KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING

Decoding Metacognitive Sensitivity from EEG using Deep Learning.

Promoter:

Prof. Jean-Marie Aerts

Prof. Kobe Desender

Department Biosystems,

Department Brein en Cognitie

Dissertation presented in

fulfillment of the requirements

for the degree of Master of Bioscience Engineering:

Human Health Engineering

Juul Vanden Abeele

September

Acknowledgements

This thesis is the cherry on the cake regarding my five-year journey studying Bio-science engineering. Starting at the University of Antwerp and after completing my bachelors joining the Catholic University of Leuven to finish my masters. This journey has been one of tremendous personal growth, till the extent that my achievement is near inconceivable to my high school teachers. I am unbelievably grateful for the facilities provided to me. Being able to pursue my thirst for knowledge in the topics I am passionate about. Without exaggeration, my major of Human Health Engineering complemented with the minor of entrepreneurship and innovation, have given me the exact tools I desired to initiate my professional career.

First, I want to extend my gratitude to my promotors: Jean-Marie Aerts and Kobe Desender. They have provided me with great guidance in this unusual adventure. I remember inquiring for a self-proposed topic in the office of Jean-Marie. The enthusiasm Jean-Marie showed was extremely motivating to pursue this challenge. Quite a challenge, I proposed a totally different project than the one you will be reading today. During the literature review in the first semester, I had a hard time defining what I wanted to achieve with the thesis. Luckily, after stumbling in the office of Kobe Desender, his expertise in a concept foreign to me instantly brought clarity to the project. After finally being able to define an objective, Jean-Marie Aerts and Kobe Desender provided me with excellent guidance and showed great availability for feedback.

Second, I want to thank Michael Briden and Narges Norouzi for providing me with the code of their research paper. This allowed me to invest time in improving the results and the application of the deep learning framework, instead of having to recreate it. Further I want to thank Annika Boldt and Nick Yeung for providing Kobe Desender and eventually me with the EEG dataset I required to pursue this project.

Third, I want to thank my family for supporting me throughout not only the last couple of intense weeks, but throughout my whole academic journey. By providing for me they have greatly contributed to helping me achieve my goals.

Last but not least, I want to thank me for putting in the work, being there for myself when I need it and pursuing this topic regardless of how unfamiliar and uncertain this adventure was. Proposing my own topic entailed a greater challenge but was undoubtably worth it.

Scientific summary

Metacognition, the ability to think about one's thinking processes, is vital for professional performance, academic achievement, and mental health. However, its ambiguous nature and subjective measurement techniques across various fields have posed significant challenges to research. Cognitive neuroscience offers a unique solution by providing objective measurements that link metacognition to brain activity, thereby establishing a ground truth. Recently, the convergence of explainable artificial intelligence (XAI) and perceptual decision-making, a subsection of metacognition within cognitive neuroscience, has led to the development of the WaveFusion framework. This innovative framework holds the potential to contribute to the unification of the fragmented metacognition research fields.

The aim of this thesis was to enhance the WaveFusion framework, an explainable deep learning model, to classify metacognitive sensitivity and confidence using EEG data. The objectives were (1) to achieve a classification accuracy of 95% for metacognitive sensitivity, (2) to improve the accuracy for metacognitive confidence to 97.5%, and (3) to identify key ambiguities and limitations in metacognition research.

This study utilized an EEG dataset with event-related potentials (ERP) response-locked for type 1 decisions. Data preprocessing addressed dataset imbalances through augmentation and balanced batch sampling. EEG samples were transformed into spectrograms and processed using the deep learning architecture comprising a Lightweight Convolutional Neural Network (LWCNN), a Squeeze and Excitation Network (SEN), and a classification network. The model was pre-trained using Subject Aware Contrastive loss (SAC) and trained with binary cross-entropy loss. SEN facilitated the models explainability by visualizing the created attention weights through topoplots, providing insights into brain areas used for classification.

The WaveFusion model achieved high classification accuracy, reaching 99.7% for metacognitive confidence and 99.1% for metacognitive sensitivity. These improvements were due to a larger selection of electrodes, response-locked ERP data, and increased dataset size. The WaveFusion model not only demonstrates high classification accuracy but also offers enhanced explainability. This allows the framework to contribute to three major ambiguities: (1) the relationship between metacognition and executive functions, (2) its connection to consciousness, and (3) the domain generality of metacognition. By leveraging the WaveFusion framework, we can overcome limitations in cognitive neuroscience research through (1) utilizing transfer learning to compare relationships, (2) employing automatic classification to investigate ecological validity, and (3) expanding the framework for multimodality to integrate insights across various fields.

Future research should focus on increasing data variability, addressing outlier performances, and improving interpretability through advanced visualization techniques to enhance the WaveFusion model’s robustness and applicability across cognitive neuroscience domains.

List of tables

List of figures

Table of contents

1. Introduction

You do not have full access to your mind and it limits your professional performance, your academic performance and your mental health [1], [2], [3], [4]. Many blame this on unchangeable factors like their innate intelligence, their circumstances or even their own consciousness. However, there is an essential factor that can be continuously developed. Metacognition, generally defined as “thinking about thinking”, a psychological mechanism that controls your cognitive functions. A seemingly foreign concept, but it plays a crucial role in your day-to-day life. Formally introduced in 1976 by John H. Flavell an American psychologist specialized in cognitive development. Metacognition has since been associated to ordinary concepts like theory of mind, self-regulated learning and delayed gratification to obscure concepts like artificial general intelligence (AGI), spirituality and vipassana meditation [5], [6], [7], [8], [9].

Unsurprisingly, metacognition is a tremendously difficult phenomenon to research accurately and define properly. Initially, a general model of metacognition was theorized, but after several decades, various fields have continued to develop their understanding based on this model independently. This separation has plagued the field of metacognition, preventing the transfer of insights and the development of broader applications. The core of the problem lies in the different assessment methods used to measure and define metacognition.

Cognitive neuroscience aims to contribute by identifying the ground truth of metacognition, linking psychological mechanisms to their respective neural activities. However, like every field, neuroscience has its own limitations. Recently, the explainable artificial intelligence (XAI) movement has intersected with cognitive neuroscience, offering potential solutions to these limitations [10], [11]. A significant development in the subsection of perceptual decision-making was the creation of an explainable deep learning framework called "WaveFusion." This framework has the potential to significantly contribute to ongoing discussions unifying the divided fields of metacognition.

Within the subsection of perceptual decision making, metacognition is investigated as metacognitive confidence and metacognitive sensitivity. In this thesis, we apply the WaveFusion framework to these two forms of metacognition and explore its position within the broader scheme of theoretical research.

2. Literature review

The literature review will clarify the WaveFusion framework's role in metacognition research. We will first examine the theoretical understanding and current challenges of metacognition in psychology. Next, we will explore cognitive neuroscience contributions to the field. Finally, we will discuss the state-of-the-art machine learning techniques for decoding metacognition from EEG.

2.1. Metacognition in psychology

Metacognition is a complex and multifaceted standalone concept and a bridge between areas like mental health and cognitive development [12]. That is why it is researched across many disciplines in psychology as seen in Figure 1. Initially developed in the context of learning with a focus on memory it was soon expanded to all aspects of cognition and psychological phenomena like problem solving, emotional regulation and social behaviour [13], [14], [15].

Its multidimensional component allows meta- to be put before any psychosocial phenomenon (e.g. metamemory, metacommunication, meta-awareness, etc). This feature risks being a blanket term, meaning it is used to describe processes that it is not capable of accurately representing [16]. The field of metacognition has suffered from its ambiguous nature [17]. In what follows we will concretely define metacognition by firstly explaining the basic mechanisms. Secondly, by going into detail by exploring the fundamental model of metacognition and shortly discussing how it is integrated, applied, and researched in areas like cognitive neuroscience.

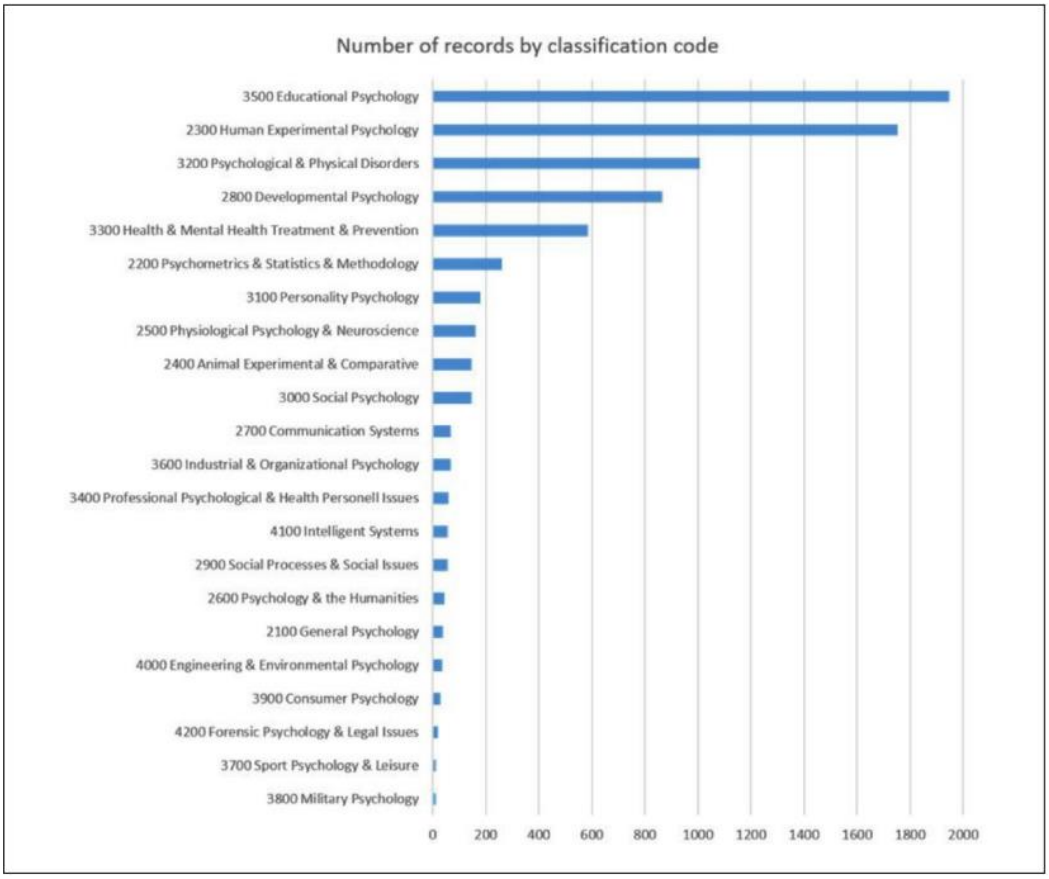

Figure 1: Number of metacognition records by classification code in the PsycINFO database [5, Fig 1].

2.1.1. The traditional mechanisms of metacognition

Metacognition is generally defined as “thinking about thinking”. The concept was introduced in 1976 by John H Flavell, considered by many as the founding father of the field [18]. Since then, the concept of metacognition has been refined through empirical research to expose the nuances of its mechanisms. This development happened largely separately for different fields [5].

For a thorough basic understanding we will discuss the following three aspects: 1) the meta-object level relationship to understand the difference between meta- and conventional cognition; 2) The relationship between consciousness and metacognition; 3) the domain interplay of metacognition as a general skill or specific to different situations.

2.1.1.1. Meta-object levels

At its core, metacognition is about how our thinking processes monitor and control themselves. It consists of two levels: the object level and the meta level. The object level is where our regular thinking happens, while the meta level is where we think

about our thinking. The meta level monitors and controls the object level (our cognition), ensuring our thinking processes are efficient and effective [19], [20]. We will dissect it in two simple steps, starting with explaining the mechanism of metacognition. Then moving on to our executive cognition and how it facilitates the meta level.

2.1.1.1.1. Mechanism of metacognition

The traditional mechanism of metacognition is seen in Figure 2. Metacognition is the monitoring and control performed from the meta level on the object level, with the object level being cognition. The mechanism of metacognition is represented by the flow of information between the meta and the object level. It is important to point out the asymmetry of the flow to understand the difference between the two levels. The meta level observes the object level, being an awareness that performs real time monitoring and receives information about the cognitive processes (the object level), while the object level receives controlling information from the meta level [19]. Here is an example to make the concept more concrete.

Imagine trying to find your way back to a hotel in a new city. You are navigating unfamiliar streets, making decisions about which roads to take. Along the way, you are 'monitoring' your cognitive actions and what is influencing your choices. Initially, you rely on the feeling of knowing the streets you have passed, but as time goes on, your confidence wanes. That is when you perform metacognitive control to switch to a different decision-making approach and pull out a map from your pocket. This map becomes your new decision-making guide, helping you find your way back to the hotel.

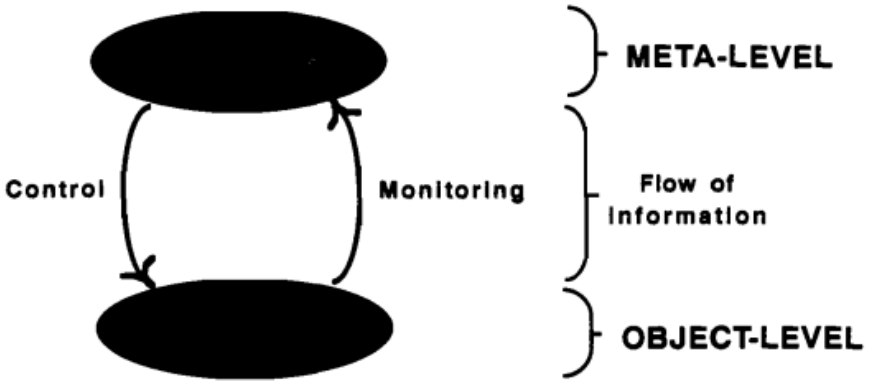

Figure 2: The theoretical mechanism of metacognition consisting of two structures (metalevel and object-level) and two relations in terms of the flow of information [19, Fig 1].

This is the most basic mechanism of metacognition, but of course it has been described more extensively to fully capture it multidimensional component. Later in the fundamental model we will fully elaborate on this. We will see that control and monitoring will fall under the term of ‘metacognitive skills’.

This is taken together with metacognitive knowledge to form the most used fundamental model, the 2-component model [21]. Metacognitive knowledge refers to the knowledge or beliefs an individual has about their own or general cognition. It is the information in Figure 2 that flows between the levels, facilitated by the mechanism of metacognitive skills.

Metacognitive knowledge and skills are acquired from multiple sources, like parents, peers and teachers [3]. It continues to develop through external acquisition, such as learning new strategies and receiving instruction, as well as through the internal feedback loop created by the experience of implementing metacognitive skills. This feedback loop involves reflecting on experiences, learning from them, and adjusting strategies accordingly, which generates more metacognitive knowledge and leads to continuous improvement and understanding [3]. Later on, we will define what metacognitive knowledge entails in detail.

2.1.1.1.2. Executive cognition

There is a fine line between executive functions and metacognition. Executive functions involve cognitive processes such as conflict resolution, error detection, inhibitory control, planning, resource allocation, and emotional control [22]. The bottom part of Figure 3 shows the executive functions in action under normal circumstances. Perceptual information serves as input, which is transformed into schemas (thoughts). The executive system then modifies these thoughts, leading to action.

For example, if you see a red traffic light while driving (perceptual information), this information is transformed into the thought that you need to stop (schema). The executive system prompts you to apply the brakes (action).

Executive functions become metacognition when they are targeted towards themselves, instead of on perceptual information and object level thoughts. For example, this occurs when you start reflecting on why you decided to apply the brakes at a red light, analysing your thought process and decision-making. The top part of

Figure 3, shows how metacognition works. Where the executive system simultaneously acts as the object level and the meta [22], [23].

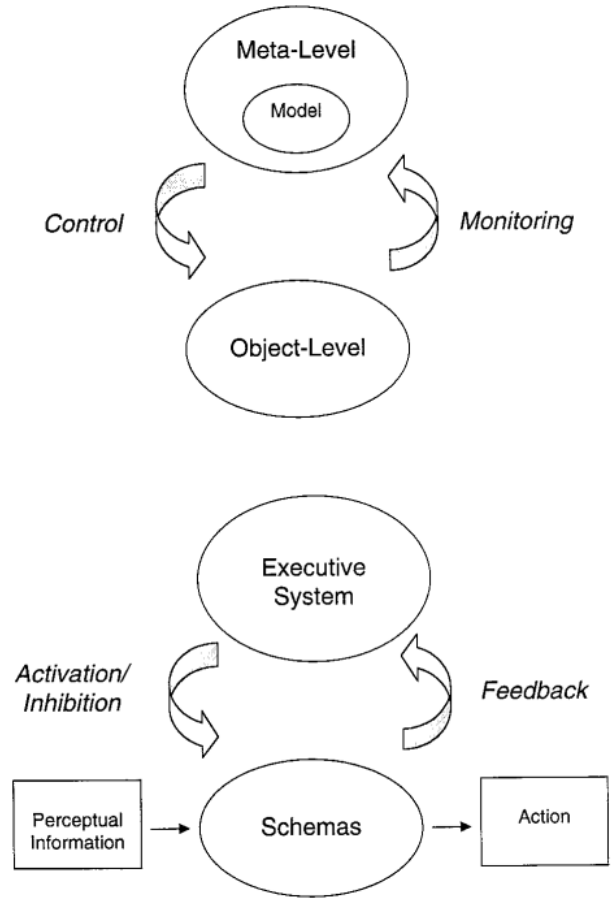

Figure 3:The top illustrates current views on metacognition and the bottom represents the views on the executive system, adapted from Nelson and Narens (1994) [22, Fig 1].

This is the fundamental relationship between executive cognition and metacognition, although it remains a topic of debate. The term 'model' in the meta level indicates that metacognition might include additional processes beyond monitoring and skills. There is no consensus on the exact relationship between executive function and metacognition, particularly regarding monitoring and information processing [20]. This uncertainty about their relationship is the first main ambiguity identified in this research.

The relationship between executive functions and metacognition has long been debated, with some initially viewing metacognition as an epiphenomenon. However, it is now generally regarded as an integral part of our cognition rather than a byproduct [24]. Furthermore, executive functions, often linked to intelligence, are highly developed in humans, distinguishing us from other animals. Interestingly, research

indicates that metacognition develops independently of intelligence and is a more significant predictor of learning and academic performance [20], [25].

2.1.1.2. Consciousness

Consciousness is a concept closely related to metacognition. There is ongoing discussion about whether metacognition is entirely conscious or if it also includes unconscious processes [3]. The definition of metacognition often influences this debate. In perceptual decision-making, metacognition is considered to include both conscious and unconscious cognitive processes [14].

It remains unknown how general metacognition is influenced by consciousness and what impact this has on resulting behaviour and subjective experience. This is particularly relevant because it plays an important role in mental disorders like schizophrenia [26]. Research into the physiological link between metacognition and consciousness is crucial for gaining a deeper understanding of their true nature [26], [27]. Thus, the relationship between metacognition and consciousness is a second significant point of ambiguity in this research.

2.1.1.3. Domain interplay

Metacognition seems to have a general and a domain specific competency that work hand in hand. These specific domains can be any cognitive tasks from estimating intercity distances, reading, spatial judgement, to solving mathematical problems [28]. It is still unknown what the metacognitive mechanism is between the interplay of the general metacognitive ability and the ability in specific domains. More granular empirical research needs to be conducted [3]. Thus, the domain interplay of metacognition is the third main point of ambiguity found in this research.

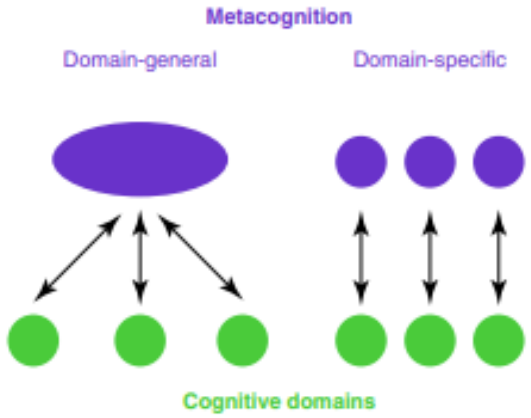

Figure 4: Domain generality or domain specificity of metacognition [29, Fig 1]

Considering the meta-object level interplay, some might wonder if an infinite loop of metacognition about metacognition could occur, essentially "thinking about metacognition." While not explicitly mentioned in the literature, this recursive loop is still regarded as just one level of meta-object level interplay.

2.1.2. The fundamental model of metacognition

Building upon the traditional mechanisms of metacognition, we can now explore the fundamental model. This model can be viewed in two ways: the 2-component model and the 3-component model.

We previously discussed the 2-component model of metacognition, which comprises metacognitive knowledge and skills. The 3-component model adds a third component: metacognitive experience. This distinction is significant because the 3-component model emphasizes the role of feelings and motivation in metacognition, which is particularly important in certain fields. In cognitive neuroscience, feelings and motivation play a crucial role in understanding metacognitive processes [14], [17].

This fundamental model can be integrated with other cognitive functions, such as memory, or applied to support specific practices, like self-regulated learning. The two most prominent extrapolated models are those used in metamemory research and self-regulated learning. [19], [30]. We will not go into detail of how these models work, as it is beyond the scope of this thesis.

2.1.2.1. The 2-component model

Introduced in 1987, the 2-component model consists of metacognitive skills and knowledge [21]. It serves as a simpler alternative to the initially introduced 3-component model. This simplicity has enabled the main components of the fundamental model to be empirically validated using the Metacognitive Awareness Inventory (MAI) questionnaire [21]. Now we will dive deeper into these two components, namely metacognitive knowledge and skills.

2.1.2.1.1. Metacognitive knowledge

Metacognitive knowledge refers to the knowledge or beliefs an individual has about their own or general cognition. It facilitates an awareness about three categories, namely the person, the task, and the strategy [23], [31]. Aside from the categories that define what it is, this knowledge is also split up into three components: declarative,

procedural and conditional knowledge components [32]. We will first discuss the categories and then move on to the subcomponents.

Three categories: person, task, strategy

Starting with the person category, it refers to all the beliefs held about one's own or others cognitive processes. The person knowledge can be divided into inter- and intra-individual differences and universal beliefs of cognition. For example, personal beliefs about the effectiveness of learning through reading versus listening, and interpersonal beliefs about varying levels of emotional sensitivity [31].

The task category refers to the nature of the cognitive activity with respect to the task. For example, knowing that the needs are often stated at the beginning of a paragraph. On the other hand, task knowledge in respect to one's own cognition can express itself as knowledge about the difficulty, abundance, familiarity, redundancy, organization, delivery method, pace, trustworthiness of the task [31].

The strategy category refers to the beliefs held with respect to the effectiveness of different cognitive strategies for achieving particular goals. For example, considering summarizing the main points in their own words as an effective learning strategy. This helps individuals regulate their cognitive activities to be more effective for specified goals [31].

The metacognitive aspect comes into play when this cognitive knowledge is used not just to perform a task but also to reflect upon and evaluate the efficacy of one's cognitive processes in achieving a cognitive goal. For example, cognitive knowledge is used to understand a text, while metacognitive knowledge will use this knowledge to evaluate how much you understand the text and if it is sufficient [23].

Three subcomponents: declarative, procedural, conditional

The three subcomponents of metacognitive knowledge: declarative, procedural and conditional can be very simply understood as the what, the how, and the why and when of metacognitive beliefs [32]. For example, knowing that reading is your preferred learning method (declarative knowledge), understanding how to effectively read and take notes to enhance learning (procedural knowledge), and recognizing why reading works best for you and when to use this method, such as when studying for an exam (conditional knowledge).

Like all knowledge, metacognitive knowledge can be inaccurate, influence cognitive activities in varying degrees or fail to be activated when necessary. It provides support in selection, evaluation, revision, and abandonment in cognitive control and influences the interpretation of your metacognitive experience [31].

2.1.2.1.2. Metacognitive skills

Metacognitive skills apply metacognitive knowledge to adapt cognition through three core skills: planning, monitoring, and evaluating. As shown in Figure 2, monitoring uses awareness to observe and obtain information about the object level and inform the meta level. Planning sets goals and outlines strategies, while evaluating assesses strategy effectiveness and makes adjustments. [23], [32], [33]. There are more regulative skills described but these are the three main skills.

2.1.2.2. The 3-component model

Introduced in 1979 by Flavell and refined in the years after, the 3-component model is made up by metacognitive knowledge, - skills and - experiences . Originally this model included a fourth component, metacognitive goals or tasks, but this has received less attention than the other three components [5], [17], [31], [34]. Metacognitive knowledge and skills generally preserve their definition across the two different models so there is no need to repeat it, we will now dive deeper into metacognitive experiences which takes feelings and judgements into account [17]. The interplay between the 3 components at the meta level and the interaction with cognition at the object level can be seen in Figure 5.

