GAZE-INDUCED QUALITY CONTROL IN GEOLOGICAL VOXEL MODELS

MSc Thesis Artificial Intelligence

written by

B. E. Zijlstra
(5774284)

Department of Information and Computing Sciences

Faculty of Science

at the *Utrecht University.*

Date of the public defense:
April 24th, 2018

Supervisors:
Dr Dr E. L. van den Broek (Utrecht University)
Dr P.-P. van Maanen (TNO)
Dr A.J. Feelders (Utrecht University, second reader)

Utrecht University

Abstract

Knowledge of the subsurface is an important aspect for a countries welfare. To gain an understanding of this, geologists use boreholes combined with interpolating methods to build statistical 3D models. Due to the stochastic properties of these models, domain experts perform a quality control procedure to find errors, which can be a time consuming endeavor.

In this thesis, we looked into methods for predicting areas of errors in GeoTOP, a geological voxel model. Firstly, we show that a previously used Attention Model performs well when we optimize parameters for each participant, but with an AUC of 0.61, the algorithm lacked finding optimal parameters for the combined participants. We show that variance among experts in assessing errors is high, making generalizing predictions hard. Secondly, we showed that entropy, the voxel models quantification of uncertainty, is not a good indicator of where errors occur. With an average AUC of 0.54, where some participants scored even under 0.5, we show that there is no relation between entropy and the assessment of experts. Finally, we introduced a Velocity-Threshold Identification (I-VT) algorithm combined with tree-based classifiers and showed that with an AUC of 0.8 over each participant, errors can be found regardless of the differences among participants. We show why finding optimal parameters for fixation algorithms is difficult due to a lack of ground truth, but despite that our new algorithm performs better and faster, allowing for real-time error predictions. These findings suggest that a geologist combined with our introduced algorithm can decrease their time spent on quality control. Furthermore, this thesis can provide as a framework for other fields with a similar problem description, such as radiologists looking for malignancies.

Contents

1 Introduction

For a country that exploits its subsurface as much as the Netherlands, a thorough geological understanding of the earth beneath is crucial. While this knowledge is mostly obscure among citizens, the Dutch society as a whole benefits immensely from applications rooted in geology. The Dutch subsurface and the Dutch part of the North Sea have been a significant contributor to the countries' wealth, due to a myriad of natural resources: natural gas, oil, geothermal energy, clean water, and pure silica [1]. Moreover, the high density of population of the Netherlands demands a good use of the limited space, making a reliable insight into the subsurface a necessity. For example, a lack of knowledge of the subsurface during the construction of tunnels and underground cables is a critical factor in failure costs [2]. Knowledge of the subsurface can also aid in agriculture since natural soils are variable in their properties [3].

This intense use of the subsurface is not without side effects, however. For example, subsiding of land is a problem that can occur due to mineral and gas distillation. Agricultural practices can contaminate shallow groundwater, threatening drink water quality. Recently, earthquakes in the province of Groningen [4] in the north of the Netherlands due to gas extraction [5] had both environmental and social impact [6], leading to a national debate on the extraction of gas. To model the risk of future earthquakes in this area, knowledge of the geological characteristics of the subsurface is needed [7].

For over a century, the knowledge accumulation of the Dutch subsurface is performed by the Geological Survey of the Netherlands (GDN) and its predecessors. At first, this task was limited to traditional maps on paper, nowadays these are replaced mainly with 3D subsurface models. While the rise of computation power gave way to sophisticated 3D models storing much more data than its predecessors, a large part of its construction is still human work. First, geologists start with collecting a substantial amount of data of the subsurface. The Data and Information of the Dutch Subsurface (DINO) of the GDN is the archive which stores thousands of samples of drillings and cone penetration tests. Secondly, these samples function as the input for interpolation, creating a model of the layers of the earth, the characteristics of the rock type, and the chemical properties. This interpolation combines geostatistical methods with preprogrammed information of the geological knowledge that domain experts provide. Thirdly, geologists perform a quality control routine to check whether areas are modeled correctly. This is a time-consuming task, and requires years of experience in the field.

In 2014, a bill by the Ministry of Infrastructure and the Environment was passed [8], stating information about the Dutch subsurface shall be registered centrally. Due to their expertise, this task is allocated to the GDN, increasing their workload substantially. This combination of the increased amount of data and a task needing lots of experience has led to an interest in finding new ways to improve the workflow of these geological models.

One of the potential areas for improvement is the quality control of the 3D models. As of now, this process is performed entirely by humans. Geologists have to assess the entire area for errors in geomodelling, mark these areas, and describe their findings. This is a laborious process since there can be numerous errors. However, advancements in Artificial Intelligence (AI) have led to the emerging of several fields, models, and systems facing similar tasks, which can

provide inspiration for automated quality control in geological models. Among these are decision support systems (DSS) [9], intelligence amplification (IA) [10], and computer-aided diagnosis (CADx) [11], human-in-the-loop (HITL) [12][13]. There is a similar principle in these fields, namely the idea of using both the benefits of a computer and of a human.

For computers to interact with humans in these frameworks, input is needed for the computer to process. Since the task of quality control is dependent on the visual senses, we look at eye movements in this thesis to look into possibilities of accelerating quality control of voxel models. Eye movements are not random [14], which means we can extract meaningful information out of our visual behavior. In turn, computers can use the cues to automate specific tasks.For example, eye behavior can function as a pointer, similar to a mouse. Using our eyes has the advantage of being much faster than operating a mouse or other media device [15][16]. Moreover, our visual attention predates our actions [17], which means, in theory, we can know even quicker which areas are of interest to the user. However, using the eyes as an input has one fundamental challenge that it needs to address: the Midas touch problem [18] [19]. Named after the famous king in Greek mythology who turned everything he touched into gold, this problem faces the same: how do we keep the gold without turning everything into it? To do this, we need to be able to make some cognitive assessment of the eye behavior, where we can separate eye behaviors leading to different cognitive behavior. In the scope of our problem, this means finding ways to separate eye movements that lead to an error in the model from eye movements that are not on errors.

In this thesis, we look into several ways where AI can provide geologists in their quality control. By doing this, we try to find an answer to our main research question: can we predict where errors in geological voxel models will occur?

First, we examine whether or not eye gaze can be a predictor for errors in voxel models. We want to make a general algorithm that can do this, which is challenging since individual differences among people concerning eye movements are large [20]. Secondly, we investigate whether model uncertainty can be a good predictor for errors in voxel models. Thirdly, we try a new approach based on fixation identification to build a faster model and to improve performance. Apart from this, we assess what conclusions we can make from these findings.

2 Related Work

The field of automating quality control in geological voxel models is not an active field of research [21], but other fields can other fields can provide inspiration for our can inspire us for finding solution. Firstly, we discuss research on geologic modelling in general, and GeoTOP in particular. Secondly, we discuss the field of eye-tracking and its advances. Thirdly, we give an overview of visual attention. Fourthly, we mention machine learning and classification and discuss several algorithms that can help with our problem space. Finally, we discuss closed-loop interaction and explain its relevance to the problem at hand.

2.1 Geological Modelling

Geologic modelling, or geomodelling is the science concerned with building computerized models of the surface and subsurface. It "consists of the set of all the mathematical methods allowing to model in a unified way the topology, the geometry, and the physical properties of geological objects while taking into account any type of data related to these objects." [22]. The goal is to portray a representation of the geological reality of the area of interest.

Geomodelling is used in studies on sustainable energy [23], mining [24], managing natural resources, gaining insight in deep foundations, predicting subsidence, reducing chances of geohazards, and many more. In these fields, an understanding of the subsurface is crucial as they are heterogeneous, meaning that each layer has its unique behavior. Furthermore, the subsurface is anisotropic, meaning it is directionally dependent. For example, each layer can have different electrical conductivity.

Building geological models is a complex task. For centuries, these maps were made in a 2D space. This lack of a third dimension made these maps hard to interpret, as 'geologists had to communicate 'complex spatial and temporal relationships as a 2-D image using standard colours and symbols.' [25] Luckily, improvements in computer technologies resulted in advance of 3D models in the last three decades [26], making these models much more sophisticated. The demand for this extra dimension in both government and industry makes intuitive sense, as geology is inherently a 3D science [27].

These models, also called geoscientific information systems (GSIS) [28], are spatial extensions of 2D maps. There are three approaches to constructing GSIS: geostatistical, cognitive modelling, and hybrid approaches where both are combined.

2.1.1 Geostatistical Models

There are several methods available in geostatistical literature, including multipoint geostatistical methods [29], transition probability indicator simulation [30], and sequential indicator simulation [31]. These stochastic methods all work by interpolating the surface based on borehole samples. One of the advantages of geostatistical methods is that the process is objective, and documentation is straightforward. However, these methods only work when there are enough (good) borehole descriptions since these statistical methods are as good as their data. Furthermore, it is difficult to describe all elements in these models in areas where much geological knowledge is needed.

2.1.2 Cognitive Models

Cognitive modelling does not try to build a 3D model explicitly but uses the implicit knowledge of a geologist to establish the intricacies of a subsurface [26]. As a consequence, cognitive models are capable of incorporating a lot of background knowledge that is often hard to replicate with geostatistical methods. Often the subsurface consists of very complex geometrical shapes, which are hard to incorporate algorithmically. A disadvantage of cognitive modelling is that the subjective nature of the method makes documentation much harder since the expert has to document all decisions for other people to understand the process. As a result, cognitive modelling is both time consuming and difficult, since expert knowledge can be hard to define explicitly. Furthermore assessing the quality of this process is also difficult as well.

2.1.3 Hybrid Models

To overcome both disadvantages, many geological models try to implement hybrid versions [32] [33]. This method combines the best of both world, by incorporating geographical knowledge in the geostatistical model, making them more realistic. Furthermore, a geologist can later check for errors in the statistical model as a quality control check. An example of a hybrid model is GeoTOP, which will provide data for this thesis.

Since most 3D geological models rely on borehole descriptions, their output is as good as the data. Relying too much on data can be problematic in building these models, as this data is often sparse [34]. Since interpolating from data is a statistical method, the output is by definition uncertain. To assess the quality of these models, many geological models try to quantify this, as 'uncertainties have a meaning' [35]. This model uncertainty is often quantified with information entropy and can be applied both on a whole model as on its parts. Information entropy is the quantified average amount of missing information in a stochastic data set, and can be calculated as follows [36]:

$$H = -\sum_{i}^{N} p_i \log p_i, \tag{1}$$

where H denotes the information entropy over N possible outcomes, and p_i is the probability of outcome i. Since the value H is the sum of the product of p_i with its logarithm, values range between 0 and 1. High entropy should be interpreted as more uncertain of the outcome. As an example, take throwing dice. If we use one fair die, the outcome of a throw is entirely random, and the entropy of the outcome (i.e., the value we throw) is 1. However, if we throw two dice, the outcome is not wholly random anymore. Here, throwing 2 or 12 is more uncommon than throwing 7 (the most random state), and the entropy is lower as a result. A lower entropy also means that we expect to make a wrong choice on the outcome of throwing one die more often than when we throw two dice.

Another challenge in geological models is the fact that layers can have complex geometries, due to structural deformations [37]. These distortions make the output of natural systems seem random at times, because the structure of the subsurface is sensitive to initial conditions. We cannot infer these initial conditions solely at the output.

2.1.4 GeoTOP

In this thesis, we use the GeoTOP model as the 3D model for our experiment. GeoTOP is a 3D schematization of the Dutch onshore, divided in 100m x 100m x 0.5m voxels. It models regions of the Netherlands, and provides among other things the lithostratigraphy of that region. Lithostratigraphy is the "logical ordering of rock units, based on lithological properties that discerns a rock unit from other rock units above, below or lateral from the observed unit."[38] The model classifies twelve different units, as can be seen in Figure 1.

The GeoTOP model constructs its voxels as follows:

1. Geological schematization of the boreholes into units that have uniform sediment characteristics using lithostratigraphical and lithological classes.

2. The model calculates top and base surfaces for each layer. These represent the bounding surfaces of the units at Formation or Member level. Each voxel now can be placed within the correct lithostratigraphical context.

3. Stochastic interpolation is used to assign a lithoclass to the voxels.[38]

The GeoTOP uses borehole descriptions as an input for the model, which are made available from the Dino-loket. A borehole description contains information about the stratigraphy and lithological properties of that area. This information then gets translated to correspond with the classes used by GeoTOP. As there are slight differences between stratigraphical units and the ones by GeoTOP, also known as geological units. Borehole descriptions provide an accurate picture of the stratigraphy and lithology, but since only 10% of the surface has a borehole description [39] (below the surface this percentage is even lower), GeoTOP uses statistical methods for predicting remaining voxels.

In the second step, stratigraphical layers get modelled by stapeling geological units onto each other with top and base rosters. By calculating the distance between them, the thickness of each layer gets calculated as well. Each of these top and base layers has a standard deviation of the height. These represent the model uncertainty, which provides an indication where errors are more likely.

Finally, stochastic interpolation determines the eight different lithoclasses. GeoTOP contains eight different lithoclasses (see Figure 1). Each voxel in the model gets a probability of each lithoclass, based on borehole descriptions, geological units, layers, and preprogrammed domain knowledge. GeoTOP subsequently picks the most likely lithoclass, and gets a model uncertainty measure for the lithoclass with a Sequential indicator simulation. Since we are primarily concerned with stratigraphy, the working of this model is beyond the scope of this thesis. For an explanation, see [40][41].

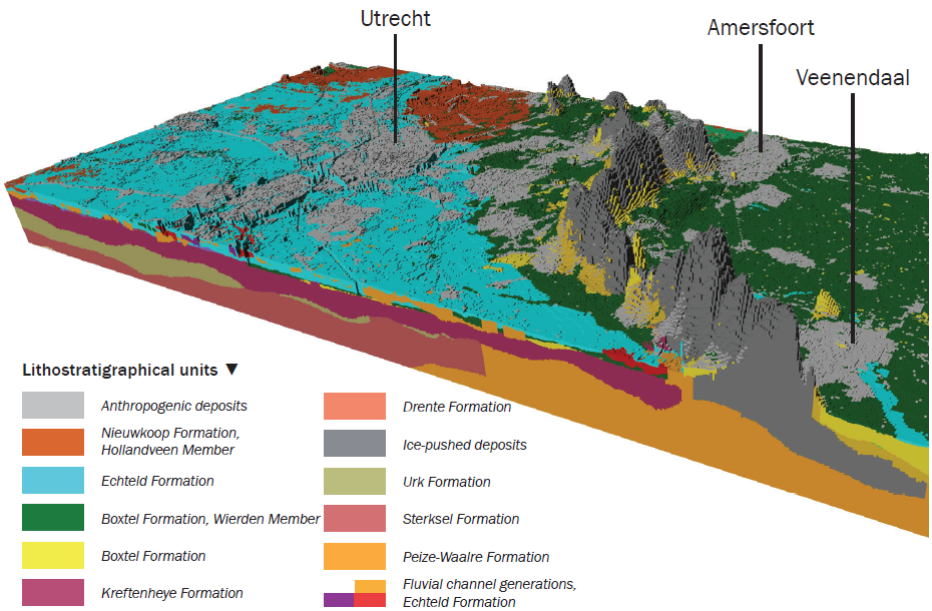


Figure 1: 3D model of the Gelderse Vallei are showing lithostratigraphic units. Adopted from [42].

Building the GeoTOP model is a labor and cost intensive process, as completing a block the size of a province takes two years [43], and the costs of building GeoTOP increased with 'two orders of magnitude' [43]. The first year is spent on data preparation and building a geological concept, the second year on writing documentation and quality control. In this thesis, we focus on the latter.

GeoTOP consists of two types of data: hard and soft. Hard data is the type of data serves as an input for the geostatistical model, i.e., the borehole descriptions mentioned earlier. The soft data in GeoTOP are geological constraints that are applied to the model [43]. These are constraints on the interpolation techniques, based on knowledge on geological plausibility. Whereas the interpolation is a stochastic method, the current geological features are entirely deterministic.

The result of this hybrid geological model is a complex interaction between borehole data, statistical modelling and decades of expert knowledge of geologists. As a result, analyzing this data is a complex task. The quality control of the geological models is often done by highly experienced experts, with an average of 7 years in the quality control of Gelderse Vallei.

2.2 Eye-tracking

In eye-tracking research, the goal is to gain meaningful information out of raw gaze points and other tracked features of the eye. Since eye movement is related to attention [44], we can gain insight into a participants thought process by following their eye movement [45]. Unfortunately, this is not a direct relationship, but since cognitive data is not readily available eye movement is the best indicator for visual attention. According to Rayner [46], we can divide

Table 1: Frequently occurring measures in eye-tracking research.

Fixation	Saccade	Mixed
Temporal		
Total fixation duration	Saccade duration	Total reading time
Gaze duration		First pass time
Average fixation duration		
Time to first fixation		
Revisited fixation duration		
Spatial		
Fixation position	Saccade length	Scanpath pattern
Fixation sequence		
Count		
Fixation count	Saccade count	
Average fixation count	Inter-scanning count	
Revisited fixation count		
Probability of fixation count		

studies into eye-movement in roughly three eras that shaped the field. The first era (1879-1920) was a period where basic eye movements were detected. The second era (1930-1958) is a period where the focus laid on application rather than understanding, unsurprisingly coinciding with the height of the behavioristic movement in psychology. During the third period (1970-1998), technological advancements led to improvements in eye movement recording systems. Furthermore, increased computational power resulted in possibilities to examine eye-tracking data much more thoroughly. After Rayner's overview, eyemovement studies continued to grow in several directions. Firstly, eye-racking devices continue to lower in price while increasing in quality and ease of use [47]. Secondly, increasing computational power allowed for more accessible analysis, now even possible on consumer hardware. Thirdly, the rise of smartphones give rise to a new field of applications, but these developments are in an early phase [48]. Fourthly, research in identifying fixations and saccades become more standardized, with overviews and introductions in eye-tracking methodology [49][50] and fixation identification [51]. Finally, machine learning techniques found its way to eye-tracking research as well [52][53].

Nowadays eye-tracking applications have found their way into many different areas, among them being measuring the effectiveness of advertisements, investigating human-machine interaction in aviation, and understanding how people solve complex tasks. The intentions of participants during an experiment have an impact on the information the experiment provides, and as a result, some experiments are more useful than others for the scope of this thesis. Applications in eye tracking can be divided in couple of ways [54], as can be seen in Figure 2. Firstly, eye-tracking systems can be either diagnostic or interactive. In the former, the systems tracks eye movement without applying changes. In the latter, eye-movement is an input to aid the user. In their turn, interactive systems can be either selective or gaze-contingent. Selective systems replace

mouse movement with the eye movement of the user, without making further assumptions. Gaze-contingent systems use knowledge about the user to process information to make more complex displays. These distinctions are conceptually important since they provide a framework to increase the understanding of how the human-computer interaction model uses eye movement. Firstly, a diagnostic eye-tracking system needs to be implemented to show how gaze behaviour can predict where experts see errors in the model. Secondly, a gaze-contingent interactive system needs to be implemented to actually automate parts of the quality control.

