How to deal with Uncertainty in Machine Learning for Medical Imaging?

Christina Gillmann *
Leipzig University

Dorothee Saur
Leipzig University, Medical Centre

Gerik Scheuermann
Leipzig University

Abstract

Recently, machine learning is massively on the rise in medical applications providing the ability to predict diseases, plan treatment and monitor progress. Still, the use in a clinical context of this technology is rather rare, mostly due to the missing trust of clinicians. In this position paper, we aim to show how uncertainty is introduced in the machine learning process when applying it to medical imaging at multiple points and how this influences the decision-making process of clinicians in machine learning approaches. Based on this knowledge, we aim to refine the guidelines for trust in visual analytics to assist clinicians in using and understanding systems that are based on machine learning.

1 Introduction and Background

Machine learning is known as the automatic generation of knowledge [28]. Since its start in the 1950s, these classes of algorithms became more and more popular in a variety of applications such as mechanical engineering, biology, and medicine [9]. This effect is strengthened by the increasing popularity of neural networks, which are a big subgroup of machine learning [11].

Especially in medical applications, machine learning becomes increasingly important as it provides the ability to predict diseases or segment organs [14]. Examples of successful uses of machine learning are lesion segmentation [17], tumor segmentation [32] and skin disease determination [30].

Still, research is centered around a further use of machine learning to improve diagnosis, drug discovery, personalized medicine, smart health records, and clinical trials. These developments can be seen as a revolution of the healthcare system induced by the use of machine learning [10, 13, 31] and are known as one of the major recent challenges in medical visualization [19].

Although the massive potential of machine learning in medical applications is known, there is a lack of transfer of such novel techniques into the clinical daily routine [51]. This is due to a variety of factors that are also coupled with legal restrictions, as shown by Maack et al. [35]. Medical software is considered to be a medical device and therefore underlies hard restrictions for real-world use in many countries. Besides these legal restrictions, machine learning approaches form a black box that is hard to interpret due to a large number of parameters that are adjusted during the learning process. Here, clinicians tend to reject these types of algorithms, as they are not able to understand the decision-making process of the neural network, but are still responsible for the decisions they make based on the provided systems [2]. This is a very specific problem in the medical domain, as the decisions of clinicians have a great effect on a patient's life. The effect is that clinicians do not desire to be directed by systems they do not fully understand.

Explainable artificial intelligence (XAI) aims to help users to understand the learning process of machine learning algorithms. Troja and Guan [52] showed a state of the art analysis of artificial intelligence in the medical context and summarized the remaining challenges. Here, they state that uncertainties in the machine learning process are an open problem that leads to the missing applicability of machine learning approaches in medical imaging. The effect of uncertainty in decision-making processes has been shown by Sacha et al [47] and guidelines to make use of visual analytics to create trust have been developed.

In this manuscript, we shed light on the general machine learning process and see how uncertainty-aware visual analytics can drive its use in medical imaging. Based on this, we aim to summarize potential sources of uncertainty in the machine learning process that will occur when applying machine learning in medical imaging. For these sources, we aim to define dependencies and check if the sources can be quantified. Further, we will revisit the guidelines formulated by Sacha et al. and refine them based on the uncertainty analysis conducted in this paper.

This paper contributes:

- Summary of the machine learning cycle in medical imaging
- Sources of uncertainty in the machine learning cycle in medical imaging
- Guideline to handle these sources of uncertainty based on visual analytics

2 The Machine Learning Pipeline in Medical Imaging

Independent from the application, machine learning is performed using a specific cycle [22], as shown in Figure 1. This cycle consist of three major parts: Data, Model and Deployment. Please note, that there exist ambiguous descriptions of the machine learning cycle. We selected the following one, as it is abstract enough to be applied in most machine learning settings in medical imaging.

Each category will be briefly explained in the following.

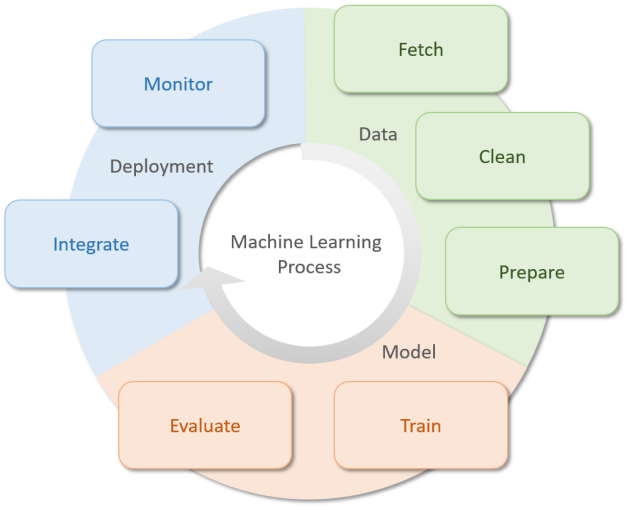


Figure 1: The machine learning process. The cycle contains 3 steps: Data, Model and Deployment, including their subcategories.