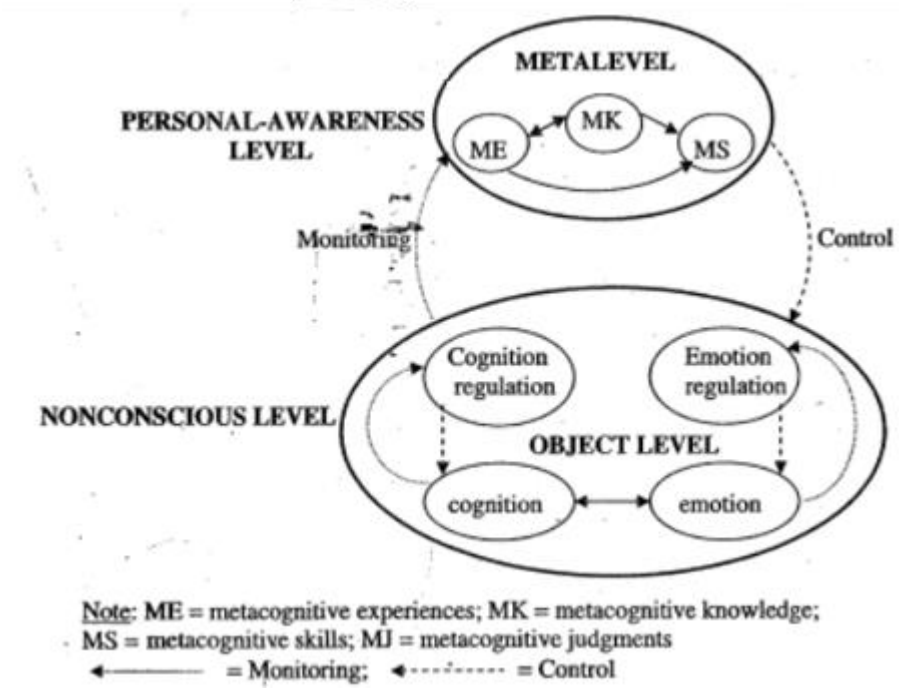

Figure 5: The meta- and object level interplay of the 3-component of metacognition [34, Fig 1].

2.1.2.2.1. Metacognitive experience

Metacognitive experiences, as seen in Figure 6, include feelings, judgements, and reactive experiences. These are similar to the monitoring and evaluative components of metacognitive skills but occur spontaneously rather than being part of a strategic plan. For example, while reading this thesis, you might “feel” confident or doubtful about your understanding. Based on this feeling, you might “judge” your comprehension level and then “react” by adjusting your learning methods if you realize you have not grasped the material well. These experiences are influenced by personal and task-related metacognitive knowledge, meaning that individual beliefs about one's abilities and the nature of the task shape our real-time metacognitive experiences.

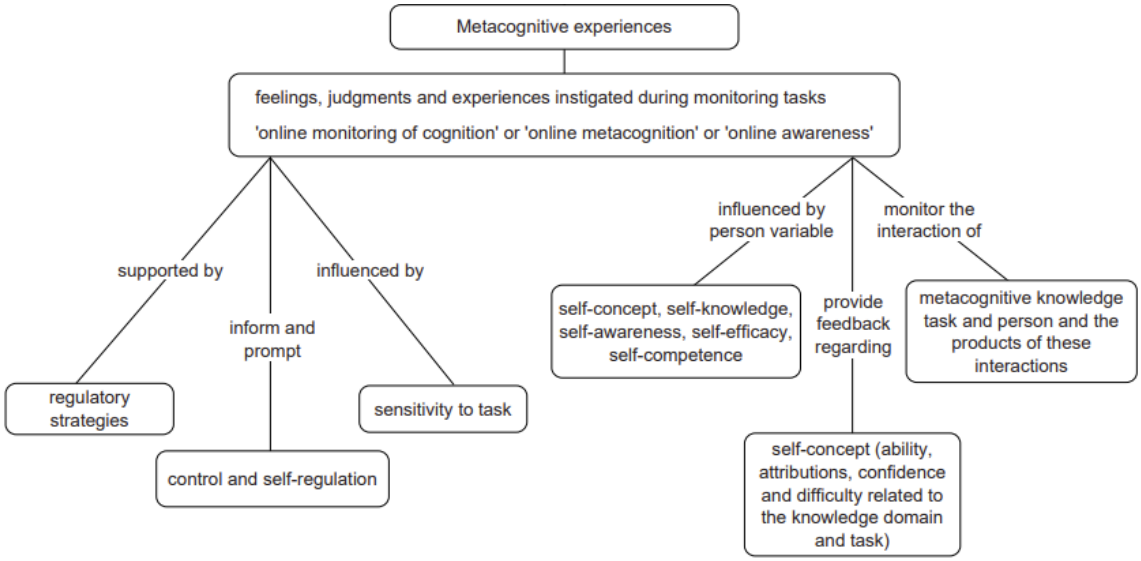

Figure 6: Amplification of metacognitive experiences in the conceptual framework of metacognition [17, Fig 9.6].

Importantly, these experiences contribute to the formation of metacognitive knowledge, creating a dynamic interplay where experience creates knowledge, which in turn shapes future experiences and guides metacognitive skills [17], [35].

In Figure 7 and Figure 8 the subcomponents of feelings and judgement are comprehensively and intuitively explained. Pay particular attention to confidence judgement, as it is especially relevant to this thesis within the field of cognitive neuroscience [17], [35]. Notably, feelings do not only include emotions but also other subjective sensations like confidence, difficulty, or satisfaction that arise during cognitive tasks.

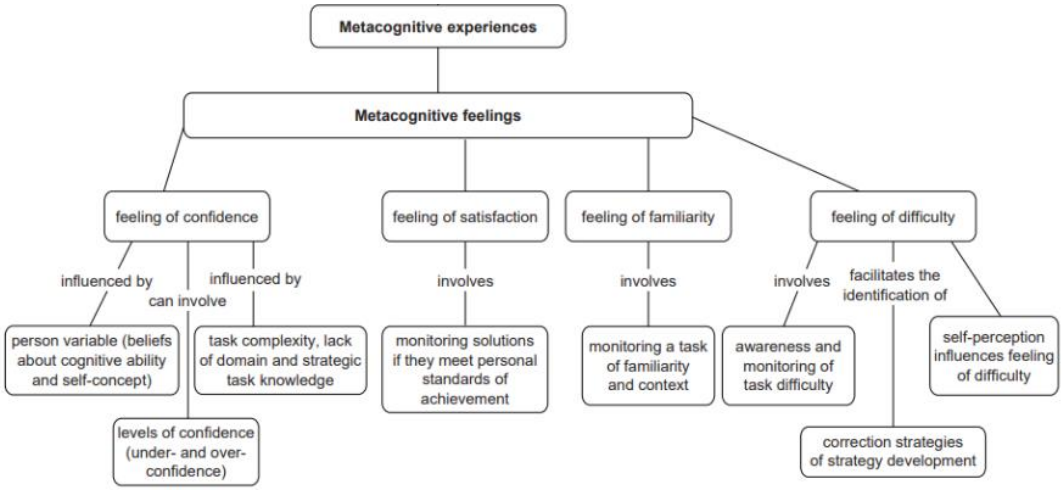

Figure 7: Amplification of metacognitive experiences, including metacognitive feelings, in the conceptual framework of metacognition [17, Fig 9.7].

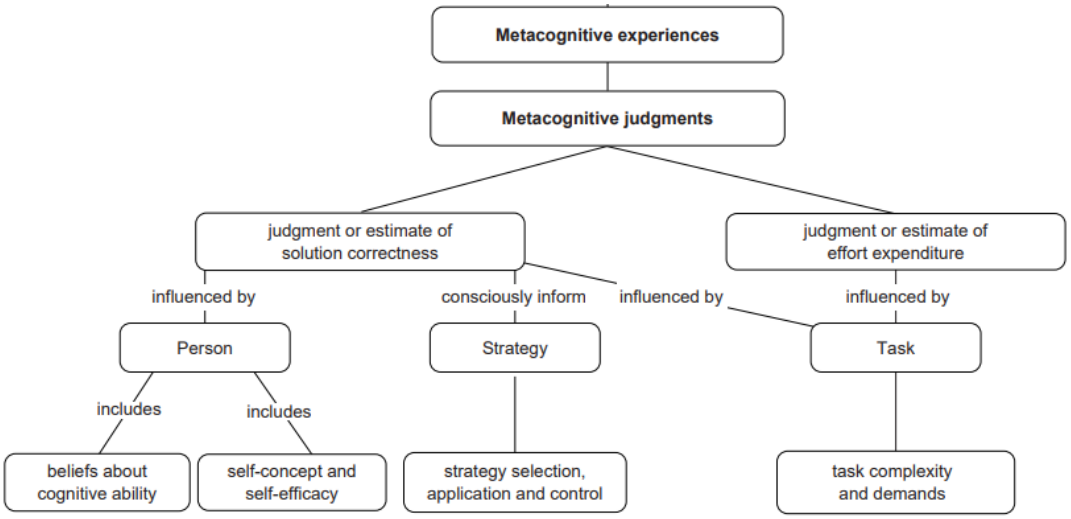

Figure 8: Amplification of metacognitive experiences, including metacognitive judgements, in the conceptual framework of metacognition [19, Fig 9.8].

2.1.2.3. Research

The multidimensionality of metacognition introduces a challenge in using consistent research methodologies within and across fields, thus makes integrating the findings challenging [36]. To get a general understanding of empirical metacognition research we will shortly discuss how we can firstly, theoretically activate metacognition, secondly which assessment methods have generally been used, and finally the differences in individuals.

2.1.2.3.1. Activating metacognition

Metacognition can be triggered by individual, social, and environmental factors [37]. This means that one's cognitive processes switches from 'thinking with' to 'thinking about' their own cognition. Theoretically there are 5 ways that contribute to making this switch [13]. Firstly, metacognition can be triggered through direct elicitation, like actively implementing a new cognitive strategy or making evaluative judgements about one's feelings and performance [37], [38]. Secondly, cognitive tasks that have a balance between high and low novelty or difficulty. This flow state balance tends to increase the likelihood of triggering metacognition, as it requires a manageable level of effort, without it being an automatic cognitive process [39]. Thirdly, metacognition often arises when individuals engage in important parts of a cognitive process that must deliver a correct result, such as problem-solving or decision-making [40]. Fourthly, the detection of errors in cognition, such as self-contradictions, prompts individuals to pause and reflect on their thought processes [37]. Lastly, the availability

of attention plays a significant role in triggering metacognition, with emotional experiences like stress and fear potentially diverting attention away from reflective thinking [39]. These various triggers underscore the dynamic and ambiguous nature of metacognition and contribute to the challenge of researching metacognition.

2.1.2.3.2. Assessment methods

Assessment methods of metacognition are made to quantify the performance of different components (metacognitive knowledge and skills) and subcomponents (E.g. Declarative knowledge, metacognitive planning). This performance is defined differently across fields like education and psychiatry [41], [42]. Assessment methods can generally be divided into off-line and on-line assessments. Off-line methods commonly take the form of questionnaires, are presented before or after the metacognitive task has been performed. On-line assessments are commonly executed by a third party that infers metacognitive performance of the candidate from thinking-out loud, eye-tracking, and behavioural observations. These assessment methods are limited by subjective biases and generalizability. The understanding of metacognition has evolved hand in hand with the evolution in assessment methods. However, the fragmentation of empirical assessment methods hinders the development of a comprehensive understanding of metacognition, leading to ambiguity [3].

2.1.2.3.3. Individual differences

Individual differences in metacognition stem primarily from three key factors. Firstly, each person possesses a uniquely developed metacognitive framework, resulting in a complex array of differences not only within individual components but also in how these components interact. For instance, variations in acquired self-beliefs embedded within one's metacognitive knowledge can lead to distinct interpretations of the same metacognitive experience. Secondly, individuals exhibit diverse emotions and motivations, which exert differential effects on their metacognitive processes. Lastly, metacognitive assessment tends to be domain-specific, complicating the interpretation by introducing confounding variables such as disorders and disabilities impacting memory, or variations in raw cognitive capacities affecting self-regulated learning. These three factors collectively contribute to individual differences in metacognition and introduce confounding elements into assessments, thereby further fostering ambiguity in understanding metacognitive processes [3], [43].

2.1.3. The ambiguities of metacognition

Empirically researching metacognition is challenging, as we can conclude from the ambiguities in activating, assessing and the impact of individual differences. This has led to a field trying to solve this problem by uncovering the ground truth definition of the mechanisms of metacognition and its activation. A field eager to contribute to the three major ambiguity points being 1) the relationship of metacognition and executive functions, 2) the relationship of metacognition and consciousness and 3) the domain interplay of metacognition. In the following chapter we will explore how cognitive neuroscience objectively identifies metacognition by linking it to the activation in our brain [44].

2.2. Metacognition in cognitive neuroscience

Central to the problem in defining metacognition is the reliance on subjective reports [16]. Cognitive neuroscience research, however, provides objective measurements, offering valuable insights that can be applied to fields such as educational science and mental health [45], [46]. Before discussing the insights obtained from fMRI and EEG research on metacognition, we will first explain the field of focus for this research project and its terminology.

2.2.1. Perceptual decision making

Perceptual decision making is a subcategory within the broader functioning of metacognition. As the name implies, it involves metacognition in the decision-making process about a perceptual task. Specifically, it refers to visual perception involved in a visual task where individuals make confidence judgements about the correctness of their decisions. First, we will specify this perceptual task. Second, we will discuss common definitions, such as how confidence judgements. This understanding is essential before exploring how the meta is encoded and decoded (read out) of the brain.

2.2.1.1. Perceptual task

The perceptual decision task used to elicit metacognition in this thesis is presented in Figure 9. This is a typical two alternatives forced choice task (2AFC), as the name implies it forces the participant to choose between two options. In this task, participants are first shown a visual stimulus, consisting of two squares, each containing a certain number of dots. The participant then decides which square, left or right, contains more

dots. Finally, the participant makes a judgement about how confident they are in the correctness of their decision, based on their visual perception of the stimulus [47].

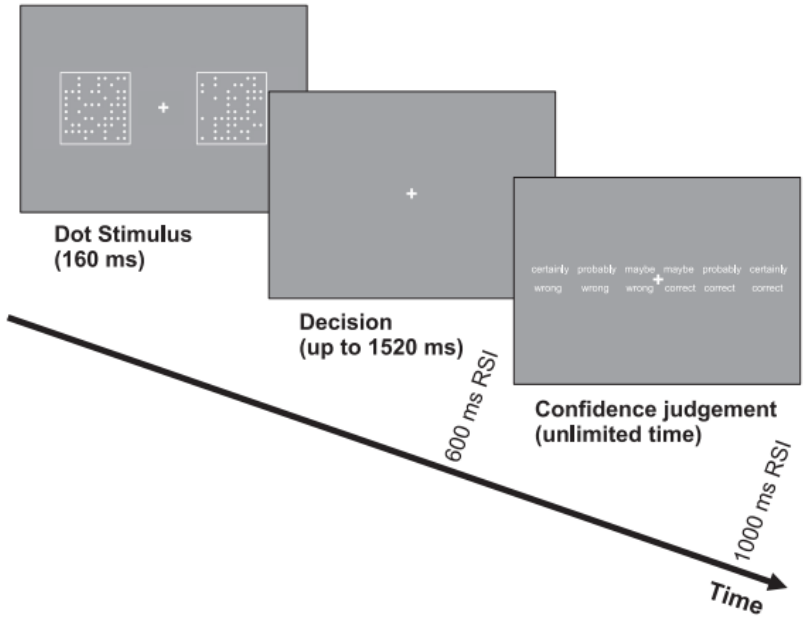

Figure 9: Summary of the task procedure. Participants first pressed a key according to the field containing more dots making a type 1 decision, then rated their confidence in their decision on a 6-point scale. RSI. [47, Fig 1]

The first decision made based on the visual stimulus is called a type 1 decision, which involves choosing either the left or right square. The confidence judgement that follows is a type 2 decision, where the participant indicates how certain they are that their choice was correct, with options such as maybe, probably, or certainly [48]. This task falls within the category of retrospective confidence judgements.

2.2.1.2. Definitions

The following definition of confidence judgement is the most important for this research project, nevertheless in Table 1 other common definitions in the field like metacognitive bias are provided. Later, metacognitive sensitivity will be discussed in more depth.

2.2.1.2.1. Confidence judgement

Previously in this literature review, we discussed the expansion from the 2-component model to the 3-component model. This expansion includes not only metacognitive skills and knowledge but also metacognitive experiences. Metacognitive experiences encompass feelings and judgements, such as confidence. Therefore, the metacognition explored within this thesis is specifically focused on the metacognitive experience of confidence related to visual perception.

Table 1: Glossary of definitions from a consensus meeting in the field of metacognitive perceptual decision making [48, Table 1].

Term	Definition
Metacognitive bias	An increase or decrease of confidence level despite basic task performance remaining constant
Metacognitive efficiency	The ability to distinguish between one's own correct and incorrect responses given a certain level of Type 1 performance
Metacognitive noise	A type of noise that affects confidence ratings but not primary decisions
Metacognitive sensitivity	The ability to distinguish between one's own correct and incorrect responses
Type 1 vs. Type 2 decisions	Type 1 decisions are about the primary task, whereas Type 2 decisions are about the quality of the Type 1 response.
Type 1 vs. Type 2 task performance	Type 1 task performance indicates how well one's choices predict stimulus identity, whereas Type 2 task performance indicates how well one's subjective ratings predict one's accuracy (i.e., metacognitive sensitivity).

In the following section, we will include broader insights on metacognition from various fields. However, the primary focus of this project will remain on the field of perceptual decision-making.

2.2.2. fMRI

Three central research questions for analysing metacognition in the field of cognitive neuroscience are: firstly, "Where in the brain do we represent the 'meta'?". Secondly, "Is 'meta' domain-specific?". Lastly, "How do we encode and read out the 'meta'?" [5]. Together, these questions aim to uncover the neural basis and functional aspects of metacognition, which are explored based on fMRI research.

Research results suggest that the neural system involved in metacognition is independent of that of decision making [49]. Several experiments display this separation through task manipulation, inhibition of neuromodulation, and neurostimulation impacting metacognition but not decision making [41], [42], [43]. The extent of correlation between the neural systems of decision making and metacognition is still a matter of research.

2.2.2.1. *Activation areas*

Figure 10 addresses the first question, “Where in the brain do we represent the ‘meta’?”. fMRI research identifies the brain areas associated the two-component model comprising metacognitive skills and knowledge with both online and offline metacognition. Overall metacognition in predominantly associated with the cerebral cortex, the largest subsection of the cerebrum which is a subsection of the forebrain. The cerebral cortex is known for the higher-level processes like language, memory, reasoning, learning, decision-making and emotion.

2.2.2.1.1. Metacognitive skills and knowledge

Metacognitive skills are primarily located in the prefrontal cortex, aligning with the regions involved in executive functions. In contrast, metacognitive knowledge is additionally associated with the posterior part of the brain, specifically the precuneus [45].

Metacognitive knowledge is typically activated in subjects by a reflective task like them rating their confidence in the perceived success of their performance in a task [45]. In the 3-component model, this is considered a metacognitive experience, where the subject evaluates their own belief about their performance.

On the other hand, metacognitive skills is closely related to the executive functions and are mainly activated with behavioural tasks such as Flanker tasks, Stroop tasks, Motion Discrimination tasks and Demand Selection tasks [45]. As discussed in the models of metacognition, this control represents the planning, monitoring and evaluating of employed strategies and resources.

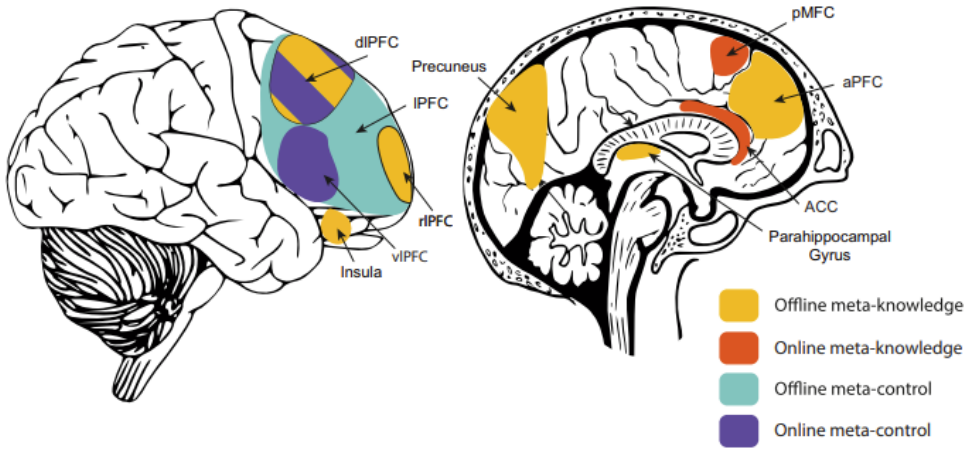

Figure 10: Brain regions associated with metacognition in the cognitive neuroscience literature. The regions are divided into online and offline metacognition. Striped are overlapping functions. [45, Fig 1]

The exact relationship between executive functions and metacognition is still up for debate. Interestingly early models in cognitive neuroscience research of metacognition were extrapolated from those of executive functions to study perception, decision-making, learning and sense of agency. Figure 11 shows how the meta was considered to be the brain regions in the prefrontal cortex (PFC) and the object level to be the posterior cortex [1].

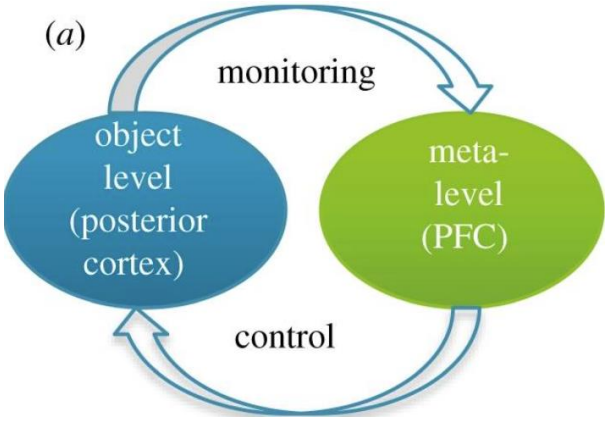

Figure 11: The brain areas mapped on the 2-component model[1, Fig 1].

2.2.2.1.2. Online and offline Metacognition

Online metacognition occurs during the execution of a task. Metacognitive skills is typically considered online because monitoring and skills generally happen quickly, without requiring reflective thinking [45]. Interestingly no example was provided nor easily found for online metacognitive knowledge.

In contrast, offline metacognition takes place during reflective breaks. Metacognitive knowledge is mostly viewed as an offline process. Judgements can be seen as offline processes of metacognitive knowledge or, more precisely, as metacognitive experiences, as they require the subject to reflect on their cognition and develop meta-representations. Offline metacognitive skills is evident in actions such as cognitive offloading, where thinking is reduced at regular intervals [45].

Brain-imaging studies suggest that the medial frontal cortex (MFC) and anterior cingulate cortex (ACC) engage in online meta-knowledge, with the ventrolateral prefrontal cortex (VLPFC) managing online metacognitive skills. In contrast, the anterior prefrontal cortex (aPFC) and precuneus, along with the lateral prefrontal cortex (lPFC), are activated when subjects engage in offline meta-knowledge and meta-control, respectively [45].

2.2.2.2. *Domain specificity*

Figure 12 addresses the question, “Is ‘meta’ domain-specific?”. fMRI research indicates that metacognition interacts differently with various cognitive functions in potentially different meta-level systems [53]. Within neuroscience, perceptual decision making and metamemory are the most extensively researched cognitive functions. The research into the domain specificity of these cognitive functions, specifically revolved around metacognitive confidence judgements. Figure 12 shows a neural overlap in metacognitive activation between these functions, with domain-specific differences highlighted in blue and red. This suggests that metacognition has both domain-general and domain-specific aspects. Overall, these areas overlap with the previous insights, but are more specified for metacognitive confidence.

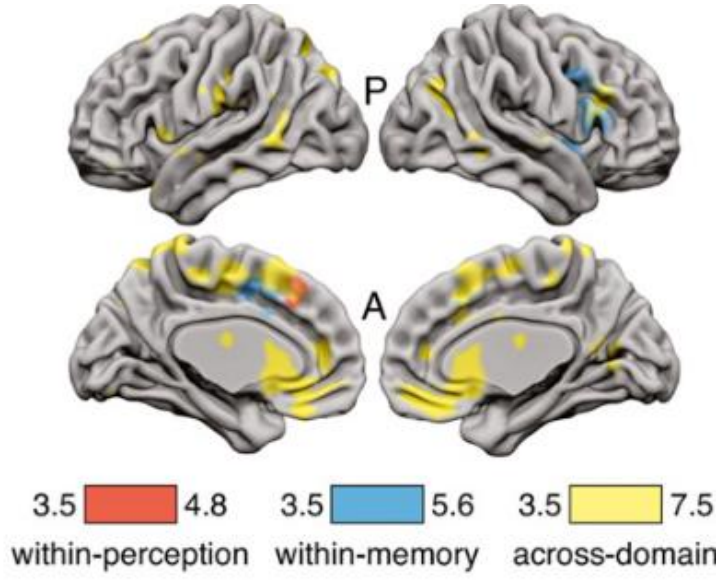

Figure 12: domain-specific patterns of confidence-related activity [29, Fig 5].

2.2.2.2.1. Domain general

Domain-general metacognitive activity has been identified in five main brain areas and are specified in Table 2. First, within the frontoparietal network, regions such as the posterior medial frontal cortex (pMFC) and the anterior prefrontal cortex (aPFC) have distinct roles. Second, the aPFC is connected to areas below the prefrontal cortex in the frontal lobe, specifically the interoceptive cortices like the dorsal anterior cingulate cortex (dACC) and the insula. Third, further below the frontal lobe in the basal ganglia, the striatum also shows activity. Fourth, medial areas in the frontal lobe, such as the ventromedial prefrontal cortex (vmPFC) and the pre-supplementary motor area (pre-SMA), are involved. Lastly, in the superior parietal lobule, the precuneus is another

region where domain-general metacognitive activity is observed [29]. Each of these areas plays a distinct role in supporting metacognitive processes, the functionalization of which we will explore later.

Table 2: Domain general brain regions of metacognitive judgements.

