedric Poyet, Emily Vertosick, and Andrew J. Vickers. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. *European Urology*, 74(2):197–203, 2018.

[37] Shahrokh F Shariat, Michael W Kattan, Andrew J Vickers, Pierre I Karakiewicz, and Peter T Scardino. Critical review of prostate cancer predictive tools. *Future Oncology*, 5(10):1555–1584, 2009.

[38] Christopher R. Porter, Eduard J. Gamito, E. David Crawford, Georg Bartsch, Joseph Charles Presti, Ashutosh Tewari, and Colin ODonnell. Model to predict prostate biopsy outcome in large screening population with independent validation in referral setting. *Urology*, 65(5):937–941, 2005.

[39] Satoshi Nitta, Masakazu Tsutsumi, Shotaro Sakka, Tsuyoshi Endo, Kenichiro Hashimoto, Morikuni Hasegawa, Takayuki Hayashi, Koji Kawai, and Hiroyuki Nishiyama. Machine learning methods can more efficiently predict prostate

cancer compared with prostate-specific antigen density and prostate-specific antigen velocity. *Prostate International*, 7(3):114–118, 2019.

[40] Zan Ke, Liang Wang, Xiang-De Min, Zhao-Yan Feng, Zhen Kang, Pei-Pei Zhang, Ba-Sen Li, Hui-Juan You, and Sheng-Chao Hou. Diagnostic Performance and Interobserver Consistency of the Prostate Imaging Reporting and Data System Version 2: A Study on Six Prostate Radiologists with Different Experiences from Half a Year to 17 Years. *Chinese Medical Journal*, 131:1666, 2018.

[41] Kensaku Kawamoto, Caitlin Houlihan, Andrew Balas, and David Lobach. Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success. *BMJ (Clinical research ed.)*, 330:765, 2005.

[42] Fritz H Schröder, Jonas Hugosson, Monique J Roobol, Teuvo L J Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Liisa Määttänen, Hans Lilja, Louis J Denis, Franz Recker, Alvaro Paez, Chris H Bangma, Sigrid Carlsson, Donella Puliti, Arnauld Villers, Xavier Rebillard, Matti Hakama, Ulf-Hakan Stenman, Paula Kujala, Kimmo Taari, Gunnar Aus, Andreas Huber, Theo H van der Kwast, Ron H N van Schaik, Harry J de Koning, Sue M Moss, and Anssi Auvinen. Screening and prostate cancer mortality: Results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *The Lancet*, 384(9959):2027–2035, 2014.

[43] Steven M. Schwartz, Kevin Wildenhaus, Amy Bucher, and Brigid Byrd. Digital Twins and the Emerging Science of Self: Implications for Digital Health Experience Design and Small Data. *Frontiers in Computer Science*, 2, 2020.

[44] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2):325–340, 2018.

[45] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation and active learning. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, pages 231–238. MIT Press, 1994.

[46] Yong Liu and Xin Yao. Negatively Correlated Neural Networks Can Produce Best Ensembles. *Australian Journal of Intelligent Information Processing Svstems*, page 10, 1997.

[47] E. K. Tang, P. N. Suganthan, and X. Yao. An analysis of diversity measures. *Machine Learning*, 65(1):247–271, 2006.

[48] Mohamed M. Elshrif and E. Fokoue. Random Subspace Learning (RASSEL) with data driven weighting schemes. *Mathematics for Application*, 2018.

[49] Kelly C. Chang, Sara Dutta, Gary R. Mirams, Kylie A. Beattie, Jiansong Sheng, Phu N. Tran, Min Wu, Wendy W. Wu, Thomas Colatsky, David G. Strauss, and Zhihua Li. Uncertainty quantification reveals the importance of data variability and experimental design considerations for in silico proarrhythmia risk assessment. *Frontiers in Physiology*, 8, 2017.

[50] Six Steps of Data Analysis Process. https://www.geeksforgeeks.org/six-steps-of-data-analysis-process/, 2021.