*e-mail: gillmann@informatik.uni-leipzig.de

2.1 Data

The data step is the first entry point in the machine learning process, aiming to build a data basis that can be used as ground truth for the machine learning process. In medicine, this mostly refers to health records such as medical images, lab results, or doctoral reports. In this area, data can often be stored distributed, or even analog, which means that a processing of data is required to make use of it. The data process can be separated into three steps: fetch, clean, and prepare.

Fetch In the step of fetching, the goal is to gather a medical image that can be used for machine learning. In the medical context, this usually includes the fetching of patient-related data such as medical images, medication plans, treatment outcomes, or demographic data. Especially in the medical context, this may also include the digitalization of data. In many countries, medical data can be acquired manually and stored offline for data security reasons. In addition, medical data is often stored distributed, which may require data fusion from different sources.

Clean The cleaning stage of the machine learning pipeline may lead to datasets that need to be excluded or completed. In the medical context, this can result in datasets that cannot be used for model training as it does not fit the given requirements. Especially in medicine, each clinic can have different data acquisition protocols that output different medical records. In addition, for example, demographic data is often ubiquitous or street names can be written differently, although they are referring to the same address. Further, in medicine images can be acquired at different time steps, but do not need to. Here, the data needs to be cleaned such that all data underlies the same requirements.

Prepare In the preparation step, data needs to be manipulated when it fits a given machine learning model. Here, several steps may be required. On one hand, transformations such as transforming all data records into the same coordinate system is a typical preprocessing task for medical image analysis. Normalization is often required as well, as medical images can be acquired using different scales on different devices.

2.2 Model

After the data acquisition step, the selected machine learning model is trained and evaluated. Usually, the gathered datasets are separated into training and testing datasets [38]. This step involved the derived dataset from the first step as an input.

Train Depending on the selected machine learning model, the model needs to be trained. Here, the training dataset is used to train the selected model based on the determined ground truth. Here, a proper model needs to be selected that fits the training dataset and the defined ground truth. In the medical context, image segmentation is an important application, where U-Nets have been developed specifically [43].

Evaluate After training the model needs to be evaluated. Here, the testing dataset is used to evaluate the performance of the trained network. Here, different metrics can be used to quantify the performance of the network.

2.3 Deployment

In the deployment phase, the goal is to provide an accessible and integrated version of the trained model. This phase separates into two steps: integrate and monitor. The challenges in the medical area to achieve deployment of these technologies have been summarized by Kelly et al. [27] and will be described in the following.

Integrate Training a machine learning algorithm is usually achieved in a very protected environment regarding the data that is used for training. Especially in medicine, many conditions can occur that may vary from the setting that has been used during the training stage. For example, if a heart is analyzed by a neural network that was only trained with uninjured rips visible in the image, an image that contains injured ribs may not output useful results. Here, it is important to investigate, if real-world conditions match the trained network. This may also include the adaptation or standardization of image acquisition techniques in the clinical daily routine.

Monitor After integration, the model needs to be monitored to check if its performance remains stable during real-world conditions. In addition, the model can be refined if the performance needs to be increased. In the area of medical imaging, this is an important issue that needs to be considered every time the image acquisition process changes.

3 Basics on Uncertainty

As we aim for an analysis of sources of uncertainty in the machine learning pipeline, we would like to give a brief background on the definition, quantification, and processing of uncertainty. Here, we provide basics on the theory that will be required in the rest of the manuscript.

3.1 Definition of Uncertainty

When a measurement a' is performed on a measurand $a \in (-\infty, \infty)$ with true value a^*. Most of the time a^* and a' differ from each other by an error $e = |a^* - a'|$. This error is the sum of various effects, like measurement inaccuracy, as some form of sensor captured the measurement. Therefore, a ground-truth a^* is needed to be able to calculate such a deviation from the real value.

The *uncertainty* of a measurement is a quantification of doubt, in particular the description of a specific *uncertainty event*, about the measurement result [23]. The uncertainty is either known, making the measurand *uncertainty-aware*, or unknown, leading to an uncertain measurand.

As stated above, uncertainty springs from various sources that are subdivided into *types* of uncertainty events, as shown in Fig. 2. Generally, uncertainty can be divided into objective uncertainty, meaning that it can be quantified, and subjective uncertainty that cannot be quantified. Objective uncertainty is further separated into epistemic uncertainty, arising from the model itself, and aleatoric uncertainty, stemming from the underlying data. Subjective uncertainty can either be rule uncertainty, treating the doubt about a rule, or moral uncertainty, dealing with the ethical correctness of a rule.

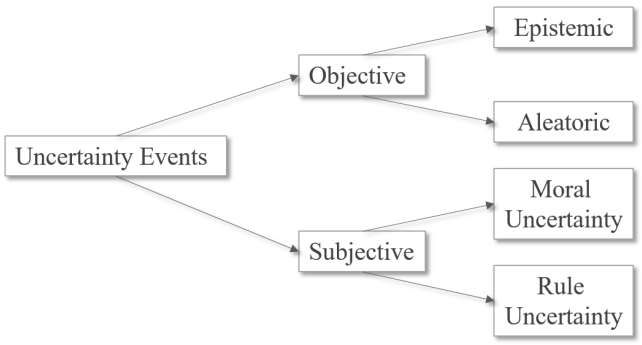


Figure 2: Types of uncertainty events as shown by Souza et al. [42]. Events can either be objective or subjective, where objective uncertainty events can be epistemic or aleatoric and subjective uncertainty events can be moral uncertainty and rule uncertainty.