Region	Specific Area	Specific Regions
Frontal Lobe	Frontoparietal Network	Posterior Medial Frontal Cortex (pMFC), Anterior Prefrontal Cortex (aPFC)
Frontal Lobe	Interoceptive Cortices	Dorsal Anterior Cingulate Cortex (dACC), Insula
Basal Ganglia	Basal Ganglia	Striatum
Frontal Lobe	Medial Areas	Ventromedial Prefrontal Cortex (vmPFC), Pre-Supplementary Motor Area (pre-SMA)
Parietal lobe	Superior Parietal Lobule	Precuneus

2.2.2.2.2. Domain specific

Firstly, domain-specific patterns for metacognitive confidence for metamemory and decision making were identified in nuanced areas of the right lateral anterior prefrontal cortex (aPFC) [29]. Furthermore, for metamemory, a distinction was found between retrospective and prospective confidence judgements. Lastly, in general mentalising, making a metacognitive judgement about someone else's performance also shows both domain-general and specific activation. These aspects highlight the domain-specificity the mechanism of metacognition portrays [54].

Answering the first two key research question provides an understanding of how metacognition is represented in the brain. The answer to the third question will be limited to strictly perceptual decision making for the relevance of this thesis.

2.2.2.3. Functionalization

The following sections provide a high-level explanation of the third question: "How do we encode and read out the 'meta'?". To address this, we need to understand the functionalization of brain areas involved in metacognition from the perspective of information processing.

Firstly, we will explore the types of information used and how they are integrated. Secondly, we explain how this information is processed for metacognitive confidence

judgements. This explanation is divided into two sections: (1) Brain area specific functions and (2) Common differentiating areas.

2.2.2.3.1. Type of information

To understand how metacognitive confidence judgements are encoded, it is essential to recognize the sources of information the brain uses. It is suggested that metacognitive judgements involve integrating three types of information, as seen in Figure 13. First, external sensory perceptual information. Second, interoceptive information, which consists of internal sensory signals from within the body, such as heart rate and breathing. Third, action information, which provides feedback on actions related to decision-making, including response strength and the fluency of response execution. This information is integrated in the central frontal cortex, while the anterior prefrontal cortex contributes by making predictions about the state of the world and the individual, leading to accurate metacognitive judgements.

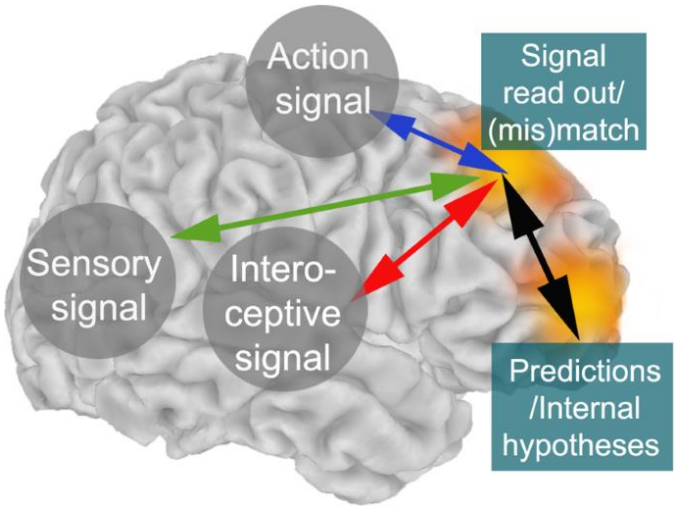

Figure 13: Sensory, interoceptive and action signals are read out in central frontal cortex. Anterior prefrontal cortex provides predictions about the "state of the world" and the "state of the decider" when a decision is made. Central frontal theta oscillations [55, Fig 9].

2.2.2.3.2. How the information is processed

Decision making confidence

Figure 14, based on fMRI research, presents a detailed understanding of how the various areas function to process the information of metacognitive confidence in perceptual decision making.

Firstly differentiating between local confidence, which is only related to the specific perceptual task performed, and global confidence, which encompasses broader self-

beliefs about one's abilities and skills across various tasks and situations [56]. Secondly, we see that the parietal brain areas are related to sensory processing. While the frontal areas being involved in metacognitive skills and generating confidence judgements.

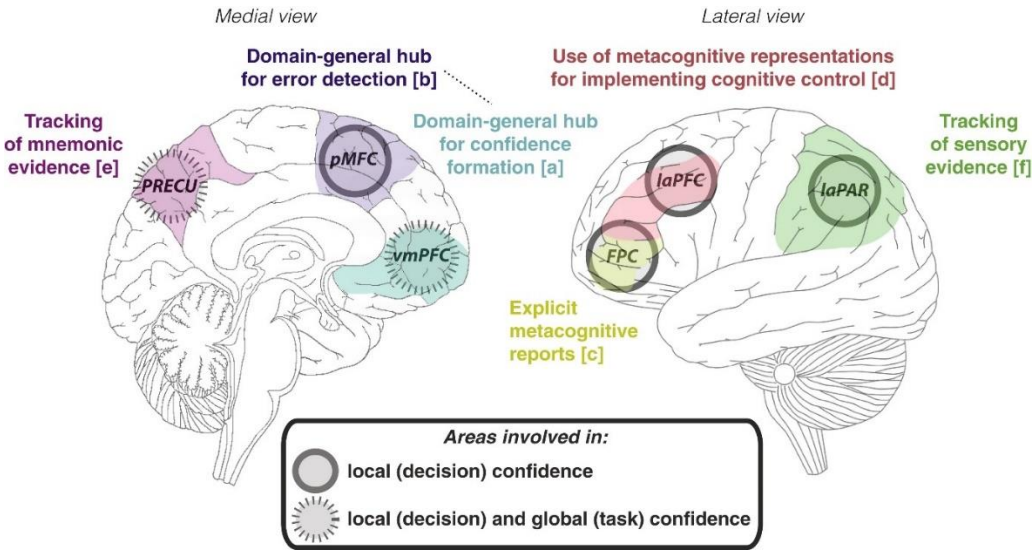

Figure 14: Neural correlates of metacognitive evaluation on a perceptual task. [56, Fig 2]

Differentiating areas: High and low confidence areas

In Figure 15, we observe that brain activity associated with high confidence is primarily found in the central and posterior regions of the brain. Conversely, activity related to low confidence is predominantly located in the frontal regions.

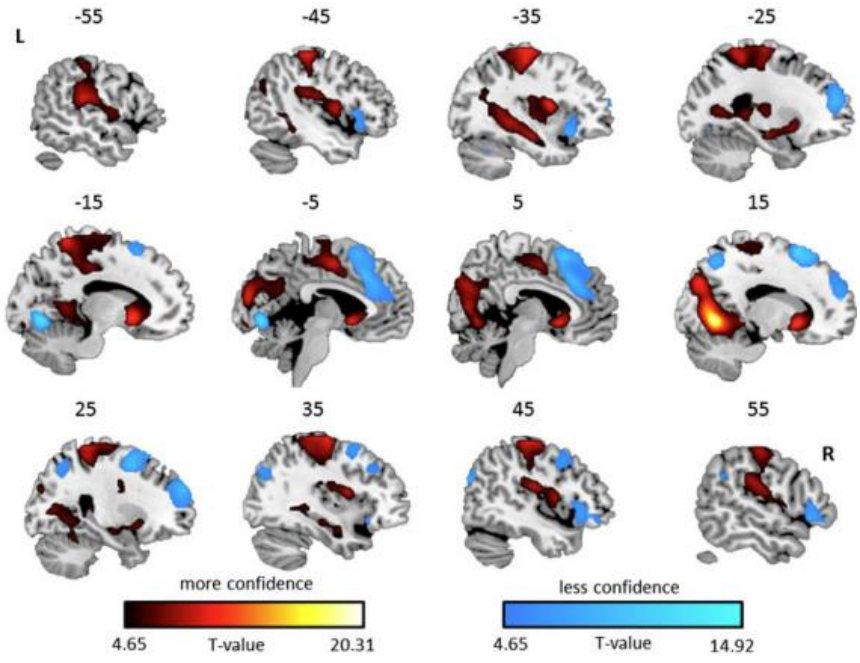

Figure 15: Significant brain regions associated with more (red) and less (blue) confidence, shown on sagittal slices with numbers above each slice representing coordinates [57, Fig 3].

2.2.3. Metacognitive sensitivity

Contrary to the intuitive belief that we have complete and accurate access to our cognitive processes, research shows that our self-assessment is not always fully aligned with reality [1]. This means that our metacognition can be inaccurate. Furthermore, the accuracy of metacognitive ability varies significantly across individuals, independent of task performance and confidence levels [58]. Understanding the neural mechanisms underlying metacognitive sensitivity is essential for comprehending failures in metacognition in conditions such as brain damage and psychiatric disorders [1]. We will first discuss the definition of metacognitive sensitivity in perceptual decision-making. Second, we will move on to the insights fMRI research obtained in metacognitive sensitivity, which is limited compared to metacognitive confidence research.

Overall domain-general metacognitive sensitivity has been linked to real-world applications such as mental health and behaviour. Individuals with psychosis-related symptoms of mental disorders show a significant reduction in metacognitive sensitivity In general, poor self-judgement and overconfidence are linked to high rates of entrepreneurial failure, global stock market crashes, the explosion of the Space Shuttle Challenger, and the nuclear accident at Chernobyl [59]. Due to its importance, metacognitive sensitivity has garnered significant interest, leading to research into training methods for improving it [60].

2.2.3.1. Definition

In perceptual decision making, the concept of metacognitive sensitivity has been developed to represent an individual's ability to distinguish between their own correct and incorrect responses [48].

Another way to understand metacognitive sensitivity is from the perspective of type 1 and 2 decision performance. Type 1 task performance represents the ability to correctly answer on the visual task. While type 2 task performance represents metacognitive sensitivity, the ability to correctly predict one’s own correct and incorrect answers in the type 1 decision [48].

2.2.3.2. *fMRI*

Although the research into metacognitive sensitivity is limited, some similar insights are found. Again, we can answer the three central questions. Firstly, “Where in the brain do we represent the ‘meta’?”. Secondly, “Is ‘meta’ domain-specific?”. Lastly, “How do we encode and read out the ‘meta’?”. From research into visual motor metacognitive sensitivity it is suggested that there are distinct brain regions for the object level performance and the metacognitive performance [61].

2.2.3.2.1. Activation Areas

To address the first question, “Where in the brain do we represent the ‘meta’?”, we will focus on activation areas of perceptual decision-making tasks with retrospective confidence judgements. An overview can be seen in Table 3.

Most prominently, increased activity in the rostral and dorsal aspects of the lateral prefrontal cortex (rlPFC and dlPFC) are found to be crucial for metacognitive sensitivity [1], [62]. Furthermore the salience network known for consisting of brain regions that evaluate the importance of internal or external stimuli, has shown to mediate metacognitive sensitivity [63]. Particularly the dorsal medial prefrontal cortex (dmPFC) and anterior insula, two critical components within the salience network that encodes self-awareness and monitors perceptual decision errors [63].

In addition to fMRI, volumetric techniques like quantitative MRI (qMRI) provide additional insights into metacognitive sensitivity. These techniques measure local grey matter myelination and iron content, revealing correlations with metacognitive sensitivity in the anterior prefrontal cortex (aPFC), precuneus, hippocampus, and visual cortices [64].

Table 3: Overview of brain areas related to metacognitive sensitivity and the respective measurement technique.

Technique	**Main Area**	**Specific Areas**
fMRI	Lateral prefrontal cortex (lPFC)	Rostral lateral prefrontal cortex (rlPFC), Dorsal lateral prefrontal cortex (dlPFC)
fMRI	Salience Network	Dorsomedial prefrontal cortex (dmPFC), Anterior insula
qMRI	Various area’s	Anterior prefrontal cortex (aPFC), Precuneus, Hippocampus, Visual cortices

2.2.3.2.2. Domain specificity

Secondly, “Is ‘meta’ domain-specific?”. Metacognitive sensitivity is also found to be domain-general and -specific. The activation areas in the brain vary for social and cognitive reasoning tasks, visuomotor tasks, and retrospective and prospective judgements [1], [61], [65]. For perceptual tasks, a domain-general aspect remains present as metacognitive sensitivity is correlated across different perceptual tasks, suggesting a task-independent mechanism underlying metacognition [66]. In general, the domain-specific mechanism of metacognitive sensitivity within perceptual decision making and their brain areas remain unclear and under researched.

2.2.3.2.3. Increased metacognitive sensitivity

Lastly, “How do we encode and read out the ‘meta’?”. This question remains largely unexplored. Further research is needed to identify the functional roles of the previously discussed brain regions in relation to metacognitive sensitivity [61]. Although there is insight into which areas are related to improved metacognitive sensitivity, this does not mean they are distinct activation regions of metacognitive sensitivity.

Firstly, the dorsolateral and anterior prefrontal cortical subregions (dlPFC and aPFC) work in conjunction with interoceptive cortices, such as the cingulate and insula, to enhance metacognitive sensitivity [1]. Furthermore higher metacognitive accuracy is associated with decreased activation in the anterior medial prefrontal cortex (amPFC) [57], [62]. Additionally, volumetric findings show that increased myelination in the right anterior prefrontal cortex (aPFC) myeloarchitecture and decreased myelination in the left hippocampus correlate with better metacognitive sensitivity [64].

2.2.4. EEG:

While fMRI provides localization of neural substrates with great spatial precision, EEG offers insight into the temporal dynamics of neural activity, albeit with less precise spatial localization. Due to these differences in neuroimaging techniques, EEG findings differ in spatial localization compared to fMRI. As EEG data is used in this thesis, we will firstly discuss the challenges that guide the data analysis approach. Secondly, we will discuss the temporal and spectral insights gained from EEG analysis for both metacognitive confidence and sensitivity.

2.2.4.1. The challenges

Starting with the three general challenges in comparing and analysing insights from EEG activation profiles of metacognition: First, metacognition has a varying temporal activation profile, as shown in Figure 16. Second, the activation profile of metacognition is task-specific, even within visual perceptual decision-making tasks, as seen in Figure 17. Third, the activation profile varies for confidence ratings about correct versus incorrect type 1 decisions, as depicted in Figure 18.

These insights originate from metacognitive confidence research. There are no temporal topographies of metacognitive sensitivity derived from the metacognitive confidence. These insights are likely transferable but not certain for metacognitive sensitivity, so they will be taken into consideration.

2.2.4.2. Metacognitive confidence

2.2.4.2.1. Temporal activity

The temporal activity of metacognitive confidence has been extensively researched. This includes identifying temporal activity both across the entire topography, encompassing all EEG electrodes, and the event-related potentials (ERP) at specific electrodes. The ERP represent the electrical activity over time at an electrode, those which are most strongly correlated with metacognition. We will begin by discussing the event-related potential (ERP) and then move on to the topographies.

ERP

In Figure 16, the event-related potential (ERP) is displayed. This shows the electrical activity, locked to the type 1 decision, in the temporal plane at a specific EEG electrode. Typically, the Pz, Cz, or Fcz electrodes are chosen to extract the ERP, as the characteristic activations for metacognition are most pronounced in the fronto-central and parietal regions of the scalp [47]. These characteristic activations include the Error-Related Negativity (ERN), Correct-Response Negativity (CRN), and Error Positivity (Pe).

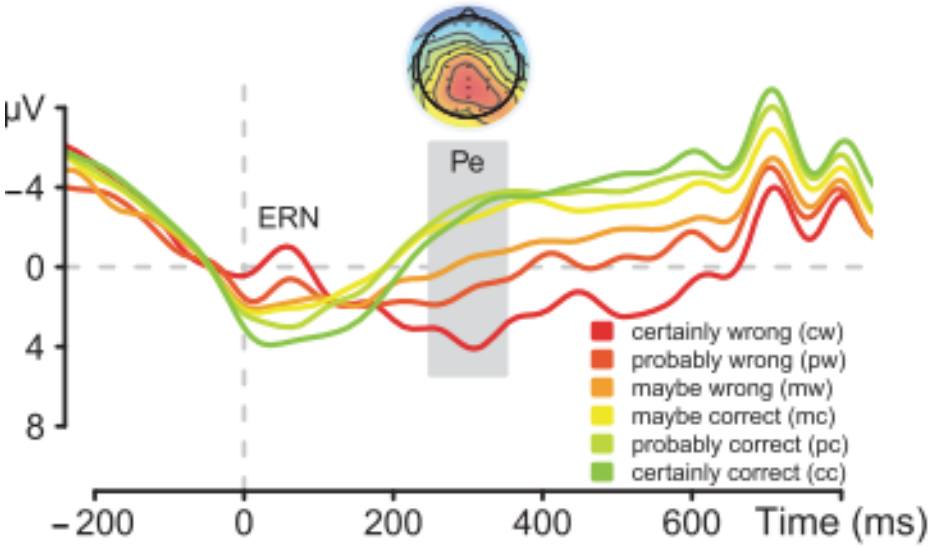

Figure 16: Response-locked event-related potential (ERP) and topography for the difference between "certainly wrong" and "certainly correct" for metacognitive confidence [47, Fig 3].

The ERN is a negative deflection in the ERP that occurs 50-100 ms after an error is made. The CRN is similar but occurs after correct responses and has a smaller amplitude [47]. The Pe is a positive deflection in the ERP that occurs 200-400 ms after an error, but the Pe is a negative deflection after a correct type 1 response [47]. Overall, the ERN and CRN are associated with general performance monitoring, a subconscious process, while the Pe is linked to error detection with conscious awareness [67]. As seen in Figure 16 the amplitudes of the ERP demonstrates a clear graded association with metacognitive confidence at the Pz electrode [47].

The Pe's time interval of 200-400 ms can be further specified to the P3 component, which occurs between 300-400 ms after the stimulus is presented. The waveform associated with the P3 component represents the metacognitive experience, specifically the confidence judgement [68].

Topography

In Figure 17, topographies are shown throughout time in a condition where a type 1 decision error was made.

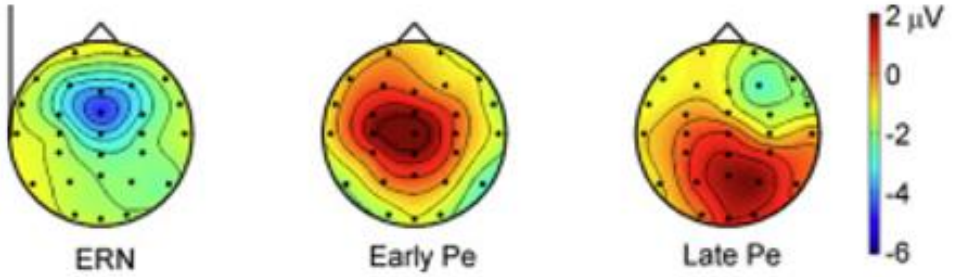

Fig. 2. Scalp topographies of ERN, early positivity and classic Pe in certain-error condition, of Eriksen flanker task.

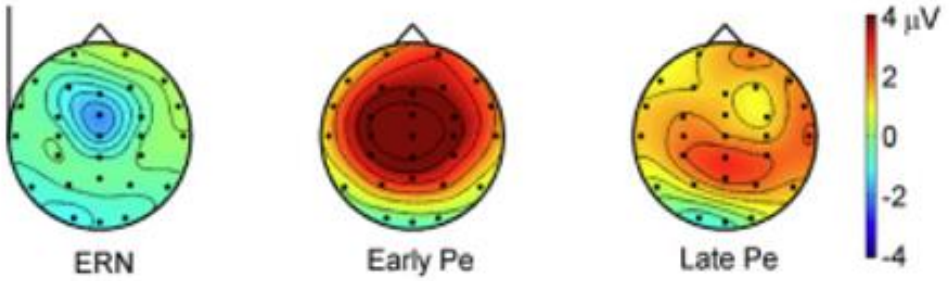

Fig. 4. Scalp topographies of ERN, early positivity and classic Pe in certain-error condition of circle discrimination task.

Figure 17: EEG topography throughout different time phases of the confidence judgement, with Fig. 2. portraying activation for the Eriksen flanker task and Fig. 4. for the circle discrimination task [69, Fig 2, 4].

The fronto-central and parietal electrodes represent the most variable activity. Starting with a general negative electrical signal at the Error-Related Negativity (ERN), which displays a slightly different topography when compared to the Correct-Response Negativity (CRN), as seen in Figure 18. Continuing to an early and late Error Positivity (Pe), showing a positive electrical signal throughout the scalp in the incorrect condition. In the correct condition this is likely a negative deflection as seen in Figure 16.

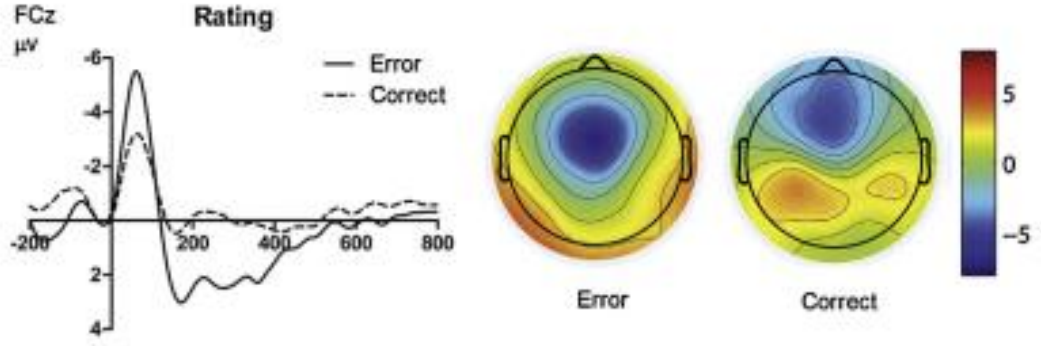

Figure 18: Topographies for the event-related potential (ERP) in the rating condition at 60 ms post-response, differentiated for erroneous and correct type 1 decisions [67, Fig 2B].

2.2.4.2.2. Spectral activity

The following insights have not been correlated strictly with metacognitive confidence but have been related to metacognitive sensitivity. Considering the mechanism of the event-related potential (ERP) we just discussed, the following insights are noteworthy for metacognitive confidence, especially considering the limited research on this aspect.

In Figure 19, a spectrogram at a frontal electrode (AFz) and occipital electrode (Oz) are shown with their respective topographies. These activities are stimulus-locked, meaning right when the visual task is presented. Note that this is different than the temporal event-related potential (ERP) insights above. The occipital electrode (Oz) and its topographies represent the neural activity right after the stimulus of the task was represented, the frontal electrode (AFz) time window starts one second after stimulus presentation. Despite the confusing time window intervals, we see that overall, around 2,5 seconds after stimulus presentation there is increased theta-band (4 – 6 Hz) activity in these electrodes. While a decrease in lower beta-band (13–20 Hz) and theta-band activity in the left and right motor channels occurred.

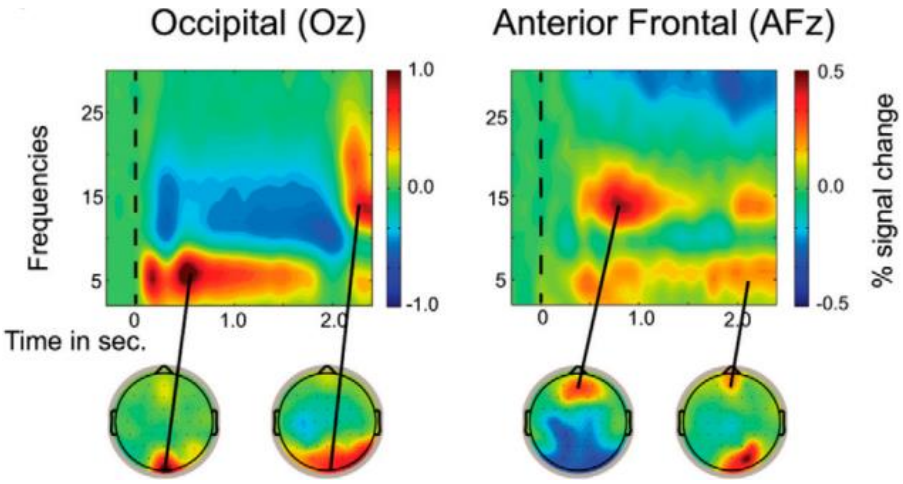

Figure 19: Time-frequency analysis of stimulus-locked neural activity at Oz and AFz electrodes [70, Fig 3C].

Furthermore, in the first second after presenting the stimulus decreased beta-band activity was found in the frontal, the left and right parietal, occipital, and motor channels was found.

Besides this indirect insight there was research that directly found pre-stimulus alpha-band (8–13 Hz) power in the posterior part of the head to be directly negatively correlated with confidence, which did not affect accuracy [71].

2.2.4.3. Metacognitive sensitivity

2.2.4.3.1. Temporal activity

The research on the neural substrates of metacognitive sensitivity using EEG is quite limited for metacognition in general and specifically retrospective judgements in perceptual decision making. This presents an opportunity for further research to explore this area in more depth.

2.2.4.3.2. Spectral activity

In Figure 20 a correlation between prefrontal theta-band activity 1,5-2,5 seconds after stimulus and metacognitive sensitivity (adequacy) is shown, but at no particular activity. This in fact represents a positive linear relationship, which is unrelated to task accuracy [70]. Further research proved that metacognitive sensitivity improved due to theta burst stimulation, again without affecting first-order decision making. Strongly suggesting the importance of theta band activity for metacognitive sensitivity [8].

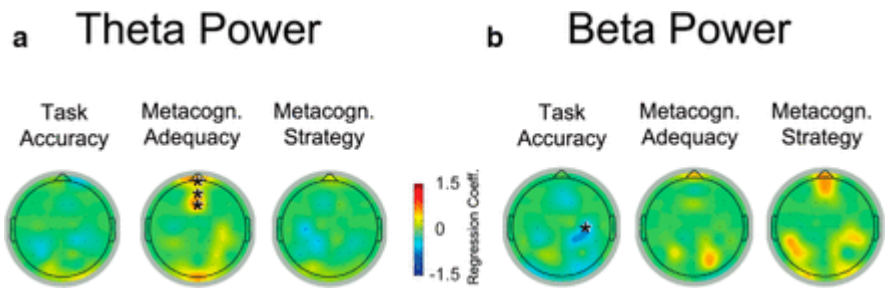

Figure 20: Multiple linear regression EEG results: late time window (1.5–2.5 s). Showcasing in particular the relationship between theta power and metacognitive sensitivity (adequacy) [70, Fig 5].