[51] Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing Predictive Models Using Electronic Medical Records: Challenges and Pitfalls. *American Medical Informatics Association*, AMIA Annual Symposium Proceedings.:7, 2013.

[52] Lars Nielsen. A Checklist for Data pre-processing before you build your Machine Learning Model. https://towardsdatascience.com/a-checklist-for-data-pre-processing-before-you-build-your-machine-learning-model-91d2d04dc53f.

[53] NIH. Prostate - Datasets - PLCO - The Cancer Data Access System. https://cdas.cancer.gov/datasets/plco/20/, 2022.

[54] Matthieu Komorowski, Dominic Marshall, Justin Salciccioli, and Yves Crutain. Exploratory Data Analysis. In *Secondary Analysis of Electronic Health Records*, pages 185–203. Springer International Publishing, 2016.

[55] Wanfu Gao, Liang Hu, and Ping Zhang. Feature redundancy term variation for mutual information-based feature selection. *Applied Intelligence*, 2020.

[56] Peter Y. Chen and P. M. Popovich. *Correlation: Parametric and Nonparametric Measures*. SAGE University Paper Series on Quantitative Applications in the Social Sciences, 2002.

[57] Jason Brownlee. Data Preparation for Machine Learning. https://machinelearningmastery.com/data-preparation-for-machine-learning/, 2022.

[58] Lei Yu and Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, page 8, 2003.

[59] Shinichi Nakagawa and Holger Schielzeth. Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews*, 85(4):935–956, 2010.

[60] Christoph Molnar. Chapter 5 Interpretable Models — Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/simple.html, 2022.

[61] Xin Jin, Anbang Xu, Rongfang Bie, and Ping Guo. Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. In Jinyan Li, Qiang Yang, and Ah-Hwee Tan, editors, *Data Mining for Biomedical Applications*, pages 106–115. Springer, 2006.

[62] U. Fayyad and K. Irani. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *IJCAI*, 1993.

[63] Jacek Biesiada and Wlodzisaw Duch. Feature Selection for High-Dimensional Data A Pearson Redundancy Based Filter. In Marek Kurzynski, Edward Puchala, Michal Wozniak, and Andrzej Zolnierek, editors, *Computer Recognition Systems 2*, volume 45 of *Advances in Soft Computing*, pages 242–249. Springer Berlin Heidelberg, 2007.

[64] Lei Yu and Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)*, page 8, 2003.

[65] Huimin Zhao, Atish P. Sinha, and Wei Ge. Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications: An International Journal*, 36(2):2633–2644, 2009.

[66] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[67] Xiulong Yang, Hui Ye, Yang Ye, Xiang Li, and Shihao Ji. Generative Max-Mahalanobis Classifiers for Image Classification, Generation and More. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, Lecture Notes in Computer Science, pages 67–83. Springer International Publishing, 2021.

[68] Kai Ming Ting, Jonathan R. Wells, Swee Chuan Tan, Shyh Wei Teng, and Geoffrey I. Webb. Feature-subspace aggregating: Ensembles for stable and unstable learners. *Machine Learning*, 82(3):375–397, 2011.

[69] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 32.41 edition edition, 2015.

[70] 978-3-319-20010-1 Kubat. *An Introduction to Machine Learning*. Springer International Publishing, 1 edition, 2015.

[71] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors. *Optimization for Machine Learning*. Neural Information Processing Series. MIT Press, 2011.

[72] Jean-Emmanuel Bibault, Steven Hancock, Mark K. Buyyounouski, Hilary Bagshaw, John T. Leppert, Joseph C. Liao, and Lei Xing. Development and Validation of an Interpretable Artificial Intelligence Model to Predict 10-Year Prostate Cancer Mortality. *Cancers*, 13(12):3064, 2021.

[73] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*, 21(1):128–138, 2010.

[74] Miroslav Kubat. Performance Evaluation. In Miroslav Kubat, editor, *An Introduction to Machine Learning*, pages 213–233. Springer International Publishing, 2015.

[75] John A. Swets, Robyn M. Dawes, and John Monahan. Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest*, 1(1):1–26, 2000.