3.2 Quantification of Uncertainty

For uncertainty events, one can determine if the event is *quantifiable*. Here, Lo and Mueller [33] defined five levels of quantifiability:

- Level 1: Complete Certainty
- Level 2: Risk without Uncertainty
- Level 3: Fully Reducible Uncertainty
- Level 4: Partially Reducible Uncertainty
- Level 5: Irreducible Uncertainty

Here, Level 1 refers to events that have a clear outcome that is not variable. For these events, there does not exist any variability in the event. Level 2 refers to uncertainty events that are fully known and quantifiable. In particular, this refers to known probability distributions of potential outcomes. Further, Level 3 refers to uncertainty events that are not completely known. Here, potential outcomes are known, but the probability distribution is not. This might be reducible when including more knowledge. Resulting from this, they can be quantified partially. In contrast, Level 4 refers to uncertainty events where neither the potential outcomes nor the probability distribution of the outcomes is fully known. Still, they might be known when more knowledge is included. At last, Level 5 describes uncertainty events that are not quantifiable in their potential outcomes, nor by their probability distributions, independent of the knowledge.

Uncertainty can be *described* throughout arbitrary approaches, where bounded uncertainty and probabilistic distributions are the most common.

For bounded uncertainty [39] there exist an interval around the measurand that can be defined as: $u_B(a) = [a' - u, a' + u]$. This description of uncertainty is chosen when it is not important how the occurrences of a measurand are distributed. Instead, it is important to know what are the limits in this variation [3].

In the case of probabilistic distribution functions [34] $u_{PDB}(a)$ the measurand usually defines the most probable location of the true value that was captured. The most prominent choice of probabilistic distribution functions are Gaussian distribution functions, but in general, any distribution can be used to express uncertainty, including generalized linear models, Poisson distribution, and count-based models [24].

3.3 Propagation and Accumulation of Uncertainty

Data is mostly propagated through mathematical operations O. These operations do not solely affect the data, but also the attached uncertainty. Besides, mathematical operations are affected by the uncertainty of their operands. This results in the need to adjust mathematical operations to be able to handle uncertainty. There exist a variety of techniques to achieve this, which are mostly inspired by error propagation [12].

The accumulation of uncertainty can in principle be achieved by arbitrary accumulation functions. Cai et al. [8] presented a survey of aggregation functions. In the machine learning process, a proper aggregation function needs to be able to properly aggregate all sources of uncertainty in the machine learning cycle and allows the user to adjust the importance of all sources of uncertainty in the machine learning cycle. This is required, as users may need to determine which sources of uncertainty are more important than others or even discard specific sources.

4 Sources of Uncertainty in the Machine Learning Pipeline

When applying machine learning to medical imaging, each step in the machine learning pipeline is affected by uncertainty and needs to be tackled [1]. Most machine learning approaches in the context of medicine make use of medical image data. This type of data has been shown to hold a high amount of uncertainty [18]. In this section, we aim to summarize the sources of uncertainty during this process.

There exist a variety of taxonomies of sources of uncertainty that are related to general uncertainty analysis and potential visualization strategies. Schunn et al [50] provided an extensive taxonomy of types of uncertainty. We use this work as a starting point and selected the sources that are relevant in the area of machine learning in medical imaging. Boukhelifa [4] et al. provided a user evaluation that revealed how important the sensemaking of uncertainty is for users. MacEachren et al [36] and Pang [40] et al provided visualization strategies for different sources of uncertainty. We aim to highlight the benefit of uncertainty-aware visual analytics throughout this manuscript.

At this point, we aim to highlight that there may occur sources of uncertainty, that are independent of the underlying domain. We still aim to list these and show which impact they have explicitly on the medical domain.

4.1 Data

In the data stage, the sources of uncertainty are mainly originating from the image processing pipeline that is executed to collect and prepare the data to train the selected machine learning algorithm, as shown in the work of Gillmann et al [18]. A summary of all sources of uncertainty in the data step is shown in Table 1. The sources will be summarized in the following.

Step	Source	Label	Level	Dependencies
Fetch	Positional Uncertainty	1.1.1	2	-
	Value uncertainty	1.1.2	2	-
	Incompleteness of data	1.1.3	2	-
Clean	Manipulation Uncertainty	1.2.1	2	1.1
	Exclusion Uncertainty	1.2.2	2	1.1
Prepare	Model Inaccuracy	1.3.1	3	1.1, 1.2
	Model Incompleteness	1.3.2	3	1.1, 1.2
	Model Parameter Uncertainty	1.3.3	3	1.1, 1.2
	Labeling Uncertainty	1.3.4	3	1.1, 1.2

Table 1: Sources of Uncertainty in the Data step. The sources are enumerated to provide consistent referencing. The level of uncertainty and the respective category are included.