2.3. Machine learning for decoding metacognition

Machine learning has been applied in various ways to study metacognition. It is used to model and gain a deeper understanding of the cognitive processes involved, estimate derivations from confidence judgements such as metacognitive efficiency, or even predict confidence judgements directly from EEG data [64], [66], [67]. We will firstly discuss the limitations of traditional approaches, then shortly explore machine learning applications, lastly dive deeper into deep learning.

2.3.1. Limitations of traditional approach

Traditional approaches to researching metacognition in cognitive neuroscience often rely on statistical correlations. These methods include identifying linear relationships or co-occurrences between brain activity and confidence levels, but also more complex computational models. While these approaches have significantly advanced our understanding of metacognition, they also have notable limitations.

The three most notable limitations identified are as follows: First, the ecological validity of the insights reviewed is unclear, making it difficult to determine how well they reflect real-life metacognition. Second, there is a need for more cross-disciplinary research to better understand the neural substrates of metacognition. Third, the distinction between the neural correlates of metacognitive skills and metacognitive knowledge remains insufficiently explored [45].

2.3.2. Machine learning applications

Machine learning is inherently different from traditional approaches. Traditional computational models are rule-based and designed to simulate patterns in systems using predefined equations and rules. In contrast, machine learning models learn these patterns directly from data.

Machine learning has been used in various ways for metacognition. Most commonly it was used to predict metacognition. On the one hand, a hierarchical Bayesian approach could predict metacognitive efficiency from confidence judgements [72]. On the other hand a multivariate classifier could predict high or low confidence directly from EEG data, with an above chance accuracy [76].

2.3.3. Deep learning

Deep learning further differentiates itself from machine learning as a subsection by being able to learn features directly from raw data without manual intervention. In contrast, other machine learning techniques require feature engineering before they can automatically learn patterns in the data. We will first explore the pros and cons of deep learning to understand its usefulness before examining specific applications.

2.3.3.1. Pros and cons

2.3.3.1.1. pros

Deep learning offers three notable advantages for metacognition research (1) ecological validation, (2) multimodality and (3) transfer learning. Starting with ecological validation, by automating the prediction of metacognition, allowing for its application in more naturalistic settings. For instance, a deep learning model trained on confidence judgements in perceptual decision-making can be utilized and compared in ecological experimental setups.

Secondly, deep learning models can integrate various types of data, such as EEG, text, and audio, facilitating the fusion of cross-disciplinary insights [77]. This multimodal integration allows for the combination of data from cognitive neuroscience and mental health surveys, for example, to predict outcomes relevant to educational sciences.

Lastly, transfer learning in deep learning allows the use of model weights from one application in a different context, reducing the need for extensive data and enabling the evaluation of relationships between different domains. For instance, weights from a model trained on metacognitive knowledge can be used to predict metacognitive skills, providing valuable insights into their interconnectedness.

2.3.3.1.2. Cons

Deep learning models face three notable disadvantages, (1) data requirement, (2) time consuming and (3) limited interpretability. Firstly, they require a large amount of high-quality data to perform effectively. This is particularly challenging in the context of confidence judgements, where imbalanced datasets necessitate data augmentation, potentially leading to generalization issues. Secondly, training deep learning models is time-consuming due to their complex architectures and the extensive computational resources required. Lastly, deep learning models suffer from limited interpretability. Although certain frameworks like Convolutional Neural Networks (CNNs) provide some degree of interpretation, they remain more abstract compared to traditional computational models.

2.3.3.2. Deep learning application

Although limited, a couple of deep learning applications have been explored within the whole of metacognition. Most notably, a transfer learning algorithm called ‘meta-

learning' uses EEG and EOG input data to make trial-by-trial confidence predictions while quickly adapting to the current subject with limited data by leveraging data from previous subjects with a 70%-80% NMSE [74]. On the other hand, a Long Short-Term Memory (LSTM) neural network model directly estimates a subject's learning confidence in an immersive VR environment using multiple data inputs such as eye gaze and controller position, achieving 85.6% accuracy [77].

Despite these interesting experiments this project focused solely on an explainable deep learning framework called WaveFusion. WaveFusion was developed for classifying and localizing neural activity in neuroscience research applications [78]. This framework has been applied to metacognition in perceptual decision-making tasks with retrospective confidence judgements, achieving a 95.7% classification accuracy of confidence judgements using EEG data as input [75]. In this research project, this framework will be expanded to additionally classify metacognitive sensitivity from EEG data. We will explore this framework in three sections, starting with the input, continuing with the processing model, and ending with the output.

2.3.3.2.1. WaveFusion framework

Input

The input for the WaveFusion model are spectrograms an example for the Pz electrode is shown in Figure 21Figure 22. These spectrograms were created for each electrode by applying the STFT to stimulus-locked EEG data of a visual task in perceptual decision making. A selection of these electrodes was made to limit the framework, only the posterior electrodes were chosen due to the importance of posterior pre-stimulus alpha band activity.

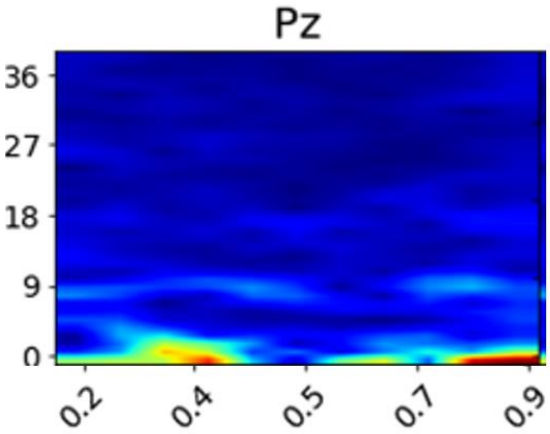

Figure 21: Spectrogram at Pz electrode used as input for the WaveFusion model [75, Fig 4].

Model structure

The input is passed through three sections in the model architecture before the classification is performed. We will discuss these sections seen in Figure 22 in more detail.

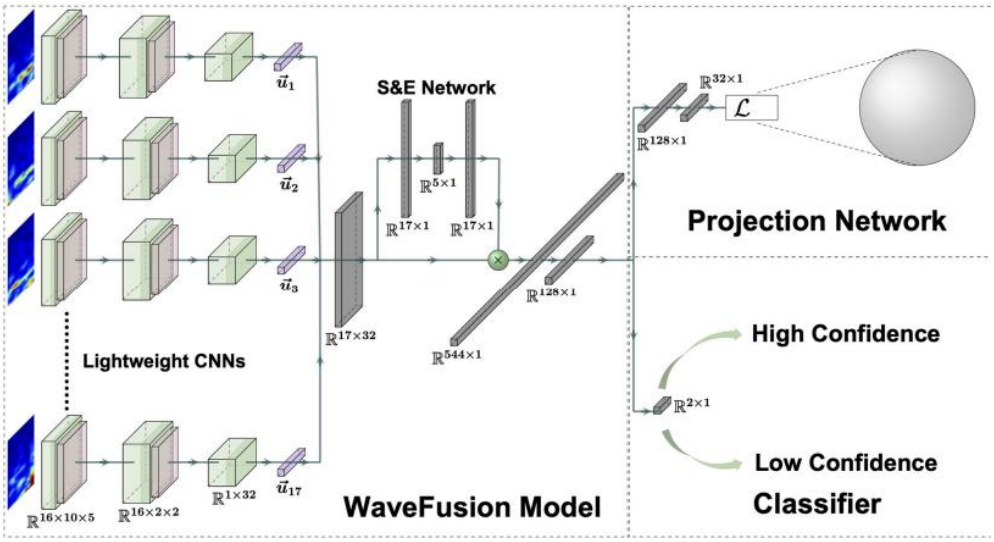

Figure 22: The WaveFusion architecture [75, Fig 1].

Firstly, there are three layers of lightweight convolutional neural networks (LWCNN) for each EEG lead, with a total of 17 electrodes in this case. These convolutional layers generate 2D feature maps from the spectrogram. Each layer processes these feature maps sequentially, learning more abstract features. To prevent overfitting, the first two layers include ReLU activation, 2x2 max-pooling, dropout with a 10% drop rate, and batch normalization. The last convolution layer also uses dropout and batch normalization, outputting a compressed feature map with more filters, each representing different variations of the learned features. The hyperparameters are detailed in Table 4

Table 4: The hyperparameters detailed for the LWCNN and SEN structure [75, Table 1].

Operation	Kernel	Strides	Padding	Count	BN?	Dropout	Nonlinearity
LWCNN: 1 × 39 × 11 input							
2D Convolution	5 × 4	2 × 1	2 × 1	16	×	0.1	ReLU
2D max-pooling	2 × 2			16	×		
2D convolution	4 × 2	2 × 1		16	✓	0.1	ReLU
2D max-pooling	2 × 2			16	×		
2D convolution	2 × 2	1 × 1		32	✓	0.1	
SEN: 17 × 32 input							
Linear	N/A	N/A	N/A	17	×		ReLU
Linear	N/A	N/A	N/A	5	×		ReLU
Linear	N/A	N/A	N/A	17	×		Sigmoid

Secondly, the compressed feature maps are forwarded to the Squeeze and Excite Network (SEN). This network, operating in an encoder-decoder mode, creates attention weights that indicate the importance of each lead. The dense encoder layer condenses the 17×1 input to 5×1, followed by ReLU activation. Then, the dense decoder layer expands the output back to 17×1. To address overfitting, the weights πi are flattened using temperature τ within the sigmoid activation function in the last linear layer of the SEN. This operation is shown in Equation 1 for the attention weights.

Equation 1: Attention weights [75, eq 1]

$$\pi_i = \frac{e^{z_i/\tau}}{e^{z_i/\tau} + 1}$$

Thirdly, the projection network creates pretrained weights for the final classification by projecting the feature maps from the LWCNN, which have been attention-focused by the SEN network and flattened. These features are sent to a unit hypersphere, and the weights are optimized based on the Subject-Aware Contrastive (SAC) loss.

The loss formula is presented in Equation 2, which uses a projection network to map representations to an embedding vector z. Here zi is the anchor, representing one sample from a particular subject, and $\tau \in R+$ is a temperature parameter. The set Q(i) contains all samples generated from the same subject and of the same class as the anchor. The set S(i) includes negative samples, which are divided into inter-subject negatives N(i)r and intra-subject negatives N(i)a. The loss function aims to maximize the similarity between the anchor and its positive samples while minimizing the similarity between the anchor and its negative samples, thereby improving the model's ability to differentiate between different classes and subjects.

Equation 2: Subject aware contrastive loss [75, eq 2]

$$\mathcal{L} = -\sum_{i \in I_s} log \left(\frac{1}{\|Q(i)\|} \sum_{q \in Q(i)} \frac{exp\left(\vec{\mathbf{z}}_i \cdot \vec{\mathbf{z}}_q / \tau\right)}{\sum_{s \in S(i)} exp\left(\vec{\mathbf{z}}_i \cdot \vec{\mathbf{z}}_s / \tau\right)} \right)$$

Finally, after pre training the weights using the subject aware contrastive loss, the weights are transferred to the classification layer and further trained to reach higher classification accuracy of high or low confidence.

Output

The WaveFusion framework provides three types of outputs. Firstly, it performs classification based on the combined information from all EEG electrodes' spectrograms, categorizing the input as high or low confidence, with an average F1 score and accuracy of 95.7%. Secondly, the attention weights, as seen in Figure 23, can be plotted to represent the importance the model assigns to each electrode for making a class prediction. Lastly, class activation maps (CAMs) highlight the importance of specific features localized in the input spectrograms for each EEG lead.

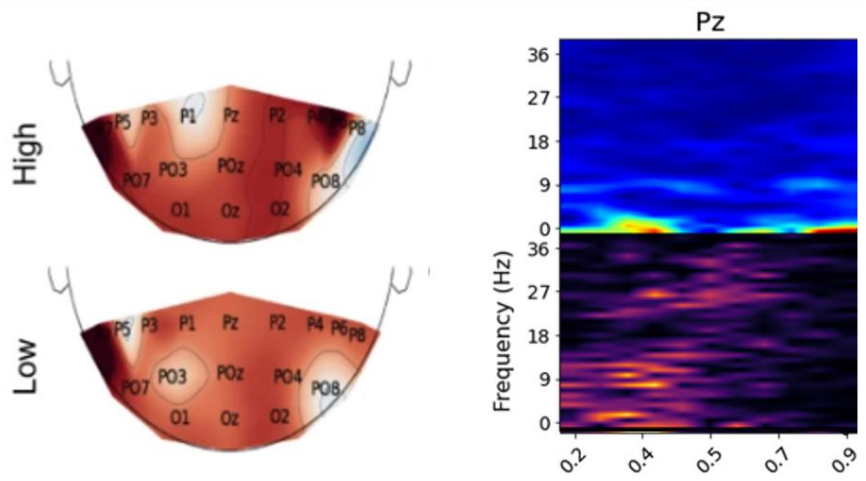

Figure 23: The visualised outputs with on the left the attention weights on a topography for both high and low confidence, and on the right the input spectrogram and it's respective class activation map for the Pz electrode in a low confidence scenario [75, Fig 3, 4].

3. Research aims and objectives

The aim of this thesis was to develop an explainable deep learning model to classify metacognitive sensitivity using EEG data, and to understand its contribution to theoretical research on metacognition. The model builds upon the WaveFusion model for classifying metacognitive confidence created by Briden and Norouzi (2023), who generously provided their code for this project. Although the python code was nearly complete, it required debugging and modifications to align with the specific objectives of this research, all code was written in python. This study utilized the response-locked event-related potential (ERP) EEG dataset from Boldt and Yeung (2015), which was

kindly provided by my co-promotor, Kobe Desender, to implement and evaluate these methods.

The specific objectives were (1) to develop the deep learning model with a classification accuracy of at least 95% for metacognitive sensitivity, and (2) to improve the metacognitive confidence classification accuracy to at least 97.5%. The code was designed to output and plot the attention weights, thereby visualizing the neural activity the model used to make its class predictions. The third objective; to identify key ambiguities and limitations in metacognition research, has been discussed in the literature review.

4. Materials & methodology

There are four main parts to this coding project, starting with the dataset, preprocessing of that data, creating the models and algorithms to process the data, and finally evaluating the model.

4.1. Dataset

The dataset used was kindly provided by my co-promotor Kobe Desender, who gave me access to the data used in Boldt and Yeung (2015). The dataset consisted of EEG data, confidence judgements and truths for 16 subjects.

The EEG data has three dimensions, first dimension (32) are the channels of the QuikCap, Neuroscan used, with the electrodes specified in the following order: FP1, FPz, FP2, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, POz, O1, Oz, O2. The second dimension (1701) is time running -500:1:1200, with time zero being the type 1 decision made. The EEG measurements are response-locked to show the event-related potential (ERP) and sampled at 1000Hz. The third dimension (800+) are the individual trials of the task procedures executed by the participants. This EEG data has been pre-processed and trial removal was performed as specified in Boldt and Yeung (2015).

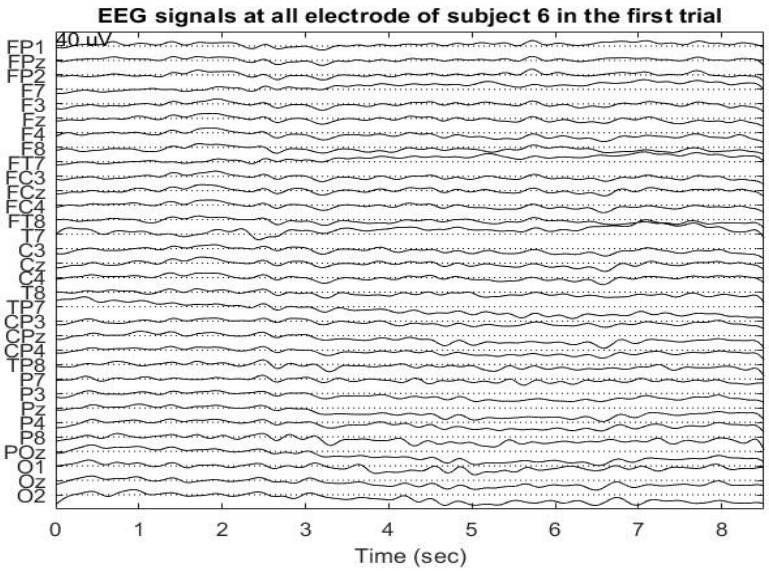

Figure 24: An example of the dataset, showcasing the signal measured across 32 electrodes for subject 6 first trial.

Furthermore, the confidence judgements are performed according to the task procedure specified in Boldt and Yeung (2015). The confidence judgements are ratings running from 1 (low) to 6 (high) and has the same length as the number of trials performed by the subject. Similarly, the truths have the same dimensions, but the values indicate if the trial was an error (1) or correct (0).

The dataset is notably imbalanced, with a predominant number of high-confidence answers that are correct. Table 5 illustrates the average number of samples per class across all subjects. There are roughly four times more samples in the high-confidence class compared to the low-confidence class. Additionally, there are about ten times more samples in the correct class than in the error class for metacognitive sensitivity. The smallest sample sizes were observed in the high-confidence and correct class for Subject 2, with only 4 answers, and in the low-confidence and incorrect class for Subject 6, with only 5 answers. The number of samples per class per subject can be seen in the table 6.

Table 5: Average number of samples per subject divided across their classes.

Average number of samples	High Confidence	Low Confidence	Total
Correct	663	112	775
error	36	38	74
Total	699	150	849

4.2. Preprocessing

Preprocessing was implemented to firstly obtain two classes for both metacognitive confidence and sensitivity. Secondly, to split the data into train and test sets. Thirdly, to generate more samples, and lastly to convert the EEG samples into spectrograms. This was done separately for every subject and for three different selection areas: the full head, frontal and posterior area. In general, the preprocessing approach was to be as consistent as possible to that of Briden and Norouzi (2023), the paper of which the deep learning model was used.

Starting with obtaining the classes, a different method was necessary for metacognitive confidence and sensitivity. Here the EEG data was already down sampled from 1000hz to 100hz to match the approach of Briden and Norouzi (2023). This down sampling maintains the integrity of the frequency activity of the brain which ranges typically between 0.5-40 Hz and reduces the size of the data to make processing more manageable. Moving on, it was important to start by defining the confidence classes before implementing the sensitivity data split. The EEG data was split based on the confidence judgements from 1-6 and were divided into low (<4) and high (≥4) confidence classes while creating separate variables for the truths to keep them aligned with their EEG samples. To create the two classes of correct and incorrect judgements for metacognitive sensitivity a division based on the truths and the confidence class was made. The correct class consisted of the high confidence class with the correct truth (0) and the low confidence class with the error truth (1), and vice versa for the incorrect class. Metacognitive sensitivity is often quantified using meta-d', which requires a batch of samples and is not suitable for individual trials [79]. To address this, we implemented a class division approach to capture metacognitive sensitivity. This approach represents instances where a subject's metacognitive sensitivity is accurate by displaying high confidence when the type 1 decision is correct and low confidence when the decision is an error. It is the opposite for the incorrect metacognitive sensitivity. The two variables of correct high and low confidence were added in variable and randomly shuffled for good measure, and vice versa for the incorrect variables.

Table 6: The amount of samples divided across their classes, including the total and derived the derived metrics, actual metacognitive sensitivity and metacognitive bias.

Dataset	High Confidence Correct	High Confidence Incorrect	Low Confidence Correct	Low Confidence Incorrect	Total	Correct trials	Metacognitive sensitivity (%)	High Confidence trials	Metacognitive Bias (%)
Subject 1	704	37	66	22	829	770	92.9	741	89.4
Subject 2	643	4	128	54	829	771	93.0	647	78.1
Subject 3	689	28	105	19	841	794	94.4	717	85.2
Subject 4	703	20	69	61	853	772	90.5	723	84.7
Subject 5	583	100	100	67	850	683	80.4	683	80.4
Subject 6	690	43	73	5	811	763	94.1	733	90.3
Subject 7	657	47	139	14	857	796	92.9	704	82.0
Subject 8	701	24	86	48	859	787	91.6	725	84.4
Subject 9	750	18	33	58	859	783	91.1	768	89.5
Subject 10	664	16	78	103	861	742	86.2	680	79.2
Subject 11	630	69	147	10	856	777	90.8	699	81.9
Subject 12	716	20	89	10	835	805	96.4	736	88.1
Subject 13	621	66	97	77	861	718	83.4	687	80.0
Subject 14	576	40	217	25	858	793	92.4	616	71.8
Subject 15	627	16	211	6	860	838	97.4	643	74.9
Subject 16	649	22	158	25	854	807	94.5	671	78.7

Secondly, the data was split into training and validation sets using the train_test_split function from scikit-learn. This division ensured that the model could learn and validate on distinct datasets, thereby enhancing the reliability and robustness of the model's performance. No test set was created due to the imbalanced data, there were not enough class two samples, being low for confidence and incorrect for metacognitive sensitivity. This imbalanced dataset would lead to too much redundant data for generating averaged samples as explained in the next step. For confidence classes a 70/30 split was performed and for sensitivity a 50/50 split was performed due to the lower number of samples in the imbalanced classes.

Thirdly, generating more samples was performed to have sufficient data for the deep learning model to train on and to stay consistent with the methodology of Briden and Norouzi (2023). More samples were created by averaging a randomly selected 10% subset of the EEG recordings. This was done separately for four variables created by splitting in the two different classes and the train and validation sets. For metacognitive confidence, a separate variable holding truth values was averaged and rounded to 0 (correct) and 1 (error) for their respective averaged EEG recording. In total five hundred samples were created for the train sets and 250 samples for each of the classes. This step not only augmented the data but also reduced noise and variability in the recordings, resulting in a more robust dataset for model training.

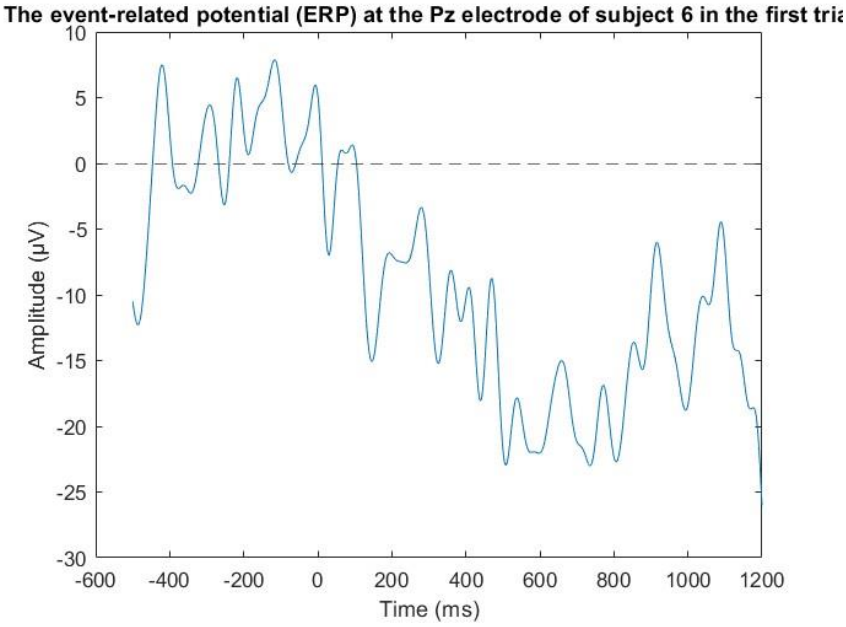

Figure 25: An example of the event-related potential (ERP) which later on is converted to a spectrogram.

Lastly, the EEG samples were converted into spectrograms. Spectrograms were again separately generated for both classes, and the training and validation. The spectrograms were created by applying the STFT using a window size of eighty datapoints, with the Hann window and a 75% overlap. These settings were specified in Briden and Norouzi (2023) and provide a good compromise in spectral and time resolution and leakage. The commonly used Hann window was not specified in the paper but was chosen in this project as again it provides a great compromise in maintaining resolution and reducing spectral leakage. The spectrograms have the dimension (41,10) with units respectively (Hertz, seconds), and are created per EEG electrode of the samples per subject. The absolute values are then taken to obtain the

magnitude of the spectrograms, in the case that only a subsection of the electrodes is selected this is done right after. Finally, the spectrograms are saved specific labelled directories, according to the class and the train or validation group. The samples themselves were name coded using the subject number, the sample number, the class label and the (augmented) truth values. The filename was the following: 'train/{group}/spectrogram_{subject}_{i}_{label}_{truth}.npy'. This careful labelling ensured the integrity and traceability of the data.

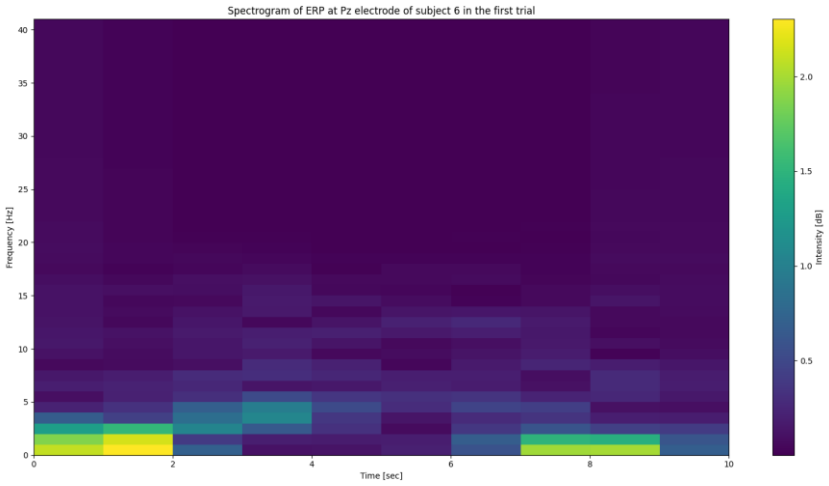

Figure 26: An example of a spectrogram used as an input for the deep learning model.