[76] Marina Skurichina and Robert P. W. Duin. Bagging and the Random Subspace Method for Redundant Feature Spaces. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 1–10. Springer, 2001.

[77] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature Selection for Classication: A Review. *Data classification: Algorithms and applications*, page 33, 2014.

[78] I. Gheyas and Leslie S. Smith. Feature subset selection in large dimensionality domains. *Pattern Recognit.*, 2010.

[79] Jianping Hua, W. Tembe, and E. Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognit.*, 2009.

[80] Chen Liao, Shutao Li, and Zhiyuan Luo. Gene Selection Using Wilcoxon Rank Sum Test and Support Vector Machine for Cancer Classification. In Yuping Wang, Yiu-ming Cheung, and Hailin Liu, editors, *Computational Intelligence and Security*, pages 57–66. Springer, 2007.

[81] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[82] L. Rocchi, L. Chiari, and A. Cappello. Feature selection of stabilometric parameters based on principal component analysis. *Medical & Biological Engineering & Computing*, 42(1):71–79, 2004.

[83] Lior Rokach. Genetic algorithm-based feature set partitioning for classification problems. *Pattern Recognition*, 41(5):1676–1700, 2008.

[84] Yongjun Piao, Minghao Piao, Cheng Hao Jin, Ho Sun Shon, Ji-Moon Chung, Buhyun Hwang, and Keun Ho Ryu. A New Ensemble Method with Feature Space Partitioning for High-Dimensional Data Classification. *Mathematical Problems in Engineering*, 2015:1–12, 2015.

[85] Christoph Molnar. 3.1 Importance of Interpretability — Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/interpretability-importance.html, 2022.

[86] K. Turner and N. Oza. Decimated input ensembles for improved generalization. *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, 1999.

[87] Marina Skurichina and Robert P. W. Duin. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.

[88] Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

[89] Jason Brownlee. A Gentle Introduction to Mixture of Experts Ensembles. https://machinelearningmastery.com/mixture-of-experts/, 2021.

[90] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[91] Dusan Popovic, Alejandro Sifrim, Jesse Davis, Yves Moreau, and Bart De Moor. Problems with the nested granularity of feature domains in bioinformatics: The eXtasy case. *BMC Bioinformatics*, 16(4):S2, 2015.

[92] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[93] Thomas G. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2):139–157, 2000.

[94] J. V. Leeuwen, J. Kittler, F. Roli, and J. V. Leeuwen. Multiple Classifier Systems. In *Lecture Notes in Computer Science*, 2001.

[95] Fabio Roli, Giorgio Giacinto, and Gianni Vernazza. Methods for Designing Multiple Classifier Systems. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, pages 78–87. Springer, 2001.

[96] D. Partridge and W. B. Yates. Engineering multiversion neural-net systems. *Neural Computation*, 8(4):869–893, 1996.

[97] Giorgio Giacinto and Fabio Roli. An approach to the automatic design of multiple classifier systems. *Pattern Recognition Letters*, 22(1):25–33, 2001.

[98] James Surowiecki. *The wisdom of crowds*. Anchor Books, nachdr. edition, 2005.

[99] M. Erp, L. Vuurpijl, and Lambert Schomaker. An overview and comparison of voting methods for pattern recognition. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, 2002.

[100] So Young Sohn and Hong Choi. Ensemble Based on Data Envelopment Analysis. In *Ensemble Based on Data Envelopment Analysis*, page 9, 2001.

[101] Jakob Hansen. Combining Predictors. *DAIMI Report Series*, 29, 2000.

[102] Jason Brownlee. Ensemble Learning Algorithms With Python. https://machinelearningmastery.com/ensemble-learning-algorithms-with-python/.

[103] L. Kuncheva, J. Bezdek, and R. Duin. Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recognit.*, 2001.

[104] Philip Chan and Salvatore Stolfo. Experiments on Multistrategy Learning by Meta-Learning. *In Proc. Second Intl. Conference on Info. and Knowledge Mgmt*, 1995.