Fetch During the fetch step, the data that is selected contains three types of uncertainty. All these uncertainties are of level 2, which means that these uncertainties are known and quantifiable. When starting the machine learning process, these sources start without dependencies.

First, the dataset can contain *positional uncertainty*. This often occurs in medical imaging datasets such as Ultrasound, where the position of the acquisition device is tracked [21]. In addition, positional uncertainty is often an issue, when multiple modalities are acquired for machine learning [48].

Next, *value uncertainty* arises in principally all acquired medical datasets. Technically, all measured values can contain uncertainty, as the measurement process is achieved by a variety of different sensors that may lead to uncertain values. Especially in medical imaging, pixel or voxel values can be affected by uncertainty caused

by the partial volume effect or voxel bleeding, which results from the reconstruction process [44].

Last, the *incompleteness of data* in medical records is a further source of uncertainty. Medical records are often acquired at specific points in time and everything that happens in between is unknown [46]. In addition, different clinics have different image acquisition devices that have varying capabilities. Here, acquisition steps may be incomplete depending on the clinic it took place.

Clean The cleaning step introduces two different types of uncertainties: manipulation uncertainty and exclusion uncertainty [6, 29]. These uncertainties are of type 2 and can usually be quantified. Unfortunately, they have dependencies with all sources of uncertainty from the prior fetch step.

First, the *manipulation of data* introduces uncertainty. If values are missing or clear outliers, a proper strategy needs to be found that completes or smoothes the data, which introduces uncertainty. Especially in medicine, this is an important step, as often many datasets need to be excluded due to prior diseases or inappropriate data collection.

In addition, the cleaning step can introduce uncertainty in the machine learning cycle as the decision if a dataset is *excluded or not* is performed based on a present metric. This can be affected by uncertainty, as it might not be clear if the metric can cover all cases that need to be excluded or if it excluded too many approaches.

Prepare In the preparation step, the sources of uncertainty mainly originate from the used algorithms that transform the collected data such that it can be processed in the selected machine learning model. Here, *model inaccuracy*, *model incompleteness*, and *model parameter uncertainty* are sources of uncertainty. Models are never able to map reality perfectly and thus introduce uncertainty. This is amplified by the fact that models cannot be complete by their definition, which also introduces uncertainty. These sources of uncertainty result are of type three which means that the uncertainty is known, but the probability distribution is not. They depend on the uncertainties that arise from 1.11 and 1.2, as the decision of models is related to the outcome of the fetching and cleaning step.

In addition, the preparation step introduces uncertainty in the machine learning pipeline while *labeling data*. Especially in medicine data is usually labeled to be used for machine learning. Unfortunately, this process is affected by uncertainty as well. This is due to the nature of medical data and flaws in the resulting labels. Often, multiple diseases can occur or doctors themselves cannot separate diseases clearly. In addition, location tasks such as determining a tumor in an organ are affected by uncertainty, as the underlying data might not give a clear separation between healthy and diseased tissue. This leads to fuzzy labels introducing uncertainty. This source of uncertainty is of type 4, as the label is usually made by a clinician and the resulting uncertainty cannot be quantified properly.

4.2 Model

In the model stage, the sources of uncertainty are manifold and mainly originate from the selected model that needs to be trained in the machine learning process. An overview of all sources can be found in Table 2. They will be explained in the following.

Both training and testing data uncertainty originate from the dataset and need to be properly separated such that the machine learning algorithm can learn features properly, allowing the testing dataset to test the learned features properly. Especially in medicine, it is important to separate the medical cases such that the model can learn all occurring conditions of patients properly. This uncertainty is of type 3 and depends on the uncertainties arising from the data step.

Train After separating the data, the machine learning model can be trained with the developed training dataset. Here, the model itself introduces *model and parameter uncertainty*, as the choice of a proper model is uncertain itself and models are not able to replicate the real world entirely. Medicine provides a variety of data that usually focuses on multiple aspects. This means that a proper algorithm for machine learning needs to be selected.

In addition, the *training uncertainty* describes, if a network is trained well enough or should be improved and to what extend. There are usually several metrics used to determine if a model needs further training. Still, these metrics are a source of uncertainty, as it is not clear if there might be a more optimal learning procedure.

Evaluate After training, the model needs to be evaluated using the test dataset. Here, the *evaluation uncertainty* is a source of uncertainty arising from the fact that evaluation is only properly possible when using proper evaluation data and setups. Upon all the possible settings, the question arises if the current chosen setup can check the performance of a machine learning algorithm.

Model evaluation is also accomplished using extitevaluation metrics. Like in the training step, these metrics are a source of uncertainty. In medicine many metrics are available, but the question is which one fits best in the given case [25].

All sources of uncertainty during training and evaluation are of level three which means that they are known, but the probability distribution is unknown. They are connected to the separation uncertainty in the respective category.