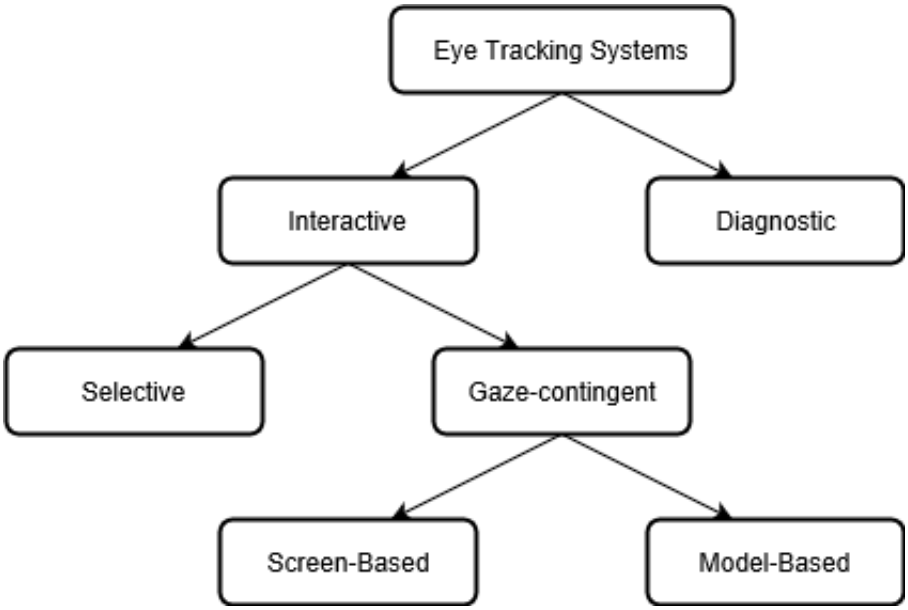


Figure 2: A hierarchy of different applications in eye-tracking. Adopted from [54].

An interesting aspect of research in eye-tracking has been the fact that it has been labelled as a promising field for over five decades [20], which is unusually long. Typically, new technologies are touted as promising very briefly, after which it either delivers its promises or get discarded as a failed endeavour. This means that on the one hand, people see the immense potential that using eye-movement for analysis has. As a French saying goes, *"Les yeux sont le miroir de l'âme"*. Since we cannot access a persons thoughts directly, knowing where they look can be one of the closest ways to address it. This relation is far from causally since there is not a one-on-one relationship between where we look at and what we think about (more on this later). On the other hand, certain obstacles proved to be too big to allow eye-tracking research to move beyond their infancy in a fast manner. Firstly, the first eye-tracking devices had reliability issues and were labour intensive, but nowadays setting up an eye-tracking experiment is less demanding. Modern eye tracking systems are often provided with software that makes set up tasks like calibration convenient. Still, not all hardware issues are solved. Goldberg and Wichansky estimates that 20% of the participants cannot be tracked reliably [55], a non-trivial amount that has many different causes.

Among these are eye tracking systems having trouble finding the center of the pupil of participants with dark iris colors, glasses or lenses inferring the signal, and people with low eyelids that obfuscate a part of the iris. Secondly, even when system tracks the eye movements correctly, the extraction of meaningful information is still a time and labour consuming process. Since eye tracking systems usually measure between 20 and 250 eye movements per second, the amount of data increases rapidly. The rise in computation power has made this process more manageable, but a consensus in methods and parameter settings are severely lacking. As a result, researchers often have to start from scratch, which makes the automating process much more time consuming. Finally, the interpretation of data has been a challenge as well. After classifying saccades and fixations, an explanation has to be made between the eye movement and the cognitive activities. There are three approaches to do this: top-down based on cognitive theory, top-down based on design hypothesis, bottom-up, or a combination of either of these three approaches. In the top-down based approach based on cognitive theory, empirically tested hypotheses will be used to infer information about the dataset. For example, a researcher can use knowledge of the correlation between pupil size and cognitive workload [56] to find places where the participant has trouble performing the task. In the top-down based on design hypothesis, a researcher can find what the effect of the experiment design has on the results. For example, does changing screen colours affect the fixation time? Finally, in the bottom-up approach no theories about eye behaviour are assumed, but rather the data is used to hypothesize a theory.

Almost all human eye movement used in eye-tracking studies consist of three types: saccadic, smooth pursuit, and fixations. Saccades are quick, simultaneous movements of both eyes between two or more fixations. They can be both voluntary and reflexive, with a duration between 10ms and 100ms [49]. These movements are too fast for the brain to process, which means that people do not perceive information during these eye movements. This process is known as saccadic suppression[46]. Fixations are the periods where the visual gaze is roughly in a single location. When our eyes locate a moving object, we call this a smooth pursuit. Since the eye-tracking data used in this thesis is on still images, we model only saccades and fixations.

2.2.1 Fixation identification algorithms

Since the properties of saccades and fixations are different, we can implement an algorithm to categorize between the two. The assumption is that since the brain can not process images during a saccade, these parts are effectively noise in the data. Unfortunately, this does not imples that each time a participant fixates on an object their mind concerns with the same object as well. The relationship between fixation, visual attention and visual attention is a highly debated area without definite conclusions as of yet.

We can find saccades by measuring the velocity of the movement of the eyes and find a threshold to categorize between fixation and saccades. According to [49], the researcher should find the threshold empirically. See 3 for a plotted example to find saccades.

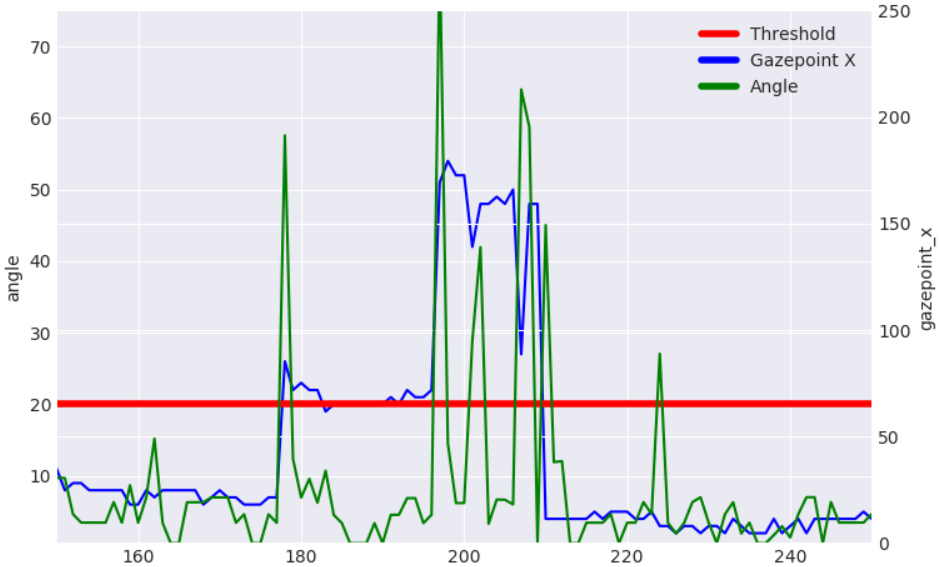


Figure 3: Sample of eye movement of participant 4 of the dataset used in this thesis, where the right y-axis denotes the position of the gaze point x through time, and the left y-axis the angle of between two time periods. The red line denotes a threshold, where values of the angle that are above the threshold are labeled as saccades and values under the threshold as fixations.

In 2000, Salvucci and Goldberg [51] proposed a taxonomy of fixation identification algorithms, where they identified five algorithms: velocity-threshold identification (I-VT), Hidden Markov model fixation identification, (I-HMMM), Minimum Spanning Tree Identification (I-MST), and Area-of-Interest Identification (I-AOI). Tafaj et al. (2012) [57] used a Bayesian Mixture Model (I-BMMM), were a Euclidean distance between gaze points was used as a metric to differentiate saccades and fixations. Santini et al. (2013) [58] proposed a Bayesian Decision Theory Identification (I-BDT) for ternary classification (i.e., smooth pursuits were labeled as well), which works on eye-tracking with smaller frame rates as well. Apart from that, they also used an expert for labeling the data as ground truth. As a result, they were able to show the performance of their algorithm and showed that the I-BDT had a precision of 95.60%, comparing favorably to the 89.57% of state-of-the-art algorithms.

Surprisingly, a consensus in ideal combinations of algorithms, parameters, which oculomotor events to identify, ideal sampling frequency, and thresholds is lacking. In fact, detecting eye movements is far from solved at the moment. Andersson et al. [59] identify five problems that researchers have to deal with, that are relevant to the scope of this thesis:

1. There is little agreement on how to evaluate eye-movement algorithms.

2. Theoretical rigor is lacking in what exactly a saccade or fixation means.

3. Most algorithms lack indications on how to select parameters.

Table 2: The effect of three visual expertise theories on features in eye-tracking data, where higher and lower indicates the performance of experts in a task in comparison with intermediates or novices.

Theory	Affected feature	Effect
Theory of long-term working memory	Fixation duration	shorter
Information-reduction hypothesis	Number of relevant fixations	higher
	Number of irrelevant fixations	lower
	Relevant fixation duration	higher
	Irrelevant fixation duration	lower
Holistic model of image perception	Time to first relevant fixation	faster
	Saccade length	longer

4. Literature on comparisons between algorithms is scarse, usually only to prove that a new algorithm performs better.

5. Despite a shared intuition between researchers about eye-movement definitions, humans labeled fixations do not have a standard ground truth.

2.2.2 Eye-tracking features

After identifying fixations and saccades, several measures can be calculated to get a better understanding of how an expert looks at geological voxel models. Lai et al. (2012) [60] summarizes commonly used eye-tracking measures in 2D applications. These measures can be drawn from fixations, saccades, or a combination and deal with either temporal aspects, spatial aspects or the number of occurrences.

The subfield of understanding how experts solve complex problems or tasks is a good fit for this thesis. Gegenfurtner, Lehtinen, and Säljö [61] provide a comprehensive overview of expertise differences in visual tasks. Their meta-analysis tested several theories about expertise in comprehension of visualizations.

Firstly, the theory of long-term working memory proposes that experts have much shorter fixation lengths. The reasoning behind this is that expertise in a subject results in qualitative changes of memory structures in the brain, which allow them to quickly encode information in long-term memory and access it during their visual task.

Secondly, the information-reduction hypothesis argues that experts are much better at separating task-relevant from task-irrelevant information, which increases the speed of completing the task.

Finally, the holistic model of image perception states that experts are able to quickly extract global information of visual stimuli. The experts can extract information from distanced and para foveal regions, and as such objects of interest do not have to be in the fovea.

These three theories can help us in assessing the level of expertise, which can be useful for thinking of implementations. Furthermore, it can give us an insight in differences in quality control.

2.2.3 The relationship between gaze, visual attention, and cognition

In this thesis, eye movement is used to infer a cognitive assessment of the stimuli, namely deciding whether an area is correct or incorrect. When we use eye movement to infer what people think, we have the underlying assumptions that gaze location and attention are related and that there is a relationship between where we look at and where we think about. Empirical evidence suggests that there is a relationship between these phenomena. Just and Carpenter [62] showed that people tend to look at the object they are thinking about in tasks such as the comparison of rotated figures, mental arithmetic, sentence verification, and memory scanning. Thomas and Lleras [63] showed that the link between eye movement and cognition work both ways: it can both show what a person thinks and diverting eye movement can influence how we think.

Unfortunately, there are phenomena where these relationships break down. One famous example that demonstrates this is Simons and Chabris [64] experiment in change blindness, a phenomenon where persons do not perceive considerable changes in objects and environment in the area of their focus. They demonstrated this by asking participants to count how often a group of persons throw a basketball towards each other. Meanwhile, a man in a gorilla suit walks between the group, a remarkable event. However, due to the participants focus on the ball, they often entirely miss the man in the suit. These findings show that when a person gazes in a direction it does not guarantee that they perceive the things in that area. There is no consensus among scientists whether participants literally do not see the gorilla, or if they do see it but the phenomenology is inaccessible. In a recent paper, Cheng differentiated several levels of seeing [65]: crowding, indexing, and attending. Crowding happens when a person sees something, but there is no attention, or even visual indexing: meaning that the object is separated from other objects in the persons vision. Indexing is an early stage in visual perception, where objects are encoded as different entities, but they are not attended by the person [66]. Finally, attending means that the object is both seen and has the attention.

Ideally, one would be able to separate these from each other, especially attending from the other levels of seeing. Just like saccadic blindness, gaze points that are not attended are primarily noise in eye tracking data. Unfortunately, methods in how to categorize these are lacking.

2.3 Visual Attention

Attention is a state where humans select a mental direction. When this attention is selected voluntary, we speak of endogenous attention. The involuntary direction of eye movement is known as exogenous attention. By definition, endogenous attention has an internal cause: our cognition directs our attention ,for example, when we have a certain task, like finding an object in a larger area, or during reading. Our endogenous attention is thus related with our goals and desires. Exogenous attention is the mental direction that is caused by stimuli from outside ourselves. Here we can think of looking at the direction of the sirens of an ambulance, but also that some colors draw our attention more than others. There are several theories as per why humans have selective attention: because of our brain's incapacity to process several visual stimuli at the same time, because cognitive incapability to have several thoughts at the same time,

or because we need a single direction to be coherent.

Conceptually, attention is an interesting phenomenon. On the one hand, the nature of it is common sense. As the American psychologist William James noted: "Every one knows what attention is."[67] However, we should divide knowledge of attention into two views: first-person and third person. Attention from a first-person view is easy to grasp, as we experience its existence every day. We know what it means to have full attention for something, and sometimes while driving, we realize we have not been attentive for a while. From a third-person perspective, explaining attention is not trivial at all, and remains an active field of research in (cognitive) neuroscience, (neuro)psychology, philosophy, computer science, and artificial intelligence. Attention also relates to other challenging philosophical concepts, for example, whether it is necessary or sufficient for consciousness and its relation to epistemology.

Visually attention modelling is primarily concerned with answering the following two questions [68]:

1. Can an agent's visual attention be predicted, given its circumstances and behavior is known?

2. Can an agents behavior or output be predicted given we know its visual attention?

To answer these questions, often a combination of behavioral cues, environmental properties and mechanism of human attention are used.

Visual attention modelling gained serious interest in the last few decades and in fields such as object recognition and detection, video summarization, and interactive computer graphics apply visual attention. Most visual attention models focus on a small aspect of human attention.

In studies in visual attention modelling, applications often use either bottom-up models or top-down models. The focus of these two is parallel with the distinction of respectively exogenous and endogenous attention.

2.3.1 Bottom-up visual attention

Bottom-up visual attention is directed focus towards visual stimuli, usually because an area in the visual field is conspicuous, meaning the area differs from other parts of the area. Bottom-up attention is automatic, reflexive, and swift. Bottom-up attentional models deal with saliency: finding areas that stand out in a picture. Areas can be conspicuous for several reasons, including color, orientation, shape, symmetry, containing humans or other animals, and including text. See Table 3 for an overview of influential bottom-up approaches. The last two decades, saliency models have become increasingly more advanced, and they have been successful in predicting fixations in free-viewing. However, when a specific task is involved their explanation power is unconvincing [69]. The reason for these results is because when tasks are involved, a vast majority of fixations are task related [70].

Table 3: An overview of bottom-up attention models

Paper	Year	Title	Stimuli/field	Brief description
[71]	1980	A feature integration theory of attention.	Cards containing letters and shapes	Focal attention and top-down processing occur separately
[72]	1987	Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry.	-	Computational architecture and elementary features that predict visual attention
[73]	1998	A Model of Saliency-based Visual Attention for Rapid Scene Analysis	Scenes of nature, pictures of art	Visual attention system based on early primate visual system
[74]	2001	Modeling the shape of the scene: A holistic representation of the spatial envelope	Images of nature and man-made objects	Introduces a holistic framework that focuses on the 'gist' of an area
[75]	2005	Saliency Based on Information Maximization	Natural images	Introducing Shannon's information measure as model
[76]	2006	Bayesian Surprise Attracts Human Attention	Video clips of different human activities	A Bayesian approach to modeling surprise
[77]	2006	Graph-based Visual Saliency	Images of nature	Bottom-up saliency model based on graph computation
[78]	2007	Predicting human gaze using low-level saliency combined with face detection.	Images of faces and same without	Adding face and text detection to saliency model to improve predicting power
[79]	2007	The Discriminant Center-surround Hypothesis for Bottom-up Saliency	Natural images and complex videos	Perceptual mechanisms are optimal in a decision theoretical sense
[80]	2007	A Nonparametric Approach to Bottom-up Visual Saliency	Natural images	Parameter free bottom-up model for saliency
[81]	2008	Paying Attention to Symmetry	Images of nature and man-made objects	Adding symmetry as a salient feature
[82]	2008	A Bayesian Framework for Saliency Using Natural Statistics	Images from natural scenes	Saliency Using Natural Statistics (SUN) model
[83]	2008	Dynamic Visual Attention: Searching for Coding Length Increments	Wide range of images and video clips	Introducing 'Ebb and flows' in attention
[84]	2008	A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network	Videos of natural scenes	Non-deterministic model to mimic human attention behaviour
[85]	2009	Decorrelation and Distinctiveness Provide with Human-Like Saliency	Artificila created examples	Using decorrelation of scales from a Principal Component Analysis
[52]	2009	Learning to Predict Where Humans Look	Large set of random images	Bottom-up saliency model that incorporates top-down features like fixations
[86]	2010	The focus of expansion in optical flow fields acts as a strong cue for visual attention	Videos of people walking or driving	Introducing Focus of Expansion (FOE) as a feature to predict saliency
[87]	2010	A Novel Multiresolutional Spatialtemporal Saliency Detection Model and its Applications in Image and Video Compression	Images and videos of nature	Using phase spectrum of quaternion Fourier transform to predict saliency
[88]	2010	Esaliency (Extended Saliency): Meaningful Attention using Stochastic Image Modeling	Natural images	Specifying saliency mathematically

2.3.2 Top-down visual attention

Top-down visual attention is much slower since it is guided by our longer-term cognitive strategies [89] such as task-related knowledge, emotions, and expectations. The relationship with these phenomena makes modelling top-down attention consequently much harder. Since task-related knowledge differ in each task, modelling top-down visual attention is task-dependent as well. Furthermore, several objects and actions are often coinciding, which makes it challenging to derive attentional conclusions based on eye movement, let alone statements about the cognitive state of the observant. From a neurological viewpoint, top-down attention is more challenging than bottom-up as well. At the moment, we know relatively much less about the neural instantiation of the top-down components of attention [90] [91].