[105] Dr M Usha Rani and G T Prasanna Kumari. A Comparison of Arbiter and Combiner Trees. *International Journal of Engineering and Innovative Technology (IJEIT)*, 1(6):6, 2012.

[106] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.

[107] Saso Deroski and Bernard enko. Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, 54(3):255–273, 2004.

[108] Michal Wozniak and Marcin Zmyslony. Combining classifiers using trained fuser – analytical and experimental results. *Neural Network World*, 20(7):925–934, 2010.

[109] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham,

Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy ai development: Mechanisms for supporting verifiable claims. https://arxiv.org/abs/2004.07213, 2020.

[110] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[111] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

[112] Andrea (AZQ) Haring. *S3-Leitlinie Prostatakarzinom*. AWMF, 2021.

[113] Conor O'Sullivan. Interpretability in Machine Learning. https://towardsdatascience.com/interpretability-in-machine-learning-ab0cf2e66e1, 2020.

[114] Christoph Molnar. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/, 2022.

[115] Koen Bruynseels, Filippo Santoni de Sio, and Jeroen van den Hoven. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Frontiers in Genetics*, 9:31, 2018.

[116] Hossein Soleimani, James Hensman, and Suchi Saria. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1948–1963, 2018.

[117] Hossein Soleimani, Adarsh Subbaswamy, and Suchi Saria. Treatment-Response Models for Counterfactual Reasoning with Continuous-time, Continuous-valued Interventions. *uai2017*, page 10, 2017.

[118] J. Masison, J. Beezley, Y. Mei, Hal Ribeiro, A. C. Knapp, L. Sordo Vieira, B. Adhikari, Y. Scindia, M. Grauer, B. Helba, W. Schroeder, B. Mehrad, and R. Laubenbacher. A modular computational framework for medical digital twins. *Proceedings of the National Academy of Sciences*, 118(20):e2024287118, 2021.

[119] Frederike Kaltheuner. *Fake AI*. Meatspace Press, 2021.

[120] Dong Wook Kim, H. Jang, K. Kim, Youngbin Shin, and S. Park. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean journal of radiology*, 2019.

[121] Chaitanya Jee Srivastava. Basics of Machine Learning - Definition, Concepts, Types of Algorithms. https://www.naukri.com/learning/articles/basics-of-machine-learning/, 2022.

[122] Michael G. Kapteyn. *PhD Thesis - Michael G. Kapteyn*. phdthesis, Massachusetts Institute of Technology - Center for Computational Science and Engineering, 2021.

E Acknowledgements

This master thesis would not have been possible without the help, support and scientific input of a large number of different people. I am therefore very grateful to them and would like to acknowledge their contribution to this work.

First of all, I would like to thank Prof. Dr. Matthias Weidemüller, head of the Quantum Dynamics group at the Physics Institute of the University of Heidelberg, for giving me the opportunity to research this exciting topic with so much personal responsibility and scientific freedom. I am very thankful for having such a motivating, fascinated and sharp-minded supervisor.

I especially want to thank Prof. Dr. Markus Hohenfellner, Medical Director of the Urological University Clinic Heidelberg, for the great cooperation and hospitable instruction into clinical processes. Additionally, I would like to thank him very much for creating this exciting, collaborative framework for my masters thesis as he is the consortium leader of the project Clinic5.1 Comprehensive Lifesciences Neural Information Computing.

I would also like to thank Prof. Dr. Björn Ommer, head of the Machine Vision & Learning group at the Ludwig-Maximilian University of Munich, for the lively discussions and helpful suggestions.

I would especially like to thank my colleague Carlos Brandl. I am very grateful for the goal-oriented team work and the joint brainstorming, through which the ideas that form the essential core of this work could emerge.

In particular, I would like to thank the team of research assistants who supported the project and thus also my work from the programming side. This includes Robert Maiwald, Fabian Egersdörfer, Patrick Leibersperger and Atanas Aleksandrov.

I would also like to thank Dr. Magdalena Görtz, Matthias Rath and Martina Heller for the support provided by the medical expertise. Additionally I thank Pingchuan Ma for his IT expertise.

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den (Datum) .....