4.3. Models and algorithms

The deep learning model in this section is based on code that originates from Briden and Norouzi (2023). It contained several scripts to load the data, create, train, and select the model. Several details were debugged, and modifications were made to train it on metacognitive sensitivity. The WaveFusion framework worked best with the subject aware contrastive loss (SAC) with a batch size of five hundred, comprising of 50% positives and 50% intrasubject negatives in the research of Briden and Norouzi (2023), so this was maintained.

The following model and algorithms can predict the class of metacognitive confidence or sensitivity. The prediction type depends on the input dataset and whether the labels are designated for confidence or sensitivity.

4.3.1. Data loading

Before the model was trained the data first had to be loaded into the python environment, and secondly the samples correctly balanced to avoid class imbalanced training.

4.3.1.1. Loading the Dataset

The data was loaded using the name coding and the file location of the samples. The file location was defined as ‘train’ or ‘val’ to organize the samples into training and validation. The labels were extracted from the filename, encoding for either metacognitive confidence or sensitivity, depending on the mode of the model.

Once the dataset was loaded, we applied data augmentation transformations to the training data. These transformations included adding pink noise, randomly dropping input signals, and corrupting the data with Gaussian noise as done in Briden and Norouzi (2023). These steps were essential for simulating real-world noise and variability, this enhanced the model's robustness and generalizability. The augmented data was then fed into PyTorch data loaders, which facilitated efficient batching and parallel processing during model training.

4.3.1.2. Balanced Batch Sampler

To address the issue of class imbalance in the training data, a custom batch sampler was employed. This sampler created balanced batches with equal representation of different classes, the superior 50/50 split mentioned earlier, ensuring that the model received a well-rounded training experience.

4.3.2. Model structure

The deep learning model structure was previously explained in Figure 27, now we will go in more depth on how this was implemented in python scripts. The model consists of four sections 1) Lightweight Convolutional Network (LWCNN) 2) Squeeze and Excitation Network (SEN), 3) WaveFusion Projection Network and 4) WaveFusion Classification Network.

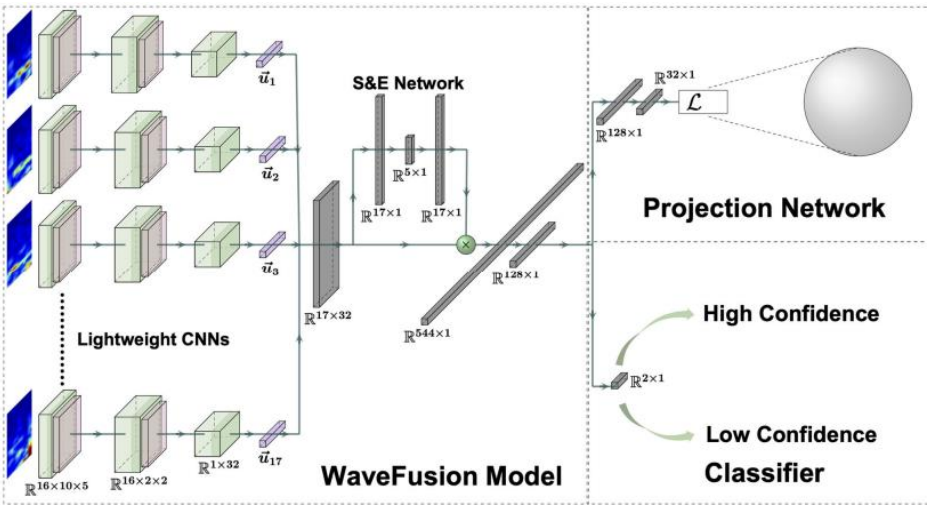

Figure 27: The WaveFusion architecture [75, Fig 1].

4.3.2.1. Lightweight Convolutional Network (LWCNN)

The model starts by processing the input spectrograms using a Lightweight convolutional neural network (LWCNN), it is called lightweight because of its relatively small and simple architecture. Each EEG electrode is convolved separately using a series of convolutional, dropout, max-pooling, and batch normalization layers as specified in Table 5. The convolutional layers capture the local temporal and spectral features of the EEG signals, creating a 2D feature map, while the dropout, max-pooling, and batch normalization layers help in regularizing the model to prevent overfitting. The output from each lead-specific CNN is then flattened and combined to form a tensor holding the learned features from all the EEG electrodes.

Table 5: The hyperparameters detailed for the LWCNN 32/12x41x10 input size.

Operation	Kernel	Strides	Padding	Count	BN?	Dropout	Nonlinearity
2D Convolution	5 × 4	2 × 1	2 × 1	16	✗	0.01	ReLU
2D Max-Pooling	2 × 2	N/A	N/A	16	✗	N/A	N/A
2D Convolution	4 × 2	2 × 1	0 × 0	16	✓	0.01	ReLU
2D Max-Pooling	2 × 2	N/A	N/A	16	✗	N/A	N/A
2D Convolution	2 × 1	1 × 1	0 × 0	32	✓	0.01	N/A

The table provided outlines the hyperparameters for the Lightweight Convolutional Neural Network (LWCNN) layers used in the model. Here is a detailed explanation of each column. The "Kernel" column indicates the size of the convolutional kernel (or

filter) used in both the convolutional layers and max-pooling layers. This kernel size defines the receptive field of the convolution operation, determining how many input pixels (or features) are considered at a time. The "Strides" column specifies the step size with which the convolutional kernel moves across the input data, influencing how much the kernel shifts at each step. The "Padding" column shows the amount of zero-padding added to the input 2D spectrograms around the borders, which helps control the spatial dimensions of the output feature maps. Lastly, the "Count" column denotes the number of output filters produced by the convolutional layer, with each filter detecting different features from the input data.

4.3.2.2. Squeeze and Excitation Network (SEN)

The model incorporates an attention mechanism to focus on the most relevant electrodes of the input data. The attention class is the Squeeze and Excitation Network (SEN) described by Briden and Norouzi (2023). The process begins with an adaptive average pooling layer that reduces the spatial dimensions of the input, forming a vector, with a value per EEG lead. The architecture then includes an encoder-decoder model, which uses convolutional layers but with a kernel of one, making it functionally the same as a fully connected layer (fc). In the encoder stage, the averaged tensor is passed through a convolutional layer (fc1), which reduces the dimensionality of the input. This layer is followed by a ReLU activation function, which introduces non-linearity and helps capture essential features. The encoded representation is further processed by batch normalization (bn), which stabilizes and accelerates the training process by normalizing the inputs of the next layer. In the decoder stage, the condensed representation is expanded back to its original size using another convolutional layer (fc2). This layer increases the dimensionality of the encoded features, preparing them for the final attention weighting. The output is then scaled by a temperature parameter using a sigmoid activation function, as seen in Equation 1, which adjusts the attention weights to ensure a balanced emphasis on all features. The sizes of these layers depend on the number of electrodes, which are equal to the input and output of the encoder-decoder model, the intermediary size varies and is set to be 25% of the input size.

Table 7: The operations within the encoder-decoder model within the squeeze and excite network (SEN).

Operation	Kernel	BN?	Nonlinearity
Avg Pooling	N/A	✗	N/A

1D Convolution (FC-like)	1	✗	N/A
ReLu and BN layer	N/A	✓	ReLU
1D Convolution (FC-like)	1	✗	Sigmoid

These attention weights are applied to the outputted feature maps of the Lightweight Convolutional Network (LWCNN). Before these feature maps are further processed in the WaveFusion projection network or the classifier, The attention-augmented features extracted undergo further processing through Relu, batch normalization, a fully connected layer. The ReLU activation function is applied to introduce non-linearity. The batch normalization stabilizes the learning process by standardizing the inputs to subsequent layers, ensuring consistent feature distributions. The output size thirty-two of the LWCNN is used for reshaping the feature distribution. The fully connected layer reduces the dimensionality of the feature maps, making the data more manageable and highlighting the most critical features for classification.

Table 8: The operations after the encoder-decoder model within the squeeze and excite network (SEN).

Operation	**Size**
ReLu	Number of electrodes * 32
Batch Normalization	Number of electrodes * 32
Fully Connected Layer	128

4.3.2.3. WaveFusion Projection Network

The WaveFusion projection network will pre-train the weights in the Lightweight convolutional neural network (LWCNN) and the Squeeze and Excitation Network (SEN) based on the Subject Aware Contrastive loss (SAC). Building on these sections the projection head module as described in Briden and Norouzi (2023) is added, consisting of an additional two fully connected layers and further refines the feature representations, applying ReLU activation function in between these two layers to capture complex patterns. This hierarchical structure enables the model to capture complex patterns across multiple EEG electrodes, significantly improving the weights pre-trained on the EEG data.

Table 9: The operations within the WaveFusion Projection Network (WFP).

Operation	**Size**
Head Module - FC Layer 1	128
ReLu	128
Head Module - FC Layer 2	32

4.3.2.3.1. Subject Aware Contrastive loss (SAC)

The Subject Aware Contrastive loss (SAC) is designed to enhance the discriminative ability of neural network models by leveraging both class and patient labels. This loss function is used to pre-train the Lightweight Convolutional Network (LWCNN) and the Squeeze and Excitation Network (SEN) within the model. The process following process in the code represents the mathematical equation seen in Equation 3. It begins by ensuring the input features have the correct shape and splitting the labels into class and patient identifiers. Masks are created to identify positive pairs (samples with the same class and patient) and used to compute similarity scores between feature vectors, which are scaled by a temperature parameter for stability. These masks are adjusted to exclude self-contrast cases, and log probabilities are computed to determine the mean log-likelihood of positive samples. The final loss, representing the contrastive loss for the batch, is averaged over all samples. This SAC loss encourages the model to bring similar samples closer in the feature space, where each dimension represents a learned feature, while pushing dissimilar ones apart. This process effectively pre-trains the LWCNN and SEN to learn discriminative and robust feature representations from the EEG signals, facilitating better classification performance in subsequent tasks.

Equation 3: Subject aware contrastive loss [75, eq 2]

$$\mathcal{L} = -\sum_{i \in I_s} log \left(\frac{1}{\|Q(i)\|} \sum_{q \in Q(i)} \frac{exp\left(\vec{\mathbf{z}}_i \cdot \vec{\mathbf{z}}_q / \tau\right)}{\sum_{s \in S(i)} exp\left(\vec{\mathbf{z}}_i \cdot \vec{\mathbf{z}}_s / \tau\right)} \right)$$

4.3.2.4. WaveFusion Classification Network

Now that the Lightweight Convolutional Network (LWCNN) and the Squeeze and Excitation Network (SEN) are pre-trained the classifier model is used to further train and validate the network. Building on the feature extraction capabilities of the previous sections, a classification head now replaces the projection head module as described in Briden and Norouzi (2023). This head consists of a fully connected layer that reduces that reduces the dimensionality of the feature maps from the SEN to 128 units, followed by a dropout layer to prevent overfitting. The final stage of the network consists of another fully connected layer that outputs the class scores, which represent probabilities of the model's predictions for each EEG input sample. The model is designed to output both the class predictions and the attention weights, providing

insight into which parts of the EEG data were most influential in making the classification decision. Originally the script did not output the attention weights but was modified to include them as an output.

Table 10: The operations within the WaveFusion Classification Network (WFC).

Operation	Output size
Fully Connected Layer	128
Dropout Layer	128
ReLu	128
Fully Connected Layer (classifier)	2

The WaveFusion Classification Network also includes methods for loading pre-trained weights and freezing or unfreezing parameters, enabling fine-tuning and transfer learning. The capability to selectively freeze or unfreeze the parameters of the model is instrumental in transfer learning. It is often beneficial to freeze certain layers of the model to preserve their learned weights while allowing other layers to adapt to the new task. This flexibility is crucial for adapting the model to different datasets or tasks while retaining previously learned knowledge. This functionality was not utilized in this project.

4.3.3. Model training

This section describes the methodology used to train the WaveFusion Classification Network, focusing on two main processes. Firstly, pre-training the model with the Subject Aware Contrastive loss (SAC), and secondly training it for classification using a standard Cross-Entropy Loss.

4.3.3.1. Pre-training

During the pre-training process, the function iterates over a specified number of epochs, each consisting of training and validation phases. In the training phase, the model parameters are updated based on the computed the Subject Aware Contrastive loss (SAC), while in the validation phase, the model's performance is evaluated on a separate validation dataset. For each batch of data, the EEG spectrogram inputs are normalized, and the optimizer's gradients are reset to zero. The model then performs a forward pass to compute the predictions and the loss. In the training phase, the loss is backpropagated, and the optimizer updates the model weights. The final loss and accuracy are calculated through the accumulation of losses and accuracies per batch

and are finally outputted by dividing the accumulated loss by the number of samples to represent the loss per sample.

The function tracks and prints the loss and accuracy for both training and validation phases. Although the model weights continuously update during training, the "best model weights" are only updated when the validation accuracy at the end of an epoch surpasses the previously recorded best accuracy. These "best model weights" are then saved and further fine-tuned.

4.3.3.2. Training for Classification

Now the pre-trained model is fine-tuned for the specific task of classification using a standard Cross-Entropy Loss. Different hyperparameters, such as dropout rates and weight decay values, were explored to find the optimal configuration.

For each combination of dropout rates and weight decay values, the function initializes a new instance of the classification model, loading the pre-trained weights into the feature extractor. In this research project the function did not use the parameter freeze for the feature extractor, allowing all previous layers such as the lightweight convolutional neural network (LWCNN) and the squeeze and excite network (SEN) to be fine-tuned during training.

The function uses the Adam optimizer and cross-entropy loss for training. It maintains histories of training and validation accuracy, as well as lists for true and predicted labels to compute performance metrics such as the accuracy, confusion matrix and F1 score.

During each epoch, the function executes forward and backward passes like the pre-training phase. Additionally, a modification was made to average the attention weights for across the final batch of the epoch with the best validation accuracy. The reason here for is that the model updates its weights gradually and the weights updated by the last batch are stored. The attention weights and the corresponding model weights are saved for the epoch that achieves the best validation accuracy during the classification phase.

The function continues training until the validation accuracy no longer improves for a specified number of epochs. The best model weights, attention weights and hyperparameters are saved and returned at the end of the training process.

4.3.4. Model selection

Model selection is done by optimizing for several hyperparameters. This optimization was performed by iterating over not only the dropout rates and weight decay values for the model fine-tuning phase but also by iterating over the pre-training weight decay, attention temperature, and contrastive temperature the values can be seen in table 11. Other fixed hyperparameters were tuned separately, like the learning rate, momentum and the pre-training and fine-tuning number of epochs. The exact values and ranges of the hyperparameters can be seen in table 11.

Table 11: The hyperparameters utilized to optimize the deep learning model

Hyperparameter	Value / Range
Attention Temperature	[27.5, 32.5, 37.5]
Supervised Contrastive Temperature	[0.01, 0.05, 0.1, 0.25]
Embedding Model Weight Decay	[0.001, 0.005, 0.0075]
Classification Model Dropout Rate	[0.5, 0.67, 0.75]
Classification Model Weight Decay	[0.001, 0.005, 0.0075]
Learning Rate	0.001
Batch Size	500
Momentum	0.0005
Contrastive Learning Epochs	2
Classification Task Epochs	5

After completing the training and fine-tuning phases, the best model configurations are identified based on the validation accuracy. The best model weights, along with the corresponding attention weights, are saved for future use. This ensures that the most effective model configuration is retained for further analysis.

4.4. Model evaluation

The trained models were subsequently utilized for further calculations to derive additional results. Initially, the accuracy was calculated on a validation set with slight data modifications. Next, the attention weights were plotted to visualize the importance the model assigned to various electrodes in making its predictions.

4.4.1. Accuracy

The prediction accuracy and F1-score were calculated for both models trained on metacognitive confidence and sensitivity. The following three steps were employed to calculate the accuracy on a modified validation dataset.

Firstly, half of the validation dataset was loaded, and several transformations were applied to this EEG data, including pink noise, drop input, and Gaussian corruption in the same way. These transformations were implemented to create more variability to challenge the model and ensure that it could perform accurately.

Secondly, the trained model with the best validation accuracy was loaded, this was done separately for the model trained on metacognitive confidence and sensitivity. These models had previously demonstrated superior performance on the validation set and was therefore selected for further analysis. The models were then employed to generate new predictions. Depending on the specific type of model used, these predictions pertained either to confidence levels or sensitivity.

Thirdly, these predictions were evaluated based on their corresponding labels to determine the accuracy and F1 score for each subject. This involved comparing the predicted values with the actual labels to assess the model's performance. The calculation of the accuracy and F1 score, both in total and per class, served as metrics to check the reliability of how well the model could generalize its predictions to the validation data with extra variability specifically for each subject.

Lastly, the actual metacognitive sensitivity from the data itself was calculated from the dataset using the amount of correct confidence judgements divided by total amount of judgements made. Likewise metacognitive bias was calculated as the percentage of high confidence trials [79]. Both values can be seen in Table 6.

4.4.2. Attention weights

To gain insight into the model's focus areas during the learning process, the attention weights are visualized on a topoplot, the actual values can be seen in the appendix 8.1. This shows the importance the model put on the different EEG electrodes to make the classification prediction. Creating this plot was done in the following four steps.

Firstly, the attention weights, which were outputted from the trained model, were extracted from the model's output files. Secondly, the loaded attention weights were baseline corrected by subtracting the minimal value from all the other attention weight values. This correction was necessary to standardize the values and facilitate create a topoplot based on only positive values. Due to the normalization method employed in the code by Briden and Norouzi (2023), all the attention weight values hovered

closely around 0.5. Baseline correction adjusted these values to ensure that any deviations from the baseline represented significant variations in attention.

Thirdly, the electrodes were selected based on the legend provided and the electrodes chosen during preprocessing. Finally, the selected electrodes and their respective corrected attention weights were plotted using standard settings. However, these plots were custom fit to the head outline using the built-in function. This customization involved adjusting the plotting parameters to align the electrode positions accurately with the head outline, providing a clear and accurate visualization of the attention distribution. The resulting topoplots offer a detailed view of the model's attentional focus and interpolates the values across different regions of the scalp.

5. Results

The results will be presented by first providing an overview of the models' performance. This will be followed by a detailed analysis of subject-specific model performance.

5.1. Performance overview

We will first explore the accuracy of the model per class and per head area used as input data. Secondly, we will show the attention weights visualized on topoplots.

5.1.1. Accuracy

Starting with an overview table containing the accuracy, F1 and the accuracy per class for both metacognitive confidence and metacognitive sensitivity classification. The table 12 also shows the impact of different head areas used as input data on the model’s predictive capability. Three modes were used, all the EEG electrodes as input data, only the frontal electrodes, meaning the first sixteen electrodes of the thirty-two, and lastly the posterior electrodes, meaning the last sixteen electrodes of the thirty-two. A bonus model was added to gain insight into the repeatability of the deep learning algorithm.

Table 12: The model performance metrics for the various head selection areas and their model type.

Head Area	Accuracy (%)	F1 (%)	Class 1 (%)	Class 2 (%)
Confidence			High	Low
Full head	99.7	99.7	99.7	99.8
Frontal	95.3	95.3	95.0	95.5
Posterior	90.3	90.6	88.0	92.8
Sensitivity			Correct	Error
Full head	98.8	98.9	97.7	99.9
frontal	99.1	99.1	98.7	99.4
posterior	98.9	98.9	98.3	99.5
Bonus			High	Low
Full head confidence (2)	99.6	99.6	99.6	99.6

Overall, the model performed best for predicting metacognitive confidence using the full head as an input data. While for metacognitive sensitivity directly the model using the electrodes from only the frontal head area performed the best. For both confidence and sensitivity for all models, class 2 had a better prediction accuracy than class 1. Notably the choice of head area was not an influential factor for sensitivity predictions, but it was for confidence where that the full head performed much better than the frontal area and likewise performed better than the posterior area. These accuracies were calculated on a modified validation dataset, while Briden and Norouzi (2023) used the accuracy from the validation mechanism built into the deep learning model. The accuracy values for the models in table 12 for this validation set can be seen in the appendix 8.2.

5.1.2. Hyperparameters

Each model was trained for a total of seven epochs: two pre-training epochs using contrastive learning and five regular training epochs using standard Cross-Entropy Loss. The training was conducted with a learning rate of 0.001, a momentum of 0.0005, and a batch size of five hundred with an equal number of samples from both classes. Table 13 displays other hyperparameters used as the combination for training the models. Interestingly, using the same hyperparameters for the head selection area does not always achieve a great result, portraying a certain inconsistency, but in general it converges to the same hyperparameters as seen for the bonus full head confidence.

Overall, the best hyperparameters for classifying metacognitive confidence were quite consistent, with only slight variations in dropout rate and weight decay. In contrast, the optimal hyperparameters for classifying metacognitive sensitivity showed more significant variability. Notably, an attention temperature of 27.5, as shown in Equation 1, was uniformly the best for training the deep learning models for classification. This value, being the lowest option, resulted in more pronounced attention weights.

Table 13: The hyperparameters used for the models.

Head Area	Embedding Model Weight Decay	Supervised Contrastive Temperature	Attention Temperature	Classifier Dropout Rate	Classifier Weight Decay
Confidence					
Full head	0.001	0.1	27.5	.75	0.005
Frontal	0.001	0.1	27.5	.5	0.005
Posterior	0.001	0.1	27.5	.5	0.001
Sensitivity					
Full head	0.0075	0.05	27.5	0.75	0.0075
frontal	0.005	0.01	27.5	0.75	0.001
posterior	0.001	0.1	27.5	0.67	0.005
Bonus					
Full head confidence (2)	0.001	0.1	27.5	.75	0.005

The embedding model weight decay is an L2 regularization parameter used in the pre-training phase. A higher value of weight decay helps prevent overfitting by penalizing larger weights. For the sensitivity model, this parameter varied, while it remained relatively low for the confidence model.

The supervised contrastive temperature in Equation 1 also varied in the sensitivity models but tended to be higher for the confidence model. Similar to its role in attention weights, the contrastive temperature in the loss function modulates the emphasis on distinguishing between the two classes. Lower contrastive temperatures increase this discrimination.

The classifier dropout rate is a regularization technique that randomly sets fractions of neuron inputs to zero during training. A higher dropout rate helps prevent overfitting by ensuring that the model does not rely too heavily on any single neuron.

Finally, the classifier weight decay, like the embedding model weight decay, is an L2 regularization parameter used during the classifier training phase. A higher value helps prevent overfitting by penalizing large weights in the classifier.

5.1.3. Attention weights

The attention weights represent the areas on the scalp, specifically the EEG electrodes that contributed the most to the predictions of the model. These attention weights are shown on topoplots, and their values are interpolated across the scalp between the areas of the EEG electrodes. The topoplots of predicting the classes of metacognitive confidence to the left and metacognitive sensitivity on the right. These topoplots are shown side by side for each of the selection areas, starting with the full head, then the frontal area and finally the posterior area. The darker the red area, the more importance the model put on the data coming from that region for making a classification prediction. The attention weight values before and after baseline correction can be seen in the appendix 8.1. Due to the large volume of data and their relatively low standalone importance, they are better visualized in the following topoplots. To test the reliability of the model an extra full head confidence topoplot was generated by creating an extra model.

5.1.3.1. Full head

Figure 28 shows the full head with the attention weights applied, as a topoplot. On the left we see the attention weights of the model trained on confidence. Which compared to the right, the attention weights of the model trained on sensitivity, is more evenly distributed from where the attention is placed. Most importance is put of the Fp electrode and the central regions C3 and Cz. For the topoplot of sensitivity the important regions are mostly concentrated in the frontal left area with the F7 and FT7. The sensitivity topoplot also shows P8 and TP8 as rather important areas, the rest of the attention is more randomly scattered with more neutral spot like P4 and FC3.

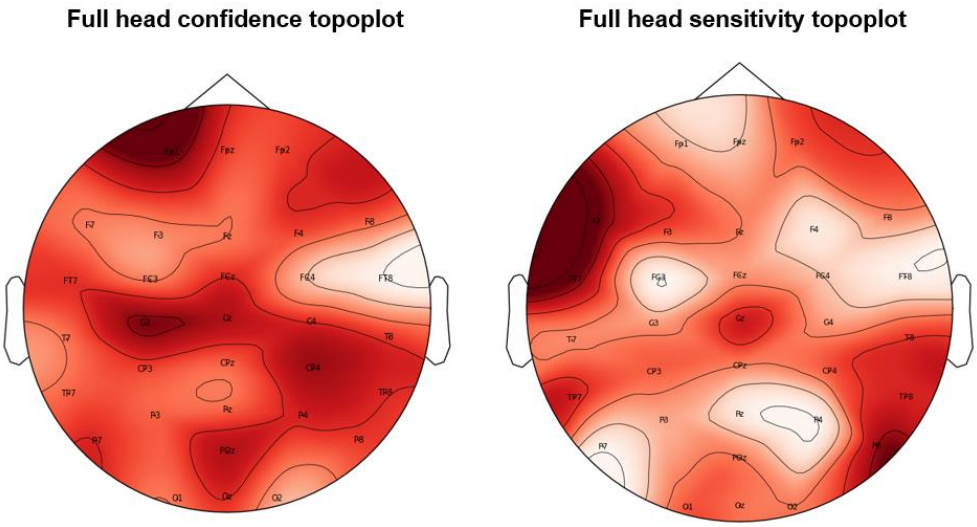

Figure 28: The topoplots visualizing the attention weights for the full head selection area, with on the left the weights for confidence and the right sensitivity.

Remarkably a model with the same hyperparameters and a slightly lower overall accuracy of 99.6% has a considerably different topoplot, as seen in the second confidence full head topoplot in Figure 29 and compared to the first confidence being the left topoplot in Figure 28.