Step	Source	Label	Level	Dependencies
Train & Evaluate	Separation Uncertainty	2.1	3	1
Train	Parameter Uncertainty	2.2.1	3	2.1
	Model Inaccuracy	2.2.2	3	2.1
	Training Uncertainty	2.2.3	3	2.1
Evaluate	Evaluation Uncertainty	2.3.1	3	2.1
	Metric Uncertainty	2.3.1	3	2.1

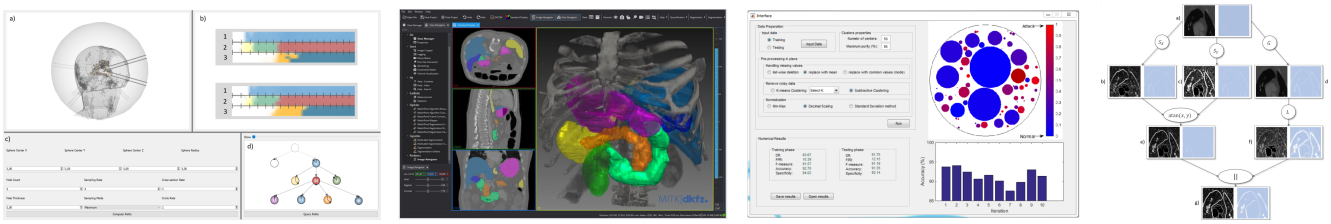
Table 2: Sources of Uncertainty in the Model step. The sources are enumerated to provide consistent referencing. The level of uncertainty and the respective category are included.

4.3 Deployment

In the deployment stage, uncertainty sources are rather inhomogeneous and can be subject to various effects. Table 3 shows an overview of these sources. They will be summarized in the following.

Data	Source	Label	Level	Dependencies
Integrate	Similarity Uncertainty	3.1.1	3	1,2
	Fitting Uncertainty	3.1.2	3	1,2
Monitor	Perceptual/ Cognitive Uncertainty	3.2.1	4	3.1
	Decision Making Bias	3.2.2	4	3.1
	Refinement Metric Uncertainty	3.2.3	3	1,2

Table 3: Sources of Uncertainty in the Deployment step. The sources are enumerated to provide consistent referencing. The level of uncertainty and the respective category are included.



(a) (b) (c) (d)

Figure 3: Examples of uncertainty-aware visual analytics in medical applications. a) Uncertainty-aware visual analytics to assist keyhole surgeries [16]. b) MITK tool with available systems functions [20]. c) Sensemaking in uncertainty-aware visual analytics [26]. d) Provenance visualization for uncertainty-aware image processing [15].

Integrate In the integration step, the uncertainty of the real-world setting is *similar* enough to fit the trained model is an important problem. Especially in medicine, where so many different conditions of patients can occur and clinics run different scanners and treatment protocols, these similarity needs to be ensured. This source of uncertainty is of level 3 and depends on the uncertainties from steps 1 and 2.

In addition, the *fitting uncertainty* describes the potential of the provided machine learning model to address the needs of the clinical environment. The daily clinical routine can be very inhomogeneous while issues might arise in the case of an emergency. Here, it is not certain if the developed machine learning approach fits the given setting. This is difficult to quantify, depending on the uncertainties arising in the data and model step.

Monitor In the monitoring step, several sources of uncertainty can occur.

First, *cognitive and perceptual uncertainty* can be introduced by the user, interpreting the machine learning results. Especially in medicine, clinicians are responsible for their decisions and therefore need to understand how machine learning algorithms make up their decisions.

Machine learning approaches might output a result not meeting the clinician's expectations of the clinician. In many cases, clinicians are left with their intuition on how to decide on a proper treatment, which has been build throughout their education and experience. In related machine learning approaches, clinicians discarded results from the data as they may not fit the expected outcome of the clinician.

Cognitive/perceptual and decision-making bias uncertainty are of level four as it is related to subjective human behavior which is hard to quantify. It is based on the uncertainty of the integration step of the deployment phase.

At last, the monitoring step requires metrics to estimate machine learning approach *refinement*. Again, metrics are a source of uncertainty as they may not be optimal to express algorithm refinement needs. As other metric-based uncertainties are of level 3. This uncertainty depends on the uncertainties of the data and model steps.

5 Guidelines, Challenges, and Examples for Medical Application

We have shown that the machine learning cycle is affected by a variety of sources of uncertainty in each step when being applied in the medical area. In the following, we will show how they can assist in providing useful visualization strategies for machine learning in medical imaging based on visual analytics. Explainable Artificial Intelligence (XAI) has been shown to assist users and developers in understanding machine learning approaches, but specific guidelines and rules are not available so far to achieve this goal.

Sacha et al. [47] provided a set of guidelines that are required to generate trust using visual analytics approaches. Namely, these are:

- Set up an uncertainty-aware visual analytics cycle (G1)
  - Quantify Uncertainties in Each Component (G1.1)
  - Propagate and Aggregate Uncertainties (G1.2)
  - Visualise Uncertainty Information (G1.3)
  - Enable Interactive Uncertainty Exploration (G1.4)
- Make the Systems Functions Accessible (G2)
- Support the Analyst in Uncertainty Aware Sensemaking (G3)
- Analyse Human Behaviour to Derive Hints on Problems and Biases (G4)
- Enable Analysts to Track and Review their Analysis (G5)

In this section, we aim to show the implications of these suggestions to the machine learning process in medical imaging. We grouped the first four guidelines by Sacha et al. into one guideline, as it can be seen as a general setup of an uncertainty-aware visual analytics cycle. For each guideline, we will summarize the guideline applied to medical applications, resulting in challenges, and give examples.