Despite these challenges, there are advancements in top-down modelling. Yarbus [45] showed that different type of questions for the observant lead to changes in gaze behavior. These findings show that we use our top-down visual attention as a 'spotlight': we zoom in on areas that are in line with our goals and tasks so that we can ignore the irrelevant areas. In tasks with a clear goal, it is possible to make inferential statements about cognitive strategies from understanding visual attention. When external referents are available in the data, research show that can we can have an idea what happens internally in the agent [92]. Navalpakkam and Itti [93] proposed a computational model for task-specific guidance of visual attention, but it requires precise specification of the task at hand. This need for a specification is a severe limitation of the current state-of-the-art top-down visual attention models. While small successes have been claimed in laboratory experiments, translating this to real-world problems seems to be too challenging at the current state of affairs.

Still, an understanding of top-down attention is essential for most applications in computational attention models. Whenever humans have a task (either internally or externally provided), top-down influences seem to dominate over bottom-up attention. For example, Zelinsky et al. [94] found that during the visual search a purely top-down model provided a better approximation of human eye behavior than a mixed model with bottom-up features. Other research suggests that visual saliency seem to have little effect on similar tasks [95] and that top-down attention often only needs one fixation. [96].

There are three major primary of top-down influences:

1. Object Features
2. Scene Context
3. Task Demand

While top-down attention is an integral part of visual attention, it "lacks principled computational top-down frameworks which are applicable to different task types" [69]. A lack of a framework makes modelling top-down attention hard, and as a result most research is focused on easy tasks for the participant instead.

In short, attention is a constant interaction between top-down and bottom-up influences. The question of how humans integrate both in tasks is still a 'central open question' [69]. While this makes the field open, it makes using these techniques non-trivial and deriving conclusions out of them troublesome.

Modelling visual attention is still in its infancy, and often assumptions have to be made because concluding evidence is lacking. For example, even though many researchers stress that attention and eye fixation are not the same, the former is ofter measured by the latter. By doing this, researches ignore covert attention fully in these models.

2.4 Machine Learning

Since voxels in GeoTOP data are either correct or incorrect, the goal of finding this is a binary classification problem. In this subsection, we briefly discuss several classifiers, their advantages and disadvantages, and a brief explanation of their workings.

Machine Learning is a field where predictions are made by generalizing from examples. The goal in using classifying algorithms is to find a way to describe the training data, in a way that can predict new data points correctly. Machine Learning has become a staple in many applications, and performance excels in problems where manual programming is not feasible [108]. This makes machine learning a justified approach, as this applies to both finding GeoTOP errors as using Eye-Tracking data.

Domingos [109] mentions a few key aspects when applying machine learning to solve problems, which can be a guide for predicting labels algorithmically. Firstly, learning by machines can be informally written in the following informal equation:

$$Learning = Representation + Evaluation + Optimization \tag{2}$$

Solving machine learning problems is dealing with these three parts. Representation means that a computer can understand the classifiers, because they are formally written. A computer can only learn the classifiers that we feed to model. This is also known as the hypothesis space. Then, after we have implemented the classifiers, we need an objective way to assess how good their predictions are. We consider this in the evaluation part. A simple function is the accuracy rate, or simply the percentage of correct guesses. However, in a later section, we show why for some problems (like quality control) this evaluation function is not ideal. After we know how to classify our data and how to identify good and bad classifiers, we search for a method to search among all possible classifiers. This part searches for the best classifier on our data, or to find a good enough solution if the data is too complicated]. Table 5 shows an overview of often used methods.

Secondly, in machine learning the most important aspect is building a classifier that can generalize. An algorithm should not only perform well on the data the model sees, but also on similar new data. The fact that the classifier does not know this data means that we do not know the function that needs to be optimized since we assess the predicting power on the test set (which the model has no access to during training).

Thirdly, data alone is not enough. Knowing about the problem to be solved can help choosing the right representation, as no classifier can beat random guessing over all possible functions [110].

Fourthly, feature engineering is a crucial part of machine learning. Often, the focus on picking the right classifier gets the most attention, but building and

Table 4: An overview of top-down attention models

Paper	Year	Title	Stimuli/field	Brief description
[45]	1967	Eye Movements During Perception of Complex Objects	Complex objects	Gaze patterns are dependent on the questions asked
[97]	1994	Where We Look When We Steer	Driving	Drivers rely on the tangent point on the inside of each curve on the road
[98]	1994	Control of Selective Perception Using Bayes Nets and Decision Theory	Object manipulation	Using a Bayesian network for representation
[99]	2001	In What Ways do Eye Movements Contribute to Everyday Activities?	Tea making	Separates types of fixations: locating, directing, guiding, and checking
[100]	2003	The E-Z Reader model of eye movement control in reading: Comparisons to other models.	Reading	Theoretical framework how identification, processing and attention co-occur
[101]	2004	Advances in Relating Eye Movements and Cognition	-	Support that we direct fixations to task-relevant relationships
[93]	2005	Modeling the Influence of Task on Attention	Natural and synthetic images	Computational model for task specific guidance of visual attention
[102]	2007	Beyond Bottom-up: Incorporating Task-dependent Influences into a Computational Model of Spatial Attention	Video games	Framework for top-down and bottom-up during tasks, with a global context
[103]	2007	Modeling Embodied Visual Behaviors	Sidewalk navigation	Adding reinforcement learning for visiomotor behavior
[104]	2008	Congruence Between Model and Human Attention Reveals Unique Signature of Critical Visual Events	Several	Modeling of the relative importance between bottom-up and top-down attention
[105]	2009	Optimal Scanning for Faster Object Detection	Natural scenes	Deriving optimal scan paths based on human eye behavior
[106]	2009	Visuomotor characterization of eye movements in a drawing task.	Sensory-motor coordination	Motor task have little influence on gaze, gaze has high influence on motor task
[107]	2011	A POMPD Model of Eye-hand Coordination		

Table 5: Overview of the three components of learning in machine learning. Adopted from [109], where italics are added to denote methods used in the scope of this thesis.

Representation	**Evaluation**	**Optimization**
Instances	*Accuracy/Error rate*	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam Search
Hyperplanes	*Likelihood*	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic Regression	*Information gain*	Unconstrained
Decision Trees	K-L divergence	Gradient descent
Sets of rules	Cost/ Utility	Conjugate gradient
Propositional rules	Margin	*Quasi-Newton methods*
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

selecting the rights features are often more important. Predicting on raw data is often not ideal, and incorporating useful features can significantly improve the prediction of the classifier. Finding features can be a difficult step however, as it requires knowledge that is specific to the domain of your problem, instead of general machine learning knowledge.

Fifthly, the fact that we can represent a function does not mean that we can learn it. For example, a decision tree cannot learn trees where there are more leaves than training samples [109]. Moreover, if the hypothesis space has too many local optima, we cannot guarantee to find the optimal function.

Finally, the most important rule is that more data is often better than better algorithms. While machine learning deals with a lot of ingenious techniques in feature engineering, classifiers, and other ways to improve the predictions, the best way to improve the performance of a classifier is to give it more data. In the next subsections, we briefly discuss a couple of conventional classifiers that we use in later sections.

2.4.1 Logistic Regression

In binary classification problems, the probability that Y is 0 or 1 is based on the input variables that can be calculated with a logistic regression. Logistic regression is a model that predicts categories based on the most probable class. The probability is calculated as follows [111]:

$$p(y = 1|x) = \frac{e^{\beta_0+\beta_1 X+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X+\cdots+\beta_p X_p}} \tag{3}$$

We fit this model by using a maximum likelihood function, a method for determining parameter values. In maximum likelihood, the method finds values that maximise the likelihood of observed data. This likelihood is calculated as follows [111]:

$$\ell(\beta_0, \beta_1, \ldots \beta_p) = \prod_{j:y_j=1} p(x_j) \prod_{j':y_{j'}=0} (1 - p(x_{j'})). \tag{4}$$

In our data, this means that we want to find values for $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_{p+1}$ that will maximize the likelihood of either a marking ($Y = 1$) or a non-marking ($Y = 0$) best.

2.4.2 Tree-based Methods

Decision trees build a model by finding splits that separate observations best. At each split, the algorithm finds a partition to make meaningful subsets that can make better predictions than the entire dataset. The goal is to find the best splits while keeping the tree as small as possible because larger trees are more likely to overfit. Splits that divide a node that consists of only one class are called leaf nodes.

To find the splits at each step, we need to quantify the performance of a split so that the tree can pick the best split at each step. Decision trees pick the best split via an impurity measure. There are different measures for doing this, each with their way to decide which split is best. Three popular impurity measures are resubstitution error, Gini-index, and information gain.

The resubstitution error is a measure that calculates the portion of instances that have an incorrect prediction if we pick the majority class at that node:

$$i(t) = 1 - \max_{\mathbf{j}} p(j|t) \tag{5}$$

While picking the majority class makes intuitive sense, there are instances where the resubstitution error picks suboptimal splits. For example, in cases where a split could lead to a leaf node, the resubstitution error sometimes picks other splits. To increase the tendency of splits towards creating leaf nodes, we decrease the value of the measure faster than in a linear fashion. The Gini-index is a measure that looks at the distribution of the labels and calculates the probability that an instance in the dataset has an incorrect label by random assignment based on the distribution. We can measure this as follows:

$$i(t) = p(0|t)p(1|t) = p(0|t)(1 - p(0|t)). \tag{6}$$

Finally, the information gain is a measure that favors smaller trees by picking splits that result in the purest nodes. We can calculate this as follows:

$$i(t) = -p(0|t)logp(0|t) - p(1|t0logp(1|t). \tag{7}$$

See Figure 4 for an example of a decision tree that uses the Gini-index. Here, we calculate the information that a random sampling generates from a node by observing its class. Section 2.1.3. mentions entropy, the average amount of information in a stochastic process. These two concepts are related since the information gain calculates the difference between the entropy of the parent node minus the weighted sum of the entropy of the child node. The information gain calculates the information that each split generates and picks the most informative split.

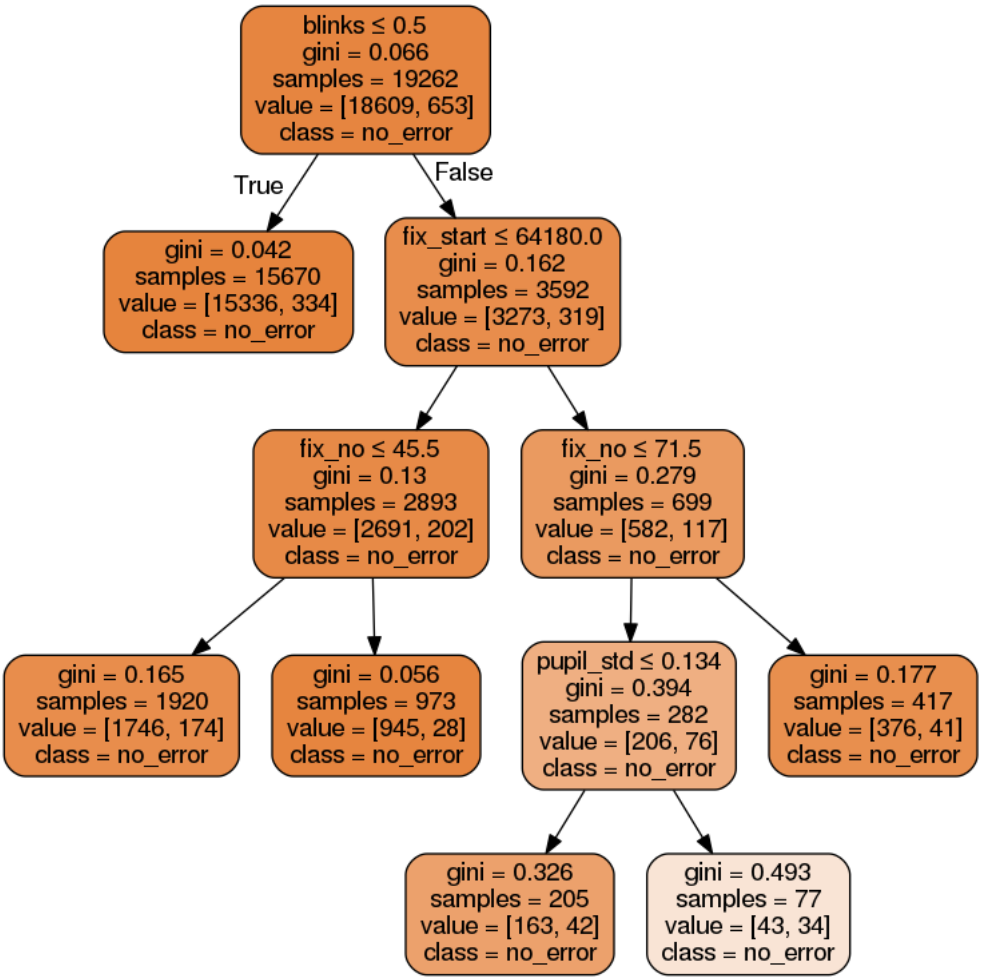


Figure 4: An example of a decision tree, with maximum depth and maximum leaf nodes of 6. This tree visualizes a fit on the preprocessed data in Subsection 4.5. Decision trees are easy to visualize, but usually, lack predicting powers.

Decision trees have several advantages compared to other machine learning techniques. First, the model is a complete white box. By visualizing the tree, even an untrained eye can see rather straightforward how the model classifies. The tree is simple to understand and to interpret, as each node is just a split based on the value of one feature. While people often struggle with statistical intuitions [112], decision trees mimic the way in which human decision making works much better. The algorithm also requires little preparation, as it can handle values that are not normalized. As splits look only at one feature at a time and consider only the order of values, normalization is not necessary.

However, decision trees have some severe limitations as well. Firstly, their performance is not among the best performing machine learning approaches. Secondly, small changes can have a massive impact on the shape of the tree. As a result, they are not very robust. The high sensitivity to changes in data means decision trees have a high variance. We want to have a tree that is a

good representation of the information of the dataset, but when small changes can alter the tree entirely the visualization becomes less useful.

Several techniques build on decision trees that can increase the performance of the model. A better performing model means decreasing the error rate. Errors can have are a combination of the following factors:

$$\text{error} = \text{bias} + \text{variance} + \text{irreducible error} \tag{8}$$

To reduce errors, a model can either reduce bias or variance (or both), as by definition you cannot prevent irreducible errors. Usually, in machine learning, we have to make a compromise between bias and variance. Having a high bias means that the algorithm is not capable of finding specific information in the data, for example, because the assumptions are too unspecified. Often simpler models suffer from a high bias and low variance since they ignore complex relations. Models that have more complex assumptions can represent the data of the training set much better. As a result, their bias is lower, but in the process, they can find 'information' in the noise. As a result, the algorithm suffers from pareidolia, where it sees faces in clouds. Increasing the sample size also reduces the variance, as more data increases the estimate of the mean of a group. As a result, the standard deviation decreases. There are several techniques to reduce error rates in decision trees: bootstrap aggregating, which tries to reduce the variance, and gradient boosting, which tries to (primarily) reduce bias.

Both techniques try to overcome the limitations of decision trees via ensemble learning. In ensemble learning, the main idea is that multiple weak learners are better than one learner. By combining some hypotheses, the idea is that together they form a better hypothesis. In bootstrap aggregating or bagging, the algorithm draws a sample with replacement from the training dataset and repeats this B times, where B denotes the number of bootstrapped data sets or trees. Each sample has the same size as the training set, which means that observations can appear several times in a bootstrapped dataset. Bagging then creates a decision tree for each sample and averages over their predictions or take a majority vote. The variance of a single tree is σ^2, which means that the mean over all trees is $\frac{\sigma^2}{B}$. This formula shows that bagging reduces variance, since increasing n will decrease the average variance. In this thesis, we use a Random Forest algorithm [113], which is a method that builds upon bagging. Just as in bagging, Random Forests build several trees with bootstrapping, but they randomly select only a selection of m predictors. Typically this is $\sqrt{p}$, where p denotes the total number of features.

Where Random Forests build multiple trees over which it averages, gradient boosting [114][115] builds trees sequentially. Instead of building new trees, Gradient Boosting uses weak hypotheses and corrects on these hypotheses in the next iteration. After each tree, the technique calculates the shortcoming of the hypotheses via gradient descent. Boosting thus uses work sequentially since the algorithm knows the information of other trees. Instead of growing one large decision tree, gradient boosting builds a small tree. These smaller trees tend to have a higher difference between observed values and estimated values, also known as residuals. Gradient boosting tries to solve this, by iteratively fitting small trees to these residuals. By making this process slower than decision trees, boosting works well on improving parts of the data where its performance

is suboptimal.

After the Gradient Boosting is finished building trees, we can also calculate the influence of each feature in the model. This can be measured by calculating the number of times a tree selects a feature for splitting. We weight this value with the squared improvement to the model of each split and average this value over all trees [116].

2.5 Closed-loop Interaction

Human-computer interface is a discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them[117]. Schomaker et al. (1995)[118] proposes a taxonomy for a basic model that we implement as a framework for this thesis. See Figure 5. As can be seen, the human agent and the computer agent are physically separated, but they communicate through an interface. There are two processes involved: perception and control. The perceptive process consists of Human Input Channels (HIC) and Computer Output Media (COM). HIC is the input that a human gets from the computer, and COM is the output from the computer to a human. The control process consists of Human Output Channels (HOC) and Computer Input Modalities (CIM).

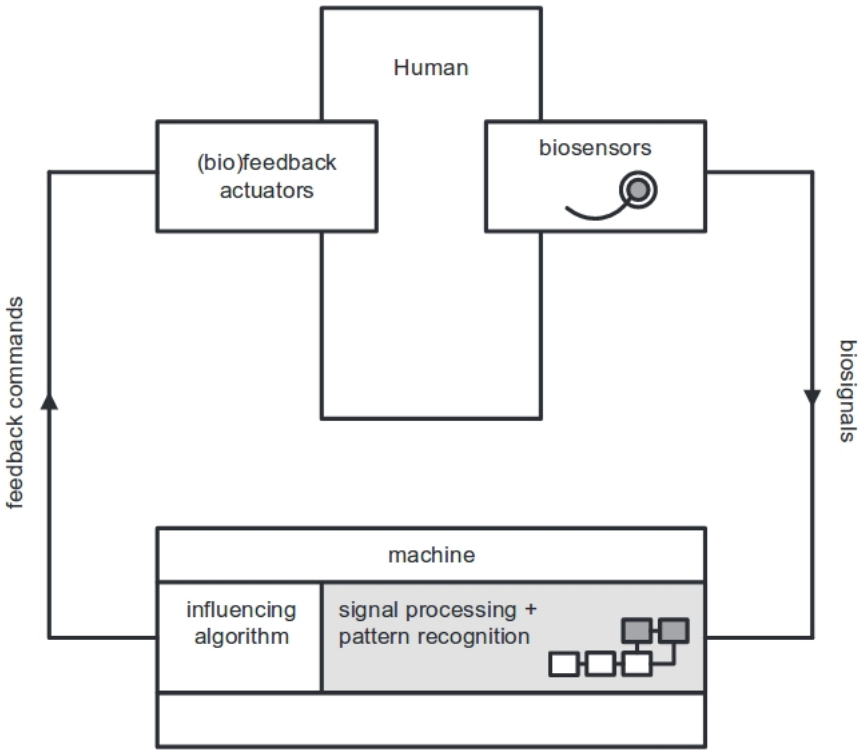


Figure 5: Basic model that shows the interaction between a human and a computer. The arrows represent the direction of the interaction, where the dotted line indicates the intrinsic perception/action loop that occurs independently of the computer. Adopted from [12].