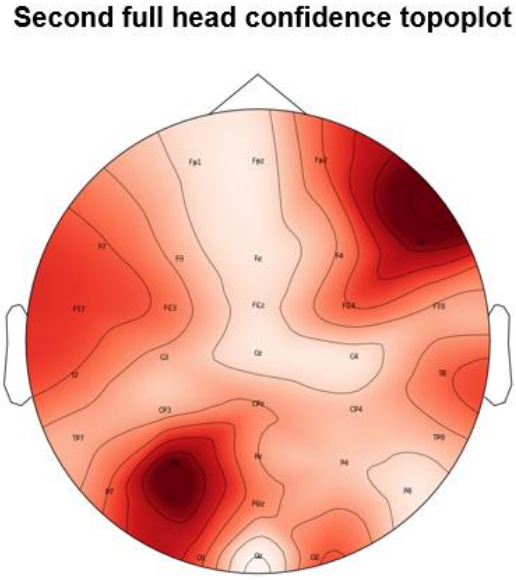

Figure 29: The topoplots visualizing the attention weights for the second confidence full head selection area.

5.1.3.2. Frontal area

Figure 30 shows the topoplot for the attention weights trained only on the first sixteen electrodes, this selection is exactly half of the thirty-two electrodes. These frontal area topoplots have more concentrated importance spots compared to the full head area. In the confidence topoplot the Fp2 and Fc4 show the most importance while C3 and

Fz are less important. The sensitivity topoplot of the frontal area shows F3, F4 and Cz as the most important electrodes, while Fcz and C3 are the least important.

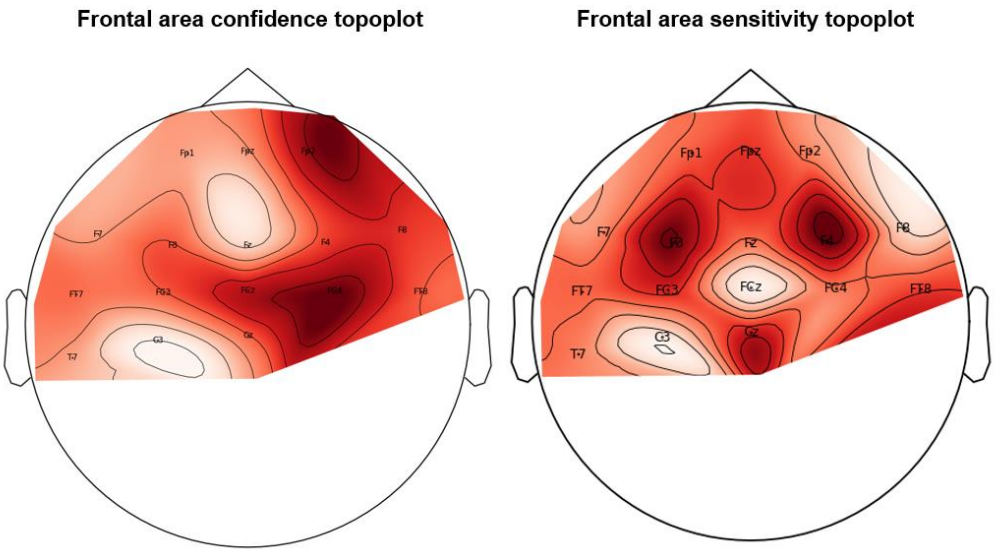

Figure 30: The topoplots visualizing the attention weights for the frontal selection area, with on the left the weights for confidence and the right sensitivity.

5.1.3.3. Posterior area

The topoplots of the posterior area are a visualization of the last sixteen electrodes of the EEG 32 measurement system. Again, the confidence topoplot has more evenly distributed importance compared to the sensitivity topoplot, with most of the importance going towards occipital Oz and left parietal electrodes Cp3, Tp7 and Pz. In the sensitivity topoplot most of the importance is directed towards the P7 lead.

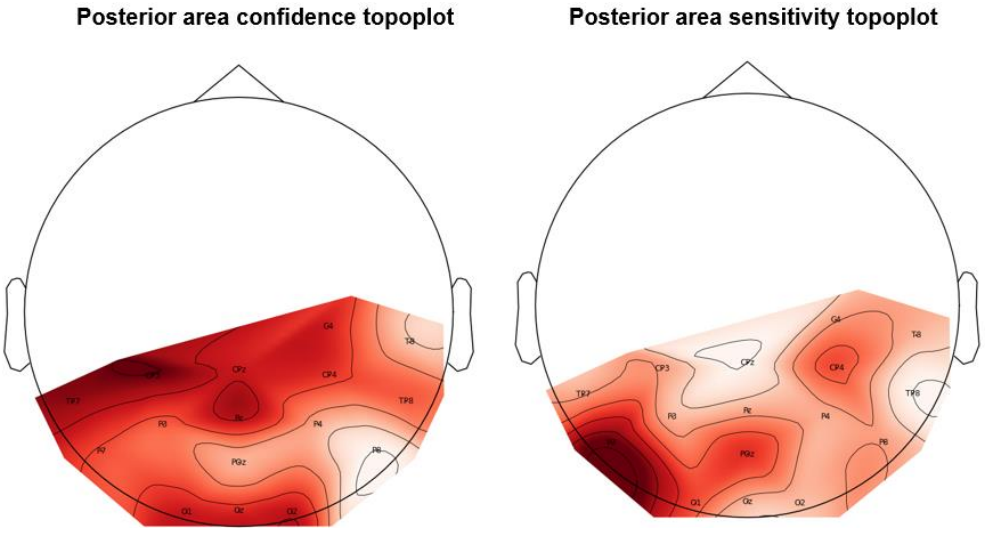

Figure 31: The topoplots visualizing the attention weights for the posterior selection area, with on the left the weights for confidence and the right sensitivity.

5.2. Subject specific performance

To gain a deeper understanding of the model's performance, we will examine its performance per subject across various head selection areas. As the F1 scores were quite similar to accuracy, as seen in Table 13, we will focus solely on accuracy. Additionally, for subjects with outlier performances, we will delve into class-specific performances to identify any underlying causes or patterns. Lastly, we will investigate the relationship between the model's performance, metacognitive bias, and actual metacognitive sensitivity using scatter plots. These results will be presented first for models trained on metacognitive confidence and then for those trained on metacognitive sensitivity.

5.2.1. Metacognitive confidence

5.2.1.1. Accuracy for each subject

In the following table, the accuracy of the model's classification of metacognitive confidence is displayed per subject for various head selection areas. The accuracy for the full head model is consistently high and near perfect for all subjects. However, when using only the frontal electrodes, the accuracy is more variable and generally lower compared to the full head accuracy. The posterior area shows even greater variability, with accuracy ranging from as high as 98% for some subjects to as low as 76% for others.

Table 14: The confidence model accuracy per subject for the various head selection areas including the second full head model.

Subject	Full head (%)	Frontal area (%)	Posterior area (%)	second full head (%)
Subject 1	99.6	87.6	83.7	99.6
Subject 2	100.0	85.7	83.5	98.8
Subject 3	100.0	100.0	97.5	100.0
Subject 4	100.0	98.7	92.4	100.0
Subject 5	100.0	97.9	95.9	100.0
Subject 6	100.0	100.0	76.6	100.0
Subject 7	100.0	98.3	98.0	100.0
Subject 8	99.6	95.3	83.9	99.6
Subject 9	100.0	98.4	98.4	100.0
Subject 10	100.0	95.8	89.3	99.6
Subject 11	100.0	97.3	98.1	100.0
Subject 12	100.0	98.0	98.0	100.0
Subject 13	98.0	89.6	82.9	98.0
Subject 14	98.8	88.7	83.2	99.2
Subject 15	100.0	97.0	89.9	99.2
Subject 16	100.0	96.2	93.4	100.0

5.2.1.2. Outlier subjects per class:

In Table 15, we examine the accuracy per class for subjects with outlying average accuracy to further analyse performance by high and low metacognitive confidence classes. Subject 13, selected from the full head selection, exhibited relatively low accuracy due to poor performance in the high-confidence class, despite perfect accuracy in the low-confidence class. For the frontal area selection, subjects 1 and 2 were chosen due to their relatively low accuracy. The low performance for these subjects was attributed to poor accuracy in the high-confidence class for subject 1 and the low-confidence class for subject 2. In the posterior area, subjects 6, 9, and 15 were notable for their low or high accuracy. Subject 6 and 15's performance was predominantly hindered by low accuracy in the high-confidence and low-confidence classes, respectively. Conversely, the high performance for subject 9 was attributed to equal performance in both classes.

Table 15: The confidence model accuracy for the outlier subject for the various head selection areas.

Confidence outlies	High class (%)	Low class (%)
Full head Subject 13	95.9	100.0
Fontal area Subject 1	75.2	100.0
Frontal area Subject 2	98.3	73.1
Posterior area Subject 6	54.6	97.6
Posterior area Subject 9	98.4	98.4
Posterior area Subject 15	97.9	81.1

5.2.1.3. Scatter plot

Scatter plots were created to explore the relationship between the model's predictive capability of metacognitive confidence and two factors: metacognitive bias and actual metacognitive sensitivity for each subject. Metacognitive bias typically reflects overconfidence, where subjects have a greater number of high-confidence trials compared to low-confidence trials, often exceeding 50%. Similarly, actual metacognitive sensitivity refers to the bias in the number of correct versus incorrect trials, with a higher sensitivity indicating more correct trials, usually above 50%.

A linear relationship between prediction accuracy and either metacognitive bias or actual metacognitive sensitivity suggests that these variables influence prediction accuracy. An R^2 value close to one or minus one indicates a strong dependence of predictive capability on the bias, while an R^2 value near zero indicates minimal dependence.

In Figure 32, confidence prediction accuracy is plotted as the dependent variable on the y-axis, while metacognitive bias is the independent variable on the x-axis. Each point represents a subject, differentiated by the various head selection areas, as shown in the legend. The results indicate only a slight influence of metacognitive bias on confidence prediction accuracy. For the full head and frontal selection areas, the R^2 values are 0.13 and 0.14, respectively, suggesting a minor influence. Conversely, the posterior area, which also had the lowest accuracy, shows an R^2 of 0.01, indicating almost no influence.

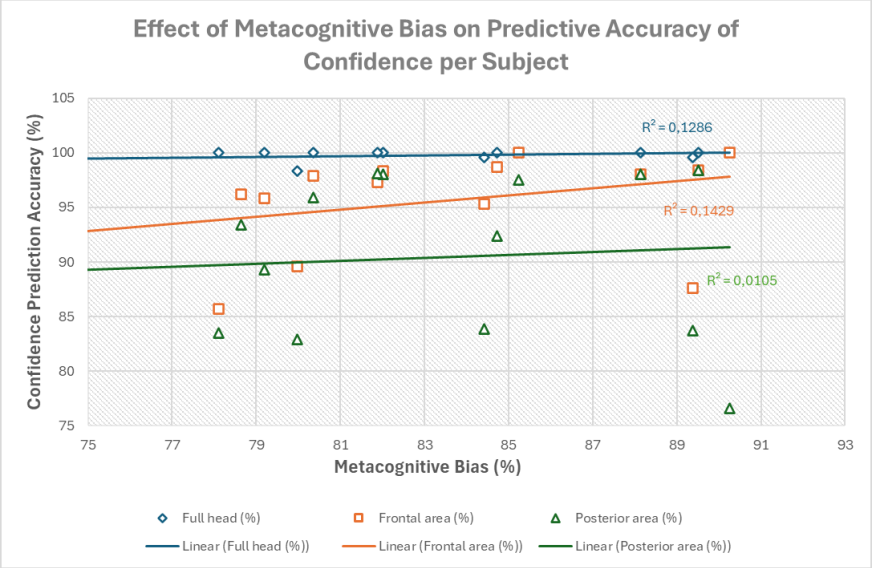

Figure 32: Scatterplot showing the relationship between metacognitive bias and the confidence predictive accuracy of the model divided for their various head selection areas.

In Figure 33, confidence prediction accuracy is the dependent variable on the y-axis, while actual metacognitive sensitivity is the independent variable on the x-axis. Each point represents a subject, categorized by different head selection areas, as indicated in the legend. The results show that there was only a slight influence of actual metacognitive sensitivity on confidence prediction accuracy for the full head selection area, which had the best performance, with an R^2 of 0.11. The frontal and posterior selection areas, with R^2 values close to zero, demonstrate no significant dependence on actual metacognitive sensitivity.

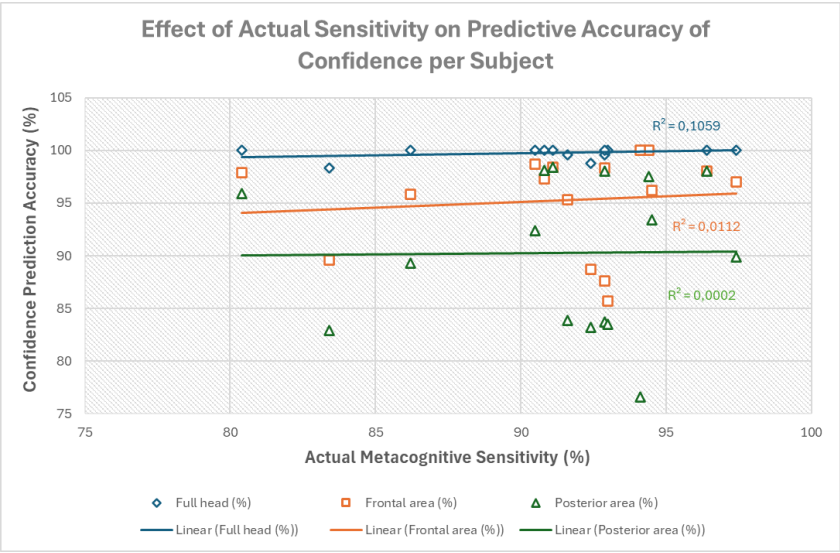

Figure 33: Scatterplot showing the relationship between actual metacognitive sensitivity and the confidence predictive accuracy of the model divided for their various head selection areas.

5.2.2. Metacognitive sensitivity

5.2.2.1. Accuracy for each subject

In the following table, the accuracy of the model's classification of metacognitive sensitivity is displayed per subject for various head selection areas. The accuracy for the full head model is consistently high and near perfect for all subjects, except for subject 10, which significantly lowered the average performance below that of the frontal area. The performance using the frontal and posterior electrodes is also consistently high, with the frontal area showing more instances of perfect accuracy compared to the posterior area. However, the posterior area has fewer scores below 98% compared to both the full head and frontal area models.

Table 16: The sensitivity model accuracy per subject for the various head selection areas.

Sensitivity	Full head (%)	Frontal area (%)	Posterior area (%)
Subject 1	100.0	100.0	99.2
Subject 2	100.0	100.0	99.6
Subject 3	100.0	99.7	100.0
Subject 4	100.0	99.6	98.3
Subject 5	97.1	96.8	98.0
Subject 6	100.0	99.1	99.6
Subject 7	100.0	98.9	99.2
Subject 8	100.0	100.0	98.8
Subject 9	99.6	98.9	98.7
Subject 10	85.5	97.8	92.8
Subject 11	100.0	96.2	99.6
Subject 12	100.0	100.0	100.0
Subject 13	99.2	99.2	99.2
Subject 14	100.0	99.6	98.7
Subject 15	100.0	100.0	100.0
Subject 16	100.0	100.0	100.0

5.2.2.2. Outlier subjects per class:

In Table 17, we examine the accuracy per class for subjects with outlying average accuracy to further analyse performance by high and low metacognitive sensitivity classes. Subject 10 was selected across all head areas due to lower performance of the model, especially for the correct class which is a trend across subjects as seen in subject 11 for the frontal area. The model trained on the full head area had an exceptionally low accuracy for subject 10 correct class having 70.9% accuracy.

Table 17: The sensitivity model accuracy for the outlier subjects for the various head selection areas.

Sensitivity outliers	Correct class	Incorrect class
Full head Subject 10	70.9	100.0
Frontal area Subject 10	95.12	100.0
Frontal area Subject 11	92.2	100.0
Posterior area Subject 10	85.9	100.0

5.2.2.3. Scatter plot

Scatter plots were now created to explore the relationship between the model's predictive capability of metacognitive sensitivity and two factors: metacognitive bias and actual metacognitive sensitivity for each subject. The same dataset used for metacognitive confidence, thus has the same distribution, is applied here. Both metacognitive bias and actual metacognitive sensitivity are skewed towards overconfidence and higher sensitivity. A linear relationship between prediction accuracy and either metacognitive bias or actual metacognitive sensitivity suggests that these variables influence prediction accuracy.

In Figure 34, sensitivity prediction accuracy is plotted as the dependent variable on the y-axis, while metacognitive bias is the independent variable on the x-axis. Each point represents a subject, differentiated by the various head selection areas, as shown in the legend. The results indicate practically no influence of metacognitive bias on sensitivity prediction accuracy, with the R^2 values being smaller than 0.10 for all head sections.

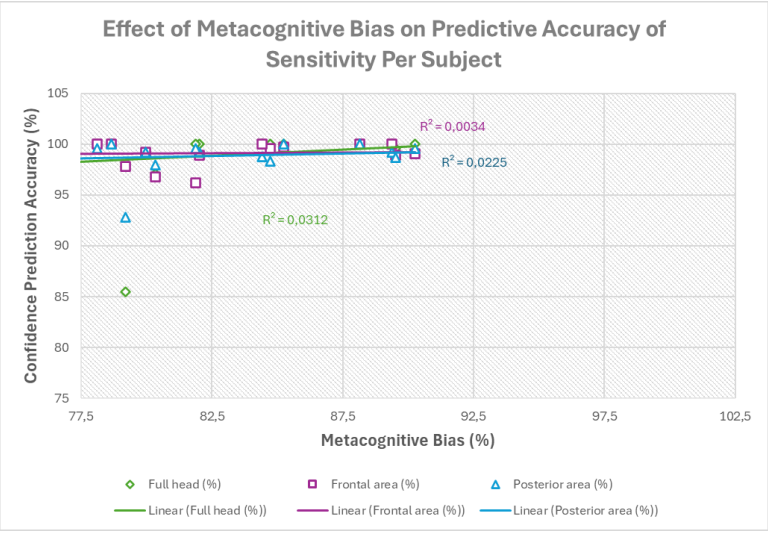

Figure 34: Scatterplot showing the relationship between metacognitive bias and the sensitivity predictive accuracy of the model divided for their various head selection areas.

In Figure 35, sensitivity prediction accuracy is plotted as the dependent variable on the y-axis, while actual metacognitive sensitivity is plotted as the independent variable on the x-axis. Each point represents a subject, categorized by different head selection areas, as indicated in the legend. The results show a slight influence of actual metacognitive sensitivity on sensitivity prediction accuracy across all head selection areas. Notably, the frontal area demonstrated the highest correlation, with an R^2 of 0.40, which is significant given its overall best predictive accuracy. The full head selection area followed with an R^2 of 0.28, and the posterior area had an R^2 of 0.21, both showing a similar classification accuracy.

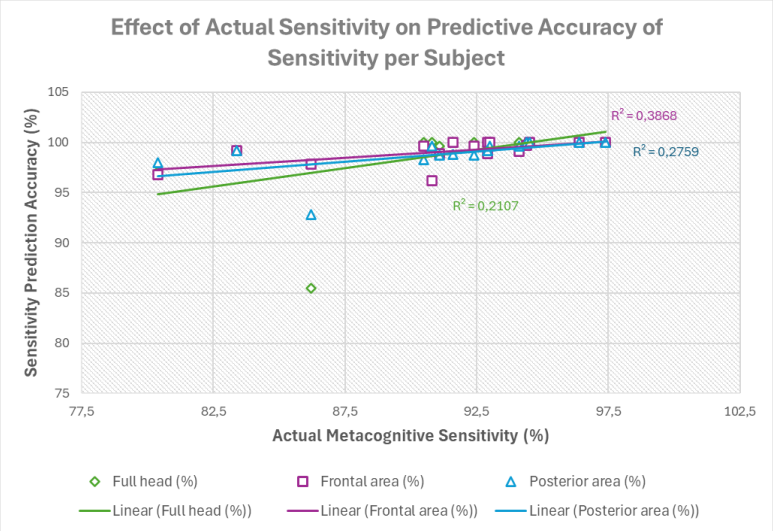

Figure 35: Scatterplot showing the relationship between actual metacognitive sensitivity and the sensitivity predictive accuracy of the model divided for their various head selection areas.

6. Discussion

The goal of this thesis was to develop and enhance a deep learning model for classifying metacognitive confidence and metacognitive sensitivity using EEG event-related potentials (ERP) across three variations of EEG electrode selections. In the following discussion, we will begin by reflecting on the methodology and the results obtained by the deep learning model. This will be followed by a discussion about the theoretical model of metacognition, concluding with suggestions for future research.

6.1. Deep learning model

To reflect on the deep learning model, we will examine three key aspects to gain a comprehensive understanding of this project: firstly, the data processing; secondly, the model performance; and finally, its explainability using the topoplots.

6.1.1. Data preprocessing

This thesis utilized the dataset from Boldt and Yeung (2015), which, as shown in Table 6, was imbalanced. This imbalance is a common issue in experimental research on metacognitive confidence [79]. The dataset contained more high confidence and correct trials, impacting the preprocessing steps. We used a standard 70/30 train/validation split for confidence, but a 50/50 split for metacognitive sensitivity to maintain data variability for the data augmentation.

In the data augmentation we averaged a random 10% of samples within each class to create five hundred training samples and 250 validation samples. However, due to the limited number of samples, this approach led to data redundancy, with duplicates of averaged samples limiting the model's ability to learn variability and generalize to other data. Although this has likely led to the higher performance of the deep learning model. This limited quantity of data electrodes us to not create a separate test set. Instead, the final performance evaluation was conducted on a noise-modified validation dataset [80]. Despite efforts to offset the effects of the imbalanced dataset through data augmentation and balanced sampling during training, these dataset limitations persisted.

The dataset included EEG measurements of response-locked event-related potentials (ERP), differing from the stimulus-locked ERP data used in Briden and Norouzi (2023). To achieve the high accuracy reported in Briden and Norouzi (2023), their methodology was closely followed, except for the differences mentioned above. Other methods, such as creating the spectrograms, were kept the same.

6.1.2. Model performance

In this thesis, we successfully reproduced and improved upon the results from Briden and Norouzi (2023). The best model for classifying metacognitive confidence achieved an impressive 99.7% overall accuracy and F1-score for the head selection area using all electrodes. Meanwhile, the best model for classifying metacognitive sensitivity

attained a 99.1% overall accuracy and F1-score for the frontal head selection area. These performance improvements were accompanied by a reduction in both pre-training and training epochs by a factor of 12.5 and 50, respectively. The main factors contributing to this enhancement are likely (1) the selection of all electrodes for training the model, and (2) the utilization of higher-quality data, including six additional subjects with an equal number of samples and the use of response-locked event-related potentials (ERP) instead of stimulus-locked ERP.

In the following sections, we will delve deeper into the model's performance for classifying confidence and sensitivity across various selection areas and individual subjects. We will then assess the hyperparameters used across the different models. Next, we will examine the class-specific performance of the models. Finally, we will explore the effect of metacognitive bias and actual metacognitive sensitivity of the subjects on the model's predictive accuracy.

6.1.2.1. confidence

The model classifying metacognitive confidence using all electrodes as data input performed significantly better than the models using selected areas, either frontal or posterior. This suggests that having more data leads to better performance or that significant metacognitive activity is present in both posterior and frontal parts. This is likely the case, as indicated by Figure 10, which shows confidence activity in the posterior-central frontal areas, particularly in the electrodes Pz, Cz, and FCz [47].

The improved performance of this model compared to Briden and Norouzi (2023) is not solely attributed to the wider selection of electrodes. Considering the posterior topoplots with similar EEG electrode selection, the validation accuracy was remarkably similar; 95.7% in their study compared to 95.2% in ours. This comparison was made using the built-in validation system in the model, disregarding the secondary validation performed in this thesis. The similar performance was achieved despite a significantly reduced number of epochs, suggesting that differences in the dataset might also have contributed to the improvement, given that preprocessing and model training methods were nearly equivalent.

Two key differences likely contributed to the performance improvement. Firstly, Briden and Norouzi (2023) used stimulus-locked EEG, while this thesis used response-locked EEG, which has been more extensively correlated with metacognitive confidence [47].

However, this aspect requires further investigation. Secondly, the increased dataset size in this thesis likely played a role. Their study included fewer subjects; ten compared to the sixteen subjects used in this thesis. Overall, using a full head selection area is the main cause for improvement. They focused on utilizing only the posterior area of the head, for which we obtained comparable results with fewer epochs. Utilizing the full scalp led to a significant improvement in the model's performance from 90.3% on the posterior area to 99.7% on the full head, using the modified validation set. This result can be recreated as seen in the bonus full head confidence model, but peculiarly it inconsistently gives a similarly result by directly inserting the same hyperparameters but does seem to converge to the same hyperparameters when training on the full hyperparameter ranges.

6.1.2.2. Sensitivity

The model classifying metacognitive sensitivity using only the frontal electrodes performed the best among the three selection areas. This superior performance, however, was largely influenced by an outlier, Subject 10, where the full head model showed particularly poor results. Despite this, the model trained on the full head area generally performed better across all other subjects. The reason for the mediocre performance on Subject 10 remains unclear, but it is specifically attributable to the correct-class instances. Notably, Subject 10 did not present similar challenges for the models classifying confidence.

6.1.2.3. Hyperparameters

Compared to Briden and Norouzi (2023), this thesis utilized considerably fewer epochs: two contrastive learning epochs for pre-training and five epochs for training the classifier, as opposed to the original 25 and 150 epochs, respectively. Despite the reduced number of epochs, high accuracy was achieved, rendering additional epochs unnecessary and avoiding the extreme computational expense associated with longer training periods. With the hyperparameters selection loops, runtimes for the lower epochs could still extend up to eight hours.