Preim and Lawonn provided an overview of visual analytics approaches in public health [41], showing that the use of visual analytics in medical imaging is a prominent example. Uncertainty-aware visual analytics is less common, mostly due to a missing workflow to generate these approaches. Examples can be found for radiation therapy [37], surgery assistance [16] and fiber tracking analysis [5]. An example of uncertainty-aware visual analytics in medical applications was given by Gillmann et al [16]. Here, a holistic tool to plan keyhole surgeries allows reviewing the probability of a surgery tunnel to affect a certain structure in the human body was provided as shown in Figure 3(a).

Still, their application to machine learning approaches is an open problem. This results from a missing generalized tool that allows exploring the design space of uncertainty visualization in medical imaging. Here, at least a library such as the visualization toolkit [49] would be beneficial to drive the development of uncertainty-aware visual analytics in medical imaging.

G1: Set up an uncertainty-aware visual analytics cycle The development of uncertainty-aware visual analytics cycles can be summarized by the four first guidelines of Sacha et al. They will be explained briefly in the following.

G1.1: Quantify Uncertainties in Each Component. In section 4, we showed that each step of the machine learning pipeline can introduce uncertainty into the machine learning process. We also showed, that not all of these sources can be quantified or completely quantified. Still, we recommend declaring all relevant sources of uncertainty in a given machine learning process, checking if they are quantifiable. For the remaining sources of uncertainty, the open challenge is to find proper quantification approaches.

To be able to review the quantified sources of uncertainty, visual analytics can be of great benefit as it may help developers and users in medicine to understand how individual steps of the image processing pipeline are affected by uncertainty.

G1.2: Propagate and Aggregate Uncertainties. When running a machine learning cycle, the sources of uncertainty are propagated and aggregate along the processing pipeline. Here, we suggest implementing proper uncertainty propagation and aggregation approaches such that users can identify the amount of uncertainty that is currently inherent in a respective component. An open challenge in this context is to determine which propagation and aggregation approaches are the most suitable in the medical context.

G1.3: Visualise Uncertainty Information. Based on G2, the amount of uncertainty inherent in each component of the machine learning process needs to be visualized to allow developers and users a quick understanding of the uncertainty. Although we showed that several uncertainty-aware visualization approaches exist, there remains an open challenge of testing which ones are the most effective in what scenario.

G1.4: Enable Interactive Uncertainty Exploration. During the machine learning cycle, multiple sources of uncertainty are introduced. A visual analytics approach that assists in understanding these sources and how they propagate and accumulate is required. Again, as in G1.3, the investigation of interaction techniques and their effectiveness is an open challenge.

G2: Make the Systems Functions Accessible Medical data analysis is not performed by the clinician itself most of the time. This can result in the rejection of a novel image processing technique as the clinician cannot follow the computations and understand how results are generated. Especially for machine learning approaches, which often act as a black box system, this does not provide trust in the made computations. Here, visualization approaches such as the MITK [20] (Medical Imaging Interaction Toolkit) are required to show medical users how computations are processed and how parameters can influence a computational result. In this tool, users are enabled to apply image operations and follow the made computations visually. An application to machine learning in medical imaging of these approaches remains an open challenge.

G3: Support the Analyst in Uncertainty Aware Sensemaking When clinicians make decisions based on a computational system they need to know how reliable and trustworthy these decisions are. Uncertainty-aware visual analytics can be of great benefit in this process as it allows clinicians to estimate the trustworthiness of the decision. Karami et al. [26] provided a visual tool for sensemaking in visual analytics, as shown in Figure 3(c). Still, this approach needs to be transferred to machine learning applications, which can be formulated as an open challenge.

G4: Analyse Human Behaviour to Derive Hints on Problems and Biases The effects of cognitive biases in the medical area are well-known [7] and subject to research. In medicine, this effect is strong as the decision on therapy is made by a clinician who is reviewing certain data that is affected by uncertainty. Here, clinicians often need to rely on their experience. Therefore, the inclusion of visualization in this area is desirable as it can indicate biases and provide visual reasoning of the run methods and the results.

G5: Enable Analysts to Track and Review their Analysis We have shown that uncertainty is inherent in any step of the machine learning pipeline and that it accumulates when running multiple steps. Here, researchers need to be able to understand this procedure with a proper visualization strategy. The understanding of how uncertainties arise and develop throughout multiple computations is a task of provenance. Xu et al. [53] provided a state of the art analysis of visualization approaches that assist in understanding provenance. Davidson et al. [45] showed the importance of provenance when considering data that is affected by uncertainty. Here, a useful combination of visualization approaches that target uncertainty in medical data is desired and describes an open challenge. An example of such approaches was given by Gillmann et al [15] showing how arbitrary image processing pipelines can be processed while reviewing the development of the uncertainty.

6 Conclusion

In this paper, we showed that the machine learning process is affected by a variety of sources of uncertainty that can affect the decision-making process of clinicians. We provided a taxonomy of uncertainties attached to each step of the machine learning cycle. Using this taxonomy we provide guidelines to make use of uncertainty-aware visual analytics while using machine learning cycles in medical applications. In addition, we provide successful examples and open challenges in this application area.

References

[1] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangooei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare, M. Panahiazar, S. Nahavandi, D. Srinivasan, A. F. Atiya, and U. R. Acharya. Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991-2020), 2020.

[2] O. Asan, E. Bayrak, and A. Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, 22:e15154, 06 2020. doi: 10.2196/15154

[3] G. Belforte, B. Bona, and V. Cerone. Bounded measurement error estimates: Their properties and their use for small sets of data. *Measurement*, 5(4):167 – 175, 1987.

[4] N. Boukhelifa, M.-E. Perrin, S. Huron, and J. Eagan. *How Data Workers Cope with Uncertainty: A Task Characterisation Study*, p. 3645–3656. Association for Computing Machinery, New York, NY, USA, 2017.

[5] R. Brecheisen, B. Platel, A. Vilanova, and B. t. Haar Romeny. Illustrative uncertainty visualization of dti fiber pathways. *The Visual Computer*, 2012.

[6] J. Broeck, S. Cunningham, R. Eeckels, and A. Herbst. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS medicine*, 2:e267, 11 2005. doi: 10.1371/journal.pmed.0020267

[7] L. P. Busby, J. L. Courtier, and C. M. Glastonbury. Bias in radiology: The how and why of misses and misinterpretations. *RadioGraphics*, 38(1):236–247, 2018. PMID: 29194009. doi: 10.1148/rg.2018170107

[8] S. Cai, B. Gallina, D. Nyström, and C. Seceleanu. Data aggregation processes: a survey, a taxonomy, and design guidelines. *Computing*, 11 2018.

[9] J. Carbonell, R. Michalski, T. Mitchell, and J. Carbonell. Machine learning: A historical and methodological analysis. *Artificial Intelligence Magazine*, 4, 08 2002.

[10] K. Feldman, L. Faust, X. Wu, C. Huang, and N. V. Chawla. Beyond volume: The impact of complex healthcare data on the machine learning pipeline. In *Towards integrative machine learning and knowledge extraction*, pp. 150–169. Springer, 2017.

[11] A.-I. Georgevici and M. Terblanche. Neural networks and deep learning: a brief introduction. *Intensive Care Medicine*, 45, 02 2019. doi: 10.1007/s00134-019-05537-w

[12] C. D. Ghilani. Statistics and adjustments explained-part 3: Error propagation. *Surveying and Land Information Science*, 64:29–34, 2004.

[13] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren. Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158:113–122, 2018. doi: 10.1016/j.cmpb.2018.01.025

[14] F. Gille, A. Jobin, and M. Ienca. What we talk about when we talk about trust: Theory of trust for ai in healthcare. *Intelligence-Based Medicine*, 1:100001, 2020.

[15] C. Gillmann, Arbelaez, P., Hernandez, J., H. Hagen, and Wischgoll, T. An uncertainty-aware visual system for image pre-processing. *Journal of Imaging*, 4(9), 2018. doi: 10.3390/jimaging4090109

[16] C. Gillmann, R. G. C. Maack, Post, T., T. Wischgoll, and H. Hagen. An uncertainty-aware workflow for keyhole surgery planning using hierarchical image semantics. *Visual Informatics*, 2(1):26 – 36, 2018. Proceedings of PacificVAST 2018. doi: 10.1016/j.visinf.2018.04.004

[17] C. Gillmann, L. Peter, C. Schmidt, D. Saur, and G. Scheuermann. Visualizing multimodal deep learning for lesion prediction. *IEEE Computer Graphics and Applications*, 41(5):90–98, 2021. doi: 10.1109/MCG.2021.3099881

[18] C. Gillmann, D. Saur, T. Wischgoll, and G. Scheuermann. Uncertainty-aware Visualization in Medical Imaging - A Survey. *Computer Graphics Forum*, 2021. doi: 10.1111/cgf.14333

[19] C. Gillmann, N. N. Smit, E. Gröller, B. Preim, A. Vilanova, and T. Wischgoll. Ten open challenges in medical visualization. *IEEE Computer Graphics and Applications*, 41(5):7–15, 2021. doi: 10.1109/MCG.2021.3094858

[20] C. J. Goch, J. Metzger, and M. Nolden. Abstract: Medical research data management using mitk and xnat. In K. H. Maier-Hein, geb. Fritzsche, T. M. Deserno, geb. Lehmann, H. Handels, and T. Tolxdorff, eds., *Bildverarbeitung für die Medizin 2017*, pp. 305–305. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.

[21] H.-E. Gueziri, M. Mcguffin, and C. Laporte. Visualizing positional uncertainty in freehand 3d ultrasound. *Proceedings - Society of Photo-Optical Instrumentation Engineers*, 9036, 03 2014.

[22] H. Hapke and C. Nelson. *Building Machine Learning Pipelines: Automating Model Life Cycles with TensorFlow*. O'Reilly Media, Incorporated, 2020.

[23] S. W. Hasinoff, F. Durand, and W. T. Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560, June 2010.