In this model, we can identify cognitive or computational components. In the case of computer agents, the processing design is known and readily available for the designer. In the case of humans, we do not have all information on this process, but rather infer this by observing human behavior (i.e., eye-movement).

For all four previously mentioned components, four epistemological levels of observation can be defined [118]:

1. Physical/physiological level. The physical level denotes the system characteristics and processes, such as resolution, delay time, and the constraint in the interaction between the agents.

2. Information theoretical level. This level denotes the characteristics of the information in the component, like entropy.

3. Cognitive level. Denotes the representation and procedures, explicitly made in syntax and semantics.

4. Intentional level. The intentional level denotes the goals and beliefs of the component and is crucial to understand for pattern recognition. Examples of these goals for the user are anomaly detection and browsing.

In a later section, we implement these components with corresponding epistemological levels.

Eye signals can provide as a signal for our machine learning task. See Figure 6 for a pattern recognition pipeline. Here, sensors of a physical system capture raw signals that are potentially useful for the task at hand. These raw signals are in turn preprocessed with filtering and artifact removal techniques. Then, the processed signals can, in turn be, synchronized and segmented based. Since the signal is often large in volume, feature extraction is often needed to separate redundant information from informative, which makes the machine learning task both faster and more robust, since it reduces the chances of overfitting. Here, the first step is finding potentially informative features with their optimal parameters. Then, in this entire pattern space, the most useful features are selected. The selected features together will form the reduced pattern space, and serves as input for the classification algorithms.

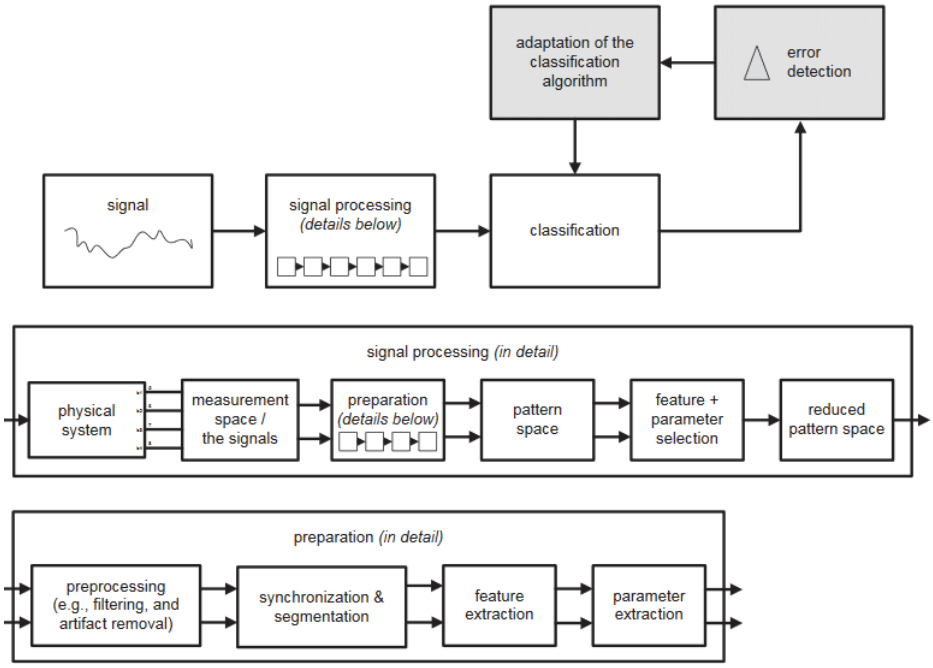


Figure 6: A pattern recognition pipeline, where raw signals are processed into a classification algorithm for decision making. Adopted from [12]

In this thesis, we will use both eye-tracking data and geological data as signals in this pattern recognition pipeline.

2.6 Previous Experiments

In van Maanen et al. (2014,[42]) an eye-tracking experiment was conducted on expert geologists of Dutch geological survey. Twelve participants, of which ten males and two females, were asked to look at a part of the GeoTOP model from Utrecht and the Gelderse Vallei, an area in the west of the Netherlands. This part consisted of a 3D space with size 20700m (x) by 24500m (y) by 50m (z), but the experts watched them in two-dimensional slices (x) with (y, z) coordinates. During each slice, participants first had to scan the image to check for errors in the GeoTOP model. When participants thought they were finished checking errors on a slice, they pressed the space bar to manually mark voxels that they assessed incorrectly. After that, another space bar continues them to the next slice. Each participant performed this routine for 70 slices in total. We use the same dataset in this thesis.

2.6.1 Eye-tracking Experiment

The slices were shown on a 1024 x 768 screen, and participants had their heads on a chin rest at 61cm from the monitor. The complete task took 70 min on average, with view intervals of 29 seconds and mark intervals of 31 seconds. For the experiment, a Tobii X50 with 50hz sample rate was used. Each trial generated two files: `gaze.csv` and `events.csv`.

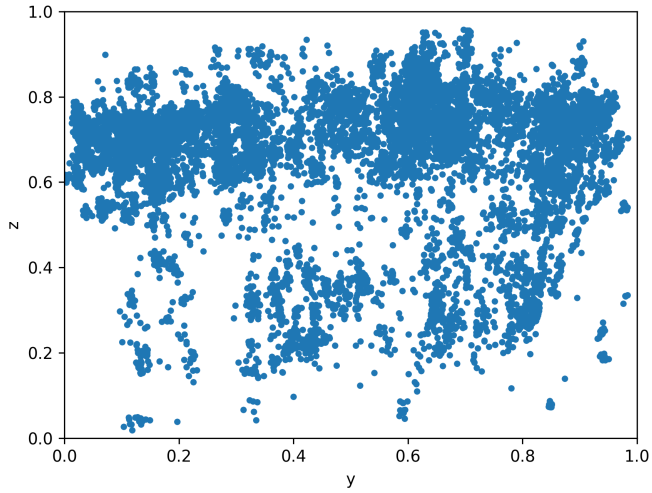


Figure 7: Gazepoints of all participants for slice 49.

The collected data consists of twelve participants (n), that for each slice (x) have gaze tracked through time t, where gaze is a vector consisting of gaze points x and y, cam x and y, distance of the pupil, pupil size and validity of both left and right eye. In total, the participants took 9 hours and 19 minutes to check the slices, excluding the time it took for them to mark the areas.

The Tobii x50 is a standalone eye tracking system, with an accuracy between 0.5-0.7. The measured gaze angles are accurate up to around 35 degrees, and has a latency of 35 ms.

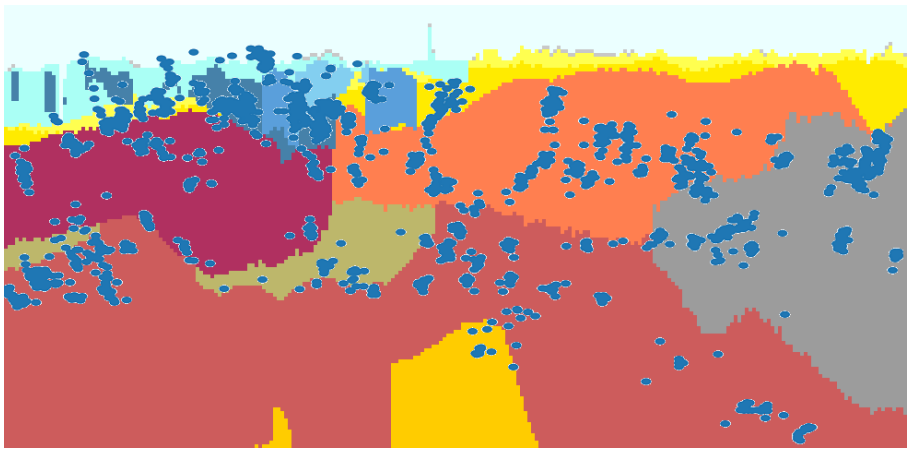


Figure 8: Gazepoints of all participants plotted on slice 10.

For each slice, participants were asked to assess whether errors in the model made by GeoTOP occurred. Since only eye behavior is measured during the view time, participants were asked to only go to the marking phase of each slice

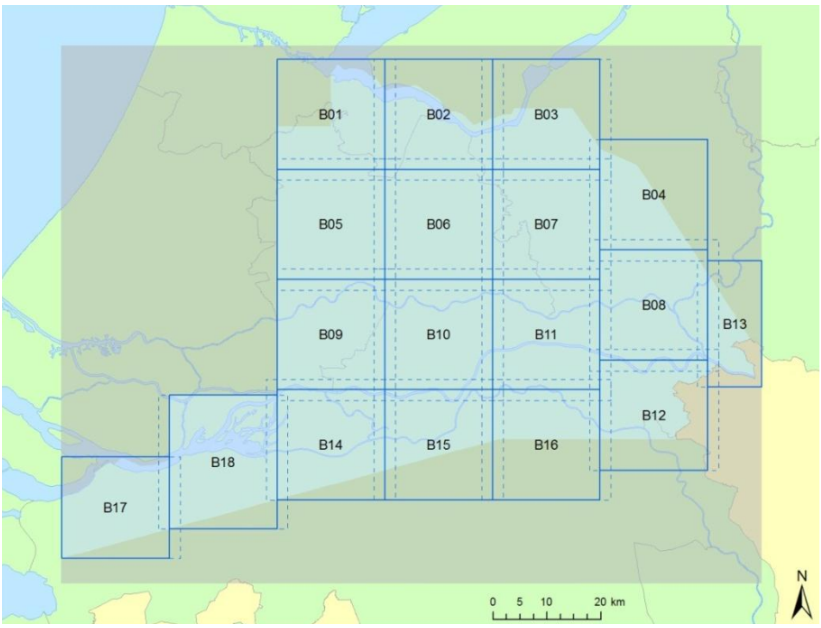


Figure 9: Overview of the Rivierengebied, an area in central Netherlands. The main area that used in this experimental set-up is the B05, but some parts of B01, B02, B05, B09, and B10 are also in this dataset. Adopted from [41].

after finishing. Each participant started with a mock slice, so that the task was fully understood before starting the experiment. They were also informed that each wrong pixel should be labeled instead of drawing circles around the area, as this would confuse evaluation metrics. While humans interpret an empty circle as one area of one error, the computer only understand pixels being wrong/right.

2.6.2 GeoTOP data

In this thesis, we use GeoTOP for our experiment on geological data. The complete GeoTOP data consists of 64 3D areas, which covers a large part of the Dutch subsurface. Each area is a dataset with rows of x, y, and z coordinates, where x denotes the direction of west to east, y from south to north, and z from the lowest measured part of the subsurface to the surface. The GeoTOP model used in this research is a part of the Rivierengebied from Utrecht and the Gelderse Vallei, an area in the west of the Netherlands. This part consisted of a 3D space with size 20700m (x) by 24500m (y) by 50m (z), but the experts watched them in two-dimensional slices (x) with (y, z) coordinates.

The largest part of the area used in this experiment is the B05 Rivierengebied, but since the slices have a small overlap, small parts of the surrounding B01, B02, B06, B09, and B10 appear in the dataset as well. See Figure 9 for the locations of these areas. After that, the x coordinates are used to divide the dataset into 70 slices, starting at 141950, with steps of -300 for each increment in x. Then for each x, each row with z and y values is used to build a 245*120 screen, creating 70 slices with y and z coordinates corresponding in shape (but not size!) of the eye tracking data.

2.6.3 Mathematically Modeling Visual Attention

In van Maanen et al. (2014)[42] a mathematical attentional model [68][119] was used to find whether visual attention can be a good predictor for manually marked errors in geological voxel models. In their experiment, eye-tracking data of twelve domain experts was used to determine the rate between attention and manually marked errors. They found that with optimal parameter settings the model was able to find 68% of he marked errors, while only 18% of the area not marked as an error drew attention. With an Area under the ROC Curve (AUC) of 0.82, this result can be classified as good. Their experiment also showed that a large percentage of the errors was found in a short period: the first 3.5% of viewing provided 54.8% of the AUC.

The areas of attention were located as follows. First, the model assumes that attention was constant:

$$A(t) = \sum_{x,y} AV(x, y, t) = 1, \tag{9}$$

where A(t) = the total amount of attention at point t, and AV(x,y,t) is the attention value for area (x,y) at time t. This assumption is not universally accepted, as some models incorporate a framework with dynamic visual attention [83]. Secondly, errors were predicted by attention exceeding a threshold:

$$e(x, y) = \begin{cases} 1 & \text{if } AV(x, y, t) \geq \alpha \text{ for some t} \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $e(x, y)$ is the estimation of an error at point (x, y).

Thirdly, since people pay more attention to the center than to the periphery of their visual space, a gaze parameter was introduced.Here the relative distance of each area (x,y) to the gaze point is used as a factor to determining the attention value of (x,y). For each new attention value of time, this is modeled as follows:

$$AV_{new}(x, y, t) = \frac{1}{1 + \gamma \cdot r(x, y, t)^2}, \tag{11}$$

where γ denotes the relative impact of distance $r(x, y, t)$ to the gaze point on the attentional state, and $r(x, y, t)$ is the Euclidean distance between the coordinates and gaze point.

Fourthly, the attentional model assumes that the amount of attention is limited, so the value of AV should be normalized to incorporate this:

$$AV_{norm}(x, y, t) = \frac{AV_{new}(x, y, t)}{\sum_{x',y'} AV_{new}(x', y', t)} \cdot (A(t) = 1), \tag{12}$$

where AV_{norm} denotes the normalized value of attention. In the normalized model, each area of (x, y) now has a value between 0 and 1.

Finally, the model assumes that gaze travels faster than attention among the grid. As a result, a decay factor is included to let attention values persist for a longer time:

$$AV(x, y, t) = \lambda^{t'-t} \cdot AV(x, y, t') + (1 - \lambda^{t'-t}) \cdot AV_{norm}(x, y, t) \tag{13}$$

Here, λ is the decay parameter that transitions the gaze into attention. Higher values of λ should be interpreted as a slow decay, while lower values indicate a faster decay.

3 Hypotheses

This thesis aims to investigate the applicability of automating tasks during quality control in geological voxel models. In 2014, previous research [42] established a link between eye-movement and errors in geological voxel models. An attentional model was implemented to map the attention to markings and showed that this method was a great way to predict errors. However, one downside of the research was the fact that the model optimized the parameters on each participant, making the model incapable of making generalized statements. Ideally, we want our model to be able to make generalized predictions without having to be trained on new participants. For this reason, we use the model to train on all participants instead of making individual participant models. Due to the fact that participants have different expertise, eye-behavior, and marking style, we hypothesize the model performance will decrease:

Hypothesis 1. *Generalizing previous attentional models on geological voxel models will decrease their predicting powers.*

Apart from the input of participants, we also look into the information provided by geological voxel models itself. GeoTOP models are partly statistical models and have an entropy value for each voxel. This entropy value represents the uncertainty about the correctness of voxels. As a consequence, we hypothesize that voxels with higher entropy are more often incorrect and voxels with a lower entropy are more often correct and vice versa. We use this information to build an a priori model. We look into the relationship between model entropy and the number of errors. Since the entropy of the GeoTOP model is an indicator of model uncertainty, we expect this to be a good predictor for finding these errors.

Hypothesis 2. *Entropy is a good predictor for finding errors in geological voxel models.*

Finally, we look into methods of analyzing eye-movement behavior. Studies have stated that processing raw gaze data into fixations can decrease complexity without losing valuable information and increase computational performance [51]. We can use these techniques to extract meaningful eye behavior features combined with machine learning classifiers to assess our last hypothesis:

Hypothesis 3. *Implementing fixation identification algorithms with a machine learning classifier will reduce computation time while improving its predictive powers.*

4 Method

In this section, we describe the several experiments we performed to test our hypotheses. Firstly, we give a description of the experimental setup of the eye-tracking research [42]. Secondly, we present an overview of the GeoTOP data that was used in the eye-tracking experiment. Thirdly, we describe methods of preprocessing our data. Fourthly, we present evaluation metrics to compare models. Finally, we test several models.

4.1 Signal Processing

In Section 2.5, a pattern recognition pipeline was discussed, which we apply here to error predictions in geological voxel models. Figure 6 gives an illustration of the order of processing.

Quality control in geological voxel models consists of several signals: the eye behavior of geology experts, information from GeoTOP itself, and the markings of the experts.

Firstly, the eye movement of experts is a signal that can be measured and used to extract relevant features from it. See ref Section 2.2 for an overview of different signals in eye movement, and behavior we can infer from it. To collect eye data, an eye-tracker collects raw signals. In this thesis, a Tobii x50 was used for measuring the experts eye behavior. The Tobii x50 has a 50hz frame rate, meaning it collects 50 data points each second. The rate at which an eye-tracker samples is a significant aspect, as the eye is in constant movement. The higher the sample rate, the more accurate picture we can have about the position of the eye at any given moment. In turn, this helps the accuracy of identifying higher-level features, like saccades and fixations. According to Nyquist-Shannon sampling theorem, the sampling frequency should be at least twice as big as the eye movements [120]. However, since the choice of eye-tracker preceded this thesis, the conclusion should be reversed: since the experiment uses a 50hz tracker, we should only consider eye behavior of 25hz and lower. Fixations are less sensitive to low sample rates than saccades [121], since they have a longer duration and less positional variation. Tracking saccades with a 50hz eye-tracking system is more problematic. There is ample support for the claim that 50hz is not a good sample rate for measuring saccades, as short saccades of 10° are not accurate under 60hz [122] and for calculating maximum saccadic velocity an eye-tracker of 300hz is needed [123]. Furthermore, data yielded by Andersson et al. [121] suggest that after 250,000 saccades, there are differences in total fixation duration time. For this reason, they suggest using event detection algorithms rather than temporal sampling.

Secondly, we can look at the geological 'signals' in the pipeline. Each of these voxels has the attribute of having a certain lithostratigraphy with an entropy value. This entropy value can be used as a predictor of error, as we expect entropy to be correlated with markings.