Certain hyperparameters were fixed across models, including the learning rate and momentum, set at 0.001 and 0.0005, respectively. While the momentum remained consistent with Briden and Norouzi (2023), the learning rate was significantly reduced by a factor of fifty. This reduction addressed oscillation around the optimal value across

the classification training epochs. The learning rate likely could be further optimized. However, due to time constraints, this was not pursued in this thesis.

The hyperparameters were optimized by selecting the best model for each combination. The attention temperature of 27.5 emerged as the preferred value across all models, being the lowest among the suggested values. This indicates that the model might have benefitted from an even lower attention temperature, which would lead to more pronounced attention weights. However, an even lower value was not used in this thesis as the prediction accuracy was already sufficiently high.

Similarly, for confidence specifically, a lower embedding model weight decay and a higher supervised contrastive temperature could be used. This would lead to a modified pre-training phase, but there is no suggestion that the pre-training model was over or underfitting. A decrease in the weight decay usually indicates underfitting, while an increase in the contrastive temperature suggests overfitting on the pre-training data. For sensitivity, the models ranged throughout these values, indicating a well-considered selection range given their good performance.

The hyperparameters for training the classifier, namely the dropout rate and weight decay, varied again throughout the selection range for both confidence and sensitivity, indicating an effective selection.

6.1.2.4. Class 2 and F1

Overall, class 2, representing the low confidence class and the incorrect sensitivity class, is more accurately classified by the models across all selection areas. This is likely due to the imbalanced dataset, where having less variability in the data makes it easier for the model to recognize recurring patterns. It remains unclear whether it is better to let the model underfit or if the model will underperform on other datasets.

On the other hand, it is remarkable that the F1 score is so close to the accuracy of the model. This is due to the high accuracy for both classes, achieved through the balanced batch sampler and data augmentation, ensuring an equal number of samples for both classes during training. This approach has led to a small number of false positives and false negatives. An imbalance in these metrics would cause a lower F1 score, indicating the bias of the model.

6.1.2.5. *Effect of metacognitive bias and actual metacognitive sensitivity*

The metacognitive bias and actual metacognitive sensitivity of the subjects effectively represent the class imbalance inherent to the experimental design for typical confidence judgements. Establishing a linear relationship between the model's predictive accuracy and these biases using scatter plots can indicate their effect on the model's performance. Unsurprisingly, there is barely any relationship between the prediction accuracy for confidence and the actual metacognitive sensitivity, and similarly, for the prediction accuracy of sensitivity and the metacognitive bias of the subjects.

However, there is a slight relationship between the prediction accuracy for confidence and metacognitive bias, as well as between the prediction accuracy of sensitivity and the actual metacognitive sensitivity of the subjects. This suggests that the higher these biases, the higher the predictive ability of the model. The R^2 values were higher for the relationship between the actual sensitivity and the prediction accuracy of sensitivity compared to confidence and its bias. This might be due to the greater class imbalance of metacognitive sensitivity compared to confidence, which on average had a ten-to-one and a five-to-one class imbalance respectively, as seen in Fleming and Lau (2014).

Remarkably, the R^2 is higher for the frontal area for both metacognitive bias and actual sensitivity, suggesting a higher association between the frontal area and metacognition compared to the posterior area. This is consistent with existing literature, as the prefrontal cortex has been largely associated with metacognitive confidence and sensitivity [1], [45], [62], [64], [65].

6.1.3. Model explainability

To gain insight into the neural activity that supports the model's performance, the attention weights are visualized using topoplots. These topoplots interpolate the attention weights per electrode across the scalp, offering a visual representation of the model's focus during classification. As seen in the two full head confidence topoplots the models produce inconsistent topoplots, which is a major limitation that needs to be addressed for dependable explainability of the deep learning model. Regardless, we will first compare the topoplots across the models and then analyse them individually, correlating them with known brain activity associated with metacognition.

Starting with the comparison of topoplots across the models, there is no resemblance between the frontal and posterior areas with the full head for both confidence and sensitivity. Similarly, there are no significant similarities between the topoplots for confidence and sensitivity within their respective selection areas.

To interpret the topoplots based on existing literature, it is important to understand that they are created based on spectrogram inputs, meaning they incorporate both time and frequency information. The attention weights used to create the topoplots are fixed values and do not vary across either time or frequency, which limits their interpretability. The attention weight represents the focus given to both classes, either for high and low confidence or for correct and incorrect sensitivity. It is crucial not to confuse the dark red areas and white areas as positive and negative electrical deflection; instead, they indicate areas of more or less importance placed on the activity of that electrode for making predictions. Lastly, the selection areas have different accuracy levels, which are high but not perfect, suggesting the potential for variation with improved accuracy.

6.1.3.1. Confidence

Starting with confidence, most of the temporal variation in neural activity seen in the fronto-central-parietal area of the scalp is consistent with previous findings in [47]. The electrodes in these locations, particularly Pz, Cz, and Fcz, have been extensively used to analyse event-related potentials (ERP) and correlate them with metacognitive confidence [47]. In addition to well-studied temporal activity, some spectral activity based on pre-stimulus EEG is related to confidence, particularly alpha-band power observed in the posterior part of the scalp, especially on the right side [71].

Given this background, the confidence topoplots in this study are not straightforward to interpret. However, some patterns can be discerned. The topoplots consistently show relative importance in the centre of the scalp. Specifically, in the Full Head Area, the most important electrodes identified are C3 and Cz. In the Frontal Area, the electrodes Fc4 and Fcz are highlighted as significant. Meanwhile, in the Posterior Area, the electrodes Cp3, Tp7, and Pz are considered significant.

These findings indicate that the model assigns importance to these fronto-central-parietal electrodes associated with the ERP of metacognitive confidence across various areas. However, it is noteworthy that while the topoplots do show these

electrodes as important, they do not always consider them the most important, instead assigning more importance to other electrodes. This discrepancy suggests that while the model is capturing some known relevant activity, it may also be identifying other patterns or regions not traditionally emphasized in the literature.

The full head area considers Fp1 the most important electrode. The full head topoplot also shows significance in the right posterior part, which might align with alpha-band frequency activity found in pre-stimulus EEG, though this is unclear as the topoplots represent response-locked activity [71]. The frontal area attaches considerable importance to the Fp2 electrode, while Fz is relatively neutral. The posterior area highlights the occipital electrodes O1, Oz, and O2, while being neutral on P8. There is no clear EEG research supporting these areas as playing a crucial role in metacognitive confidence. Moreover, it is unclear whether the model assigns importance to these areas due to variation in activity or the lack thereof, across both time and frequency.

The emphasis on the frontal Fp1 and Fp2 electrodes and the occipital electrodes might be explained through fMRI research. Although the differences between fMRI and EEG measurement techniques prevent direct comparison, some insights can be gleaned. Activity in the frontopolar cortex (FPC) and underlying ventromedial prefrontal cortex (vmPFC) might have influenced these channels [56]. Meanwhile, the occipital channels capture activity from the visual sensory cortex. High and low confidence have been associated with the right and left areas of the visual cortex, respectively. However, this has been attributed to the setup of the task, where judgements of low confidence are directed to the left and high confidence to the right [57]. However, this region seems to be important for discriminating high and low confidence in the posterior area.

There is a slight overlap between the posterior topoplots of this thesis and those of Briden and Norouzi (2023), particularly around the P08 and P8 electrodes being relatively unimportant. Further comparison is challenging due to differences in EEG measurement systems and slightly different posterior head area selections.

6.1.3.2. Sensitivity

EEG research on metacognitive sensitivity is quite limited, with the main insight being a relationship with prefrontal theta activity, particularly for the stimulus-locked ERP.

However, this differs from the response-locked event-related potential (ERP) used in this thesis [70]. Therefore, relating this insight to the full head and frontal area topoplots is challenging.

fMRI research indicates that activity in the frontal areas is most correlated with metacognitive sensitivity [1], [62], [63], [64]. Although this does not directly translate to EEG, it suggests that the frontal area might be more important. Interestingly, the frontal area model did achieve the highest accuracy, but this was predominantly due to the mediocre performance of the full head model on subject 10. The full head model itself does not indicate particular importance in the frontal area, mostly highlighting the F7 and FT7 areas on the left frontal side of the scalp.

There is little to no basis for further interpretation of the frontal and posterior head areas from the topoplots, given the current understanding and the specific nature of the data used in this thesis.

6.2. Theoretical model of metacognition

Since its introduction in 1976 by John F. Flavell, metacognition generally defined as "thinking about thinking" has garnered significant attention. The two- and three-component models have served as foundational frameworks for further research across various fields [5]. However, the use of these models has been hindered by their ambiguous nature [17]. Different fields have developed their own definitions and measurement techniques, leading to a divide between disciplines and a lack of transferability of insights [5].

The field of neuroscience aims to enhance the research on metacognition by offering objective measurements based on brain activity using techniques like fMRI and EEG [45], [46]. This approach has significantly advanced our understanding of metacognitive confidence. However, it faces three main challenges. Firstly, the ecological validity of the findings is often limited, as most studies are conducted under lab-controlled conditions due to the experimental setup. Secondly, there is a need for deeper insight into metacognitive skills. Lastly, there are issues with integrating and transferring insights from neuroscience to other fields [45].

Explainable deep learning models, as discussed in this thesis, offer three key advantages. Firstly, Automation of predicting metacognition can enhance ecological

validity by allowing 'lab-trained' models to predict metacognition in varied experimental setups. Secondly, transfer learning further allows the application of model weights to different contexts, aiding in the evaluation of relationships. Lastly, these models can integrate multiple types of data, such as EEG and eye-tracking, which can help combine insights across domains. These advantages give deep learning models the potential to ignite discussions on key issues in metacognition research.

However, deep learning models also have notable limitations. They require substantial amounts of data to train effectively, which can be difficult to obtain. Furthermore, their complexity often results in limited interpretability. Therefore, it is crucial to complement deep learning approaches with other research methods to gain a fuller understanding of metacognition.

6.2.1. Explainability of the fundamental model

In several fields, metacognition is suggested to have a hierarchical mechanism for its domain specificity, as illustrated in the following image from educational sciences. Similarly, in cognitive neuroscience, metacognition is proposed to have a hierarchical structure [29]. This hierarchical mechanism refers to a layered domain generality that can be divided into specific use cases. In Figure 36, it is categorized as follows: General Metacognitive Ability (GMA) at the top, a range of broad metacognitive skills (BMS) in the middle layer, and several specific metacognitive skills (SMS) at the bottom layer [81].

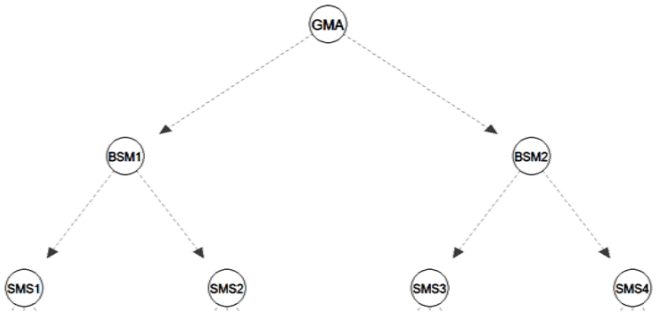

Figure 36: the hierarchical model of metacognition (adapted) [81, Fig 7]

In the cognitive neurosciences, perceptual tasks can be considered as a broad metacognitive skill (BMS), with decision making and memory tasks representing a specific metacognitive skill (SMS). It would be interesting for further research to utilize the explainable deep learning model to be able to predict and differentiate

metacognition in memory and decision making. And see if the model depends on a domain general and specific mechanism to do so.

Although speculative, this mechanism could be extrapolated across various fields, potentially offering a novel approach to re-examine or partially validate a fundamental model based on domain-general metacognition. This approach allows for the theoretical research of metacognition to be re-evaluated and discussed, integrating insights from multiple disciplines. It could reconsider various theoretical elements, such as the methods for activating metacognition and the contents of metacognitive knowledge.

6.2.2. The WaveFusion framework for the theoretical research of metacognition

This thesis improved and applied the WaveFusion framework for classifying high and low metacognitive confidence and correct and incorrect metacognitive sensitivity, while providing insights into the neural activity utilized to make these classification predictions. This framework can specifically contribute to the theoretical understanding of metacognition in the following three ways.

Firstly, a robust WaveFusion model trained on metacognition in perceptual decision-making can be used to classify confidence or sensitivity in the event-related potentials (ERP) of other cognitive functions like consciousness, executive functions, or other domains of metacognition such as metamemory [82], [83], [84]. This approach could provide insight into the ambiguous relationship between metacognition and executive functions, consciousness, and the domain generality of metacognition.

Secondly, using the trained model can enhance the ecological validity of metacognition research. By generating predictions, the model allows for greater freedom in experimental design, without requiring explicit judgements. For example, using VR technology to simulate real-life environments, as done in Yudong Tao et al (2020).

Lastly, expanding the framework for multimodality can integrate various assessment methods across different domains of metacognition, potentially improving the model's accuracy across domains. Domain-specific attention weights could visualize the correlation of brain activity between different domains of metacognition.

The theoretical debate on metacognition is complex, and the suggestions above are concise and do not address the full complexity of the issue. However, they are meant to spark the creativity of future researchers and initiate an exchange between the model's results and traditional research methods. To obtain qualitative insights in the suggestions above, a robust model needs to be created, several recommendations for improvement are provided in the following section.

6.3. Future research

Although the models in this thesis demonstrate impressive performance, several limitations need to be addressed in future research to further enhance their robustness and interpretability.

Firstly, data variability is a critical area for improvement. A more balanced dataset will significantly impact data augmentation by reducing data redundancy in the smaller class. Utilizing a more balanced dataset, without relying on data augmentation and balanced batch sampling, will provide clearer insights into the effects of metacognitive bias and actual metacognitive sensitivity on the model's performance. Furthermore, an increase in data variability, such as more subjects and various visual tasks, will provide the model with more of a challenge to achieve high accuracy but will also be more representative. Additionally, further research should isolate the differences between response-locked and stimulus-locked event-related potentials (ERPs) to better understand their respective contributions to the model's performance.

Secondly, a more detailed analysis of the model's performance might uncover additional insights. Investigating why subject 10 is an outlier for sensitivity but not for confidence is necessary for improved model performance and understanding the differences between metacognitive confidence and sensitivity. Systematically varying the electrode selection area could aid in understanding their respective importance. Hyperparameter optimization, such as using a lower attention temperature for all models and adjusting the embedding model weight decay and supervised contrastive temperature for confidence, could further enhance the models.

Thirdly, Improving the reliability of the attention weights is crucial for ensuring the explainability of the deep learning model. One approach to achieve this is by averaging the attention weights across multiple runs or possibly converging the model after more

epochs. Furthermore, interpretability utilizing class activation maps or variable topoplots is worth exploring. While topoplots are useful, class activation maps may offer better interpretability as they can highlight the most important regions contributing to the model's predictions for both time and frequency, as shown in Briden and Norouzi (2023). On the other hand, the interpretability of topoplots can be improved by making them vary over time or frequency instead of one fixed value for both. Another improvement could be by differentiating topoplots for high and low confidence as was done in Briden and Norouzi (2023).

Lastly, enhancing explainability through traditional research into the neural correlates of metacognitive sensitivity is crucial. The current model's explainability is limited by the lack of such traditional research. The insights from the deep learning model could be used to guide further traditional neuroscience research.

7. Conclusion

The goal of this thesis was to develop the explainable WaveFusion deep learning model to classify metacognitive sensitivity and confidence from EEG data, and to understand how it can contribute to the theoretical research of metacognition. The specific objectives were (1) to develop the WaveFusion deep learning model with a classification accuracy of 95% for metacognitive sensitivity, (2) improve its classification accuracy to 97.5% and (3) identify the main limitations in the theoretical research of metacognition.

In terms of model development, the WaveFusion deep learning model successfully achieved a classification accuracy of 99.7% for metacognitive confidence using all electrodes, and 99.1% for metacognitive sensitivity using the frontal head selection area. These results surpass the initial objectives, and the F1-scores demonstrate the model’s robustness for both the low and high confidence classes and the incorrect and correct sensitivity classes.

The interpretation of these results reveals several key insights. Firstly, the use of a full head selection area enhances the model's performance compared to using only frontal or posterior electrodes. This suggests that metacognitive activity is distributed across both posterior and frontal regions, and utilizing a broader range of data inputs improves classification accuracy. Further performance enhancements compared to

Briden and Norouzi (2023) 95.7% is likely due to the increased dataset size with more subjects and the use of response-locked event-related potentials (ERP) instead of stimulus-locked ERP.

The implications for theoretical research can be profound, as overcoming the following obstacles can reevaluate or validate the fundamental model of metacognition. The WaveFusion model not only achieves high classification accuracy but also provides explainability through topoplots, which visualize the neural activity associated with its predictions. This allows the framework to contribute to three major ambiguities: (1) the relationship between metacognition and executive functions, (2) its connection to consciousness, and (3) the domain generality of metacognition. By leveraging the WaveFusion framework, we can overcome limitations in cognitive neuroscience research through (1) utilizing transfer learning to compare relationships, (2) employing automatic classification to investigate ecological validity, and (3) expanding the framework for multimodality to integrate insights across various fields.

Future research should focus on addressing the limitations identified in this thesis to make the model more robust. Improving data variability by using more balanced datasets and exploring the differences between response-locked and stimulus-locked ERPs could provide deeper insights into metacognitive processes. Detailed analysis of outlier subjects and systematic variation of electrode selection areas will help understand the importance of different brain regions. Furthermore, optimizing hyperparameters and improving the reliability of the interpretability through advanced techniques like averaging attention weights across multiple equivalent models and class activation maps to enhance the model's robustness and utility in theoretical research.

In conclusion, this thesis successfully enhanced the WaveFusion deep learning model for classifying metacognitive sensitivity and confidence, achieving high classification accuracy, and providing insights into the neural mechanisms underlying its predictions. This model is a prime example of how explainable artificial intelligence (XAI) might be able to contribute to theoretical research, with potential to aid in the discussion for unifying the divided fields of metacognition research.

8. Appendix

8.1. Attention weights

8.1.1. Pre-correction

8.1.1.1. Full head

Table 18: Full head confidence attention weights pre-correction.

Electrode	Attention Weight
Fp1	0.5027132
Fpz	0.50015855
Fp2	0.5001319
F7	0.49898297
F3	0.49866417
Fz	0.49916422
F4	0.5000009
F8	0.49955645
FT7	0.50048214
FC3	0.49905646
FCz	0.50066346
FC4	0.49767974
FT8	0.49616224
T7	0.49932697
C3	0.5024128
Cz	0.501616
C4	0.50064427
T8	0.5003597
TP7	0.4995896
CP3	0.50015205
CPz	0.49965382
CP4	0.5019871
TP8	0.5007266
P7	0.50063294
P3	0.4995569
Pz	0.49951604
P4	0.5009426
P8	0.5003627
POz	0.5015538
O1	0.49941555
Oz	0.5007155
O2	0.49837905

Table 19: Full head sensitivity attention weights pre-correction

Electrode	Attention Weight
Fp1	0.49762455
Fpz	0.49729618
Fp2	0.50039280

F7	0.50452197
F3	0.50039990
Fz	0.49925244
F4	0.49592975
F8	0.49870750
FT7	0.50410295
FC3	0.49526978
FCz	0.49866170
FC4	0.49726573
FT8	0.49550372
T7	0.49897105
C3	0.49842006
Cz	0.50199044
C4	0.49841338
T8	0.50125260
TP7	0.50182563
CP3	0.50015910
CPz	0.49925184
CP4	0.50039740
TP8	0.50185126
P7	0.49560010
P3	0.49809796
Pz	0.49592748
P4	0.49533340
P8	0.50325000
POz	0.49926195
O1	0.50030550
Oz	0.49970454
O2	0.49931952

8.1.1.2. Frontal area

Table 20: Frontal area confidence attention weights pre-correction.

Electrode	**Attention Weight**
Fp1	0.49262312
Fpz	0.4936609
Fp2	0.5118585
F7	0.49364752
F3	0.5005596
Fz	0.487411
F4	0.5026649
F8	0.50394183
FT7	0.49821824
FC3	0.49994797
FCz	0.5095788
FC4	0.5151598
FT8	0.50041217
T7	0.4925592

C3	0.48222366
Cz	0.499707

Table 21: Frontal area sensitivity attention weights pre-correction.

Electrode	Attention Weight
Fp1	0.500296
Fpz	0.5007079
Fp2	0.49982914
F7	0.49988323
F3	0.5018503
Fz	0.49990508
F4	0.5019784
F8	0.49891722
FT7	0.5001888
FC3	0.5008733
FCz	0.49828672
FC4	0.49996555
FT8	0.5005929
T7	0.49984854
C3	0.49852243
Cz	0.50108933

8.1.1.3. Posterior area

Table 22: posterior area confidence attention weights pre-correction.

Electrode	Attention Weight
C4	0.50019515
T8	0.4996794
TP7	0.5003658
CP3	0.5004566
CPz	0.50021094
CP4	0.5001204
TP8	0.4999967
P7	0.50003886
P3	0.50000006
Pz	0.5002832
P4	0.4997696
P8	0.4994602
POz	0.4997359
O1	0.5001962
Oz	0.5001234
O2	0.50022036

Table 23: Posterior area sensitivity attention weights pre-correction.

Electrode	Attention Weight
C4	0.50000006

T8	0.49999997
TP7	0.50000006
CP3	0.50000006
CPz	0.49999958
CP4	0.5000004
TP8	0.49999958
P7	0.500001
P3	0.49999994
Pz	0.5
P4	0.5
P8	0.50000006
POz	0.50000054
O1	0.5000004
Oz	0.5
O2	0.49999994

8.1.2. Baseline corrected

8.1.2.1. Full head

Table 24: Full head confidence attention weights baseline corrected.

Electrode	**Attention Weight**
Fp1	0.00655097
Fpz	0.00399631
Fp2	0.00396967
F7	0.00282073
F3	0.00250193
Fz	0.00300199
F4	0.00383866
F8	0.00339422
FT7	0.00431991
FC3	0.00289422
FCz	0.00450122
FC4	0.0015175
FT8	0.00000000
T7	0.00316474
C3	0.00625056
Cz	0.00545377
C4	0.00448203
T8	0.00419748
TP7	0.00342736
CP3	0.00398982
CPz	0.00349158
CP4	0.00582486
TP8	0.00456434
P7	0.00447071
P3	0.00339466
Pz	0.0033538
P4	0.00478035

P8	0.00420046
POz	0.00539154
O1	0.00325331
Oz	0.00455326
O2	0.00221682

Table 25: Full head sensitivity attention weights baseline corrected.

Electrode	Attention Weight
Fp1	0.002355
Fpz	0.002026
Fp2	0.005123
F7	0.009252
F3	0.005130
Fz	0.003983
F4	0.000660
F8	0.003438
FT7	0.008833
FC3	0.000000
FCz	0.003392
FC4	0.001996
FT8	0.000234
T7	0.003701
C3	0.003150
Cz	0.006721
C4	0.003144
T8	0.005983
TP7	0.006556
CP3	0.004889
CPz	0.003982
CP4	0.005128
TP8	0.006581
P7	0.000330
P3	0.002828
Pz	0.000658
P4	0.000064
P8	0.007980
POz	0.003992
O1	0.005036
Oz	0.004435
O2	0.004050

8.1.2.2. Frontal area

Table 26: Frontal area confidence attention weights baseline corrected.

Electrode	Attention Weight
Fp1	0.01039946
Fpz	0.01143724

Fp2	0.02963486
F7	0.01142386
F3	0.01833597
Fz	0.00518733
F4	0.02044126
F8	0.02171817
FT7	0.01599458
FC3	0.01772431
FCz	0.02735516
FC4	0.03293613
FT8	0.01818851
T7	0.01033553
C3	0.00000000
Cz	0.01748335

Table 27: Frontal area sensitivity attention weights baseline corrected

Electrode	**Attention Weight**
Fp1	0.00200927
Fpz	0.0024212
Fp2	0.00154242
F7	0.00159651
F3	0.00356358
Fz	0.00161836
F4	0.00369167
F8	0.0006305
FT7	0.0019021
FC3	0.0025866
FCz	0.0
FC4	0.00167882
FT8	0.00230616
T7	0.00156182
C3	0.00023571
Cz	0.00280261

8.1.2.3. Posterior area

Table 28: posterior area confidence attention weights baseline corrected.

Electrode	**Attention Weight**
C4	0.00073496
T8	0.0002192
TP7	0.0009056
CP3	0.00099638
CPz	0.00075075
CP4	0.00066021
TP8	0.0005365
P7	0.00057867
P3	0.00053987

Pz	0.00082299
P4	0.00030941
P8	0.0
POz	0.0002757
O1	0.00073603
Oz	0.00066319
O2	0.00076017

Table 29: posterior area sensitivity attention weights baseline corrected.

Electrode	**Attention Weight**
C4	4.7683716e-07
T8	3.8743019e-07
TP7	4.7683716e-07
CP3	4.7683716e-07
CPz	0.0000000e+00
CP4	8.3446503e-07
TP8	0.0000000e+00
P7	1.4305115e-06
P3	3.5762787e-07
Pz	4.1723251e-07
P4	4.1723251e-07
P8	4.7683716e-07
POz	9.5367432e-07
O1	8.3446503e-07
Oz	4.1723251e-07
O2	3.5762787e-07

8.1.3. Bonus confidence full head area

8.1.3.1. Pre baseline correction

Table 30: Second full head confidence attention weights pre-correction.