[24] T. M. Hegel, S. A. Cushman, J. Evans, and F. Huettmann. Current state of the art for statistical modelling of species distributions. In *Spatial complexity, informatics, and wildlife conservation*, pp. 273–311. Springer, 2010.

[25] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa. On evaluation metrics for medical applications of artificial intelligence. *medRxiv*, 2021. doi: 10.1101/2021.04.07.21254975

[26] A. Karami. A framework for uncertainty-aware visual analytics in big data. In *CEUR Workshop Proceedings*, vol. 1510, pp. 146–155. CEUR Workshop Proceedings, 2015.

[27] C. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 12 2019. doi: 10.1186/s12916-019-1426-2

[28] M. Kubat. *An Introduction to Machine Learning*. Springer Publishing Company, Incorporated, 1st ed., 2015.

[29] C. Lee and H.-J. Yoon. Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, 36:3–11, 03 2017. doi: 10.23876/j.krcp.2017.36.1.3

[30] H. Li, Y. Pan, J. Zhao, and L. Zhang. Skin disease diagnosis with deep learning: a review, 2020.

[31] S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017. doi: 10.1016/j.visinf.2017.01.006

[32] Z. Liu, L. Chen, L. Tong, F. Zhou, Z. Jiang, Q. Zhang, C. Shan, Y. Wang, X. Zhang, L. Li, and H. Zhou. Deep learning based brain tumor segmentation: A survey, 2020.

[33] A. Lo and M. T. Mueller. Warning: Physics envy may be hazardous to your wealth! *IO: Regulation*, 2010.

[34] D. P. Loucks and E. van Beek. *An Introduction to Probability, Statistics, and Uncertainty*, pp. 213–300. Springer International Publishing, Cham, 2017.

[35] R. G. C. Maack, D. Saur, H. Hagen, G. Scheuermann, and C. Gillmann. Towards Closing the Gap of Medical Visualization Research and Clinical Daily Routine. In C. Gillmann, M. Krone, G. Reina, and T. Wischgoll, eds., *VisGap - The Gap between Visualization Research and Visualization Software*. The Eurographics Association, 2020. doi: 10.2312/visgap.20201107

[36] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005. doi: 10.1559/1523040054738936

[37] D. Maleike, J. Unkelbach, and U. Oelfke. Simulation and visualization of dose uncertainties due to interfractional organ motion. *Physics in medicine and biology*, 51:2237–52, 06 2006.

[38] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[39] C. Olston and J. D. Mackinlay. Visualizing data with bounded uncertainty. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, INFOVIS '02, p. 37. IEEE Computer Society, USA, 2002.

[40] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13, 10 1996. doi: 10.1007/s003710050111

[41] B. Preim and K. Lawonn. A survey of visual analytics for public health. *Computer Graphics Forum*, 39:1–35, 11 2019. doi: 10.1111/cgf.13891

[42] R. Rocha Souza, A. Dorn, B. Piringer, and E. Wandl-Vogt. Towards a taxonomy of uncertainties: Analysing sources of spatio-temporal uncertainty on the example of non-standard german corpora. *Informatics*, 6(3), 2019. doi: 10.3390/informatics6030034

[43] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[44] O. G. Rousset, Y. Ma, and A. C. Evans. Correction for partial volume effects in pet: Principle and validation. *Journal of Nuclear Medicine*, 39(5):904–911, 1998.

[45] S. Roy. Provenance and uncertainty. 2012.

[46] S. Rässler, D. B. Rubin, and E. R. Zell. 19 incomplete data in epidemiology and medical statistics. In C. Rao, J. Miller, and D. Rao, eds., *Epidemiology and Medical Statistics*, vol. 27 of *Handbook of Statistics*, pp. 569–601. Elsevier, 2007. doi: 10.1016/S0169-7161(07)27019-1

[47] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2016. doi: 10.1109/TVCG.2015.2467591

[48] M. Schlachter, T. Fechter, M. Jurisic, T. Schimek-Jasch, O. Oehlke, S. Adebahr, W. Birkfellner, U. Nestle, and K. Bühler. Visualization of deformable image registration quality using local image dissimilarity. *IEEE Transactions on Medical Imaging*, 35(10):2319–2328, 2016.

[49] W. Schroeder, K. Martin, and B. Lorensen. *The Visualization Toolkit*. Kitware, 2006.

[50] C. Schunn and J. Trafon. *The psychology of uncertainty in scientific data analysis*. 11 2012.

[51] P. Shah, F. Kendall, S. Khozin, R. Goosen, J. Hu, J. Laramie, M. Ringel, and N. Schork. Artificial intelligence and machine learning in clinical development: a translational perspective. *npj Digital Medicine*, 2(1):1–5, July 2019. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Computer science;Translational research Subject_term_id: computer-science;translational-research. doi: 10.1038/s41746-019-0148-3

[52] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–21, 2020. doi: 10.1109/tnnls.2020.3027314

[53] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch. Survey on the analysis of user interactions and visualization provenance. *Computer Graphics Forum*, 39(3):757–783, 2020. doi: 10.1111/cgf.14035