Finally, the markings of the experts are explicit indications of their cognitive assessment of voxels, indicating the expert thinks the lithostratigraphy of the voxel is either correct or incorrect. We use this as ground truth for the classifier, where $Y = 0$ for correct voxels and $y = 1$ for incorrect voxels.

Ideally, we would like these experts to be reasonably robust in their agreements among each other. When raters tend to agree on the value of Y, we can extract valuable information from this. For example, in these situations, we can identify when some participants fail to see an incorrect marking and can hypothesize that this is due to fatigue. We can also assess which errors are easy, and which are more difficult. High agreement among experts can also help us in the process of creating a single artificial expert, and help us identifying when experts make mistakes.

Unfortunately, participants in assessing the GeoTOP model were in general not in agreement with each other. When looking at the markings, we must conclude that finding a ground truth between the participants is non-trivial. In

Table 6: Inter-rater agreement of pixels between all participants. Values are calculated with Cohen's Kappa, where a 1 means a perfect agreement between participants and 0 means two participants do not agree more than random assigning labels. As we can see, the inter-rater agreement is low, making it hard to create an 'average' expert that looks like the participants.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	-	0.04	0.04	0.03	0.03	0.02	0.03	0.05	0.03	0.05	0.04
1	0.04	-	0.19	0.09	0.05	0.01	0.07	0.14	0.02	0.04	0.01
2	0.04	0.19	-	0.08	0.03	0.01	0.05	0.12	0.02	0.03	0.03
3	0.03	0.09	0.08	-	0.08	0.06	0.09	0.13	0.10	0.05	0.09
4	0.03	0.05	0.03	0.08	-	0.06	0.05	0.07	0.09	0.07	0.07
5	0.02	0.01	0.01	0.06	0.06	-	0.05	0.04	0.15	0.09	0.12
6	0.03	0.07	0.05	0.09	0.05	0.05	-	0.06	0.06	0.08	0.04
7	0.05	0.14	0.12	0.13	0.07	0.04	0.06	-	0.06	0.06	0.05
8	0.03	0.02	0.02	0.10	0.09	0.15	0.06	0.06	-	0.09	0.12
9	0.05	0.04	0.03	0.05	0.07	0.09	0.08	0.06	0.09	-	0.07
10	0.04	0.01	0.03	0.09	0.07	0.12	0.04	0.05	0.12	0.07	-

order to assess the level of agreement between experts, we calculate inter-rater agreement with a Cohen's kappa coefficient (k). This is calculated as follows [124]:

$$k = \frac{p_0 - p_c}{1 - p_c}, \tag{14}$$

where p_0 is the relative observed agreement, and p_c the probability of chance agreement. By doing this, we can measure the agreement between experts while taking into account that they agree with each other based on chance. In Table 6 we see that the inter-rater agreement is low among experts.

This low inter-rater agreement shows an added complexity in this problem space. Apparently, even though these experts have the same global task, the execution of their role is different. One of the explanations is that most of these experts have a specific domain of expertise, and as a result is the part that they look at most. A second explanation is that it shows the difficulty of the task at hand. If finding errors would be a trivial task, we would see a much higher inter-rater agreement. The fact that these are low means that experts see different errors, some unobservable to others. It also means that we cannot use the agreement of experts to make a conclusive division between errors and non-errors. Earlier we mentioned that the targets of each participant provide a ground truth, but ideally we would like to assess their performance as well. This way, we would compare the expert markings to a 'true' correct/incorrect array. However, if the experts lack any real agreement, then the subsequent analysis will yield spurious results [125].

Table 7 adds further evidence in the difference between expert markings. As we can see, some participants have over 20 times as much markings as other participants. These large differences in labeling can be problematic, since this can favor classifiers heavily to the participants with the highest number of marking pixels.

Table 7: Overview of marking behavior of all participants. The total markings is the total number of pixels where $y = 1$, average markings is the number of connected pixels with value 1, and average marking size the average amount of pixels adjacent to each other. This table indicates clearly how different participants mark, as participant 2 has over 28 times as many markings as participant 5.

Participant	**Total markings**	**Average markings**	**Average marking size**	**Number of empty slices**
0	6603	41	162	0
1	22340	19	1187	4
2	23923	22	1088	0
3	5969	23	256	0
4	3006	10	308	4
5	846	4	207	24
6	2489	5	454	10
7	16677	34	495	0
8	2447	9	287	1
9	3420	6	614	4
10	1977	12	171	1

The marking dataset consists of 111945 ones and 362096055 zeros, respectively 0.03% and 99.97%. These differences are fairly consistent between both participants and slices, with standard deviations of 0.00034 for participants and 0.00031 for slices. See figure for the entire distribution. 81.47% of the pixels coordinates of the screen has no markings in any of the (slice,participant) combinations. One issue we can infer from the number of markings of errors in the data is the fact that the division between errors and non-errors favors non-errors, which makes the classes imbalanced. The problem of imbalanced data sets generally applies in situations where the majority class is large and the minority class is small [126]. Imbalanced datasets provide problems for many machine learning algorithms and evaluation metrics, making this a non-trivial aspect of our data. Traditional machine learning algorithms tend to favor the majority class, which can lead a poor performance for predicting the minority class. As we want to find the errors in the data set, this is a significant challenge.

There are a couple of ways to deal with this problem, with the most popular being undersampling, oversampling, and weighting samples. Undersampling is the process of reducing the amount of the majority class samples in the training data to correct the bias [127]. Contrastingly, oversampling is a technique that adds minority samples to reach to the same balance as in undersampling [128]. An example of these techniques is SMOTE [129], where synthetic minority instances are created by using k-nearest neighbors on the feature values. Finally, the bias towards the majority class can also be reduced by adding weights to the minority class, making errors on these samples more costly. As a result, choosing majority class always would reduce the model performance.

Class imbalances can be problematic for training models, since predicting the majority will give us a 99.97% accuracy rate. One possible solution is to add weight to the mean squared error, in this case the positive class label. As

a result, predicting markings as non-markings will be punished more severely, which will increase the cost of predicting only majority classes.

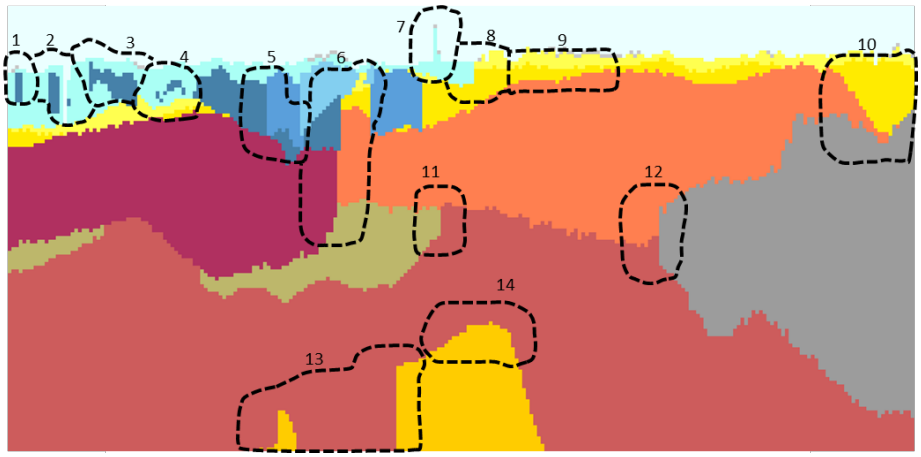


Figure 10: An example of a slice used in the questionnaire performed on participants after the eye-tracking experiment. The experts were asked to indicate whether they thought the shading was an error, the shape was correct, how conspicuous, and what kind of error it was.

Apart from the markings of all slices, a follow-up questionnaire was performed on all experts, where they were showed six of the 70 slices to answer questions about the nature of errors. Each slice contained between twelve and fifteen marks, where a region was considered a possible error if a certain threshold was met of expert marks. See Figure 10 for an example of an image that was used in this questionnaire. The participant answered a series of questions on each marking to give a better understanding of the nature of these marking. They indicated whether it was an error, the level of certainty they had, the difficulty of finding the error, the severeness of the error, whether the error was conspicuous, whether the error was marked correctly, and the type of error. The questionnaire indicated eleven types of errors: the base of the unit is too pointy, the base of the unit is too irregular, the top of the unit is too irregular, the lateral transition between the units is too sharp, a certain unit is not present or underrepresented, a certain unit is overrepresented, the unit is lying too deep, the unit is lying too high, the order of units is wrong, the shape of the units is wrong, or other. See Appendix C for the entire questionnaire.

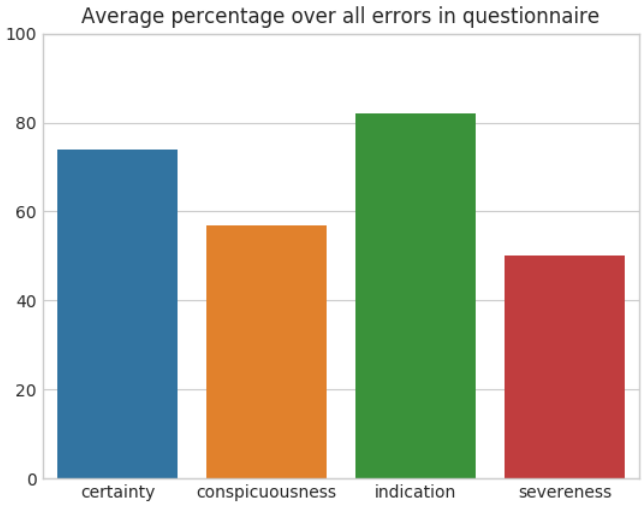


Figure 11: Average of values given by participants over all errors used in the questionnaire. This bar plot shows that the average severeness of errors is very low, and that they were hard to find.

The six slices in the questionnaire can provide a ground truth, since all experts have checked these markings. See Figure 11 for the average percentages of the levels of severeness, certainty, conspicuity, and indication. This figure shows clearly that the average severeness of an error is very low (50%). Even though the questionnaire only treats around 8% of the data, we can already see that there is a difference of opinion of the experts concerning the errors. The existence of disagreements is a thought-provoking find, since if the experts showed more agreeableness among each other, the difference in inter-rater agreement could be due to experts not seeing some markings. The results of the questionnaire show that this is not the case, because even after being shown the areas they still disagree. Figure 12 shows a distribution of how each participant perceives areas as errors. Here we see an apparant disagreement among experts, as most of the areas are either really an error or really not. The results of this make finding an average error rather challenging, and suggests that using each participant's own ground truth is the best we can do. These disagreements also show again the difficulty of both interpreting the GeoTOP data and interpreting the expert's performance.

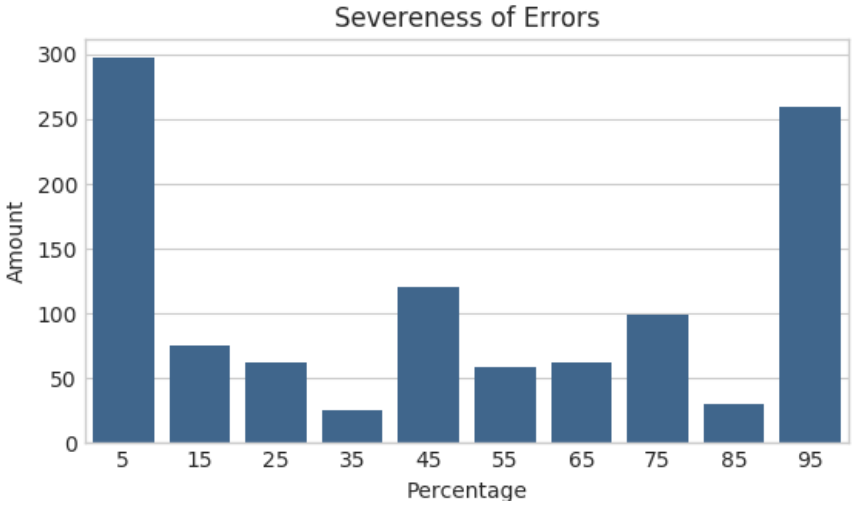


Figure 12: Distribution of severity assessment by experts of each error showed in the questionnaire. Here we can see that the errors follow an extreme bimodal distribution: most are either very unlikely or very likely to be an error.

4.1.1 Preparation

To prepare the data, we preprocess on the `gaze.csv` and `events.csv` of each participant. The preparation of the dataset consists of several stages. First, data preprocessing removes noise and unreliable data. Secondly, data segmentation creates segments of the data to make analysis simpler. Thirdly, we extract meaningful features out of raw signals. And finally, parameter values for these features can be found.

We preprocess the eye behavior of each participant in the `gaze.csv` file. Since the eye-tracker takes samples at a fixed interval, sometimes participants are blinking during an interval. These blinks are removed from the data, as participants do not have visual input during blinks. The Tobii x50 makes blink removal trivial since these occurrences have a validity of 0 and their coordinates of gaze points are a negative integer.

Then, gaze points at areas of the screen that are outside of the image are removed as well. The experiment is performed on a $1024 * 768$ screen, with the geological slices being $980 * 480$. The image offset of x is 22 and the offset of y is 144. For both x and y, the gaze points less than offset or more than the screen size minus offset are removed. After this, we resize the images. The GeoTOP model consists of slices with size $245*120$, but the experiment enlarged each voxel to 4 by 4 pixels. As a result, all gaze points of 4 adjacent in either horizontal or vertical direction refer to the same voxel. For this reason, we resize the images, and it has the added advantage that it decreases computation time and file sizes dramatically. The gaze behavior is resized by dividing all gaze points by 4. After this, gaze point coordinates are divided by the offset, making all gaze points in the range of either 0-245 for x and 0-120 for y. The markings in `events.csv` are resized as well, in the same way.

After the preprocessing, gaze and marking data gets segmented, separating

data for each slice. The `events.csv` of each participant consists of 4 columns: `timestamp`, `event_type`, `click_x`, and `click_y`. The event type is an integer that can be used to interpret the type of action. See 18 in Appendix B for an overview. For each row where `event_key` is 4, we use the value of `click_x` to indicate the slice number in a new column `slice`. Rows with an empty value of this column (i.e., with another event type) get interpolated. Each slice now gets a matrix of size $245 * 120$, where markings get a value of 1 and the rest 0. Then, we use `timestamp` to find the timeframes of each slice and use that to segment `gaze.csv` as well. Then, all gaze data and target data get stored in a multidimensional array, with the first dimension being slice number, the second participant number and the following for storing either gaze data or a $245 * 120$ matrix with the markings.

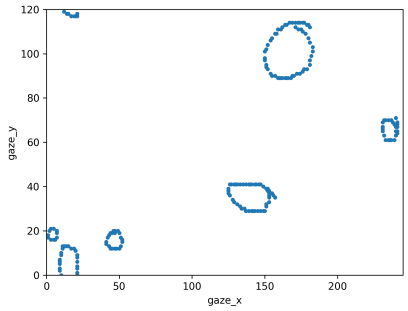
Figure 13: The markings of all participants plotted on slide 10.

The GeoTOP areas were also preprocessed, using only the voxels that were looked at in the experiment. For this several GeoTOP files were used: B01, B02, B05, B06, B09, and B10. See Figure 9 for the locations of these areas. Each file had rows with values for x, y, and z. For each file, we used the voxels with a value of x between 120950 and 141950, with a step size of 300. For the y values, we use between 444550 and 468950 (step size 100). All values of z are in the data available. Now each row had a unique value of x, y, and z, which were transformed in matrices as well. For the values of entropy and lithostratigraphy, we made a matrix with dimensions x, z, y. Note that these values correspond to the Tobii output of slice, y, and x.

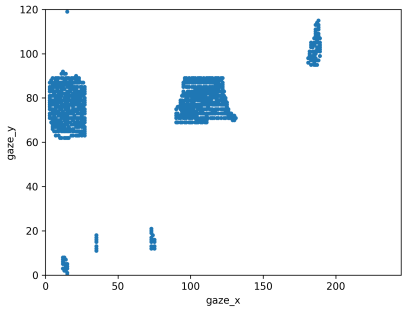
4.1.2 Dilation

When evaluating a model, a straightforward method is by checking which pixels are correctly predicted corresponding to the markings. For this reason, it is important that participants mark the entire area that they consider to be erroneous since if they mark only half evaluation metrics will underperform. Another advantage of filling the entire area is that in this case, it will make the dataset less imbalanced, which is beneficial for the performance of machine learning algorithms.

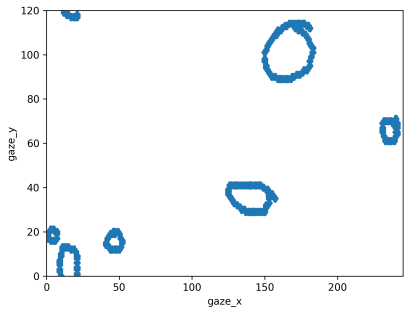
Figure 14: Effect of dilation on the markings of experts. Slices on the left are all slice 14 of participant 0, slices on the right are slice 16 of participant two.This demonstrates how dilation can help removing errors in filling the marked areas.



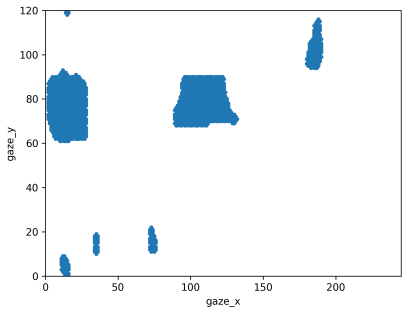
(a) Slice 14, participant 0, no dilation



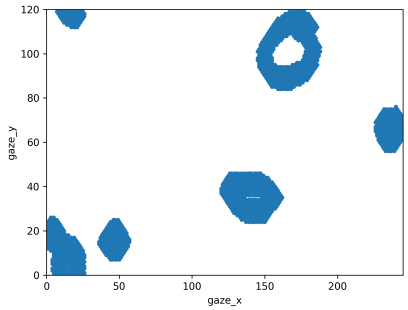
(b) Slice 16, participant 2, no dilation



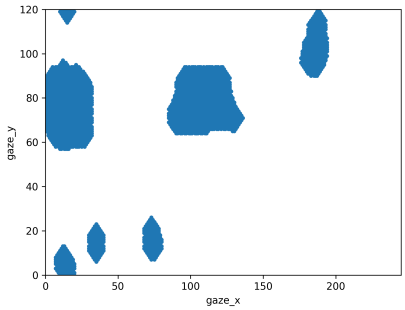
(c) Slice 14, participant 0, dilation with 1 iteration



(d) Slice 16, participant 2, dilation with 1 iteration



(e) Slice 14, participant 0, dilation with 5 iterations



(f) Slice 16, participant 2, dilation with 5 iterations

Analyzing these markings show that errors occur, however. Often some areas have small spaces of non-markings, and some participants encircled areas instead of filling them in. See Figure 14a for an example of these incorrect labeling. We use dilation to correct these markings. Dilation is an operator in mathematical morphology, where foreground areas (i.e., markings) are gradually enlarged to fill in holes in the area.

Dilation can be operated as follows [130]:

$$A \oplus B = x|(B)_x \cap A \neq \emptyset, \tag{15}$$

for the sets A and B, where $(B)_x$ denotes the translation by $x = (x_1, x_2)$.

We implement this by performing the `ndimage.morphology.binary_dilation` function of the SciPy library [131]. The amount of growth of the object depends on the structuring element, where the common elements are $4-n$ or $8-n$ [132]. In the former, pixels grow on horizontal and vertical neighbors, and the latter also on diagonal adjacent neighbors. In this thesis, we use a $4-n$ operation, which can be calculated as follows [130]:

$$C_4(n) = \{(x, y) \in \mathbb{Z}^2 : |x| + |y| \leq n\}, \tag{16}$$

where n denotes the number of iterations. We use 6 iterations, which empirically seem to correspond best to better markings. We use this dilated target data as an extra dataset, to compare whether this leads to better results. See Figure 14 for an example of this dilation. This figure illustrates well how dilation can lead to better markings. In the right subfigures, we see that the participant struggled with filling in pixels inside an object. Dilation helps to fill these empty pixels, without distorting the borders too much. In the left figure, we see an even more problematic image. Here, the participant circled the areas without marking the inner area. In some of these cases, dilation is not enough to fill the areas without some other areas sticking to each other.

After preprocessing eye-tracking and GeoTOP data, we can extract features and parameters for classification. Since we test three different models, these will be discussed separately in the following subsections.

4.2 Evaluation Metrics

To be able to compare the tested models, we implement the same technique for measuring the effectiveness of each method. This section explains the chosen performance metric with motivation so that the different models can be compared and we can derive a conclusion from this comparison. One of the most frequently used methods in pattern recognition and machine learning is the Receiver operating characteristics (ROC) graph, a technique for visualizing, organizing and selecting classifiers based on their performance [133]. ROC curves are commonly used in binary classifiers and shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) for various hit rating settings. The ROC curve is a good alternative for just using classification accuracy, since the former work better on unbalanced datasets than the latter. For example, whenever the distribution of a binary classification problem is 99%-1%, picking solely the majority class will result in a 99% accuracy. However, the ROC would be able to show that the classifier is poor by showing that the FPR is 1.

Here the TPR is calculated as follows:

$$True\ positive\ rate = \frac{true\ positives}{all\ positives}, \tag{17}$$

and the FPR:

$$False\ positive\ rate = \frac{false\ positives}{all\ negatives}. \tag{18}$$

The ROC graph is a two-dimensional graph with the TPR on the Y-axis and the FPR on the X-axis. By plotting the rates of different thresholds, we can see the relative trade-off between benefits and costs. The diagonal line where $y = x$ shows the performance of randomly guessing the correct class. The more a point is near the upper left corner, the better that model is at exploiting information of the dataset.

After plotting the ROC curves of each model, we can go one step further by measuring the Area Under the ROC Curve (AUC). The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Values of the AUC vary between 0.5 and 1, where 0.5 means the classifier does not perform better than random, and 1 indicates a perfect classifier. However, it should be noted that AUC weights FPR and FNR equally. As a result, it could be the case that a model has a lower AUC than another, but still could be favored due to the distribution of FPR and FNR. In our case, this is a domain specific preference.

In the performance measure, we measure the AUC of each model with the others. For each model, we measure the AUC of each fold in a 10-fold cross validation, which allows us to compare each fold. Each fold consists of randomly selected 63 train slices and 7 test slices. We use this fold size because it has proven to be empirically successful [134], and because more folds will lead into problems in slices where no markings are present. The motivation for comparing folds instead of the entire model is because we see whether a certain model consistently outperform others or if variation in folds is more important.

4.3 Generalized attentional Model

Firstly, we will use the gaze data of participants to implement the attentional model [42] of section 2.3.

We separate the training and test data in each fold. During the training phase, we use the gaze data of the training set as input and the normalized markings as target data. The motivation for normalizing the targets is that now both matrices have a sum of one, and we can now easily calculate the distance between the predictions and targets with a Mean Squared Error.

The model tunes two parameters for minimizing this distance: γ and λ. As previously mentioned, the gamma parameter denotes how visual attention is spread among the pixels from the gaze location. Since calculating the γ for each gaze point is computationally intensive, we calculate the values beforehand. So for each γ we calculate the distance matrices, which is a two-dimensional array of $((screensize_x * 4), (screensize_y * 4))$. Then, at each point t a we use a subset of the matrix, where the gaze points correspond with the center of the matrix. The attention at time t is then combined with the decay attention, where the division is calculated with Equation 13.

This model returns an attention matrix for each (participant, slice) combination in the test set, where the sum of each combination equals to 1. These can be compared to matrices of the same sizes with markings. Since the dataset contains a huge imbalance between classes, we expect scoring metrics to be a problematic part of the research. For this reason, we test Mean Squared Error

(MSE) with different weights of the data, and AUC. The MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}((y_i - \hat{f}(x_i))^2 * w_i), \tag{19}$$

where at each ith observation $\hat{f}(x_i)$ is the prediction that function $\hat{f}$ gives and w_i the weight of i. Since 1s are underrepresented in our data set, we experiment with different weights sizes for each i where $y_i = 1$. Different weight sizes of 1, 10, 100 and 1000 are measured. After finding the best γ and λ values of the training set, we can use the prediction matrix of normalized attention to find the best α value to separate markings from non-markings, respective values of 1 and 0.

We test two methods for finding optimal parameters: a linear search and by minimization of the cost function. For linear search, values of γ were used between 0 and 120, and λ between 0 and 20. For the minimizing function, the SciPy package was `optimize.minimize` was used [131]. The minimization used a LBFGS-B algorithm [135] [136]. The bounds corresponded with the values of the linear search.

4.4 Entropy Model

Before implementing other eye-tracking models, we now look whether we can find errors in the model a priori. Since these models are built using statistical methods, we can use this data as input for a model to find markings of experts as well. The motivation to test this model first is straightforward: there is much more data already available. Where the eye-tracking experiment covers only one area, a priori data of over 100 regions is available. Furthermore, only a small sample of the 3d model is in the experiment, whereas data is available for every voxel. Conducting eye-tracking experiments is both time and cost expensive, so trying to find alternatives is a good idea. Should we be able to find a good fit using only GeoTOP data, automating quality control would be much easier.

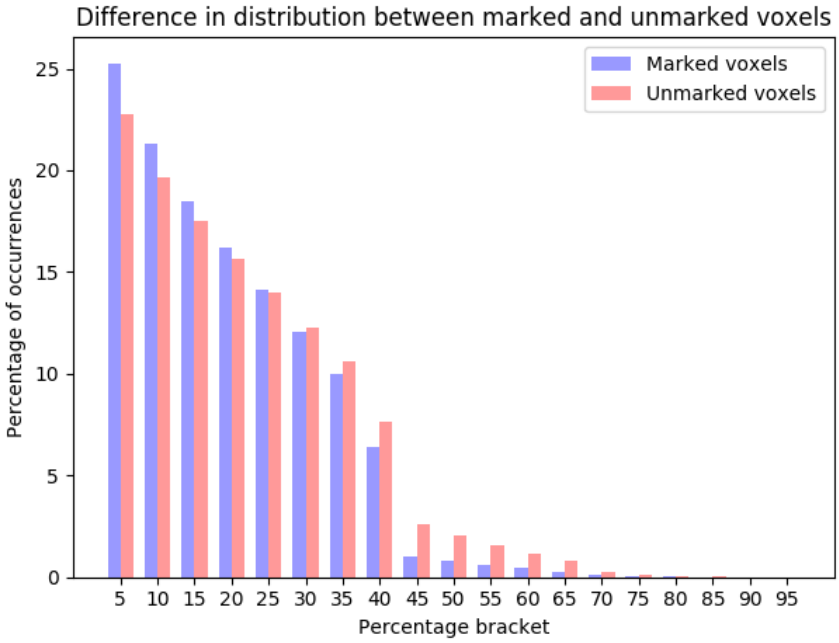


Figure 15: Difference in the distribution of entropy levels between the marked voxel unmarked voxels. The x-bar represents normalized entropy levels between 0 and 1 in blocks of 5% and the y-axis the percentage of occurrences.

4.5 Velocity-Threshold Identification (I-VT)

The Velocity-Threshold Identification algorithm is a spatial algorithm that classifies gaze points in eye-tracking data as either fixations or saccades based on whether the velocity between two consecutive data points surpasses the threshold.

Picking the right fixation detection algorithm is non-trivial, as there is no general consensus on which algorithm to pick from the wide range of possible algorithms and parameters. For this thesis, we use the I-VT for several reasons: its relatively ease of implementation, it has a clear guideline for picking the right parameter values, its performance is among the best[59], and because studies suggest event detection algorithms perform better with sampling rates of only 50hz.

For each data point, the visual angle is measured as follows [49]:

$$V = 2 \arctan \left(\frac{S}{2D} \right), \tag{20}$$

where S denotes the Euclidean distance between gazepoint x and gazepoint y of two data points, and D the distance in cm between the participant and the screen.

Then, values per second are measured, and when these values surpass V under the threshold are labeled as a saccade. We interpret uninterrupted rows of fixations as part of one fixation; the rest are one or more subsequent saccades.

Algorithm 1 Velocity-Threshold Identification. Adapted from [51]

```
function CLASSIFY GAZEPOINTS                    ▷ As saccades or fixations
    for gazepoint_t do
        gazedistance ← Euclidean distance(t, t − 1)
        angle ← 2arctan(gazedistance/2 ∗ eyedistance)
        velocity ← angle ∗ 1000/(t−(t−1))
        if velocity > velocity threshold then
            gazepoint_t ← saccade
        else
            gazepoint_t ← fixation
    return gazepoints

function REMOVE BLIPS                   ▷ Group fixations close in time or space
    for fixation_n (f_n) do
        if angle(f_n, f_{n−1}) < angle threshold then
            fixation_{n−1} ← merge(f_n, f_{n−1})
        else if offset_n − end_{f−1} < time threshold then
            fixation_{n−1} ← merge(f_n, f_{n−1})
    return fixations
```

We measure the location of each fixation group by calculating the centroid of the group, which is measured as follows:

$$C_x = \frac{x_1 + x_2 + \cdots + x_k}{k}, C_y = \frac{y_1 + y_2 + \cdots + y_k}{k} \tag{21}$$

Where K is a set of k fixation points consisting each of a x and y coordinate.

Then, fixations that are very near in time and location will be merged. Evinger et al. [137] stated that fixations within 75 ms and 0.5° should be merged, to account for respectively blinks and microsaccades.

After merging fixations, we remove saccades in the dataset and measure centroids of fixations again. For each fixation, the following features get extracted: fixation centroid, fixation area, fixation time, mean pupil size, pupil size standard deviation, fixation time point, and the participants experience level.

There is research on the link between pupil size and cognition [56] [138] [139], that suggest increased pupil size correlates with a high mental workload. For this reason, we use this as a feature. Blinks are also associated with workload [140] [141], hence these are extracted as well. In Section 2.2 we discussed some other aspects of expert knowledge, which motivated extracting the other features.

After extracting the features with the IV-T algorithm, we now have a (42293, 14) matrix, where each row corresponds to a unique fixation. As we can see, there are considerable differences in fixations that lead to an error (relevant) and fixations that don't lead to an error (irrelevant). Firstly, we see that relevant fixations tend to be more than twice as long than irrelevant fixations. This finding is in line with other eye-tracking studies about the visual behavior of experts in complex tasks. Mean fixation lengths can be deceiving, since outliers can influence these values easily. However, Figure 17 shows that the differences are fairly robust.

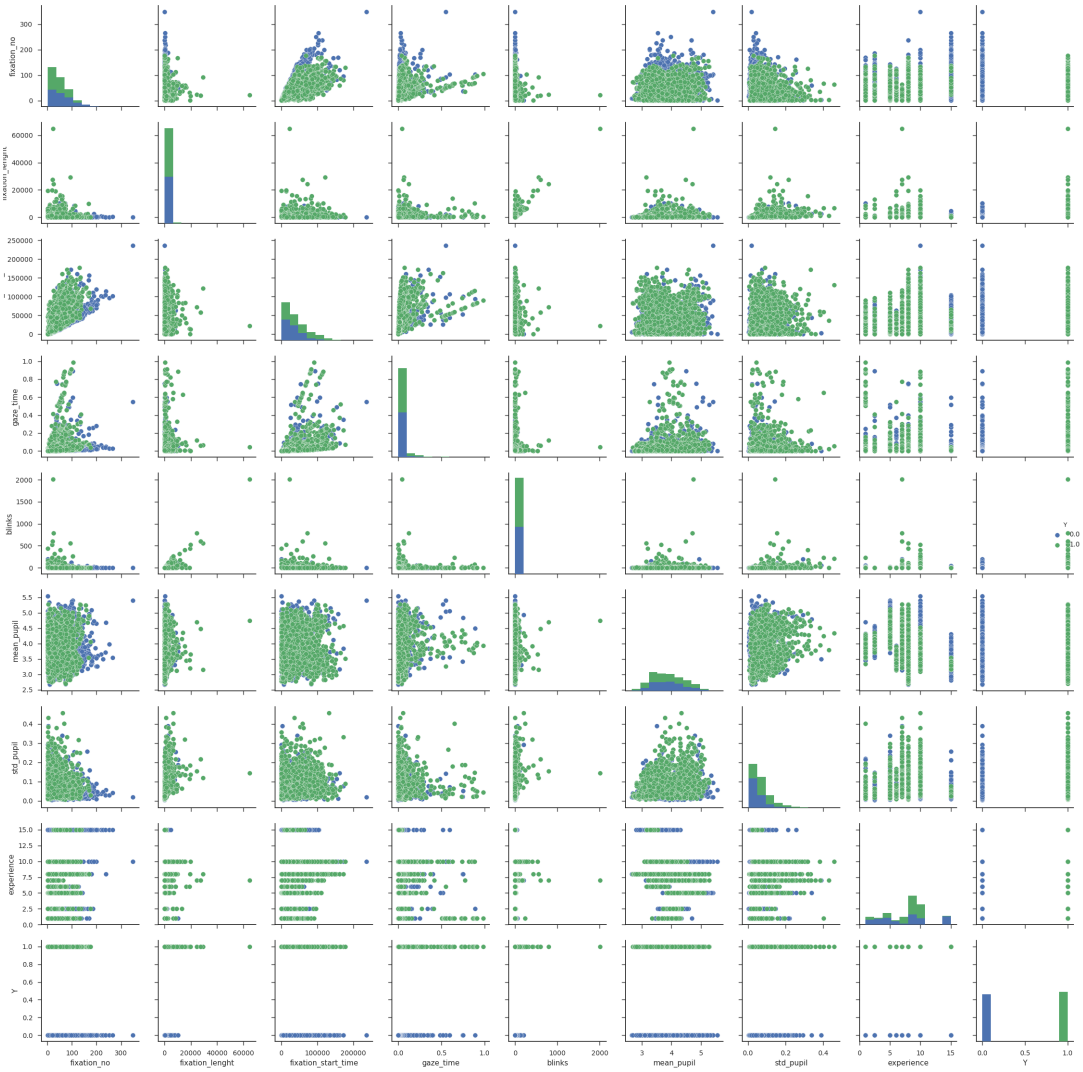


Figure 16: Scatterplot of the relationship between Eye behaviour features and information about participants.

Table 8: Differences in relevant and irrelevant fixations of both the normal markings and the dilated markings.

	No Dilation		**Dilation**	
	0	1	0	1
fixation number	59.47	45.89	60.08	47.75
fixation length	525.35	1277.97	494.72	1137.76
fixation start time	33487.46	46676.11	32951.08	44214.01
gaze time in slice	0.03	0.06	0.03	0.05
blinks during fixation	2.35	12.57	1.99	10.05
mean pupil size	3.92	3.87	3.92	3.89
pupil std	0.05	0.09	0.05	0.08

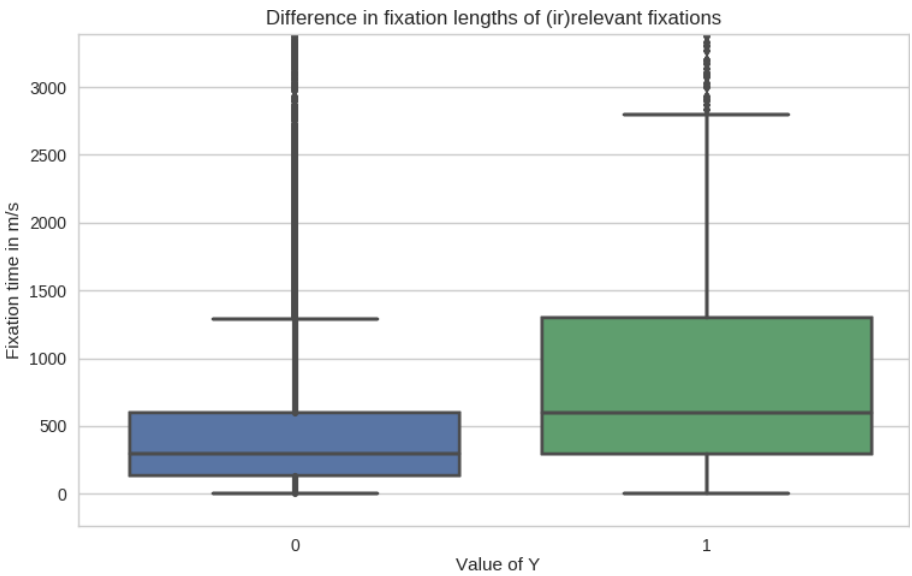


Figure 17: Difference in fixation times between fixations in areas with label error versus fixation times in areas that have a label correct. We can infer from this that fixation lengths could be useful for classifying errors, since the mean of class 1 is higher than the Q3 of class 0.

In Figure 17, we plotted the differences in fixation lengths between relevant and irrelevant fixations. Here we can see clearly that the distribution of fixation lengths with the current settings is asymmetric, with a positive skew. While this distribution could be because sometimes people fixate for long periods on one area, a more sensible explanation would be that this behavior is due to parameter settings. The value of the Finding a good value of the threshold is finding a balance between the length of the fixations and the number of fixations. The lower the threshold value, the more fixations with a lower average fixation time and vice versa.

In Figure 16 we plotted the difference of the relevant and irrelevant markings, to see whether we can find patterns. As we can see, some of these features show promising differences. See Table 8 for an overview of differences in values. Scatterplots are a great tool for visualizing the problem, before using classifiers. By plotting the relationship between features, we can quickly see whether some feature combinations allow for easy separation between relevant and irrelevant fixations. When there is a separation between the two classes, this means that they have a decision surface that allows for classification and gives an indication whether this is possible at all. As we can see in Figure 16, some of the features we extracted have a clear decision space. For example, we see that `fix_no` and `fix_start_time` show a difference between relevant and irrelevant fixations: relevant fixations tend to be at later stages. Another feature combination that looks promising is `fix_no` and `std_pupil`. .

These features can be useful for predicting target areas, but apart from that, they can also give insight into the way experts look at geological data. In Table 2, we show the difference in saccadic and fixation behaviour between experts

and novices. By extracting these features, we can test whether these theories hold for this data set.

Table 9: Extracted features with their descriptions.

Feature	**Definition**
Fixation length	Time between the onset and offset of the fixation
Fixation start time	Minimum size of 'timestamp' of the fixation
Fixation time point	Time point of the fixation, a continuous value where 0 denotes the start of the slice and 1 the end.
Centroid X	Centroid of the horizontal gaze of the fixation
Centroid Y	Centroid of the vertical gaze of the fixation
Blinks	Time of blinks in the fixation
Pupil mean	Mean size of the pupil during the fixation
Pupil std	The standard deviation of pupil size during fixation
Experience	The experience of the participant

5 Results

5.1 Mathematical attention model

First, we minimized the cost function of the mathematical attention model, to find optimal parameter settings. For each participant, we used their gaze data as input with the corresponding markings as output. This method proved to be unsuccessful, as we could not find optimum parameters. In fact, it turned out that the function did not converge at all and barely changed from initial parameters.

After the optimization function, we performed a linear search on the mathematical attention model. By using a wide range of parameters, we thought to get a clearer picture of the behavior of the parameters γ and λ on the AUC. However, there seemed to be no correlation between this either. See Figure 10 for the results over each fold in the 10-fold cross-validation. These results show that using the same parameters for each participant does not lead to good results. One explanation for this the high variance between participants, as we have shown in Table 6. Furthermore, in Table 7 we have shown that the number of pixels suffer from high variance either, as some participants mark over ten times as much pixels as others. When using individual parameter settings we can overcome these obstacles. For example, a participant that uses smaller markings can score better with a smaller γ, and a participant that uses smaller timeframes for spotting errors can use a higher λ.

Table 10: AUC values of each fold.

Fold	Lambda	Gamma	AUC
0	0.78	108.5	0.62
1	0.01	0.927	0.64
2	0.96	0.96	0.58
3	0.78	108.5	0.60
4	0.011	0.847	0.59
5	0.04	0.06	0.57
6	0.94	0.35	0.58
7	5.5	1.5	0.63
8	0.011	0.847	0.63
9	0.74	0.05	0.61

Table 11: AUC values of each fold after dilating the target data

Fold	Lambda	Gamma	AUC
0	0.78	108.5	0.59
1	0.01	0.927	0.60
2	0.89	0.13	0.54
3	0.78	108.5	0.55
4	0.011	0.847	0.54
5	0.04	0.06	0.55
6	0.96	0.53	0.56
7	0.1	11.7	0.56
8	0.87	0.04	0.54
9	0.81	0.08	0.55

5.2 Entropy

Apart from using gaze as input, we can also look at the model itself. Since entropy is a quantification of the model uncertainty, we can expect to find more errors at higher entropy voxels. The most promising variables are `model uncertainty lithoclass` and `model uncertainty lithostratigraphy`, since we hypothesize that a higher uncertainty of the voxel corresponds to whether an expert considers one to be wrong. Interestingly, these differences turned out to be marginal. See Figure 15 for an overview of the distribution of entropy levels between correctly labeled and incorrectly labeled voxels.

Table 12: AUC of the entropy data, where we use the targets of each participant.

Participant	AUC
0	0.578
1	0.425
2	0.423
3	0.561
4	0.483
5	0.689
6	0.452
7	0.528
8	0.632
9	0.631
10	0.667

Since the values of entropy are already between 0 and 1, an AUC can be measured for classifying voxels based on their entropy. Since each voxel has been labeled by each participant, every participant has its classifier. See Table e12 for the results. The results show that entropy is not a good predictor for labels of experts. Interestingly, participants 1 and 2, who had the largest amount of markings, scored worst. Participant 5, the best performer, had many slices without markings.

Table 13: Percentage of each lithostratigraphy occurrence in areas that are marked versus areas that are unmarked.

Lithostratigraphy	Label	Marking	Non-Marking
none	-999	0.140	0.167
Anthropogenic deposits	1000	0.000	0.003
Echteld Formation	1070	0.000	0.003
Nieuwkoop Formation, Hollandveen Member	1090	0.003	0.018
Naaldwijk Formation	1100	0.000	0.001
Nieuwkoop Formation, Basisveen laag	1130	0.001	0.002
Echteld Formation	2010	0.004	0.036
Boxtel Formation, Laagpakket van Wierden	3030	0.006	0.014
Boxtel Formation	3100	0.021	0.030
Kreftenheye Formation, Laag van Wijchen	4000	0.000	0.001
Krefenheye Formation, Laagpakket van Delwijnen	4010	0.338	0.244
Drente Formation	5000	0.006	0.026
Land Ice stowed units	5020	0.000	0.012
Urk Formation	5060	0.208	0.122
Sterksel Formation	5070	0.234	0.271
Peize and Waalre Formation	5120	0.032	0.019
Stroombaan geneation A, Echteld Formation	6000	0.000	0.001
Stroombaan generation A, Naaldwijk Formation	6100	0.003	0.012
Stroombaan generation C, Echteld Formation	6200	0.000	0.003
Stroombaan generation D	6300	0.001	0.016

Furthermore, we looked into the differences between markings in each lithostratigrapy, but these were marginal as well. See Table 13 for an overview.

5.3 Classification

The fixation dataset produced by the I-VT is not sufficient for predicting on its own but the output can serve as a basis for classification. In the following subsections, we show the results of three of these classifiers. As the scope of this thesis was on using feature extraction based on domain knowledge, we use off-the-shelf classifiers.

5.3.1 Logistic Regression

First, we use the `linear_model.LogisticRegression` of the scikit-learn package [142] for our logistic regression. We perform a 10-fold cross-validation, where each fold consists of 63 train slices and 7 test slices. The total number of observations (i.e. fixations) is 42293. Our motivation for 10 folds is because lower values would increase bias, while higher will increase variability. Since some slices do not contain any markings, we found that for example one-versus-all cross-validation have large outliers in performance. The other parameters that were used were a l2 penalty, a tolerance for stopping criteria of 0.001, and we used a weight based on the class imbalance. For example, when the training data consists of 28 times more irrelevant than relevant fixations, we weighted the relevant with a factor of 20. This resulted in an average AUC of 0.682, with

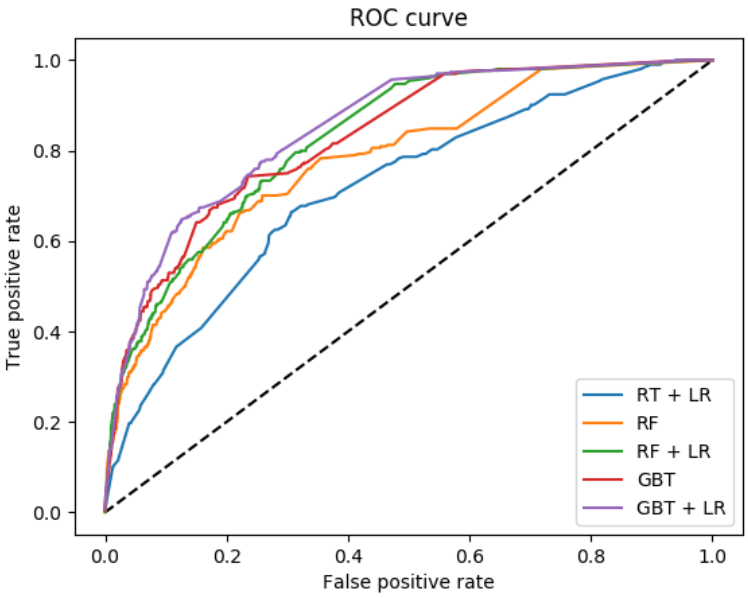


Figure 18: ROC curve of several tree-based methods.

a standard deviation of 0.02. These results are not convincing, but show that an easy implementation work for predicting the relevant areas. Furthermore, we can use this to find optimal parameter values. See Table 14 for finding the best time range for fixation merging. Since the literature gives parameter values for merging fixations between 75ms-425ms, we test the effect in this range.

Table 14: Effect of merging nearby fixations within a certain time frame, where the left column denotes the amount of milliseconds between the end time of $fixation_n$ and the begin time of $fixation_{n+1}$.

Merge fixations $<m/s$	**AUC test set**
75	0.675
110	0.672
145	0.680
180	0.673
215	0.675
250	0.643
285	0.676
320	0.643
355	0.656
390	0.670

5.3.2 Tree-based methods

Since the results of logistic regression suggest room for improvement, it makes sense to see whether other prediction models fare better. For this reason, we implement a Random Forest and a gradient boost. See Figure 18 for results. As we can see, these models perform better than logistic regression. Gradient Boosting and a general Random Forest stick out as best performing. With respectively 0.780 and 0.779, their results were almost identical. However, we also performed a one-versus-all Cross-validation, where the Random Forest performed better. Furthermore, the performance of the Random Forest is more robust: the standard deviation over folds is 0.045 and 0.119 with Random Forest versus 0.049 and 0.176 in the Gradient Boost with respectively 10-fold and 1-versus-all cross-validation. Testing per slice shows the high variance, indicating finding meaningful fixations is much harder in some slices than others. After this, we ran another test with both a test and validation set, in order to see whether the algorithms did not overfit to the hyperparameters. Here, we performed a cross validation with ten folds, each 7 slices in both test and validation set. With an average of 0.801 for the Random Forest and 0.800 for Gradient Boosting, the results were almost identical.

Table 15: Influence of each feature in the classification using Gradient Boosting.

Feature	**Influence**
Pupil mean	0.127
Pupil std	0.126
Fixation start time	0.124
Gaze x	0.108
Gaze y	0.098
Slice	0.094
Blinks	0.086
Fixation in time	0.084
Fixation length	0.077
Fixation number	0.076

After this, we measured feature importance to see how the model makes decisions. See Table 15 for an overview. It is interesting to see that the mean and standard deviation of the pupil size are the most important features, as previous research did not use these features.

6 Discussion and Future Research

In this section, we point out a few directions for further research. As we have discussed in Section 2, computer-aided quality control entails a lot of interesting fields of study, many of them active and progressing. There are numerous ways of advancing current research.

Firstly, we can extract more features. We have shown that several common eye tracking features deal with saccades, but due to low sample rate saccades were unavailable to extract. Future research with a higher sample rate can lead to extracting features like saccade lengths, saccade duration, inter-scanning count, and scan path patterns. These can potentially improve the prediction power of the current model. Apart from eye movement, using mouse behavior as a feature can also be an indicator of interesting areas in a voxel model [143].

Secondly, a focus on using the correct classifiers will presumably lead to a higher AUC in the IV-T model. Since we used off the shelve classifiers, we hypothesize that there is still room for optimization in this area. For example, Ganganwar [144] mentions performances of several classifiers like a modified support vector machine with good performances on unbalanced datasets. Developing a robust classifier for increased performance could be an interesting future research opportunity.

Thirdly, general research in eye-tracking can lead to better performance and understanding of the problem space. Currently, there is no ground truth in fixation algorithms; the parameters should be inferred empirically. Further research could use hand labeling to find these parameters. The eye-tracking data in this research can provide a useful data set for testing fixation algorithms, but this is a time-consuming endeavor, performed by trained people in labeling fixations. For example, Hessels et al. [145] stated that 40 min of eye-tracking data took around 3 hours of labeling. Since the scope of automating geological voxel models is rather a niche subject, using eye-tracking experts to replace geological experts will not result in a time saving procedure. However, we showed that even a simple algorithm was very successful in predicting these labels.

Fourthly, a move from 2D to 3D in computer-aided predictions [146] could increase saved time even further. Current research uses 2D slices, but the geological voxel models are 3D . Research on predicting errors in 3D models could be an interesting pursuit, as we do not have to make samples of slices, but can use the entire model.

Fifthly, the questionnaire performed on six slices gives valuable information on how wrong an error was and the type of error. Having this information on all eye-tracking data could advance research immensely. When we have an idea of how wrong each marked area is, we can use this as a better ground truth. Having a ground truth can be used to research whether cognitive workload can lead to wrong markings, and can help with training the experts as well. Findings in the field of radiology could potentially be useful, due to having similar goals [147][148][149]. The type of error can provide interesting information as well since we can study whether some experts focus more on certain types of errors and we can find whether some errors occur only in some parts.

Sixthly, while current models can separate relevant from irrelevant fixations, ideally, the model would be able to spot different objects within a fixation. Sometimes several semantically dissimilar groups of markings occur within one

fixation. Mathe and Sminchisescu (2013) [150] were able to find semantically different areas by implementing a Hidden Markov Model that classifies Areas of Interest.

We showed that you can use eye behavior in a pattern recognition pipeline to predict the location of errors. Since the feature selection plus classification performs faster than the total amount of gaze time we can do predictions in real-time. Future research should implement software that uses an eye-tracker to predict these errors. Apart from errors, the quality control of GeoTOP consists of the following information per marked error:

1. The error number
2. Date
3. Status ('Open','Solved','Do not solve')
4. Borehole number where the error occurred
5. Stratigraphic unit on which the error occurs
6. A description of the error

We can incorporate most of these points easily in a software tool, but the description of the error is non-trivial. Some initial work has been done on these via a questionnaire, where for each error the participants were asked to indicate the type of error. See Table 19 for this questionnaire.

One possible method is to decompose the multiclass problem into several binary classification tasks. For example, a margin-based binary learning algorithm [151] has been used after decomposition, or it can be solved via error-correcting output codes [152]. Another possibility is using a Support Vector Machine [153].

Despite the challenging nature of this problem, we hypothesize this can significantly increase the benefit of automating parts of quality control, since this would automatize the annotation part completely.

7 Conclusions

We gave an overview of research and methods in computer aided quality control in geological voxel models. The multifaceted nature of the subject makes this an interesting and complex problem, incorporating ideas that go beyond just geomodelling. We tested three models, each with a hypothesis.

Conclusion 1. *The Attentional Model lacks generalizing power*

While in previous research [42] showed good results, applying this model for a generalized model was less convincing. Since the participants vary too much, a model predicting on pixel level will suffer in results due to marking differences.

Conclusion 2. *Entropy is not a good predictor for finding errors in the GeoTOP model*

Surprisingly, entropy turned out to be not a good predictor for errors. Since by definition entropy should say something about model uncertainty, we expected this to see a clear overlap between entropy and expert assessment. As this is not the case, changes should be made to the entropy model.

Conclusion 3. *Fixation identification algorithms can be used for real-time prediction in GeoTOP models*

We showed that using knowledge of eye movements can produce a good predicting model for markings in GeoTOP models. Even without access to a ground truth for fixations, our model was able to perform well on this dataset. With simple classifiers, an AUC of 0.78 with a 10-fold cross-validation was achieved. Further research on feature extraction and optimized classifiers will increase this performance even further.

References

[1] Wim Westerhoff and Wim Dubelaar. Beneath the Netherlands - Geological Survey of the Netherlands. Brochure, 2013.

[2] M Korff. Case studies and monitoring of deep excavations. *Proceedings of the 9th International Symposium on Geothechnical Aspects of Underground Construction in Soft Ground*, 2017.

[3] Tamer Elkateb, Rick Chalaturnyk, and Peter K Robertson. An overview of soil heterogeneity: quantification and implications on geotechnical field problems. *Canadian Geotechnical Journal*, 40(1):1–15, 2003.

[4] Bernard Dost and Dirk Kraaijpoel. The august 16, 2012 earthquake near huizinge (groningen). *KNMI Scientific Report. Royal Netherlands Meteorological Institute (KNMI), Utrecht, The Netherlands*, 2013.

[5] AG Muntendam-Bos and JA De Waal. Reassessment of the probability of higher magnitude earthquakes in the groningen gas field. *State Supervision of Mines Publishing, The Hague-Leidschenveen*, 2013.

[6] Nick Van der Voort and Frank Vanclay. Social impacts of earthquakes caused by gas extraction in the province of groningen, the netherlands. *Environmental Impact Assessment Review*, 50:1–15, 2015.

[7] Pauline P Kruiver, Ane Wiersma, Fred H Kloosterman, Ger de Lange, Mandy Korff, Jan Stafleu, Freek S Busschers, Ronald Harting, Jan L Gunnink, Russell A Green, et al. Characterisation of the groningen subsurface for seismic hazard and risk modelling. *Netherlands Journal of Geosciences*, 96(5):s215–s233, 2017.

[8] Ministry of Infrastructure and the Environment. Wet basisregistratie ondergrond (BRO) no 30/2015, 2015. https://zoek.officielebekendmakingen.nl/stb-2015-362.html.

[9] Daniel J Power, Ramesh Sharda, and Frada Burstein. *Decision support systems*. Wiley Online Library, 2015.

[10] Mary M Poulton. Neural networks as an intelligence amplification tool: A review of applications. *Geophysics*, 67(3):979–993, 2002.

[11] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[12] E. L. van den Broek. *Affective Signal Processing (ASP): Unraveling the mystery of emotions*. PhD thesis, Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, The Netherlands, 2011.

[13] Lorrie Faith Cranor. A framework for reasoning about the human in the loop. *UPSEC*, 8(2008):1–15, 2008.

[14] Roger PG Van Gompel. *Eye movements: A window on mind and brain*. Elsevier, 2007.

[15] Colin Ware and Harutune H Mikaelian. An evaluation of an eye tracker as a device for computer input2. In *Acm sigchi bulletin*, volume 17, pages 183–188. ACM, 1987.

[16] Linda E Sibert and Robert JK Jacob. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 281–288. ACM, 2000.

[17] Michael F Land and Sophie Furneaux. The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 352(1358):1231–1239, 1997.

[18] Päivi Majaranta and Kari-Jouko Räihä. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 15–22. ACM, 2002.

[19] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer, 2014.

[20] RJ Jacob and Keith S Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3):4, 2003.

[21] Timothy L Nyerges, David M Mark, Robert Laurini, and Max J Egenhofer. *Cognitive aspects of human-computer interaction for geographic information systems*, volume 83. Springer Science & Business Media, 2012.

[22] Jean-Laurent Mallet. *Geomodeling*. Oxford University Press, 2002.

[23] Malti Goel. Sustainable energy through carbon capture and storage: role of geo-modeling studies. *Energy & Environment*, 23(2-3):299–317, 2012.

[24] EJ Sides. Geological modelling of mineral deposits for prediction in mining. *Geologische Rundschau*, 86(2):342–353, 1997.

[25] A Keith Turner. Challenges and trends for geological modelling and visualisation. *Bulletin of Engineering Geology and the Environment*, 65(2):109–127, 2006.

[26] Flemming Jørgensen, Rasmus Rønde Møller, Lars Nebel, Niels-Peter Jensen, Anders Vest Christiansen, and Peter BE Sandersen. A method for cognitive 3d geological voxel modelling of aem data. *Bulletin of Engineering Geology and the Environment*, 72(3-4):421–432, 2013.

[27] Holger Kessler and Stephen Mathers. The past, present and future of 3d geology in bgs. *Journal Open University Geological Society*, 27(2):13–15, 2006.

[28] A. Keith Turner. Challenges and trends for geological modelling and visualisation. *Bulletin of Engineering Geology and the Environment*, 65(2):109–127, May 2006.

[29] Luis Manuel de Vries, Jesus Carrera, Oriol Falivene, Oscar Gratacós, and Luit Jan Slooten. Application of multiple point geostatistics to non-stationary images. *Mathematical Geosciences*, 41(1):29, 2009.

[30] Steven F Carle and Graham E Fogg. Transition probability-based indicator geostatistics. *Mathematical geology*, 28(4):453–476, 1996.

[31] Clayton V Deutsch, Andre G Journel, et al. Geostatistical software library and user's guide. *New York*, 119:147, 1992.

[32] Katherine R Royse. Combining numerical and cognitive 3d modelling approaches in order to determine the structure of the chalk in the london basin. *Computers & Geosciences*, 36(4):500–511, 2010.

[33] Erik R Venteris. Three-dimensional modeling of glacial sediments using public water-well data records: An integration of interpretive and geostatistical approaches. *Geosphere*, 3(6):456–468, 2007.

[34] Qiang Wu, Hua Xu, and Xukai Zou. An effective method for 3d geological modeling with multi-source data integration. *Computers & Geosciences*, 31(1):35–43, 2005.

[35] J Florian Wellmann and Klaus Regenauer-Lieb. Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models. *Tectonophysics*, 526:207–216, 2012.

[36] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[37] Philippe Calcagno, Jean-Paul Chilès, Gabriel Courrioux, and Antonio Guillen. Geological modelling from field data and geological knowledge: Part i. modelling method coupling 3d potential-field interpolation and geological rules. *Physics of the Earth and Planetary Interiors*, 171(1-4):147–157, 2008.

[38] Jan Stafleu and Denise Maljers. Geotop - 3d modelling of the shallow subsurface.

[39] J Stafleu and CW Dubelaar. Product specification subsurface model geotop. Technical report, version 1.3, Tech. Rep, 2016.

[40] Amilcar Soares. Geostatistical estimation of multi-phase structures. *Mathematical geology*, 24(2):149–160, 1992.

[41] J Stafleu, D Maljers, FS Busschers, JL Gunnink, J Schokker, RM Dambrink, HJ Hummelman, and ML Schijf. Geotop modellering. *TNO report*, 10991, 2012.

[42] Peter-Paul van Maanen, Freek S Busschers, Anne-Marie Brouwer, Michiel J van der Meulen, and Jan BF van Erp. Quality control of geological voxel models using experts' gaze. *Computers & geosciences*, 76:50–58, 2015.

[43] D Maljers, J Stafleu, MJ Van der Meulen, and RM Dambrink. Advances in constructing regional geological voxel models, illustrated by their application in aggregate resource assessments. *Netherlands Journal of Geosciences*, 94(3):257–270, 2015.

[44] Martin Shepherd, John M Findlay, and Robert J Hockey. The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology Section A*, 38(3):475–491, 1986.

[45] Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

[46] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[47] Joseph Coyne and Ciara Sibley. Investigating the use of two low cost eye tracking systems for detecting pupillary response to changes in mental workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 37–41. Sage Publications Sage CA: Los Angeles, CA, 2016.

[48] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. *arXiv preprint arXiv:1606.05814*, 2016.

[49] Andrew T Duchowski. *Eye tracking methodology: Theory and practice*. Springer, 2017.

[50] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[51] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[52] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[53] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):105–119, 2010.

[54] Andrew T. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4):455–470, Nov 2002.

[55] Joseph H Goldberg and Anna M Wichansky. Eye tracking in usability evaluation: a practitioners guide. *To appear in: Hyönä*, 2002.

[56] Marc Pomplun and Sindhura Sunkara. Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI*, volume 2003, 2003.

[57] Enkelejda Tafaj, Gjergji Kasneci, Wolfgang Rosenstiel, and Martin Bogdan. Bayesian online clustering of eye movement data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 285–288. ACM, 2012.

[58] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 163–170. ACM, 2016.

[59] Richard Andersson, Linnea Larsson, Kenneth Holmqvist, Martin Stridh, and Marcus Nyström. One algorithm to rule them all? an evaluation and discussion of ten eye movement event-detection algorithms. *Behavior research methods*, 49(2):616–637, 2017.

[60] Meng-Lung Lai, Meng-Jung Tsai, Fang-Ying Yang, Chung-Yuan Hsu, Tzu-Chien Liu, Silvia Wen-Yu Lee, Min-Hsien Lee, Guo-Li Chiou, Jyh-Chong Liang, and Chin-Chung Tsai. A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational research review*, 10:90–115, 2013.

[61] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, 23(4):523–552, 2011.

[62] Marcel Adam Just and Patricia A Carpenter. Eye fixations and cognitive processes. *Cognitive psychology*, 8(4):441–480, 1976.

[63] Laura E. Thomas and Alejandro Lleras. Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review*, 14(4):663–668, Aug 2007.

[64] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.

[65] Tony Cheng. Attention, fixation, and change blindness. 2017.

[66] Zenon W Pylyshyn. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1):127 – 158, 2001. Objects and Attention.

[67] Charles James, Zachary Daniel, and Charles JZ Daniel. The principles of psychology. 1890.

[68] Tibor Bosse, Peter-Paul van Maanen, and Jan Treur. Simulation and formal analysis of visual attention. *Web Intelligence and Agent Systems: An International Journal*, 7(1):89–105, 2009.

[69] Ali Borji, Dicky N Sihite, and Laurent Itti. What/where to look next? modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(5):523–538, 2014.

[70] John M Henderson and Andrew Hollingworth. High-level scene perception. *Annual review of psychology*, 50(1):243–271, 1999.

[71] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[72] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[73] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[74] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[75] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

[76] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[77] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[78] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, pages 241–248, 2008.

[79] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in neural information processing systems*, pages 497–504, 2008.

[80] Wolf Kienzle, Felix A Wichmann, Matthias O Franz, and Bernhard Schölkopf. A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems*, pages 689–696, 2007.

[81] Gert Kootstra, Arco Nederveen, and Bart De Boer. Paying attention to symmetry. In *British Machine Vision Conference (BMVC2008)*, pages 1115–1125. The British Machine Vision Association and Society for Pattern Recognition, 2008.

[82] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008.

[83] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in neural information processing systems*, pages 681–688, 2009.

[84] Derek Pang, Akisato Kimura, Tatsuto Takeuchi, Junji Yamato, and Kunio Kashino. A stochastic model of selective visual attention with a dynamic bayesian network. In *Multimedia and expo, 2008 ieee international conference on*, pages 1073–1076. IEEE, 2008.

[85] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil. Decorrelation and distinctiveness provide with human-like saliency. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 343–354. Springer, 2009.

[86] Masaki Fukuchi, Naotsugu Tsuchiya, and Christof Koch. The focus of expansion in optical flow fields acts as a strong cue for visual attention. *Journal of Vision*, 9(8):137–137, 2009.

[87] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE transactions on image processing*, 19(1):185–198, 2010.

[88] Tamar Avraham and Michael Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):693–708, 2010.

[89] Charles E Connor, Howard E Egeth, and Steven Yantis. Visual attention: bottom-up versus top-down. *Current biology*, 14(19):R850–R852, 2004.

[90] Joseph B Hopfinger, Michael H Buonocore, and George R Mangun. The neural mechanisms of top-down attentional control. *Nature neuroscience*, 3(3):284, 2000.

[91] Maurizio Corbetta, J Michelle Kincade, John M Ollinger, Marc P McAvoy, and Gordon L Shulman. Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature neuroscience*, 3(3):292, 2000.

[92] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[93] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

[94] Gregory Zelinsky, Wei Zhang, Bing Yu, Xin Chen, and Dimitris Samaras. The role of top-down and bottom-up processes in guiding eye movements during visual search. In *Advances in neural information processing systems*, pages 1569–1576, 2006.

[95] John M Henderson, James R Brockmole, Monica S Castelhano, and Michael Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements*, pages 537–III. Elsevier, 2007.

[96] Wolfgang EinhÃ, Ueli Rutishauser, Christof Koch, et al. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of vision*, 8(2):2–2, 2008.

[97] Michael F Land and David N Lee. Where we look when we steer. *Nature*, 369(6483):742, 1994.

[98] Raymond D Rimey and Christopher M Brown. Control of selective perception using bayes nets and decision theory. *International Journal of Computer Vision*, 12(2-3):173–207, 1994.

[99] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

[100] Erik D Reichle, Keith Rayner, and Alexander Pollatsek. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476, 2003.

[101] Mary M Hayhoe. Advances in relating eye movements and cognition. *Infancy*, 6(2):267–274, 2004.

[102] Robert J Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[103] Nathan Sprague, Dana Ballard, and Al Robinson. Modeling embodied visual behaviors. *ACM Transactions on Applied Perception (TAP)*, 4(2):11, 2007.

[104] Robert Peters and Laurent Itti. Congruence between model and human attention reveals unique signatures of critical visual events. In *Advances in neural information processing systems*, pages 1145–1152, 2008.

[105] Nicholas J Butko and Javier R Movellan. Optimal scanning for faster object detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 2751–2758. IEEE, 2009.

[106] Ruben Coen-Cagli, Paolo Coraggio, Paolo Napoletano, Odelia Schwartz, Mario Ferraro, and Giuseppe Boccignone. Visuomotor characterization of eye movements in a drawing task. *Vision research*, 49(8):810–818, 2009.

[107] Tom Erez, Julian J Tramper, William D Smart, and Stan CAM Gielen. A pomdp model of eye-hand coordination. In *AAAI*, 2011.

[108] Tom M Mitchell. Does machine learning really work? *AI magazine*, 18(3):11, 1997.

[109] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[110] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[111] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[112] Daniel Kahneman and Amos Tversky. On the study of statistical intuitions. *Cognition*, 11(2):123–141, 1982.

[113] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[114] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[115] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[116] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

[117] Gaurav Sinha, Rahul Shahi, and Mani Shankar. Human computer interaction. In *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on*, pages 1–4. IEEE, 2010.

[118] L Schomaker. A taxonomy of multimodal interaction in the human information processing system, 1995. Met lit. opg.

[119] Tibor Bosse, Rianne Van Lambalgen, Peter-Paul van Maanen, and Jan Treur. A system to support attention allocation: Development and application. *Web Intelligence and Agent Systems: An International Journal*, 10(1):1–17, 2012.

[120] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[121] Richard Andersson, Marcus Nyström, and Kenneth Holmqvist. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research*, 3(3), 2010.

[122] JT Enright. Estimating peak velocity of rapid eye movements from video recordings. *Behavior Research Methods, Instruments, & Computers*, 30(2):349–353, 1998.

[123] M Juhola, V Jäntti, and I Pyykkö. Effect of sampling frequencies on computation of the maximum velocity of saccadic eye movements. *Biological Cybernetics*, 53(2):67–72, 1985.

[124] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[125] Steven E Stemler. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4):1–19, 2004.

[126] Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies.

[127] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA, 1997.

[128] Marcus A Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, pages 2–1, 2003.

[129] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[130] E. L. van den Broek, Th. E. Schouten, P. M. F. Kisters, and H. C. Kuppens. Weighted Distance Mapping (WDM). In N. Canagarajah, A. Chalmers, F. Deravi, S. Gibson, P. Hobson, M. Mirmehdi, and S. Marshall, editors, *Proceedings of the IEE International Conference on Visual Information Engineering (VIE2005)*, pages 157–164, Glasgow, United Kingdom, 2005. Wrightsons - Earls Barton, Northants, Great Britain.

[131] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001-2018. [Online; accessed December 1st. Version 1.0.0].

[132] Azriel Rosenfeld and John L Pfaltz. Distance functions on digital pictures. *Pattern recognition*, 1(1):33–61, 1968.

[133] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[134] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

[135] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[136] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[137] Craig Evinger, Karen A Manning, and Patrick A Sibony. Eyelid movements. mechanisms and normal data. *Investigative ophthalmology & visual science*, 32(2):387–400, 1991.

[138] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1477–1480. ACM, 2004.

[139] Miguel A Recarte and Luis M Nunes. Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, 9(2):119, 2003.

[140] JA Veltman and AWK Gaillard. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5):656–669, 1998.

[141] Bin Zheng, Xianta Jiang, Geoffrey Tien, Adam Meneghetti, O Neely M Panton, and M Stella Atkins. Workload assessment of surgeons: correlation between nasa tlx and blinks. *Surgical endoscopy*, 26(10):2746–2750, 2012.

[142] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[143] Vignesh Raghunath, Melissa O Braxton, Stephanie A Gagnon, Tad T Brunyé, Kimberly H Allison, Lisa M Reisch, Donald L Weaver, Joann G Elmore, and Linda G Shapiro. Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. *Journal of pathology informatics*, 3, 2012.

[144] Vaishali Ganganwar. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47, 2012.

[145] Roy S Hessels, Diederick C Niehorster, Chantal Kemner, and Ignace TC Hooge. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior research methods*, 49(5):1802–1823, 2017.

[146] Eui Chul Lee, Jin Cheol Woo, Jong Hwa Kim, Mincheol Whang, and Kang Ryoung Park. A brain–computer interface method combined with eye tracking for 3d interaction. *Journal of neuroscience methods*, 190(2):289–298, 2010.

[147] Lindsey Cooper, Alastair Gale, Iain Darker, Andoni Toms, and Janak Saada. Radiology image perception and observer performance: how does expertise and clinical information alter interpretation? stroke detection explored through eye-tracking. In *Medical Imaging 2009: Image perception, observer performance, and technology assessment*, volume 7263, page 72630K. International Society for Optics and Photonics, 2009.

[148] A Van der Gijp, CJ Ravesloot, H Jarodzka, MF van der Schaaf, IC van der Schaaf, Jan PJ van Schaik, and Th J Ten Cate. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education*, 22(3):765–787, 2017.

[149] David J Manning, SC Ethell, and Tim Donovan. Detection or decision errors? missed lung cancer from the posteroanterior chest radiograph. *The British journal of radiology*, 77(915):231–235, 2004.

[150] Stefan Mathe and Cristian Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in neural information processing systems*, pages 1923–1931, 2013.

[151] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research*, 1(Dec):113–141, 2000.

[152] Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286, 1995.

[153] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

Appendices

A Voxel attributes

Table 16: Specification of all attributes that a voxel in the GeoTOP model contains.

#	Attribute	Description
1	Model surface version	Indicates the version of the model's area
2	X coordinate	The X coordinate calculated from the center of the voxel
3	Y coordinate	The Y coordinate calculated from the center of the voxel
4	Z coordinate	The Z coordinate in meters in relation with NAP, measured from the center of the voxel
5	Geological Unit	Refers to the geological unit of the voxel
6	Most probable lithoclass	Refers to the the lithoclass the voxel is colored with
7	Probability anthropogenic deposits	Probability of the voxel being filled with anthropogenic deposits, where 0 denotes low probability and 1 high probability
8	Probability clay	Same for this class
9	Probability sandy clay and clayey sand	Same for this class
10	Probability fine sand	Same for this class
11	Probability coarse sand	Same for this class
12	Probability medium grained sand	Same for this class
13	Probability gravel	Same for this class
14	Probability peat	Same for this class
15	Model uncertainty geological unit	The degree in which the model is capable of predicting the correct where 0 denotes low uncertainty and 1 high uncertainty
16	Model uncertainty lithoclass	The degree in which the model is capable of predicting the correct where 0 denotes low uncertainty and 1 high uncertainty

B Eye-tracking Output

Table 17: The column output of gaze data of `gaze.csv` after each trial.

Data	Description
Number	A sequential serial number given to the gaze point
Time	The timestamp, in ms, for this gaze point from the start of the recording
Screen X (left eye)	The horizontal position of the gaze point, measured in pixels from the left
Screen Y (left eye)	The vertical position of the gaze point, measured in pixels from the top
Cam X (left eye)	The horizontal location of the pupil in the camera image, on a scale from 0 to 1
Cam Y (left eye)	The vertical location of the pupil in the camera image, on a scale from 0 to 1
Distance (left eye)	The distance from the camera to the left eye, in mm
Pupil (left eye)	The size of the pupil, in mm
Code (left eye)	The validity of the gaze data
Screen X (right eye)	The horizontal position of the gaze point, measured in pixels from the left
Screen Y (right eye)	The vertical position of the gaze point, measured in pixels from the top
Cam X (right eye)	The horizontal location of the pupil in the camera image, on a scale from 0 to 1
Cam Y (right eye)	The vertical location of the pupil in the camera image, on a scale from 0 to 1
Distance (right eye)	The distance from the camera to the right eye, in mm
Pupil (right eye)	The size of the pupil, in mm
Code (right eye)	The validity of the gaze data

Table 18: Event data columns of `event.csv` files for each recording.

Event	Event Key	Data 1	Data 2
Showslide	4	Slide number	-
Hideslide	5	Slide number	-
Keyboard	3	ASCII code for key pressed	-
LMouseButton	1	X mouse coordinate	Y mouse coordinate
RMouseButton	2	X mouse coordinate	Y mouse coordinate

C Expert Questionnaire

Table 19: Questions asked during the questionnaire. For six out of the 70 slies, experts were asked to answer these questions about marked errors in the slice.

Questions		
How wrong is the marked area?	0 = not wrong	100 = completely wrong
How sure are you are you of the error?	0 = unsure	100 = completely sure
How easy is the error to spot?	0 = not easy	100 = very easy
How important is the error?	0 = unimportant	100 = very important
How conspicuous is the error?	0 = very inconspicuos	100 = very conspicuos
Is the error marked correctly?	0 = incorrect	100 = correct
What is the error type?		
The base of the unit is too pointy	0 = incorrect	100 = correct
The base of the unit is too irregular	0 = incorrect	100 = correct
The top of the unit is too irregular	0 = incorrect	100 = correct
The lateral transition between the units is too sharp	0 = incorrect	100 = correct
A certain unit is not present or underrepresented	0 = incorrect	100 = correct
A certain unit is overrepresented	0 = incorrect	100 = correct
The unit is lying too deep	0 = incorrect	100 = correct
The unit is lying too high	0 = incorrect	100 = correct
The order of units is wrong	0 = incorrect	100 = correct
The shape of the units is wrong	0 = incorrect	100 = correct