Electrode	**Attention Weight**
Fp1	0.5027132
Fpz	0.50015855
Fp2	0.5001319
F7	0.49898297
F3	0.49866417
Fz	0.49916422
F4	0.5000009
F8	0.49955645
FT7	0.50048214
FC3	0.49905646
FCz	0.50066346
FC4	0.49767974
FT8	0.49616224
T7	0.49932697

C3	0.5024128
Cz	0.501616
C4	0.50064427
T8	0.5003597
TP7	0.4995896
CP3	0.50015205
CPz	0.49965382
CP4	0.5019871
TP8	0.5007266
P7	0.50063294
P3	0.4995569
Pz	0.49951604
P4	0.5009426
P8	0.5003627
POz	0.5015538
O1	0.49941555
Oz	0.5007155
O2	0.49837905

8.1.3.2. Post baseline correction

Table 31: Second full head confidence attention weights baseline corrected.

Electrode	**Attention Weight**
Fp1	0.00655097
Fpz	0.00399631
Fp2	0.00396967
F7	0.00282073
F3	0.00250193
Fz	0.00300199
F4	0.00383866
F8	0.00339422
FT7	0.00431991
FC3	0.00289422
FCz	0.00450122
FC4	0.0015175
FT8	0.00000000
T7	0.00316474
C3	0.00625056
Cz	0.00545377
C4	0.00448203
T8	0.00419748
TP7	0.00342736
CP3	0.00398982
CPz	0.00349158
CP4	0.00582486
TP8	0.00456434
P7	0.00447071
P3	0.00339466

Pz	0.0033538
P4	0.00478035
P8	0.00420046
POz	0.00539154
O1	0.00325331
Oz	0.00455326
O2	0.00221682

8.2. Non modified validation accuracy

Table 32: The validation accuracy of the models across various selection areas

Model	Accuracy (%)
Full head confidence	99.74
Frontal electrodes confidence	96.60
Posterior confidence	95.21
Full head sensitivity	99.26
Frontal electrodes sensitivity	99.71
Posterior sensitivity	99.65
Second full head confidence	99.92

9. Bibliography

[1] Stephen M. Fleming, S. M. Fleming, and R. J. Dolan, 'The Neural Basis of Metacognitive Ability', *Philosophical Transactions of the Royal Society B*, vol. 367, no. 1594, pp. 1338–1349, May 2012, doi: 10.1098/rstb.2011.0417.

[2] Ruth Colvin Clark and R. C. Clark, 'Metacognition and Human Performance Improvement.', *Performance Improvement Quarterly*, vol. 1, no. 1, pp. 33–45, Oct. 2008, doi: 10.1111/j.1937-8327.1988.tb00005.x.

[3] M. V. J. Veenman, B. van Hout-Wolters, and P. Afflerbach, 'Metacognition and learning: conceptual and methodological considerations', *Metacognition and Learning*, vol. 1, no. 1, pp. 3–14, Mar. 2006, doi: 10.1007/s11409-006-6893-0.

[4] Marianne Hohendorf and Markus Bauer, 'Metacognitive sensitivity and symptoms of mental disorder: A systematic review and meta-analysis', *Frontiers in Psychology*, vol. 14, Feb. 2023, doi: 10.3389/fpsyg.2023.991339.

[5] E. Norman *et al.*, 'Metacognition in Psychology', *Review of General Psychology*, vol. 23, no. 4, pp. 403–424, Oct. 2019, doi: 10.1177/1089268019883821.

[6] Mitsuo Kawato, M. Kawato, Aurelio Cortese, and A. Cortese, 'From internal models toward metacognitive AI.', *Biological Cybernetics*, Oct. 2021, doi: 10.1007/s00422-021-00904-7.

[7] Athanasios Drigas, A. Drigas, Eleni Mitsea, and E. Mitsea, 'The Triangle of Spiritual Intelligence, Metacognition and Consciousness', *Int. J. Recent Contributions Eng. Sci. IT*, vol. 8, no. 1, pp. 4–23, Mar. 2020, doi: 10.3991/ijes.v8i1.12503.

[8] Alexander Soutschek *et al.*, 'Frontopolar theta oscillations link metacognition with prospective decision making.', *Nature Communications*, vol. 12, no. 1, pp. 3943–3943, Jun. 2021, doi: 10.1038/s41467-021-24197-3.

[9] A. Bhargav and S. Srivastava, ‘Effectiveness of Vipassana Meditation on Meta Cognition and Resilience in Patients of The Major Depressive Disorder : A Pre-Post Study’, vol. 31, no. 33, pp. 277–289, May 2020.
[10] Jean-Marc Fellous *et al.*, ‘Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation’, *Frontiers in Neuroscience*, vol. 13, pp. 1346–1346, Dec. 2019, doi: 10.3389/fnins.2019.01346.
[11]Fakhirah Badrulhisham, Esther Pogatzki-Zahn, Daniel Segelcke, Tamás Spisák, and Jan Vollert, ‘Machine learning and artificial intelligence in neuroscience: A primer for researchers’, *Brain, behavior, and immunity*, Nov. 2023, doi: 10.1016/j.bbi.2023.11.005.
[12] Thomas O. Nelson, T. O. Nelson, Louis Narens, and L. Narens, ‘Why investigate metacognition’, pp. 1–25, Jan. 1994.
[13] John H. Flavell and J. H. Flavell, ‘Speculations about the nature and development of metacognition’, pp. 21–29, Jan. 1987.
[14] Stephen M. Fleming, S. M. Fleming, R. J. Dolan, and C. D. Frith, ‘Metacognition: computation, biology and function’, *Philosophical Transactions of the Royal Society B*, vol. 367, no. 1594, pp. 1280–1286, May 2012, doi: 10.1098/rstb.2012.0021.
[15] Jonathan Freedman, J. L. Freedman, Thomas K. Landauer, and T. K. Landauer, ‘Retrieval of long-term memory: “Tip-of-the-tongue” phenomenon’, *Psychonomic science*, vol. 4, no. 8, pp. 309–310, Aug. 1966, doi: 10.3758/bf03342310.
[16] P. Georghiades, ‘From the general to the situated: three decades of metacognition’, *International Journal of Science Education*, vol. 26, no. 3, pp. 365–383, Feb. 2004, doi: 10.1080/0950069032000119401.
[17] Pina Tarricone and P. Tarricone, ‘The Taxonomy of Metacognition’, Feb. 2011, doi: 10.4324/9780203830529.
[18] J. H. Flavell, ‘Metacognitive aspects of problem solving’, pp. 231–235, Jan. 1976.
[19] T. O. Nelson and T. O. Nelson, ‘Metamemory: A Theoretical Framework and New Findings’, *Psychology of Learning and Motivation*, vol. 26, pp. 125–173, Jan. 1990, doi: 10.1016/s0079-7421(08)60053-5.
[20] C. M. Roebers, ‘Executive function and metacognition: Towards a unifying framework of cognitive self-regulation’, *Developmental Review*, vol. 45, pp. 31–51, Sep. 2017, doi: 10.1016/j.dr.2017.04.001.
[21] Gregory Schraw, G. Schraw, Rayne Sperling Dennison, and R. S. Dennison, ‘Assessing metacognitive awareness’, *Contemporary Educational Psychology*, vol. 19, no. 4, pp. 460–475, Oct. 1994, doi: 10.1006/ceps.1994.1033.
[22] D. Fernandez-Duque, J. A. Baird, and M. I. Posner, ‘Executive attention and metacognitive regulation.’, *Consciousness and Cognition*, vol. 9, no. 2, pp. 288–307, Jun. 2000, doi: 10.1006/ccog.2000.0447.
[23] Jennifer A. Livingston and J. A. Livingston, ‘Metacognition: An Overview.’, Jan. 2003.
[24] A. L. Brown and A. L. Brown, ‘Knowing when, where, and how to remember ; A problem of metacognition’, vol. 1, Jan. 1978.
[25] Marcel V. J. Veenman, M. V. J. Veenman, Marleen A. Spaans, and M. A. Spaans, ‘Relation between intellectual and metacognitive skills: Age and task differences’, *Learning and Individual Differences*, vol. 15, no. 2, pp. 159–176, Jan. 2005, doi: 10.1016/j.lindif.2004.12.001.

[26] C. D. Frith, ‘Consciousness, (meta)cognition, and culture’, *Quarterly Journal of Experimental Psychology*, vol. 76, no. 8, pp. 1711–1723, Apr. 2023, doi: 10.1177/17470218231164502.
[27] Asher Koriat and A. Koriat, ‘Metacognition and consciousness.’, May 2007, doi: 10.1017/cbo9780511816789.012.
[28] Gregory Schraw *et al.*, ‘Does a general monitoring skill exist’, *Journal of Educational Psychology*, vol. 87, no. 3, pp. 433–444, Sep. 1995, doi: 10.1037//0022-0663.87.3.433.
[29] M. Rouault *et al.*, ‘Human metacognition across domains: insights from individual differences and neuroimaging.’, *Polymer Engineering and Science*, vol. 1, Aug. 2018, doi: 10.1017/pen.2018.16.
[30] E. Panadero, ‘A Review of Self-regulated Learning: Six Models and Four Directions for Research’, *Front. Psychol.*, vol. 8, p. 422, Apr. 2017, doi: 10.3389/fpsyg.2017.00422.
[31] J. H. Flavell, ‘Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry.’, *American Psychologist*, vol. 34, no. 10, pp. 906–911, Oct. 1979, doi: 10.1037/0003-066X.34.10.906.
[32] Gregory Schraw, Gregory Schraw, David Moshman, and David Moshman, ‘Metacognitive theories’, *Educational Psychology Review*, doi: 10.1007/bf02212307.
[33] G. Schraw, ‘Promoting general metacognitive awareness’, *Instructional Science*, vol. 26, no. 1, pp. 3–16, Mar. 1998, doi: 10.1007/978-94-017-2243-8_1.
[34] A. Efklides, ‘Metacognition: Defining its facets and levels of functioning in relation to self- and co-regulation’, *European Psychologist*, Jan. 2008, doi: 10.1027/1016-9040.13.4.277.
[35] A. Efklides, ‘Metacognitive experiences in problem solving: Metacognition, motivation, and self-regulation.’, Jan. 2001.
[36] A. Zohar and A. B. David, ‘Paving a clear path in a thick forest: a conceptual analysis of a metacognitive component’, *Metacognition and Learning*, vol. 4, no. 3, pp. 177–195, May 2009, doi: 10.1007/s11409-009-9044-6.
[37] Young Rae Kim *et al.*, ‘Multiple levels of metacognition and their elicitation through complex problem-solving tasks’, *The Journal of Mathematical Behavior*, vol. 32, no. 3, pp. 377–396, Sep. 2013, doi: 10.1016/j.jmathb.2013.04.002.
[38] Regina Boulware-Gooden *et al.*, ‘Instruction of Metacognitive Strategies Enhances Reading Comprehension and Vocabulary Achievement of Third-Grade Students.’, *The Reading Teacher*, vol. 61, no. 1, pp. 70–77, Sep. 2007, doi: 10.1598/rt.61.1.7.
[39] M. Ann Dirkes and M. A. Dirkes, ‘Metacognition: Students in charge of their thinking’, *Roeper Review*, vol. 8, no. 2, pp. 96–100, Nov. 1985, doi: 10.1080/02783198509552944.
[40] Marvin S. Cohen, M. S. Cohen, Jared Freeman, J. T. Freeman, Steve Wolf, and S. Wolf, ‘Metarecognition in Time-Stressed Decision Making: Recognizing, Critiquing, and Correcting’, *Human Factors*, vol. 38, no. 2, pp. 206–219, Jun. 1996, doi: 10.1177/001872089606380203.
[41] Nesrin Öztürk, N. Ozturk, and Nesrin Ozturk, ‘Assessing Metacognition: Theory and Practices’, *International Journal of Assessment Tools in Education*, vol. 4, no. 2, pp. 134–148, Mar. 2017, doi: 10.21449/ijate.298299.
[42] Vassilis Martiadis, Enrico Pessina, Fabiola Raffone, Valeria Iniziato, Azzurra Martini, and Pasquale Scognamiglio, ‘Metacognition in schizophrenia: A practical overview of psychometric metacognition assessment tools for researchers and

clinicians', *Frontiers in Psychiatry*, vol. 14, Apr. 2023, doi: 10.3389/fpsyt.2023.1155321.
[43] Philip H. Winne and P. H. Winne, 'A metacognitive view of individual differences in self-regulated learning', *Learning and Individual Differences*, vol. 8, no. 4, pp. 327–353, Jan. 1996, doi: 10.1016/s1041-6080(96)90022-9.
[44] Stephen M. Fleming, S. M. Fleming, and C. D. Frith, 'The cognitive neuroscience of metacognition', *Springer Berlin Heidelberg*, vol. 9783642451904, pp. 1–407, Dec. 2014, doi: 10.1007/978-3-642-45190-4.
[45] D. Fleur, B. Bredeweg, and Wouter Van Den Bos, 'Metacognition: ideas and insights from neuro- and educational sciences', *npj Science of Learning*, 2021, doi: 10.1038/s41539-021-00089-5.
[46] Shaohan Jiang *et al.*, 'Metacognition and mentalizing are associated with distinct neural representations of decision uncertainty', *PLOS Biology*, vol. 20, no. 5, pp. e3001301–e3001301, May 2022, doi: 10.1371/journal.pbio.3001301.
[47] Annika Boldt, A. Boldt, A. S. Boldt, Nick Yeung, and N. Yeung, 'Shared Neural Markers of Decision Confidence and Error Detection', *The Journal of Neuroscience*, vol. 35, no. 8, pp. 3478–3484, Feb. 2015, doi: 10.1523/jneurosci.0797-14.2015.
[48] Dobromir Rahnev *et al.*, 'Consensus Goals in the Field of Visual Metacognition', *Perspectives on Psychological Science*, pp. 174569162210756–174569162210756, Jul. 2022, doi: 10.1177/17456916221075615.
[49] Lirong Qiu *et al.*, 'The neural system of metacognition accompanying decision-making in the prefrontal cortex', *PLOS Biology*, vol. 16, no. 4, pp. 1–27, Apr. 2018, doi: 10.1371/journal.pbio.2004037.
[50] T. U. Hauser, M. Allen, N. Purg, M. Moutoussis, G. Rees, and R. J. Dolan, 'Noradrenaline blockade specifically enhances metacognitive performance', *eLife*, vol. 6, May 2017, doi: 10.7554/elife.24901.
[51] Paolo Di Luzio, Luca Tarasi, Juha Silvanto, Alessio Avenanti, and V. Romei, 'Human perceptual and metacognitive decision-making rely on distinct brain networks', *PLOS Biology*, vol. 20, no. 8, pp. e3001750–e3001750, Aug. 2022, doi: 10.1371/journal.pbio.3001750.
[52] P. Grimaldi, Hakwan Lau, H. Lau, and M. A. Basso, 'There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making.', *Neuroscience & Biobehavioral Reviews*, vol. 55, pp. 88–97, Aug. 2015, doi: 10.1016/j.neubiorev.2015.04.006.
[53] A. P. Shimamura, 'Toward a Cognitive Neuroscience of Metacognition', *Consciousness and Cognition*, vol. 9, no. 2, pp. 313–323, Jun. 2000, doi: 10.1006/ccog.2000.0450.
[54] A. G. Vaccaro and S. M. Fleming, 'Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements', *Brain and Neuroscience Advances*, vol. 2, p. 239821281881059, Jan. 2018, doi: 10.1177/2398212818810591.
[55] M. E. Wokke, D. Achoui, and A. Cleeremans, 'Action information contributes to metacognitive decision-making.', *Scientific Reports*, vol. 10, no. 1, pp. 3632–3632, Feb. 2020, doi: 10.1038/s41598-020-60382-y.
[56] Tricia Seow *et al.*, 'How local and global metacognition shape mental health', *Biological Psychiatry*, vol. 90, no. 7, pp. 436–446, May 2021, doi: 10.1016/j.biopsych.2021.05.013.
[57] Pascal Molenberghs *et al.*, 'Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study', *Social Cognitive and Affective*

Neuroscience, vol. 11, no. 12, pp. 1942–1951, Jul. 2016, doi: 10.1093/scan/nsw093.
[58] Stephen M. Fleming *et al.*, 'Relating Introspective Accuracy to Individual Differences in Brain Structure', *Science*, vol. 329, no. 5998, pp. 1541–1543, Sep. 2010, doi: 10.1126/science.1191883.
[59] Paul J. Healy, D. A. Moore, Don A. Moore, and P. J. Healy, 'The Trouble With Overconfidence', *Psychological Review*, May 2007, doi: 10.1037/0033-295x.115.2.502.
[60] Jason M. Carpenter *et al.*, 'Domain-general enhancements of metacognitive ability through adaptive training', *bioRxiv*, p. 388058, Aug. 2018, doi: 10.1101/388058.
[61] Indrit Sinanaj, I. Sinanaj, Yann Cojan, Y. Cojan, Patrik Vuilleumier, and P. Vuilleumier, 'Inter-individual variability in metacognitive ability for visuomotor performance and underlying brain structures', *Consciousness and Cognition*, vol. 36, pp. 327–337, Nov. 2015, doi: 10.1016/j.concog.2015.07.012.
[62] Stephen M. Fleming, S. M. Fleming, J. Huijgen, and R. J. Dolan, 'Prefrontal Contributions to Metacognition in Perceptual Decision Making', *The Journal of Neuroscience*, vol. 32, no. 18, pp. 6117–6125, May 2012, doi: 10.1523/jneurosci.6489-11.2012.
[63] Cuizhen Liu, Keqing Wang, and Rongjun Yu, 'The neural representation of metacognition in preferential decision-making', *Human Brain Mapping*, 2024, doi: 10.1002/hbm.26651.
[64] Micah Allen *et al.*, 'Metacognitive ability correlates with hippocampal and prefrontal microstructure.', *NeuroImage*, vol. 149, pp. 415–423, Apr. 2017, doi: 10.1016/j.neuroimage.2017.02.008.
[65] Pascal Molenberghs *et al.*, 'Understanding the minds of others: A neuroimaging meta-analysis', *Neuroscience & Biobehavioral Reviews*, vol. 65, pp. 276–291, Jun. 2016, doi: 10.1016/j.neubiorev.2016.03.020.
[66] Chen Song *et al.*, 'Relating inter-individual differences in metacognitive performance on different perceptual tasks', *Consciousness and Cognition*, vol. 20, no. 4, pp. 1787–1792, Dec. 2011, doi: 10.1016/j.concog.2010.12.011.
[67] R. Grützmann, T. Endrass, J. Klawohn, and N. Kathmann, 'Response accuracy rating modulates ERN and Pe amplitudes', *Biological Psychology*, vol. 96, pp. 1–7, Feb. 2014, doi: 10.1016/j.biopsycho.2013.10.007.
[68] Kobe Desender *et al.*, 'The temporal dynamics of metacognition: Dissociating task-related activity from later metacognitive processes', *Neuropsychologia*, vol. 82, pp. 54–64, Feb. 2016, doi: 10.1016/j.neuropsychologia.2016.01.003.
[69] Aslihan Selimbeyoglu *et al.*, 'What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision', *Clinical Neurophysiology*, vol. 123, no. 6, pp. 1158–1167, Jun. 2012, doi: 10.1016/j.clinph.2011.10.037.
[70] Martijn E. Wokke, M. E. Wokke, Axel Cleeremans, A. Cleeremans, K. Richard Ridderinkhof, and K. R. Ridderinkhof, 'Sure I'm Sure: Prefrontal Oscillations Support Metacognitive Monitoring of Decision Making', *The Journal of Neuroscience*, vol. 37, no. 4, pp. 781–789, Jan. 2017, doi: 10.1523/jneurosci.1612-16.2016.
[71] Jason Samaha, J. Samaha, Luca Iemi, L. Iemi, Bradley R. Postle, and B. R. Postle, 'Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy', *bioRxiv*, p. 089425, Nov. 2016, doi: 10.1101/089425.

[72] Stephen M. Fleming and S. M. Fleming, ‘HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings.’, *Neuroscience of Consciousness*, vol. 2017, no. 1, p. 007, Jan. 2017, doi: 10.1093/nc/nix007.
[73] Christopher L. Hewitson, Naser Al-Fawakhiri, and Samuel Mcdougle, ‘Metacognitive Judgments during Visuomotor Learning Reflect the 3 Integration of Error History’, 2023.
[74] Christoph Tremmel *et al.*, ‘A meta-learning BCI for estimating decision confidence’, *Journal of Neural Engineering*, Jun. 2022, doi: 10.1088/1741-2552/ac7ba8.
[75] M. Briden and N. Norouzi, ‘Toward metacognition: subject-aware contrastive deep fusion representation learning for EEG analysis’, *Biol Cybern*, vol. 117, no. 4–5, pp. 363–372, Jul. 2023, doi: 10.1007/s00422-023-00967-8.
[76] K. Desender, P. R. Murphy, A. Boldt, T. Verguts, and N. Yeung, ‘A post-decisional neural marker of confidence predicts information-seeking’, *bioRxiv*, p. 433276, Oct. 2018, doi: 10.1101/433276.
[77] Yudong Tao *et al.*, ‘Confidence Estimation Using Machine Learning in Immersive Learning Environments’, *Conference on Multimedia Information Processing and Retrieval*, pp. 247–252, 2020, doi: 10.1109/mipr49039.2020.00058.
[78] Michael Briden and Narges Norouzi, ‘WaveFusion Squeeze-and-Excitation: Towards an Accurate and Explainable Deep Learning Framework in Neuroscience.’, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2021, pp. 1092–1095, Nov. 2021, doi: 10.1109/embc46164.2021.9630605.
[79] S. M. Fleming and H. C. Lau, ‘How to measure metacognition’, *Front. Hum. Neurosci.*, vol. 8, Jul. 2014, doi: 10.3389/fnhum.2014.00443.
[80] Et Al. Nand Kumar, ‘Enhancing Robustness and Generalization in Deep Learning Models for Image Processing’, *Power system technology*, 2023, doi: 10.52783/pst.193.
[81] Cristiano Mauro Assis Gomes, C. M. A. Gomes, Hudson Golino, H. Golino, Igor Gomes Menezes, and I. G. Menezes, ‘Predicting School Achievement Rather than Intelligence: Does Metacognition Matter?’, *Psychology*, vol. 2014, no. 9, pp. 1095–1110, Jul. 2014, doi: 10.4236/psych.2014.59122.
[82] Jona Förster, J. Förster, Mika Koivisto, M. Koivisto, Antti Revonsuo, and A. Revonsuo, ‘ERP and MEG correlates of visual consciousness: The second decade.’, *Consciousness and Cognition*, vol. 80, pp. 102917–102917, Apr. 2020, doi: 10.1016/j.concog.2020.102917.
[83] Michelle Downes, M. Downes, Joe Bathelt, J. Bathelt, Michelle de Haan, and M. de Haan, ‘Event-related potential measures of executive functioning from preschool to adolescence’, *Developmental Medicine & Child Neurology*, vol. 59, no. 6, pp. 581–590, Jun. 2017, doi: 10.1111/dmcn.13395.
[84] E. F. Chua, D. Pergolizzi, Rachel Weintraub, and R. R. Weintraub, ‘The Cognitive Neuroscience of Metamemory Monitoring: Understanding Metamemory Processes, Subjective Levels Expressed, and Metacognitive Accuracy’, pp. 267–291, Jan. 2014, doi: 10.1007/978-3-642-45190-4_12.

Use of Generative Artificial Intelligence (GenAI) – Form to be completed

Student name: Juul Vande Abeele
Student number: r0955305

Please indicate with "X" whether it relates to a course assignment, to the BIG-project or to the master's thesis:

X This form is related to my master's thesis.
Title master's thesis: Decoding Metacognitive Sensitivity from EEG using Deep Learning.
Promoter: Jean-Marie Aerts, Kobe Desender

O This form is related to a BIG-project.
Title BIG-project: …
Promoter: …

O This form is related to a course assignment.
Course name: …
Course code: …

Please indicate with "X":

- o I did not use GenAI tools.

X I did use GenAI tools. In this case specify which one (e.g. ChatGPT/GPT4/...): ChatGPT

Please indicate with "X" (possibly multiple times) in which way you were using it:

X As a language assistant for reviewing or improving texts you wrote yourself, provided that the model does not add new content. In this case, the use of GenAI is similar to the spelling and grammar check tools we already have today, so you do not need to explicitly mention using GenAI for this).

X As a search engine to get initial information on a topic or to make an initial search for existing research on the topic. (This way of gathering information is similar to using an ordinary search engine when working on anassignment. As a student, you are responsible for checking and verifying the absence and correctness of references. Therefore, after this initial search, look for scientific sources and conduct your own analysis of the source documents. Interpret, analyse and process the information you obtained; don't just copy-paste it. If you then write your own text based on this information, you do not have to mention you used GenAI.)

- o **To generate text blocks.** (If you do copy-paste text blocks of GenAI output, you have to cite your GenAI sources and quote them, i.e. you clearly state that the item was created via GenAI by citation/reference.)
- o **To generate graphs or figures.** (If you do copy-paste graphs/figures of GenAI output, you have to cite the GenAI sources and quote them, i.e. you clearly state that the item was created via GenAI by citation/reference.)

X To generate some code as part of a larger assignment. (Watch out, this can only be done if the teacher/promotor explicitly allows it.)

- o **Other** (Contact the teacher of the course or the supervisor of the thesis or BIG project. Explain how you comply with article 84 of the examination regulations. Explain the usefulness or added value of using GenAI.)
- o

Further important guidelines and remarks:

The faculty follows the KU Leuven policy regarding responsible use of GenAI. This form is an aid towards transparency about the use of GenAI by the student which is essential. Irresponsible and non-transparent use of GenAI can be considered an irregularity and can be sanctioned. Students who consider to use GenAI should inform themselves through the university website concerning the additional guidelines (How to correctly quote and refer to GenAI? What is (not) allowed? Tips and points of attention for responsible use):
https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence