

Uncertainty Quantification in Machine Learning for Biosignal Applications - A Review

Ivo Pascal de Jong^{a,*}, Andreea Ioana Sburlea^a, Matias Valdenegro-Toro^a

^a*Department of Artificial Intelligence, Bernoulli Institute, University of Groningen, Nijenborg 9, 9747 AG, Groningen, The Netherlands*

Abstract

Uncertainty Quantification (UQ) has gained traction in an attempt to improve the interpretability and robustness of machine learning predictions. Specifically (medical) biosignals such as electroencephalography (EEG), electrocardiography (ECG), electrooculography (EOG), and electromyography (EMG) could benefit from good UQ, since these suffer from a poor signal-to-noise ratio, and good human interpretability is pivotal for medical applications. In this paper, we review the state of the art of applying Uncertainty Quantification to Machine Learning tasks in the biosignal domain. We present various methods, shortcomings, uncertainty measures and theoretical frameworks that currently exist in this application domain. We address misconceptions in the field, provide recommendations for future work, and discuss gaps in the literature in relation to diagnostic implementations as well as control for prostheses or brain-computer interfaces. Overall it can be concluded that promising UQ methods are available, but that research is needed on how people and systems may interact with an uncertainty-model in a (clinical) environment.

Keywords: Uncertainty Quantification, Bayesian Neural Networks, Biosignals, EEG, ECG, EOG, EMG, BCI

1. Introduction

Standard Machine Learning (ML) systems such as Random Forests, SVMs, and Neural Networks typically produce single-point estimates for their classification task. Such single-point models neglect alternative predictions that are consistent with the training data, and therefore give an inadequate estimate of the uncertainty of a prediction. As a result, they may give overconfident but completely inaccurate predictions, which induces skepticism and hinders the implementation of Machine Learning methods in clinical settings [1]. Uncertainty Quantification (UQ) attempts to address this problem by adapting Machine Learning systems to also predict a measure of confidence for a given prediction. Over the past years this has been gaining traction in Computer Vision [2], but it is still only lightly explored in Machine Learning tasks that focus on Biosignals.

Applications using biosignals can gain particular benefits from uncertainty quantification. Their signals are sensitive to artifacts that could corrupt the prediction of a Machine Learning system in unexpected ways. Uncertainty Quantification methods may help here by recognizing that the data is corrupted and indicate increased uncertainty.

Another argument for the importance of Uncertainty Quantification is that the human interpretation of the signal requires substantial time investment. Automating this

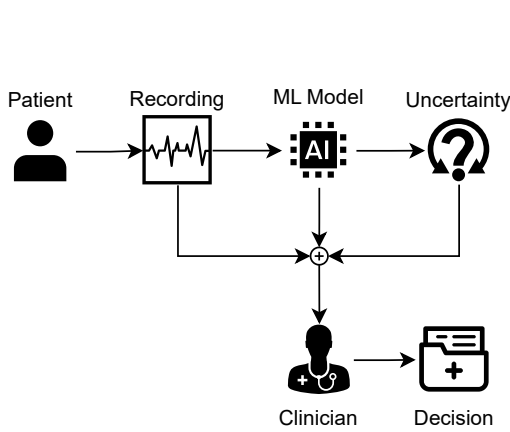
work with a Machine Learning model requires UQ to indicate when the model does not know and minimise misclassifications. To give an order of scale to the human effort: sleep scoring a patient's EEG recording of an overnight stay will typically take a neurologist about two hours [3]. A Machine Learning system that can automatically classify the majority of the overnight stay with high confidence while identifying the parts that it is uncertain on may reduce this.

Figure 1 shows various roles uncertainty estimation can play in a biosignal Machine Learning system. The primary use cases are to improve transparency of predictions for a decision support system, or to make independent classifications only when it is likely to be correct. Additionally, uncertainty estimates may be used in various ways to improve the predictions of a Machine Learning model, and it may even be used to determine when additional medical tests are needed. The interactions with a clinical system put specific expectations on uncertainty estimation for biosignal applications that do not arise in other application domains.

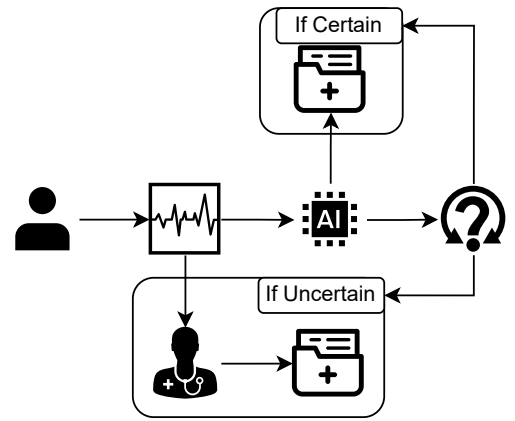
With the value that this direction of research can bring this review attempts to identify how Uncertainty Quantification methods should be used in biosignal applications. Answering this question directly is impossible, but by investigating and critically assessing the way research is currently being conducted we provide some adjustments to the current directions and suggest new avenues to be explored in the future. Moreover, we provide an overview of currently common methods as an entryway for researchers

*Corresponding author

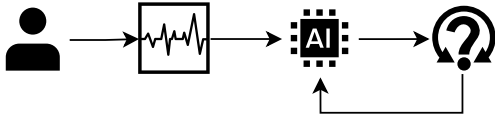
Email address: ivo.de.jong@rug.nl (Ivo Pascal de Jong)



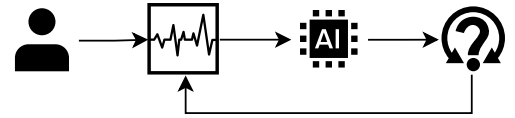
(a) Uncertainty for Decision Support. The uncertainty, prediction and data are available for the clinician. This requires interpretable uncertainties.



(b) Uncertainty for Rejection. The model will make a decision if it is highly certain and likely to be correct. Otherwise, the data is given to the clinician. This requires uncertainty that can separate highly accurate predictions, and reduces diagnostic workload on the clinician.



(c) Uncertainty for improved ML. Uncertainty may be used in methods to improve the classification performance. This puts no direct restrictions on the uncertainty.



(d) Uncertainty for extra recordings. Additional or alternative tests may be run when the model is uncertain. The predicted uncertainty should align with a clinician's uncertainty.

Figure 1: Different positions of Uncertainty Quantification in medical biosignal interpretation. Different ways of using uncertainty put different constraints on the predicted uncertainties. These designs are not mutually exclusive and uncertainty estimation may deliver multiple benefits.

new to the topic of UQ in Biosignal processing, together with a simplified end-to-end guide for implementing, applying and evaluating uncertainty.

In the rest of this section we explain how the literature review was performed to offer some usability, and we end the section with a thorough explanation of what uncertainty is. In Section 2 we discuss different methods for quantifying uncertainty. For each method we specifically discuss the relation to biosignal tasks, and we discuss niche methods that were used in biosignal tasks but that are otherwise not considered in general Uncertainty Quantification review papers.

In Section 3 we address misconceptions and confusion we observed about how a numerical measure of uncertainty should be extracted from a predicted distribution generated by some of the uncertainty quantification methods. We discuss the different uncertainty measures encountered, and give clear recommendations. While this topic is discussed a bit in the most cited review on Uncertainty Quantification [2], we provide a more explicit overview and comparison, including insights from recent research on uncertainty measures.

Then, Section 4 we describe different ways uncertainty has been used in the biosignal domain. We discuss how the

choice of use case is important as it affects what properties it should have and how it should be evaluated. This is not always apparent. By giving these guidelines we intend to make it clearer for authors and reviewers what uncertainty is useful for and how that should be evaluated.

We conclude our review paper with two sections that aim to progress the research on uncertainty in biosignals. In Section 5 we provide a guideline on how uncertainty quantification may be added to a biosignal classifier, and in Section 6 we discuss open research challenges for applying uncertainty in biosignals. Those challenges focus specifically on the interaction of an uncertainty estimating model with the environment in which it is deployed, specific properties of biosignal data, and more broadly how uncertain Machine Learning behaves in a clinical setting.

1.1. Search Method

To ensure reproducibility we used a systematic review. A first search had a higher level structure of $((\text{Uncertainty Quantification} \wedge \text{Machine Learning}) \vee \text{Bayesian Neural Networks}) \wedge \text{Biosignals}$. However, it was found that a line of research [4]–[6] uses the term “Bayesian Neural Networks” erroneously to describe classical Neural

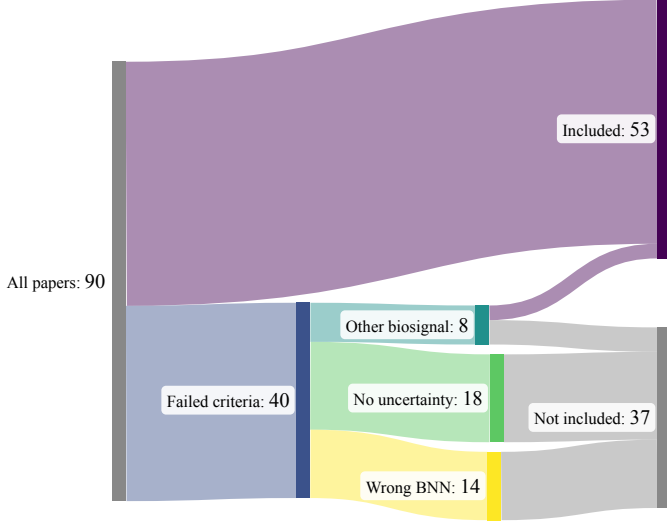


Figure 2: The flow of papers that were covered in the systematic literature search, divided by exclusion criteria.

Networks trained with Bayesian Regularization [7]. A second search was performed without the Bayesian Neural Networks disjunction.

To ensure good coverage of the review various synonyms and abbreviations were used for each term. Specifically for the Machine Learning term several Neural Networks methods were used, and various Machine Learning models such as SVM, Random Forest and Fuzzy Logic. For the application domain we searched on the following terms: EEG, ECG, EOG, EMG, BCI and fNIRS. The choice of these terms was selected for the consistent modality, as each of them covers data from a set of time series from different locations.

Works that did not discuss uncertainty in Machine Learning for one of the listed biosignals were excluded from the review. The two searches were applied to the databases: Web of Science, Scopus, IEEE Xplore and PsycINFO. Manual filtering by abstract and title resulted in a total of 90 papers, of which 50 met the criteria. 14 papers used the Bayesian Neural Networks term erroneously, 18 did not look at the predictive uncertainty of an ML model, and 8 papers did not concern a relevant biosignal. Another three papers looked at different biosignals, but were kept due to their interesting application of uncertainty quantification. The number of included and not included papers as well as their exclusion criteria are visualised in Figure 2. The search covers studies before 2024.

Figure 3 shows an overview of the results from this search. It shows that from 2018 to 2023 there has been an increase in the use of Uncertainty Quantification. This shows a growing interest in applying Uncertainty Quantification to the biosignal domain.

1.2. Fundamentals of Predictive Uncertainty

Before going into the specific Machine Learning models that can quantify uncertainty for a given prediction, it

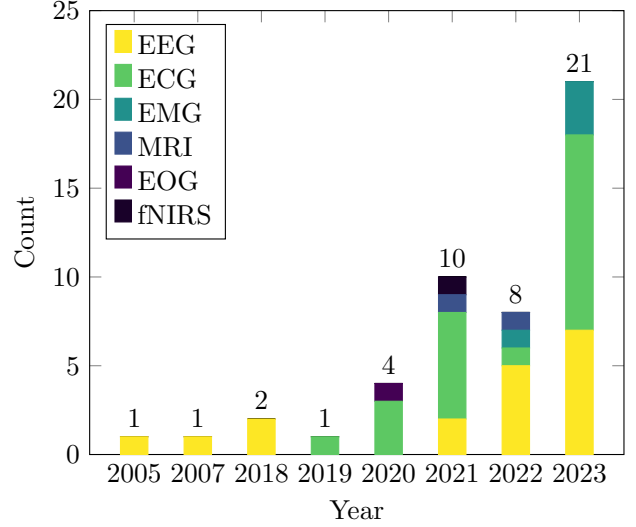


Figure 3: Histogram of the number of papers per year using Uncertainty Quantification for each Biosignal. Overall this shows an increase in the number of Biosignal papers using Uncertainty Quantification. This shows an increase in the popularity of Uncertainty Quantification methods.

is important to first understand what uncertainty really entails. Hüllermeier *et al.* [8] explains how predictive uncertainty can arise from two conceptual sources: aleatoric uncertainty and epistemic uncertainty. In the biosignal literature various definitions are used, some of which are incomplete. We give a thorough and exact definition of both and add clarifications.

Aleatoric uncertainty¹ is the uncertainty that comes from stochasticity in the true function $f : X \rightarrow y$ from which dataset D is sampled. This means that aleatoric uncertainty arises when the optimal function given infinite samples still does not perfectly predict y .

From this definition follows that aleatoric uncertainty cannot be reduced by having a better model, and that humans also cannot give better predictions. Even with arbitrarily many training samples, the aleatoric uncertainty will not decrease. Aleatoric uncertainty is commonly simplified to either label noise (such as imperfect annotations) or sensor noise in the inputs. Artifacts that destroy the underlying signal such as disconnected leads or signal clipping cause aleatoric uncertainty at the inputs.

Epistemic uncertainty² (also known as model uncertainty) is the uncertainty that comes from not knowing the true function $f : X \rightarrow y$. The learned model f^θ may not match the true model due to model misspecification, limited approximation quality, or limited training samples.

Under epistemic uncertainty a better model or a better human expert would be able to make a more accurate prediction. Epistemic uncertainty may arise when a model

¹ Aleatoric is derived from the Latin word "alea", meaning "dice" or "chance".

² Epistemic is derived from the Greek word "episteme", meaning "knowledge".

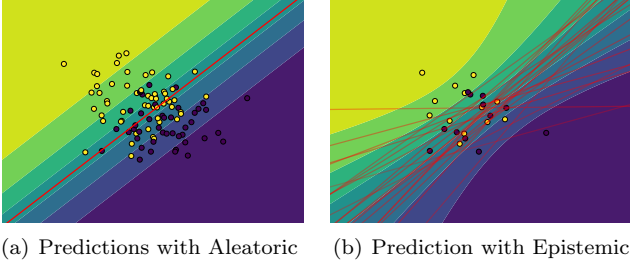


Figure 4: Predictions with aleatoric or epistemic uncertainty in the 2D feature space of a binary classification task. The dots represent training samples, and the background colour the uncertain predictions. In Figure 4(a) aleatoric uncertainty is shown as two classes for which the distribution of the data overlaps, resulting in uncertain predictions in the background. Epistemic uncertainty is shown in Figure 4(b) where due to limited data it is not clear which decision boundary (red) is the true decision boundary.

is applied to data that is different from the data it was trained on, which is referred to as out-of-distribution [9]. Unlike aleatoric uncertainty, epistemic uncertainty does decrease with an increase in training samples.

Artifacts that obscure the signal such as baseline drift or line noise make learning the true function harder, but not impossible. Therefore, these are sources of epistemic uncertainty. The second cause of epistemic uncertainty for biosignals is insufficient (diverse) training samples. If a classifier is to be applied on different people, different hardware, or in different contexts this introduces generalisation error, which is caused by epistemic uncertainty.

The distinction between aleatoric and epistemic uncertainty is made clear in Figure 4, which shows how aleatoric and epistemic uncertainty arise in classification. In this case we see that in the area of feature space where both classes occur, aleatoric uncertainty arises. Epistemic uncertainty arises as the model cannot perfectly learn the distribution of the classes in feature space.

Van Gorp *et al.* [10] emphasises the need for this distinction in sleep stage classification, although this need also applies to other areas. They explain how aleatoric uncertainty should be addressed differently than epistemic uncertainty. If there is high aleatoric uncertainty for a given ECG sample, theoretically there would be no use in having a clinician review the same ECG for a second opinion as they would not be able to give a better prediction. Instead you should consider getting another recording, or collecting additional information. Larsen *et al.* [11] for example proposes to run SPECT-MPI tests only when an ECG classifier is uncertain to create a multi-stage classifier. In practice, because aleatoric uncertainty is estimated by an imperfect model it is very possible that a clinician would be able to make a better prediction.

For epistemic uncertainty more (relevant) training data, better models, or having the samples interpreted by a clinician can improve the quality of a diagnosis.

1.2.1. Limitations of Aleatoric and Epistemic Uncertainty

In Section 2 we will discuss how aleatoric and epistemic uncertainty can present differently in some ML methods, and discuss methods that claim to be able to separate them. However, we first want to highlight the limitations of estimating aleatoric and epistemic uncertainty.

The primary limitation is that we currently cannot adequately quantify aleatoric and epistemic uncertainty separately in classification. In later sections we will introduce methods for quantifying aleatoric and epistemic uncertainty, but theoretical arguments [12], observations [13] and experimental demonstrations [14] have shown that there are interactions between aleatoric uncertainty and epistemic uncertainty in classification. Mucsányi *et al.* [13] has shown that predictions of aleatoric and epistemic uncertainty are highly rank correlated, Wimmer *et al.* [12] has shown that under high aleatoric uncertainty current methods will not be able to predict epistemic uncertainty, and de Jong *et al.* [14] shows that this problem extends to multiple datasets, UQ methods, and uncertainty measures. While estimates of aleatoric and epistemic uncertainty may be useful, with the current method we cannot trust that a prediction of a certain kind of uncertainty is truly attributable to that specific uncertainty for classification. This makes the idea of different actions for different kinds of uncertainty as proposed in [10] infeasible with the current methods.

Additionally, specifically in biosignal applications we should explicitly consider the role of preprocessing. Using fewer features or more aggressive filtering trades epistemic uncertainty for aleatoric uncertainty from the model’s perspective. We therefore need to be aware and explicit in what we define as our learning task for which disentangled uncertainty is estimated.

Since the aleatoric-epistemic perspective is only a perspective on uncertainty there are other ways to look at uncertainty. Some of these alternatives fit into the aleatoric-epistemic framework, but others do not. For example, in Section 2.6 we discuss Prior Networks, where the epistemic uncertainty is split into *model uncertainty* and *distributional uncertainty*. Meanwhile Bishop *et al.* [15] makes a distinction between *discriminative* and *generative* models, where the former learns a decision boundary between the classes, and the latter learns the class likelihood in feature space. Under these generative models samples with low likelihood for either class may be considered uncertain. However, this does not intuitively fit into either aleatoric or epistemic uncertainty.

1.2.2. Uncertainty in Terms of Evidence

One alternative perspective on uncertainty is discussed in the literature. Lin *et al.* [16], distinguishes between uncertainty from *vacuity* and from *dissonance*. This comes from the domain of Subjective Logic [17]. Here, vacuity is the absence of evidence for a prediction. Dissonance arises from conflicting evidence. Lin *et al.* [16] describes these in a context of evidence-based Machine Learning.

Similar to the aleatoric and epistemic uncertainty one can use this distinction to make decisions on how to improve the quality of a model.

This perspective is much less explored in the biosignal literature but warrants further research as it may be more suitable for interpretation by clinicians than the aleatoric-epistemic perspective.

2. Methods for Uncertainty Quantification

As most of the development of Uncertainty Quantification methods happens in the field of Computer Vision [2], it is no surprise that the Machine Learning models for which Uncertainty Quantification is defined are models that perform well in Computer Vision. As a result we find most works build on Neural Networks. Specifically, this review found many Convolutional and Recurrent Neural Networks. An overview of the type of different Neural Network types is given in Figure 5.

With the vast majority of models being Neural Networks, the Uncertainty Quantification methods are also mostly intended for Neural Networks. An overview of the most common methods covered is given in Table 1. This gives a quick reference of the most important properties, but how the methods work and how they specifically relate to biosignals is discussed below. At the of this section Table 1 gives a complete list of each method and the reviewed papers that use them.

This section mostly discusses Neural Networks methods for Uncertainty Quantification, as this is most extensively studied. First, the concept of Bayesian Neural Networks is explained, including the range of different implementations. Bayesian Neural Networks are the current standard for Uncertainty Quantification, and they lend themselves well to interpretation through the lens of aleatoric and epistemic uncertainty. Next, we will discuss some other common Uncertainty Quantification methods such as Variational Autoencoders [18], Evidential Deep Learning [19] and Gaussian Process Regression [20]. We also discuss post-hoc uncertainty calibration methods [21], and end this section with a list of the less established and novel methods for uncertainty quantification that have been used for biosignals. Altogether, this section gives a complete overview of the Uncertainty Quantification methods that are used in the biosignal application domain.

2.1. Notation for Softmax Uncertainty

Standard Neural Networks give point-estimate predictions for a given sample. In regression, this prediction is a scalar with no indication of uncertainty or expected error. However, in classification with standard Neural Networks a Softmax activation function is often used such that the prediction is given as

$$p(y=c|x, \theta) = \frac{\exp(f_c^\theta(x))}{\sum_c \exp(f_c^\theta(x))}. \quad (1)$$

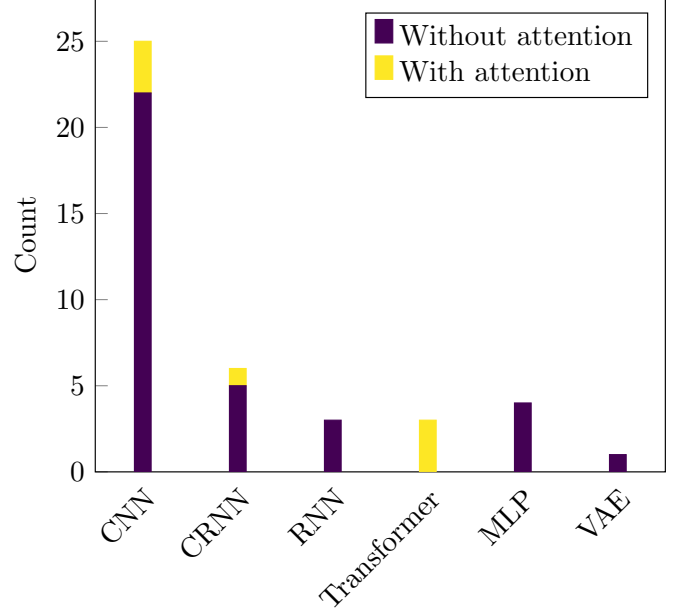


Figure 5: Popularity of various Neural Network architectures in this review. Models with at least one convolutional or recurrent layer are respectively labeled CNN or RNN. Models with both are labeled as CRNN. Yellow indicates models with attention layers.

Where f_c^θ predicts the logits for a given input x , as parameterized by θ . To ease notation we introduce the predicted probability of a class c as

$$p_c := p(y=c|x, D), \quad (2)$$

which in the case of a standard Neural Network with parameters θ learned on dataset $D = \{\mathbf{X}, \mathbf{y}\}$ is approximated with $p_c = p(y=c|x, \theta)$.

Before going into how uncertainty is modelled in Bayesian Neural Networks, it is important to be aware that predicting class probabilities, rather than directly predicting a class label already quantifies uncertainty. However, it only quantifies aleatoric uncertainty and neglects epistemic uncertainty, making it overconfident under epistemic uncertainty.

We found a common misconception in the literature that normal Neural Networks cannot estimate predictive uncertainty. Using the predicted class probabilities they can, but possibly not very well. Therefore, applications of Bayesian Neural Networks for estimating uncertainty should consider an equivalent normal Neural Network as a baseline to justify the added complexity and computational cost.

2.2. Heteroscedasticity in Classification and Regression

The above formulation for Softmax gives simple estimates for aleatoric uncertainty in the standard approach for classification with Neural Networks. Such class probabilities are standard in classification tasks, but not in regression. Standard regression models will only predict the best value, and not give any indication of uncertainty. In

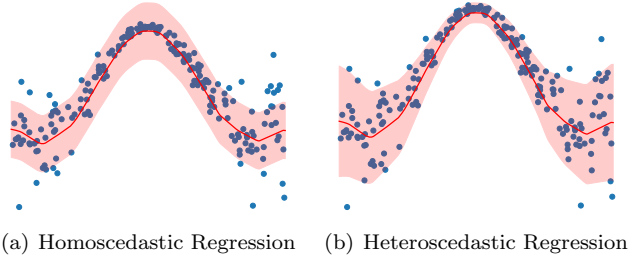


Figure 6: Modelling homoscedastic and heteroscedastic uncertainty in regression. The dots indicate the training samples, the red line the regression prediction, and the red shaded areas the 95% confidence interval based on either Mean Absolute Error (homoscedastic) or predicted absolute error (heteroscedastic). In both cases the data has heteroscedastic noise, but the predictions change on whether the models assume homoscedasticity or heteroscedasticity.

those models, uncertainty can still be derived from performance metrics like the Mean Squared Error. This assumes *homoscedastic uncertainty*, where the risk of error is uniform throughout the feature space.

To be able to distinguish between more and less difficult samples we need to consider *heteroscedastic uncertainty*. In regression this can be done by having a second prediction that estimates the variance as described in Section 2.5, but also by estimating a lower and upper bound [22], or estimating intervals without assuming any distribution [23]. Alternatively, heteroscedastic uncertainty may be estimated with Quantile Regression methods [24], [25], where regression lines are learned for higher and lower quantiles. Figure 6 shows the difference between homoscedastic and heteroscedastic uncertainty estimation in regression. It shows that if some parts have more or less noise in the output, then homoscedastic uncertainty estimation averages these out, whereas heteroscedastic uncertainty estimation maintains the distinction

In Figure 7 we show a classification problem with heteroscedastic uncertainty. The white dots represent one class, and the black dots another. At the cluster on the left these are well separated, with low uncertainty, but at the cluster on the right these overlap with high uncertainty. A simple multi-layer perceptron with Softmax shows increased uncertainty (bright background) where the class distributions overlap.

2.3. Bayesian Neural Networks

Given a starting point of aleatoric uncertainty with softmax, we move towards quantifying epistemic uncertainty with Bayesian Neural Networks. The foundational difference is the way both methods look at learning the parameters. In the standard Neural Network the parameters θ are learned from the space of all possible sets of parameters Θ to minimize a loss function $\mathcal{L}(\theta, D)$. The loss function primarily measures the error between the predictions and the annotated ground truth. Under Bayesian Neural Networks, instead of considering a single optimized set of parameters θ , we consider a distribution of all possible

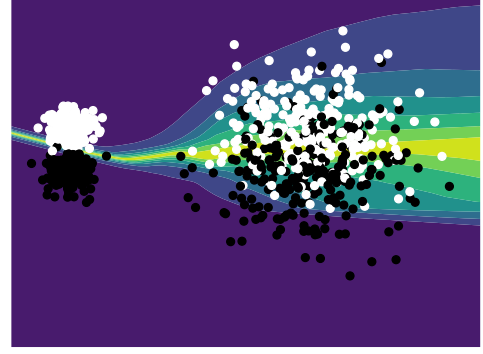


Figure 7: Heteroscedastic uncertainty in classification. The white and black dots represent samples from different classes. The bright areas in the background represent the predicted heteroscedastic uncertainty from softmax.

sets of parameters in Θ . Since some parameters are more likely under dataset D than others, we also consider the likelihood of each set of parameters. This results in the integral

$$p_c = \int \underbrace{p(y=c|x, \theta)}_{\text{Aleatoric}} \underbrace{p(\theta|D)}_{\text{Epistemic}} d\theta. \quad (3)$$

From this the epistemic uncertainty as the probability distribution of the parameter vector $p(\theta|D)$ also becomes apparent.

Some approximations of Bayesian Neural Networks such as MC-Dropout [26] and Deep Ensembles [27] are based on this equation. They sample multiple parameter vectors θ which are all trained to maximise $p(\theta|D)$ through e.g. the negative log-likelihood. From each parameter vector predictions are made. The disagreement between these predictions now captures epistemic uncertainty.

To complete the picture of the Bayesian Neural Network, we take the dataset D as Random Variables $\{X, Y\}$ and deconstruct the posterior $p(\theta|D)$ with Bayes theorem as

$$p(\theta|X, Y) = \frac{p(Y|X, \theta)p(\theta)}{p(Y|X)}. \quad (4)$$

The evidence term $p(Y|X)$ is intractable³. Fortunately, it is constant for a given dataset, so we can optimize θ only on the likelihood and the prior. The likelihood is determined by the model fit to the data and may be computed through a loss function. The prior $p(\theta)$ can be selected to match assumptions about the modelling task.

The rest of this section explains different ways in which Bayesian Neural Networks are approximated to be computationally feasible. For each method we will provide a conceptual understanding, and show the specific limitations in how they might affect biosignal applications.

³This would result in an integral for each parameter of the Neural Network such that $p(Y|X) = \int_{\theta_1} \int_{\theta_2} \dots \int_{\theta_D} p(Y|X, \theta) d\theta_1 d\theta_2 \dots d\theta_D$ where D represents the number of dimensions of the parameter vector θ .

2.3.1. MC-Dropout

Dropout [28] has been a prominent regularization method in Deep Learning applications. During training with dropout, some nodes have a probability p to be dropped (i.e. activation set to 0). This adds noise to the training procedure and has been thoroughly shown to be an effective regularizer.

Normally, the dropout is removed during inference to prevent dropping important information. MC-Dropout (Monte Carlo Dropout) [26] keeps this dropout during inference, and uses multiple predictions to make sure all important information is sampled.

Dropout can be considered as a special probability distribution over parameter vectors, because dropping a node is equivalent to setting all the incoming or outgoing weights of that node to zero. With this we can think about the sampling of MC-Dropout as sampling from an unusual probability distribution over weights. Due to the training process, each of these samples is optimized to be as-likely-as-possible. When we then make predictions with MC-Dropout, it is approximately sampling from $p(\theta|D)$. The predictions that they make are samples from the predictive distribution in Equation 3.

A commonly considered advantage of MC-Dropout is the simplicity with which it can be applied to a Deep Learning model. Many Deep Learning architectures are already trained with dropout, so MC-Dropout can easily be applied without even re-training the model. The big disadvantage however is that it takes many forward passes⁴ for the MC-Dropout to capture the predictive distribution, making inference computationally expensive.

MC-Dropout is therefore easy to apply to architectures that have dropout that have been proven effective, but at the cost of added inference cost. For ECG or EEG monitoring the 100-fold increase in inference cost can be prohibitive.

2.3.2. Deep Ensembles

Although there are many ways to do ensembling in Machine Learning, the idea of a Deep Ensemble as an approximation of a Bayesian Neural Network takes the form of several independently trained Neural Networks following the same architecture and trained on the same data [27]⁵.

A Deep Ensemble may be interpreted as a small set of samples of the parameter distribution $p(\theta|D)$ [33]. Each of these samples is trained to the data, so each sample should reflect a parameter vector with high posterior probability.

Remarkably, with only a limited number of models⁶ we can achieve an acceptable approximation of the weight dis-

tribution $p(\theta|D)$. This keeps the inference cost cheap compared to MC-Dropout, but performing the training several times and storing several models in memory may be expensive. Particularly for applications with model personalization such as in EEG-based BCIs the additional training time can be prohibitive [34].

Much like MC-Dropout, ensembles are conceptually simple, and intuitive to reason about. It aligns with human analogies where when all the models/people disagree, then there is a lot of (epistemic) uncertainty. Contrastingly, situations where all models/people agree must be very certain.

Xia *et al.* [29] shows that ensembles represent epistemic uncertainty under distributional shifts better than MC-Dropout, and that the accuracy of the predictions is also better. They do this on various Biosignal classification tasks such as auditory COVID-19 classification, respiratory abnormality detection and heart arrhythmia detection. By providing various forms of dataset shift, they concur with findings from computer vision and language models [35], suggesting that Deep Ensembles may be better at presenting epistemic uncertainty under dataset shifts. In Computer Vision, Deep Ensembles are considered to have state-of-the-art performance for a wide range of Uncertainty Quantification tasks [13], and from the results of Xia *et al.* [29] it is reasonable to expect that this extends to biosignal applications.

For biosignal applications that do not use Deep Learning alternative ensembling strategies are needed to ensure diversity. Larsen *et al.* [11] uses pseudo-bootstrapped [36] ensembles of a logistic regression classifier. In this strategy each model is trained on a subset of the training data to maintain a spread of plausible models. Pseudo-bootstrapping is a viable alternative to achieve ensembling for biosignal applications that use linear classifiers.

2.3.3. Variational Inference

In variational inference (VI) the intractable posterior distribution $p(\theta|X, Y)$ is approximated with a simpler distribution $q_\omega(\theta)$. A possible approximation through $q_\omega(\theta)$ might say that each weight is a Gaussian distribution with a mean and a variance. The goal is then to optimize the parameters ω for the high-dimensional Gaussian, so that it is similar to the true posterior. With this, we can then sample models from $q_\omega(\theta)$ to predict class probabilities according to Equation 3.

In order to make a good approximation of the posterior, VI needs to minimize the Kullback-Leibler divergence (KL-divergence) between the approximate distribution $q_\omega(\theta)$ and the true distribution $p(\theta|X, Y)$. The KL-divergence measures the distance between two distributions. In this case it is given as

$$KL(q_\omega(\theta) || p(\theta|X, Y)) = \int_{\Theta} q_\omega(\theta) \log \frac{q_\omega(\theta)}{p(\theta|X, Y)} d\theta. \quad (5)$$

This minimization task still contains the posterior distribution term $p(\theta|X, Y)$ which is intractable as discussed in

⁴ $T = 50$ is recommended, but anywhere from $T = 10$ to $T = 1001$ may be used. [26], [29], [30]

⁵Originally Deep Ensembles were introduced as a non-Bayesian method for UQ [27], but it has since been shown that it can be considered as a very coarse approximation of a BNN [31], [32].

⁶As an example: Lakshminarayanan *et al.* [27] uses an ensemble of 5 models.

Equation 4. By rearranging the KL-divergence into the evidence lower bound (ELBO) we instead get the maximization task [2]:

$$\text{ELBO}(\omega) := \int_{\Theta} q_{\omega}(\theta) \log p(Y|X, \theta) d\theta - KL(q_{\omega}(\theta) || p(\theta)) \quad (6)$$

The prior chosen for $p(\theta)$ may still be defined by the modeller, and can have an impact on the quality of the model. For the purposes of transfer learning, this prior may even be a learned distribution on another dataset (see [37]).

While Variational Inference is a better approximation of a Bayesian Neural Network than Ensemble-based methods, it is often much more expensive to train and do inference on. Moreover, implementing it introduces many new decisions to make. The form of the posterior approximation needs to be chosen, as well as the prior for its parameters. Moreover, measuring the evidence lower bound requires Monte-Carlo sampling from the approximated posterior. The number of samples to use is a balance between computational cost per epoch, and the stochasticity of the gradient descent.

Having many Bayesian layers in a Deep Bayesian Network can cause the loss to become numerically unstable. This instability has made Variational Inference less popular in Computer Vision as they use very large models, but it is not such a big problem for biosignal applications due to the smaller models.

2.4. Variational Autoencoders

A Variational Autoencoder [18] is a specific type of neural network architecture. It has an encoder which receives a high-dimensional input x and encodes it into a lower dimensional latent distribution $p(z|x, \theta)$. It does so by predicting a mean and a variance for each dimension of the latent distribution, from which latent representations $z \sim p(z|x)$ can be sampled. A decoder network then reconstructs the encoding back into the original dimensionality of the input to achieve $x' = f_{\theta'}(z \sim p(z|x, \theta)) \approx x$.

The VAE model is trained to minimize the difference between the input x and the reconstructed output x' . As a result, the latent distribution $p(z|x, \theta')$ should be a lower-level representation of the salient features that exist in the data. This works under the concept of *manifold learning* where many of the points on the high-dimensional input have near-zero likelihood, and that actually a lower-dimensional manifold should be able to capture the distribution of the actual data.

VAEs were originally intended as generative unsupervised learning models, and were not invented with Uncertainty Quantification in mind. However, because the latent representation is a distribution which can be sampled from, researchers have constructed various methods to extract uncertainty from that stochasticity. Belen *et al.* [38] uses a trained VAE on a dataset of segments of ECG with and without expert annotated atrial fibrillation. They then use the sampled latent representations as

input for a multi-layer-perceptron to do the classification task as

$$p(y = c | p(z|x, \theta), \theta'). \quad (7)$$

This results in a distribution of probabilities, of which the variance is used to measure aleatoric uncertainty.

Van De Leur *et al.* [39] apply Principal Component Analysis to get a 2-dimensional visualization of the latent space as a method for interpretability for ECG classification. They show how various diagnoses would show in the latent representation, so that a sample on the boundary of two classes, or far away from any known classes can be qualitatively assessed as uncertain. This shows unique opportunities for using VAEs for uncertainty.

The primary downside to using VAEs for Biosignal analysis is that it imposes specific architecture constraints. A lot of the biosignal literature relies on established architectures that are known to perform well in similar tasks, but those cannot easily be turned into VAEs. Additionally, they are not as extensively studied as Bayesian Neural Networks and their uncertainty quantification performance and weaknesses are not well established.

2.5. Heteroscedastic Uncertainty Quantification

In contrast to the previous methods which rely on stochasticity to quantify uncertainty, there is also a set of methods that aim to directly predict uncertainty as part of the model training task. The most intuitive form of this is heteroscedastic uncertainty quantification for regression [40]. In these models, the Neural Network not only attempts to learn a predicted regression value, but it has a separate output for the predicted error. This results in a prediction, paired with a measure of aleatoric uncertainty. Taking $\mu_{\theta}(x)$ as the predicted mean and $\sigma_{\theta}^2(x)$ as the predicted variance for a sample, the predicted value \hat{y} is given as

$$\hat{y} = \mathcal{N}(\mu_{\theta}(x), \sigma_{\theta}^2(x)). \quad (8)$$

Such a model is then trained with a loss function that optimizes both the predicted mean and the variance. The Gaussian Negative Log-Likelihood loss

$$\mathcal{L}_{NLL}(y_{true}, x) = \frac{\log \sigma_{\theta}^2(x)}{2} + \frac{(\mu_{\theta}(x) - y_{true})^2}{2\sigma_{\theta}^2(x)} \quad (9)$$

is the simplest, but alternatives have been proposed [40].

Vranken *et al.* [41] and Jin *et al.* [42] combine this concept with Bayesian Neural Networks to get separate predictions of aleatoric and epistemic uncertainty for ECG and EEG classification. This approach is not used often in the biosignal literature, but it has been shown that for some datasets it can give better out-of-distribution detection performance [14].

2.6. Evidential Deep Learning

Evidential Deep Learning [19] offers a computationally affordable alternative to Bayesian Neural Networks where the distribution of the predictions is captured by

a c dimensional Dirichlet distribution parameterised by $\alpha_c \in [0, 1]$, which are predicted by a Neural Network. This setup therefore predicts a distribution of probabilities in a single forward pass.

Sensoy *et al.* [19] proposed EDL to look at uncertainty from the perspective of the Dempster-Shafer Theory of Evidence (DST) instead of the aleatoric-epistemic approach. In this approach the parameter α_c gives the amount of *evidence* for that class.

The uncertainty can then be defined into *vacuity* and *dissonance* [16], [43]. Vacuity is the absence of evidence causing uncertainty. Like standard Neural Networks with a Softmax activation function, Evidential Machine Learning assumes that exactly one class must be the ground truth. The absence of evidence for any of the classes would then result in a form of uncertainty referred to as vacuity. The opposite uncertainty is *dissonance*, which occurs when the model has found evidence for multiple classes, which is not in line with the assumption of mutual exclusivity.

Prior Networks Malinin *et al.* [44] are another approach form of Evidential Deep Learning. It uses the same setup of predicting a Dirichlet distribution but interprets it as an alteration of the aleatoric-epistemic uncertainty. Under the Bayesian Neural Network framework we consider the uncertainty due to generalization error, such as when the model is evaluated under out-of-distribution data, as part of the epistemic uncertainty. Prior networks add the term *distributional uncertainty* to Equation 3. This then gives

$$p_c = \int \int \underbrace{p(y=c|\mu)}_{\text{aleatoric}} \underbrace{p(\mu|x, \theta)}_{\text{distributional}} \underbrace{p(\theta|D)}_{\text{epistemic}} d\mu d\theta. \quad (10)$$

Evidential Deep Learning has shown good performance on out-of-distribution detection tasks, but has theoretical and practical limitations when representing epistemic uncertainty [45]. The reviewed literature generally does not compare EDL methods with BNN methods for biosignal applications. A thorough investigation of uncertainty quantification should consider both top-performing methods for BNNs and EDL methods for biosignal tasks. From the current literature, it can only be established that EDL methods give better uncertainty estimates than standard Neural Networks [16], [46] on EMG grasp classification and that its uncertainty indeed goes up with noise for myocardial infarction detection under noisy ECG [47].

2.7. Gaussian Process Regression

Gaussian Process Regression [48] is a non-parametric regression method that considers epistemic uncertainty. It assumes a Gaussian prior over the dependent variable Y . It also assumes that the samples in the training data D are drawn without measurement error. This leaves uncertainty in the regression between and outside training samples, and gives more certainty at points close to the training samples.

As more training samples are collected, the epistemic uncertainty will decrease. The assumption that data are

drawn without measurement error does naturally lead to an inability to capture aleatoric uncertainty.

Gaussian Process Regression is suitable for biosignal applications due to the smaller datasets, but because most tasks are classification tasks it does not see a lot of use. Current implementations on EMG [49] and ECG [20] demonstrate it as an effective method in combination with physics-informed simulation models.

2.8. Post-hoc Calibration

Post-hoc calibration methods [21] look at uncertainty only in terms of the predicted probability for each class, and addresses how this may deviate from the observed probability. A class prediction with $p = 0.75$ should be correct 75% of the time, but this does not hold for standard softmax classification. Post-hoc calibration methods aim for an optimal calibration such that

$$p(y=c|p_c) = p_c. \quad (11)$$

Various methods for post-hoc probability calibration methods exist [21]. Temperature Scaling is the simplest method of post-hoc calibration, which determines the *softness* of the Softmax function. It does so by introducing a hyperparameter τ to get the scaled Softmax function

$$p(y=c|x, \theta) = \frac{\exp(\frac{f_c^\theta(x)}{\tau})}{\sum_c' \exp(\frac{f_{c'}^\theta(x)}{\tau})}. \quad (12)$$

Post-hoc calibration methods cannot provide better separation between correct and incorrect predictions, and do not account for epistemic uncertainty. They only ensure that the probabilities are appropriately scaled, which is important when those probabilities need to be interpreted by a clinician [52].

2.9. Non-standard UQ Methods

Above, a selection of common and well-studied methods for Uncertainty Quantification is discussed. This does not cover all the UQ methods that were encountered in the review. Below we continue the description of uncertainty quantification methods with some non-standard methods encountered in the reviewed literature to provide an exhaustive presentation of UQ research on biosignals.

Biosignal classification often uses smaller models and smaller datasets than computer vision, which makes it suitable for unique Uncertainty Quantification methods that are not standard in other domains. We critically assess these methods below.

2.9.1. Bayesian Model Averaging with Reversible-Jump MCMC

Schetinin *et al.* [53] attempted to classify EEG artifacts using a method based on Bayesian Model Averaging. They use Markov-Chain Monte Carlo to sample changes to a decision tree. These changes are any of 4 types: adding

Table 1: A simplified overview of the different UQ methods discussed in 2. Computational cost of UQ methods is qualitatively grouped into 3 classes. *None* has negligible added computational cost. *Small* has some added computational cost e.g. due to slower convergence or training steps being more computationally expensive. *Large* indicates substantial increase in computational cost, such as 5 times the training cost, or 50 times the inference cost.

Method	Model Agnostic	Epistemic UQ	Aleatoric UQ**	Training Cost	Inference Cost
MC-Dropout [26]	NN only	✓		None	Large
Ensembles [27]	✓*	✓		Large	Small
Variational Inference [50]	NN only	✓		Large	Large
Variational Autoencoder [18]			✓	Small	Large
Evidential Machine Learning [19], [44]	✓	✓***	✓	None	None
Gaussian Process Regression [51]		✓	✓	Small	Small
Post-hoc calibration [21]	✓		✓	None	None

*Requires bootstrapping [36] for non-stochastic training procedures. May perform poorly without local minima.

**Aleatoric uncertainty may still show in classification with Softmax.

***Not faithful epistemic uncertainty [45].

a split in the tree, removing a split in the tree, changing the variable a split is focused on, or changing the rule of a split. These changes are accepted or rejected based on the likelihood given the data. This consists of how well a given change improves the training classification, as well as how likely it is given a set prior.

As a measure of uncertainty the authors consider the entropy in the leaf nodes. The authors showed that subtracting a non-stationary component from the power of the subdelta band improved the accuracy of their model, but since the dataset is not specified and no other models are shown it is not possible to assess the quality of the model, nor the resulting entropies.

Another reviewed work also used the entropy of the leaf-nodes in a decision tree as a measure of uncertainty, but this also lacked interpretation [54].

2.9.2. Majorization-Minimization and Hierarchical Bayesian Modelling

Bekhti *et al.* [55] compares Majorization-minimization and Hierarchical Bayesian Modelling and shows how they are fundamentally the same. Unlike the majority of works found in this review which try to learn an arbitrary function $p(y|x, D) = f_\theta(x)$, this work starts with the assumption that observed EEG recordings X are a linear combination of underlying sources G connected through a known linear forward propagation matrix G , with some Gaussian noise E such that $M = GX + E$. This results in a multi-task regression where we need to learn an optimal matrix X that minimizes the E . Without considering regularization this results in the optimization

$$\hat{X} = \arg \min_X \frac{1}{2} \|M - GX\|_F^2. \quad (13)$$

Majorization-minimization solves this by taking a random initialization, fitting a Taylor expansion to the cost function at that point, and then using the X^t that minimizes

that Taylor expansion as the next initialization. To avoid overfitting $l_{2,p}$ -norm regularization is used. This has the added benefit of promoting sparse solutions.

They are able to show that the full maximum a posteriori estimate of a Hierarchical Bayesian Modelling approach can be re-derived as a Majorization-Minimization optimization problem. From this insight, the authors propose a method of sampling multiple initialization for the MM optimization, resulting in multiple sparse solutions to the inverse problem.

Using the multiple sparse solutions, together with how well they minimize the objective function, the authors are able to present various source attributions to an observed EEG or MEG signal, together with a measure of how (un)certain each solution is. This gives a more complete insight into the source of a given signal.

2.9.3. Bayesian Moderated Outputs

Based on Mackay [56], Mohamed *et al.* [57] compare Bayesian Moderated Outputs to a standard Multi-Layer Perceptron for the task of epileptic activity classification in sleep EEG recordings. The concept of Bayesian Moderated Outputs is that instead of having a single optimal parameter vector $\hat{\theta}$, a more robust method will have a Gaussian distribution of parameters around an optimum $\Theta = \mathcal{N}(\hat{\theta}, s^2)$. The hypothesis is that the mean prediction over these different models provides a better representation of the predicted probability.

Unfortunately, this did not lead to apparent better performance than a maximum-likelihood trained Multi-Layer Perceptron [57]. This was observed by using a rejection threshold of 0.9 for both models. The Bayesian Moderated Outputs did achieve slightly higher accuracy (up to 1 percent-point), but at the cost of rejecting up to 15 percent-point more samples from classification.

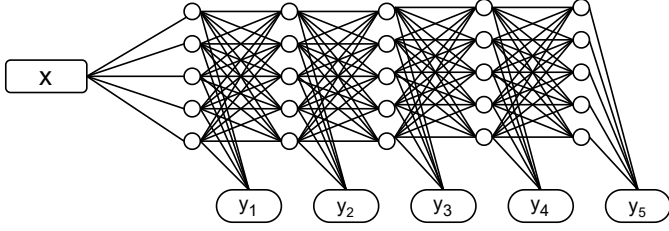


Figure 8: A diagram showing the concept of Early Exit Ensembles [60]. There is a shared backbone network, from which Exit Branches make predictions. Each exit branch makes an independent prediction. The distribution of predictions may be interpreted similar to a normal ensemble.

2.9.4. Neural Stochastic Differential Equations

Wabina *et al.* [58] propose a novel method called Neural Stochastic Differential Equations to learn an electrical conductivity model of the head based on MRI. Such conductivity models can be used to inform the forward propagation of EEG signals as referred to in Section 2.9.2.

They use a class of Deep Neural Networks proposed in [59], which includes a split block consisting of a drift and a diffusion network to consider the Neural Network as a Stochastic Differential Equation. The drift network continues to attempt to optimize predictions, while the diffusion network predicts a heteroscedastic amount of Gaussian noise. The noise should be minimal for samples in the training distribution, and maximal for out-of-distribution samples. The result of the SDE-block can be sampled and passed through a final block of dense layers to reach a distribution of regression predictions. The complete Neural Network proposed is called SDE-Net.

An experiment on the Single Individual volunteer for Multiple Observations across Networks (SIMON) MRI dataset showed that SDE-Net outperformed Bayesian methods. However, the effect of epistemic uncertainty on the spread of the predictions and SDE-Net’s ability to capture epistemic uncertainty is not investigated, so the results may be explained by better estimation of aleatoric uncertainty alone.

2.9.5. Early Exit Ensembles

As a quasi-ensembling method Campbell *et al.* [60] propose Early Exit Ensembles. Early exit ensembles work by taking any deep neural network and adding various *exit* branches to points of the network as illustrated in Figure 8. Each exit will have a global pooling operation and 2 dense layers. The idea is that each exit branch will try to learn to do the classification task (as an ensemble), but depending on the location on the *backbone* architecture they may learn on either lower or higher level features.

Like normal ensembling methods, the disagreement between the various classifiers corresponds to epistemic uncertainty. The advantage compared to normal ensembling is that the large amount of weight sharing can reduce the computational cost of training and inference, as well as the size of the model. The ways in

which constructing an Early Exit Ensemble from an existing architecture affects the quality of the predicted uncertainty is an interesting avenue for research, which may be partly inspired by what is already known about early-exit neural networks (see [61], [62]).

The quality of uncertainty estimates of Early Exit Ensembles is still unknown, but it is a promising avenue for inference and fine-tuning on edge devices [34] and real-time monitoring of ECG.

2.9.6. Reconstruction Error

Martinez *et al.* [63] look at how to reconstruct an ECG signal based on bioimpedance recordings. Bioimpedance can be much easier to record, but also difficult for cardiologists to interpret. They propose a method where an Autoencoder regression model uses the biosignals to construct the ECG morphology, but without correct amplitudes. Then a second autoencoder uses this amplitude-invariant data, and the original bioimpedance to reconstruct the ECG.

Since the amplitude-corrected data should have the same morphology as the predicted ECG, any differences in morphology can be attributed to a generalisation failure of the second autoencoder. Thus, the authors measure the Pearson correlation between the amplitude invariant and amplitude-corrected data as a measure of uncertainty.

They show that this uncertainty indeed correlates with the translation quality, but a thorough comparison with more established UQ methods is still needed.

2.9.7. Fuzzy Logic

The systematic search found three works that rely on methods from Fuzzy Logic. Fuzzy Logic relies on the concept of a Fuzzy Set with a fuzzy membership function. This gives a probabilistic notion of a set where an element can have partial membership to a set or multiple sets.

The fuzzy membership function may be defined in different ways, based on (fuzzy) unsupervised clustering [64], (fuzzy) classification [65] or even as a composite of other fuzzy membership functions [66].

The fuzzy memberships can be interpreted directly as predictions [66], but more complex setups are also possible. Sovatzidi *et al.* [64] uses them to construct a Fuzzy Cognitive Map: A directed probabilistic graph that offers an explainable decision support system [67] for diagnosis. Liu *et al.* [65] instead uses it to combine predictions from multiple modalities using the Dempster-Shafer Theory of Evidence. They show that their proposed Weighted Fuzzy Dempster-Shafer Framework (WFDSF) can fuse predictions from different modalities to achieve better predictive performance than either modality alone.

Fuzzy Logic allows a lot of freedom for the modeller to design probabilistic systems, which is relevant for biosignal analysis where we have limited datasets but do have prior knowledge on how a decision should be made. We find that the proposed works have good reason for their design and

show improved task performance, but a systematic evaluation of predictive uncertainty under such Fuzzy Logic systems is still missing.

2.9.8. Assumed Density Filtering

Duan *et al.* [68] applies a more computationally affordable method for modelling data uncertainty called Assumed Density Filtering (ADF). Whereas Bayesian Neural Networks model a distribution for each weight, ADF takes a single-point solution for the weights, but has a distribution for the activations.

This is achieved by modelling the input as a Gaussian distribution around the single-point input features such that

$$\mathbf{z} = \mathcal{N}(x, \sigma^2). \quad (14)$$

Passing this as the input to a Neural Network results in distributions for each activation. Each activation is modelled by a mean and variance, where the variance corresponds to the uncertainty. This ultimately results in a mean (prediction) and variance (uncertainty) in the output. This method is intended to correspond to aleatoric uncertainty caused by uncertain inputs. For biosignals this may represent sensor noise. Combined with a Bayesian Neural Network as done by Duan *et al.* [68] provides explicit modelling for both uncertainty of the model, and uncertainty of the biosignal recording. In other Uncertainty Quantification literature the input uncertainty is largely ignored [69], [70], but it may be particularly relevant for noisy biosignals.

They demonstrate that this gives better uncertainty estimates for a BCI task than many alternative methods including Deep Ensembles [27], MC-Dropout [26] and Direct Uncertainty Quantification [71], showing that this is a promising direction.

2.9.9. Data Uncertainty Learning

As a method for aleatoric uncertainty, Data Uncertainty Learning [72] models uncertainty as a distribution in an embedding such that

$$p(z|x) = \mathcal{N}(x; \mu, \sigma^2 I). \quad (15)$$

Here a Neural Network learns an embedding as a Gaussian distribution. This method holds similarities to a Variational Autoencoder, as both methods learn a Gaussian distributed representation of the input. However where a VAE normally has structural symmetry between the encoder and decoder, Data Uncertainty Learning has the embedding as the penultimate layer. For Data Uncertainty Learning the decoder is then replaced with a shallow classifier.

Deng *et al.* [73] applied this method to predict seizures from EEG. The uncertainty in the embedding should then capture the uncertainty that is in the EEG recording. Since the uncertainty is modelled in a deep embedding it may represent more nuanced uncertainty in the EEG signal that relates directly to the task.

Although Deng *et al.* [73] do not give a thorough evaluation of the uncertainty, they do show that the modelling of uncertainty improves the classifier as compared to a deterministic equivalent, with minimal additional computational cost. They find that wrong predictions indeed are more likely to have high uncertainty, but a thorough evaluation and comparison with alternative methods is still needed.

2.9.10. Miscellaneous methods

We encountered three more uncertainty quantification methods, but they were sufficiently rare that they do not fit into the presented narrative. The first of these is Adaptive Stochastic Gradient Hamiltonian Monte Carlo [97], which Chetkin *et al.* [98] uses for Motor Imagery classification. This Bayesian Neural Network method does not assume a parametric distribution over each weight, but uses a Markov Chain to converge to the posterior distribution. They found that this worked better than an ensemble when applied to ShallowConvNet [107], but there was no statistically significant difference when applied to EEGNet [108]. Zhang *et al.* [99] uses it for knee torque regression based on EMG and finds it gives comparable prediction and uncertainty estimation to Gaussian Processes. From these results we suggest that this method may be feasible for small Neural Networks common in biosignal applications, but there's no strong evidence of increased performance to warrant the added computational cost of training compared to Deep Ensembles.

To deal with the large amount of data in the Temple University Hospital Seizure Corpus (TUSZ) [109] dataset, De Rooij *et al.* [102] used Kalman Filters to solve the least squares adaption of SVMs. Rather than optimizing the SVM for epilepsy classification against the whole dataset at once, they consider parts of the dataset to continually learn the parameters of the SVM. Since Kalman Filters allow for some uncertainty, this method should capture model uncertainty. However, the authors do not go into detail on how well the uncertainty quantification performs.

Lastly Li *et al.* [106] investigated Trust Scores [105] under dimensionality reduction. Trust Scores assign uncertainty based on disagreement between a proposed model and a kNN-based classifier, where high disagreement indicates high uncertainty. However, they found that for some dimensionality reduction methods the uncertainty was not monotonically increasing with the precision, indicating a potential risk when implementing Trust Scores in a classification pipeline.

2.10. Recommendations for UQ methods

We conclude from the analysis of UQ methods that Deep Ensembles and MC-Dropout are the best established, and that Deep Ensembles may be considered state-of-the-art for estimating epistemic uncertainty. The review found relatively little comparative analysis, specifically we find that comparing a computationally expensive Bayesian Neural

UQ Method	EEG publications	ECG publications	Other biosignal
MC-Dropout [26]	Epilepsy [60], [74], [75], Sleep [76], Motor Imagery BCI [68], [77], Denoising [42]	Emotion [30], Respiration [78], Arrhythmia [29], [41], [52], [79]–[83], Anxiety [84]	EOG: Ataxia [85], MRI: Focal Cortical Dysplasia [86]
(Deep) Ensemble [27]	Motor BCI [68]	Arrhythmia [29], [41], [79], [80], [87], [88], CRT response [11]	
Variational Inference [89]	P300 BCI [90], Motor BCI [91]	Arrhythmia [29], [41], [92],	fNIRS: Motor BCI [93]
Softmax [94]	Sleep [95], Epilepsy [96]	Arrhythmia [29], [96]	
Variational Autoencoders [18]		Arrhythmia [29], [38], [39], Modality Translation [63]	
Evidential Deep Learning [19]		Myocardial Infarction [47]	EMG: Hand movement [16], [43]
Post-hoc calibration [21]	Sleep [76]	Arrhythmia [29], [79]	
Gaussian Process [48]	Motor BCI [68]	Heart Modelling [20]	EMG: Knee torque [49]
Heteroscedastic UQ [40]	Motor BCI [68], Denoising [42]	Arrhythmia [41]	
Early Exit Ensemble [60]	Epilepsy [34], [60]		
Hamiltonian Monte Carlo [97]	Motor BCI [98]		EMG: Knee torque [99]
Fuzzy Sets [67]	Depression [64]	Arrhythmia [66]	
Bayesian Model Averaging [100]	Sleep [53]		
Hierarchical Bayesian Modelling	Inverse Problem [55]		
Entropy in Decision Tree Leafs		Arrhythmia [54]	
Bayesian Moderated Output [56]	Epilepsy [57]		
Direct Uncertainty Learning [72]	Epilepsy [73]		
Kalman Filters [101]	Epilepsy [102]		
Deep SVDD [103]	Epilepsy [75]		
Neural SDE [59]			MRI: Inverse Problem [58]
Assumed Density Filtering [104]	Motor BCI [68]		
DUQ [71]	Motor BCI [68]		
WFDSG [65]	Drowsiness [65]		EOG: Drowsiness [65]
Trust Scores [105]		Arrhythmia [106]	

Table 2: Reviewed papers and their Uncertainty Quantification methods.

Network against a standard single-point Neural Network for uncertainty estimation is necessary for all implementations.

Early Exit Ensembles are not yet well established and require further investigation, but they may prove as a more computationally affordable alternative to Deep Ensembles.

Post-hoc calibration gets little attention in the presented research, but may be valuable for addressing overconfidence and ensuring clinically interpretable predictive probabilities. We encourage future work to combine Bayesian Neural Networks with post-hoc calibration.

3. Uncertainty Measures

Most uncertainty quantification methods (e.g. BNNs, EDL) when applied in classification tasks produce a distribution over class probabilities. However, upon reviewing the biosignal literature we found that papers are inconsistent or non-specific about how to extract scalar measures of uncertainty from them⁷. We critically review the existing *Uncertainty Measures* in relation to theoretical expectations of aleatoric and epistemic uncertainty, and provide strong recommendations on how to extract uncertainty measures from a distribution of predicted probabilities. An overview is given in Table 2.

The theoretical analysis considers whether the measure captures aleatoric uncertainty, epistemic uncertainty or both. It relies on the notion that epistemic uncertainty is represented by disagreement between model predictions.

Figure 9 shows how aleatoric and epistemic uncertainty interact [12], [14]. These plots are generated by taking 3 Gaussian distributions to represent predicted logits. 100.000 samples are taken from these logits and passed through the Softmax function. The closeness to each vertex represents the predicted class probability. This provides an intuition of how aleatoric and epistemic uncer-

tainty may present as predicted class probabilities. It becomes apparent that under high epistemic uncertainty, determining aleatoric uncertainty becomes difficult. The idea that some measures purely represent aleatoric uncertainty and others purely represent epistemic uncertainty is only theoretic.

3.1. Class Probability

The standard method for measuring uncertainty in Neural Networks is the predicted Softmax probability of a classification. An epilepsy classifier that gives the diagnosis of epilepsy with $p = 0.55$ is less certain than if it gives the diagnosis with $p = 0.97$.

This uncertainty measure captures aleatoric uncertainty. However, softmax probabilities are infamously overconfident in single-point neural networks, even when using a proper scoring loss function [21].

When multiple forward passes are made with a BNN the class probability is determined by the average of all forward passes. With T as the number of forward passes and \bar{c} as the max probability class of the average probabilities ($\bar{c} = \arg \max_c T^{-1} \sum_t p_c$) we define the class probability as:

$$\mathbb{P}(p) \equiv T^{-1} \sum_T p_{\bar{c}} \quad (16)$$

Or in a shorthand:

$$\mathbb{P}(p) \equiv \bar{p}_{\bar{c}} \quad (17)$$

For approximations of Bayesian Neural Networks we can assume that the logits increase in variance as the epistemic uncertainty increases. The Softmax function pushes high logits down into a $[0, 1]$ range, while lower logits are shifted less. As such, logits from a distribution with high variance will result in less confident probabilities. Figure 10 visualizes this effect.

This shows that the average class probabilities $\mathbb{P}(p)$ will show more uncertainty under increased epistemic uncertainty. Therefore, it is a measure that combines aleatoric and epistemic uncertainty. This explains why the average probability $\mathbb{P}(p)$ of a BNN is less overconfident than a single-point Neural Network [76], [80].

⁷In regression the literature is more consistent: Variance from either aleatoric or epistemic methods indicate the source of uncertainty, as shown in Jin *et al.* [42]. Whether this separates aleatoric and epistemic uncertainty correctly in practice is still unknown [13].

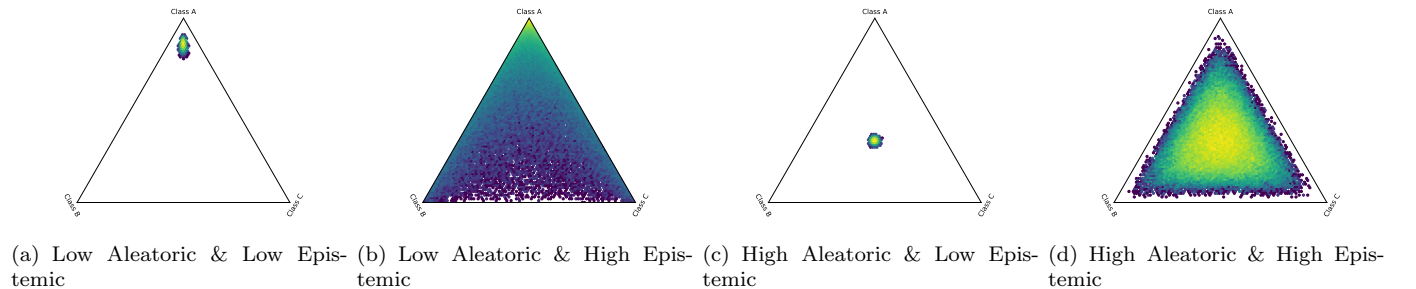


Figure 9: Simplexes presenting different types of uncertainty. Epistemic uncertainty is presented by increased variance in the logits. Aleatoric uncertainty is presented by decreasing the difference between the means of the logits between classes. The points represent softmax probabilities determined by logits following a multivariate Gaussian $\mathcal{N}(\mu, \sigma^2)$. For high aleatoric uncertainty we set $\mu = [10, 10, 10]$, whereas for low we use $\mu = [10, 8, 8]$. For high epistemic uncertainty we set $\sigma^2 = [2, 2, 2]$, whereas for low we use $\sigma^2 = [0.01, 0.01, 0.01]$.

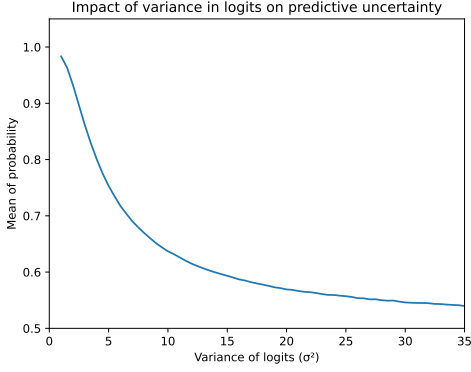


Figure 10: Mean class probability decreases for higher variance in the logits. This illustration assumes binary classification. The logits are distributed as $\mathcal{N}([4, 0], [\sigma^2, \sigma^2])$. Only the first class is shown.

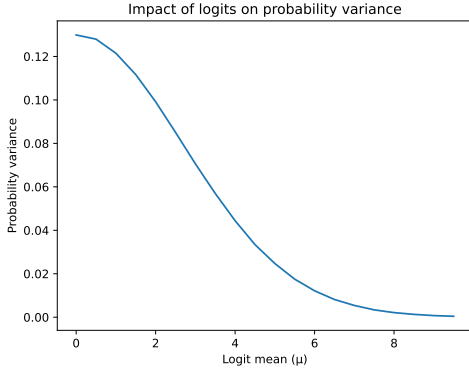


Figure 11: Probability variance decreases as the logit mean increases. This illustration was generated by taking logits as $\mathcal{N}([\mu, 0], [2, 2])$. Here we see that probability variance (used as a measure of epistemic uncertainty) becomes smaller when aleatoric uncertainty decreases.

3.2. Variance

Several papers consider the variance or standard deviations of the class probabilities as a measure of uncertainty [30], [52], [76], [84], [85], [87]. This should represent epistemic uncertainty as it measures disagreement between model samples.

Under multi-class classification it can be unclear which variance should be computed. Some implementations measure the variance over each class and either present all those variances to clinicians [52] or as features to another Machine Learning model [85]. One may also present only the variance of the predicted class as a measure of epistemic uncertainty, or the average variance over multiple classes.

To be specific, this leaves two possible scalar measures for probability variance under multi-class predictions:

$$\mathbb{V}_{\bar{c}}(p) = T^{-1} \sum_t (p_{ct} - \bar{p}_{\bar{c}})^2 \quad (18)$$

$$\mathbb{V}(p) = C^{-1} \sum_c T^{-1} \sum_t (p_{ct} - \bar{p}_c)^2 \quad (19)$$

A similar effect as shown in Figure 10 occurs when applying variance uncertainty measures to the class probabilities. Figure 11 illustrates that the difference in the mean of the logits increases (less aleatoric uncertainty) the variance of the class probabilities decreases. As a result a decrease in aleatoric uncertainty can present as a perceived decrease in epistemic uncertainty. Future works should consider using the variance of the logits as described in [110] to get a more independent measure of epistemic uncertainty.

3.3. Predictive Entropy

Predictive entropy measures the total amount of uncertainty over the probabilities of all classes. This is also a method commonly used for single-point Neural Networks. It is functionally equivalent to class probability for a binary classification task, but for more classes it also considers the amount of uncertainty remaining in the other classes.

Predictive Entropy⁸ is given as:

$$\mathbb{H}_{\text{pred}}(p) = - \sum_c \bar{p}_c \log \bar{p}_c \quad (20)$$

Variations of this include normalizing the entropy by dividing it by $\log(C)$ or taking $1 - \mathbb{H}_{\text{pred}}$ to get a confidence measure instead of an uncertainty measure [95].

Because Predictive Entropy and Class Probability both measure the combination of aleatoric and epistemic uncertainty, they can be expected to have similar behaviour. Predictive Entropy gives a well-supported approach to deal with multi-class classification, but Class Probability is likely to be more interpretable by a clinician.

3.4. Disentangling Entropy

By capturing the total uncertainty, predictive entropy responds to both aleatoric and epistemic uncertainty. I.e. it is high when aleatoric uncertainty is high, or when epistemic uncertainty is high. It may be desirable to disentangle these un.

The mutual information between a model's parameters ω and a new labelled sample $\{x, y\}$ gives the amount of information gained by knowing the label of that sample, relative to what was already known by the model's parameters. Since this may be considered equivalent to epistemic uncertainty [111] we get an intractable epistemic uncertainty measure:

$$I(\omega, y|D, x) = H[p(y|x, D)] - \mathbb{E}_{p(\omega|D)} H[p(y|x, \omega)] \quad (21)$$

This can be approximated by sampling from the posterior distribution:

$$\mathbb{I}(p) \approx \mathbb{H}_{\text{pred}}(p) + T^{-1} \sum_t \sum_c p_{ct} \log p_{ct} \quad (22)$$

⁸While the current work strictly defined this as predictive entropy, some works refer to this simply as entropy. Expected Entropy will sometimes also simply be referred to as entropy or Shannon entropy. In this work we consistently keep these distinct.

Table 2: An overview of different Uncertainty Measures that capture predictive uncertainty/confidence from a distribution over probabilities.

Name	Formula	Intuition	Ale UQ	Epi UQ
Class Probability [74]	$\mathbb{P}(p) = \bar{p}_{\bar{c}}$	Mean probability of predicted class	✓	✓
Predictive Entropy [77]	$\mathbb{H}_{\text{pred}}(p) = -\sum_c \bar{p}_c \log \bar{p}_c$	Uncertainty in mean prediction	✓	✓
Probability Variance [84]	$\mathbb{V}_{\bar{c}}(p) = T^{-1} \sum_t (p_{\bar{c}t} - \bar{p}_{\bar{c}})^2$	Variance of the predicted probability		✓
Expected Entropy [29], [111]	$\mathbb{H}_{\mathbb{E}}(p) = -T^{-1} \sum_t \sum_c p_{ct} \log p_{ct}$	Average uncertainty for each prediction	✓	
Mutual Information [77]	$\mathbb{I}(p) \approx \mathbb{H}_{\text{pred}}(p) - \mathbb{H}_E(p)$	Information gain from new sample		✓
Margin of Confidence [77]	$\mathbb{M}(p) = T^{-1} \sum_t p_{\bar{c}t} - \max_{c' \neq \bar{c}} p_{c't}$	Average distance to second class	✓	?

We consider some number of forward passes $t \in T$. We denote some number of classes $c \in C$. A given probability for a class c on pass t is then p_{ct} . The average probability of a class c over all passes T is denoted \bar{p}_c . To denote the highest probability class after averaging over T we use \bar{c} . Lastly, $f_{\bar{c}}$ is the number of passes in T where $p_{\bar{c}t} = \max_c p_{ct}$.

These terms can be reordered as shown by Mukhoti *et al.* [112] into:

$$\underbrace{\mathbb{H}_{\text{pred}}(p)}_{\text{total}} \approx \underbrace{\mathbb{I}(p)}_{\text{epistemic}} - \underbrace{T^{-1} \sum_t \sum_c p_{ct} \log p_{ct}}_{\text{aleatoric if ID}} \quad (23)$$

We consider the latter part the *Expected Entropy*, which is a measure of aleatoric uncertainty.

This disentangling of Predictive Entropy into Mutual Information and Expected Entropy is well established in Computer Vision literature [12], but we found surprisingly little traction for biosignal applications. Only Zhang *et al.* [83] used this set of complementary measures, though using only Predictive Entropy is more common [29], [34], [47], [73], [80], [91], [92], [95].

3.5. Margin of Confidence

Lastly, Milanes-Hermosilla *et al.* [77] proposes the Margin of Confidence as an intuitive uncertainty measurement. This ad-hoc measure looks at the average distance between the probability of the predicted class and the class with the next highest probability. Note that while the predicted class is taken over the average from the forward passes $\bar{c} = \arg \max_{c \in C} \bar{p}_c$, the second-highest is chosen on each sample. This means that in some forward passes, the second-highest probability $\max_{c' \neq \bar{c}} p_{c't}$ is actually higher than the probability of the predicted class $p_{\bar{c}t}$.

In its full form the Margin of Confidence is given as:

$$\mathbb{M}(p) = T^{-1} \sum_t p_{\bar{c}t} - \max_{c' \neq \bar{c}} p_{c't} \quad (24)$$

Milanes-Hermosilla *et al.* [77] used the Margin of Confidence to separate correctly and incorrectly classified predictions. They found that the Margin of Confidence had a greater Bhattacharyya distance between the correctly and incorrectly classified predictions than Mutual Information, Predictive Entropy and Probability Variance, but replications with other models, UQ methods and data are needed.

3.6. Recommendations for Uncertainty Measures

We find that the uncertainty measures used in the biosignal literature are often ad-hoc, lack thorough argumentation and are sometimes underspecified. We argue that future work should always specify how they measure uncertainty to ensure reproducibility.

In Computer Vision the established method of uncertainty measures for classification is using Predictive Entropy, Expected Entropy and Mutual Information. While this has substantial limitations [12]–[14], we find that it is currently the best approach. This gives a measure of total uncertainty, epistemic uncertainty and aleatoric uncertainty, though we caution that this disentanglement cannot be fully trusted. Instead, they should be used as *best estimates*, rather than true predictions. For regression the aleatoric variance or the epistemic variance would be a best estimate [42].

Additionally, we believe that the class probability and the class variance are easy to interpret by clinicians, and are therefore most suitable when uncertainty estimates are used in a decision support system.

4. Uncertainty Use Cases

When applying Uncertainty Quantification to a biosignal application there should always be some purpose to the uncertainty estimation. Different ways of using uncertainty put different expectations on it, and the way uncertainty is used in biosignals comes with some special considerations.

It also comes with different ideals for which kind of uncertainty should be used. We provide an overview of which (theoretical) uncertainty measure is most fitting for which use case in Table 3. While there are no guarantees that measures for epistemic uncertainty only predict epistemic uncertainty, a paper should at least include the appropriate uncertainty for the appropriate task. We found that this is not always well understood in the biosignal literature, so this overview may help authors and reviewers.

Table 3: Various Uncertainty Use Cases grouped in their required type of uncertainty. In general, methods that need either aleatoric or epistemic uncertainty may still do well with a mixture of both. Rejection is split into rejection when the data is in-distribution (ID), or out-of-distribution (OOD), relative to the training data.

Aleatoric	Epistemic	Both
Feature	Active Learning	Interpretability
Rejection (ID)	Model Pruning	Social Bias
	Data Augmentation	Soft Voting
	Rejection (OOD)	

4.1. Rejection Methods

The most common use for estimating uncertainty is to be able to not make a prediction when the likelihood of that prediction being wrong is too high. 34% of papers in this review use a measured uncertainty to reject difficult samples from the testing data.

Below we highlight how this impacts evaluation, the choice of uncertainty measures, and implementation in a biosignal context.

4.1.1. Evaluating Rejection methods

A common technique used to evaluate uncertainty quantification for rejection methods is setting a threshold against uncertainty and observing an increase in accuracy and a decrease in coverage [16], [30], [43], [57], [76], [95]. This framework considers uncertainty as a tool to improve classification performance, instead of having uncertainty as an inherent goal. While some works set a single threshold against uncertainty [43], [57], [76] we recommend a range of thresholds [16], [30], [95], as the right balance between coverage and accuracy is typically not well established and a comparative analysis after a single threshold is not possible.

Instead, coverage-accuracy plots as visualised in Figure 12 may be used to assess the reject-performance of a model. By going over all possible thresholds, this plot shows the options for balancing coverage and accuracy, which may be used for comparing models.

The alternative framework is to consider uncertainty as a classification task, where the goal is to classify whether a prediction will be correct or incorrect [16], [47], [77], [79], [80], [91]. This results in the common classification metrics area under the ROC-curve [113], but specifically for rejection.

We recommend using the accuracy/coverage curve to explain the possible behaviour a classifier with rejection can exhibit, whereas a separate task-ROC and a rejection-ROC may give better insight into the individual components.

The results from Jahmunah *et al.* [47] indicate a limitation of rejection with standard classification metrics. Their results show that even with large noise for an ECG classification task, the ECG is sometimes guessed correctly even

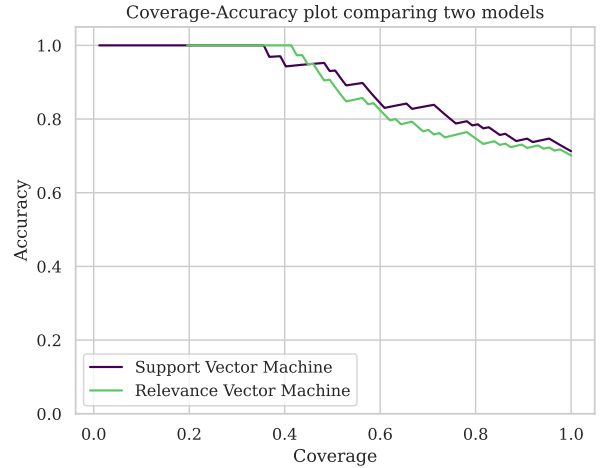


Figure 12: Example plot showing the trade-off between coverage and accuracy for two EEG Motor Imagery classifiers. The plot shows that both models have very similar accuracy without rejection (coverage at 1.0), but that the Support Vector Machine has a better accuracy-coverage trade-off.

when the model should be uncertain. Therefore, considering uncertainty as a classification task will inflate the number of false negatives and thus underestimate the rejection performance.

4.1.2. Choice of Uncertainty Measure

Both aleatoric and epistemic uncertainty can contribute to a risk of predictions being wrong, so total uncertainty would theoretically be optimal. However, in practice it may be that a measure of only aleatoric or epistemic uncertainty might work better. We recommend considering measures of aleatoric, epistemic and total uncertainty and seeing which performs best.

Most works pick one uncertainty measure and do not actively compare them. Only Fiorillo *et al.* [76] made such a comparison, by considering both average class probability and probability variance as the uncertainty to use for rejection. They found the accuracy improved most under rejection with average class probability, across multiple datasets. However, it may be possible that other measures work better when there is more epistemic uncertainty involved.

4.1.3. The Rejected Samples

In rejection methods it is worth contemplating what happens to the samples that are rejected. van Gorp *et al.* [10] suggests that under epistemic uncertainty a clinician could re-assess the data, while under aleatoric uncertainty a re-recording of the electrodes would be needed instead [38], or even alternative tests [11]. However, we caution that current predictions of aleatoric and epistemic uncertainty are not sufficiently separable to implement such systems [14].

Implementations where predictions are made and used in real-time require a well-considered behaviour for rejected cases. Machine Learning with rejection should consider how the rejected samples impact the larger clinical diagnosis system and what the outcome will be for patients who are rejected by the classifier.

4.2. Uncertainty for Interpretability

Uncertainty is sometimes proposed as a method to alleviate the black-box problem of Neural Networks [95]. By presenting uncertainty a model is able to show that a given prediction may not be correct, which can make the clinician more confident in trusting the certain predictions from a Machine Learning system.

Determining what good communication of a quantified uncertainty is can be difficult.

Research on scientific visualization of uncertainty is available [114], [115], but is not interweaved with the reviewed literature and does not demonstrate how to present different measures of uncertainty. Specifically clinical interpretation of uncertainty is critical, as it may affect the quality of a diagnosis or the adoptability of Machine Learning methods. For some ECG applications time-sensitivity is given as a factor affecting manual diagnosis [47], so the interpretation of an uncertain prediction may be subject to time constraints in such cases.

In standard classification tasks an accepted way of presenting a quantified uncertainty is by reporting an accurate class probability. A predicted class probability that accurately corresponds to the true probability of a class (even under epistemic uncertainty) can be mathematically interpreted and gives a well-defined and well-understood measure of uncertainty. Expected Calibration Error (ECE) has been used to capture this goal in a metric [29], [60], [74].

ECE is a common method for evaluating uncertainty quantification. It measures the difference between a predicted probability for a classification and the actual observed probability on a validation set [116]. This is often visualised with a calibration plot as shown in Figure 13. While Expected Calibration Error measures directly the correspondence between predicted probability and true probability, it does not show what is the cause. A consistently over-confident or under-confident classifier can both have a bad ECE. We recommend additionally looking at Net Calibration error [117], which measures the over/under confidence in isolation.

ECE is only defined for classification problems. For regression problems a similar method exists called ENCE [118]. In this method similar bins are made, but instead of comparing accuracy with predicted probability, it compares root mean-squared error to the predicted root mean variance. This can also be evaluated with plots similar to the calibration plot in Figure 13.

Currently, we recommend ECE as a metric to evaluate predicted probabilities when those probabilities are to be

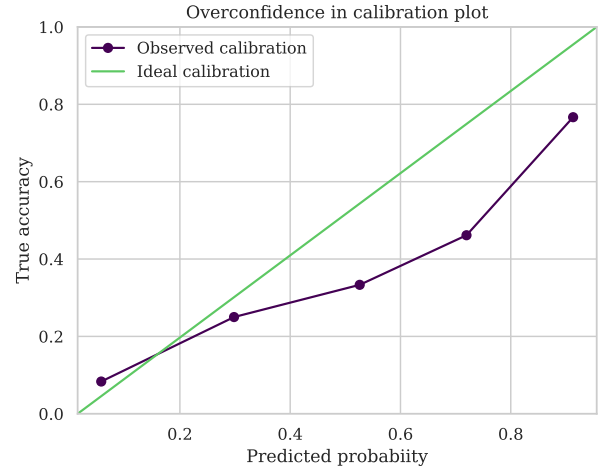


Figure 13: Example calibration plot for an EEG Motor Imagery classifier. The predicted probability is consistently higher than the true probability of being correct. This means the model is overconfident. Note that a plot like this is only reliable with sufficient samples with different predicted probabilities.

interpretable to a clinician, but research on how clinicians interact with predicted uncertainty is lacking. It may be that uncertainties are easier to interpret if presented as natural language as in Mendoza *et al.* [82], which would result in different evaluation criteria. Additionally, it may be necessary to evaluate uncertainty on data that is recorded with the same hardware, in the same clinic and by the same people as where it would be implemented, as this may introduce more epistemic uncertainty.

4.2.1. Visualizations of Uncertainty

Interpretation of uncertainty may be further improved by having an appropriate visualization that aligns with the specific biosignal task.

We generalised the reviewed approaches into three different categories that are shown in Figure 14. We found visualisations that makes a distinction between the prediction and the (epistemic) uncertainty, visualisations that show all possible predictions (e.g. as a histogram) to show uncertainty while leaving the prediction implicit, and visualisations that offer insight into a sample without explicitly quantifying the prediction or the uncertainty. The reviewed visualisations do not have theoretical or empirical arguments for their design, but by defining the framework we offer some grasp on otherwise varying visualisations.

The specific design details vary depending on the exact task and application context. We discuss those details and variations found in the reviewed literature below.

Gill *et al.* [86] uses a CNN with MC-Dropout to classify lesional voxels in patients with focal cortical dysplasia. The results are presented by a map of class probability voxels (predictive uncertainty) and a separate map of probability variance voxels (epistemic uncertainty). This gives an explicit separation between the prediction and the

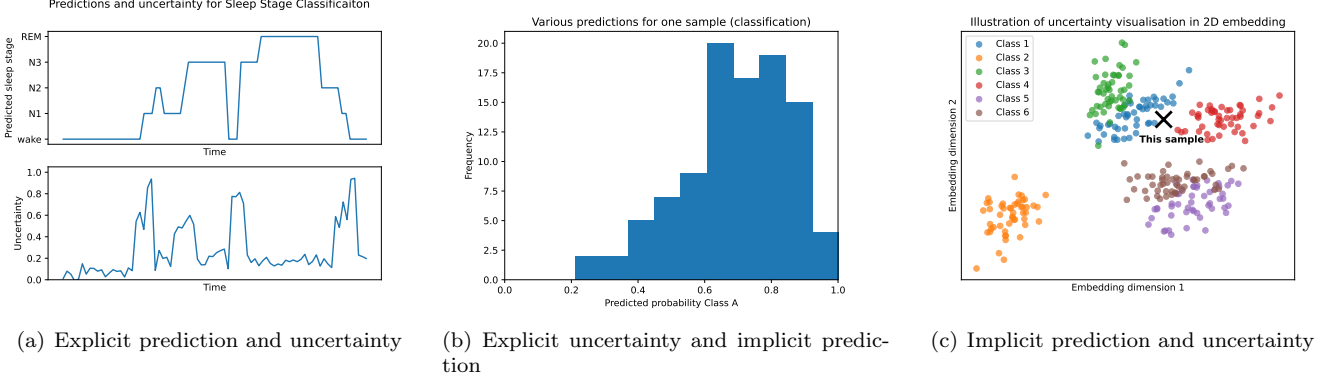


Figure 14: Three different general approaches to visualising uncertainty. The first plot specifically shows a separate prediction and a degree of uncertainty. The second plot explicitly shows the uncertainty, but leaves the predicted probability to be determined by the user. The third plot shows an embedding of the sample, but does not explicitly give a prediction nor an uncertainty.

epistemic uncertainty. It is then up to the user to combine these two sources of information.

Bekhti *et al.* [55] proposes a Markov Chain Monte Carlo (MCMC) approach to solve the inverse problem. The MCMC sampling results in multiple sparse solutions, where the agreement between solutions is interpreted as uncertainty. By presenting a heatmap of the source localization solutions on 3D brain renderings they allow the reader to interpret the level of uncertainty based on the relative density and the total spread of solutions. This also allows readers to involve their prior knowledge about neuroanatomy implicitly by contrasting the certainty of the predictions against prior knowledge.

Phan *et al.* [95] shows a method to support EEG-based sleep classification. They show a timeseries of the predictive entropy, the stacked class probabilities and the classifications above each other. To improve readability they highlight the parts where confidence drops below a given threshold. This is used to show how uncertainty is highest during stage transitions. Representing uncertainty over the timeseries, in combination with the original signal will let a Machine Learning system work as an effective decision support system for biosignal analysis.

A more generalisable method is given by Costabal *et al.* [20], who present a histogram of the whole distribution of class probabilities. This allows readers to intuitively assess central tendencies, spread and skew, and generalises naturally to visualisations for regression.

An exceptionally interesting approach to dealing with the interpretation of uncertainty is suggested by Van De Leur *et al.* [39], where a VAE embedding of an ECG is reduced to 2 dimensions using Principal Component Analysis. A cardiologist is presented with the embeddings of known diagnoses. This allows them to determine a measure of uncertainty based on a more fluid notion of vacuity, dissonance, aleatoric or epistemic uncertainty. By not trying to quantify uncertainty, but instead allowing the cardiologist to assess uncertainty, this method aims to make a diagnosis more interpretable.

4.3. Uncertainty as an Instrumental Goal

All other usecases of uncertainty we found were using uncertainty to improve some other task, such as Active Learning, and pruning in a more complicated classifier pipeline. For these cases, evaluating the uncertainty specifically has limited relevance as it is not an output from the system. Instead, the impact of uncertainty should be measured based on how it helps with the downstream task. For these aspects, the specific relation to biosignals is somewhat limited, as these may also be applicable to other tasks. However, we highlight these as this topic gets little attention in literature reviews focused on methodology.

We outline the setups that use uncertainty to improve some other outcome to demonstrate possible setups, and to illustrate the usefulness of uncertainty.

4.3.1. Uncertainty as a Feature

The most direct use of uncertainty is as a feature for subsequent Machine Learning tasks. For example, Stoean *et al.* [85] attempts to detect presymptomatic spinocerebellar ataxia using electrooculography. They observe the saccadic eye movements in healthy, sick, and presymptomatic participants. Healthy participants show a sudden eye movement with nearly instant acceleration and deceleration. Sick participants can show more chaotic movement with slower acceleration and speed. Presymptomatic participants can show a decrease in control, speed and rate of acceleration. Since there is a lot of variation between participants and each saccade, 85 saccades are recorded for each participant, and classified with an ensemble of Deep Neural Networks using MC-Dropout. The 3 class probabilities and the 3 class standard deviations for all 85 saccades were used for a decision tree classifier. The system was able to classify sick and healthy participants quite well, and performed acceptably at classifying presymptomatic participants.

When uncertainty is used as a feature for another Machine Learning model the constraints of what a good uncertainty is are loosened. Uncertainty may be expressed

with multiple uncertainty measures, and over or underconfidence will not have an impact on the system.

4.3.2. Uncertainty to Control Social Bias

As fairness and negative social biases are a growing concern in Machine Learning, Zanna *et al.* [84] present a rather unique usecase for uncertainty quantification. They propose a Multi-Task Learning method using Uncertainty Quantification to reduce social bias while classifying periods of anxiety from ECG features. The bias mitigation strategy uses a separate output that attempts to classify whether the samples belong to a person from an unprivileged demographic group.

The model is trained for 100 epochs, with the weights being saved every 5 epochs. After training, the model with the highest average epistemic uncertainty (probability variance) on the demographic-classification and the lowest average uncertainty on the anxiety-classification is selected. The model performing poorly at demographic classification should not have features in the latent representation to capture demographic classification. The authors showed that this minimized bias, but this did come at a loss in model performance.

While this method is still somewhat ad-hoc, it paves the way for future methods in minimizing social bias through uncertainty quantification. Future research may focus on forms of adversarial training, so that an anxiety model will try to optimize the anxiety classification while under an ongoing constraint of having no features that may be used to infer the demographic class. The different effects of aleatoric and epistemic uncertainty are also worth exploring here.

4.3.3. Bayesian Active Learning

Under Active Learning training samples are iteratively selected by the epistemic uncertainty that the model has about that sample [119]. These methods are proposed for situations where insufficient labelled training data is available, and manual labelling of data is expensive. Active Learning starts with a model trained on very little data, and observes the uncertainty it has on the unlabelled data. The most uncertain samples are then manually labelled by an *Oracle*: a system that produces the ground truth labels. This Oracle can be the expert annotations, but may also be additional (expensive) testing to establish a better ground truth. We found three different ways this is used for uncertainty.

Wabina *et al.* [58] compared their Neural Differential Equation approach to a BNN trained with Active Learning. BNNs that use Active Learning can use their (epistemic) uncertainty to indicate about which samples they are uncertain. Remarkably, the best performance was actually observed by predictive entropy (total uncertainty rather than Mutual Information (epistemic uncertainty), presumably due to poor uncertainty disentanglement.

Vavaroutas *et al.* [96] instead uses Active Learning to guide their Data Augmentation process for ECG and EEG

classification tasks. They add Data Augmentations to an existing dataset, theorise a scenario with unlabelled samples, and use Active Learning to achieve acceptable performance with only 20 annotated samples. We believe that this is a promising direction, but that care should be taken to ensure that if augmentations need to be annotated by clinicians, then those augmentations should be specifically designed to maintain the integrity of a realistic signal. Clinicians might not be able to annotate a horizontally flipped ECG.

Lastly, Fawden *et al.* [34] uses a method similar to Active Learning to reduce the size of the dataset to train the model on. They show that this reduces the computational cost of transfer learning, which may be important for edge devices where the biosignal recording may be privacy sensitive and must be used to train a model locally.

Overall we find that Active Learning is a promising avenue, although more work is needed to understand the downstream impact of using Bayesian Active Learning.

4.3.4. Miscellaneous use cases for uncertainty

Two works propose novel ways to use uncertainty for Brain-Computer Interfaces. As part of their UNCER model, Duan *et al.* [68] uses uncertainty to assess the quality of data augmentation. They consider data augmentation as a method to reduce uncertainty to unseen corruptions.

For a P300 speller Ma *et al.* [90] look at model uncertainty, not only in terms of how it affects predictive uncertainty, but also in what it says about the model. They argue that weights with a poor signal-to-noise ratio are redundant. With this method they were able to prune 75% of the weights without decreasing the F1 score. In the single-point model any amount of pruning would result in a (slight) decrease in F1 score.

Additionally, Ma *et al.* [90] used the predicted probability for a special soft-voting strategy. In P300 spellers each letter is flashed several times, and a classifier tries to identify a P300 wave. By using the probability of a P300 wave their Bayesian CNN outperformed an equivalent single-point model. This strategy of voting with probabilities, rather than with discretised predictions is similar to Soft Voting in Machine Learning ensembles.

4.4. Recommendation for use cases

We advise future work on applications of Uncertainty Quantification to specify what purpose of uncertainty estimation they are considering. One may consider a rejection scenario, a decision support system, or using uncertainty as an instrument to achieve some other goal.

If the goal of uncertainty is to achieve good rejection, appropriate evaluation should use the accuracy/coverage curve or consider rejection as a classification task and present an ROC-curve. Reporting results with a single threshold is not sufficient, as it cannot be interpreted. We also find that it is well established that rejection gives some

benefit, so studies should focus on comparing methods to achieve the most benefit, or investigate how to deal with rejected samples in a clinical setting.

Works focusing on uncertainty to improve interpretability can evaluate their methods using ECE, NCE and the Brier score, though a rejection ROC-curve may be a good addition. Foundational research on how clinicians interpret predictive uncertainty, how to communicate it, and how to visualise it are also needed.

When uncertainty is used as an intermediary, for example for pruning, Active Learning, or as a feature for another model, there are fewer constraints to the uncertainty quantification, and the evaluation does not need to be as thorough. Instead, evaluation should look at how uncertainty estimates affect the downstream task performance.

5. Guidelines for Adding Uncertainty Quantification

The review covered various methods for obtaining quantified uncertainties and presented methods which people have been using uncertainty for. Based on these findings, we aim to conclude a guideline on how to implement uncertainty quantification for a Machine Learning task on biosignal data. There is no singular solution or decision tree that works best for all cases. Nonetheless, we provide an outline below of decisions to make for researchers using a Machine Learning system for a biosignal task that are interested in using Uncertainty Quantification. These instructions should be taken with a critical eye and may be subject to disagreement. Still, it provides a starting point from which further methodologies may be constructed.

We start by considering the cost of adding Uncertainty Quantification to a Machine Learning task. After this the first step will cover the uncertainty quantification methods, which is mostly guided by your choice of Machine Learning model and computational constraints. Second is the choice of uncertainty measure, which is chosen on the constraints of the uncertainty usecase. The last step is the evaluation. Depending on the uncertainty usecase, different evaluation methods align best with the specific goal. Lastly, we discuss some sanity checks to validate that the uncertainty quantification works as intended.

5.1. Choice of Uncertainty Quantification Method

Knowing when your model’s predictions are likely to be wrong, and a hint of why they might be wrong, can be quite valuable. However, there is always a price to pay.

For MC-Dropout and Deep Ensembles this price is computational cost. MC-Dropout requires many forward passes, so the cost of inference might increase 100 times. Deep Ensembles require training several models, which means training cost may increase 5 times. At inference, this also requires having enough memory for 5 models.

However, these methods do not result in a decrease in model accuracy. MC-Dropout converges to roughly

the same prediction that a single-point model would have made after 100 forward passes [110], and ensembles are well established at improving model accuracy [120].

Methods that optimise a model for uncertainty (such as Variational Inference, Prior Networks, Evidential Machine Learning and Variational Autoencoders) are at risk of decreased model accuracy. Since the model is now optimised towards two tasks simultaneously, this may have a negative effect on the predictive performance. However, this is not guaranteed as multi-task learning leverages a similar mechanism to improve predictive performance [121].

Post-hoc calibration does not directly have a substantial computational cost, nor does it directly affect the model predictions. However, doing post-hoc calibration requires data to do the calibration on, which generally cuts into the data available for training or testing.

We generally recommend trying Deep Ensembles, MC-Dropout and a standard Neural Network and comparing their performances for the task at hand. If a five-fold increase in training cost or a 100-fold increase in inference cost a prohibitive only either Deep Ensembles or MC-Dropout may be a viable starting point. If well-calibrated uncertainties are a requirement, we recommend adding a post-hoc calibration method such as temperature scaling.

When the computational cost is a large constraint, one might try Evidential Deep Learning or Early Exit Ensembles to further reduce computational cost.

If the base-model of choice is not a Neural Network there is little previous work available to build on. We recommend implementing Bayesian methods for standard Machine Learning models such as Bayesian Logistic Regression, Bayesian Linear Discriminant Analysis and Relevance Vector Machines as explained by Prince [122], or doing bootstrap ensembling [11]. These methods have the ability to incorporate epistemic uncertainty, which is otherwise neglected.

5.2. Choice of Uncertainty Measure

There is fairly limited literature on regression with biosignals [20], [49], [58], [99], but we recommend from our experience two measures of uncertainty for regression: the variance of the prediction, or the 95% Confidence Interval. Measures of variance may be well suited for rejection systems, as they present a scalar uncertainty that can be thresholded against. Confidence Intervals may be preferable for human interpretation as they give a notion of likely possible values.

For classification problems the current state-of-the-art is more conflicting. For rejection the predictive entropy, expected entropy, or mutual information may all be good options. While they theoretically correspond with total, aleatoric and epistemic uncertainty in practice this is not straightforward and we recommend trying all three.

Alternatively, the Gaussian Logits disentangling gives a predicted probability, aleatoric variance and epistemic

variance, but this is less established in the current literature. Further research comparing these two methods of disentangling uncertainty for biosignals is needed.

For uncertainty to be interpreted by people (clinicians or users) a (well-calibrated) class probability is easiest to interpret. Epistemic uncertainty may be represented by the class variance, but would ideally be incorporated into a more uniform probability distribution.

When the purpose of the uncertainty is an intermediary multiple measures may be observed and combined with dimensionality reduction methods as needed. However, we expect that a combination of aleatoric, epistemic and mixed uncertainty measures will perform best.

5.3. Evaluating Uncertainty Quantification

Whenever Uncertainty Quantification is considered as a tool to improve the outcome of a larger system, rather than as its own end-goal, the evaluation methods may need to be adjusted to the purpose for which uncertainty is used. Below take in each section a given uncertainty usecase, and discuss how to evaluate the uncertainty quantification for that usecase.

5.3.1. Rejection

If uncertainty is used in order to reject difficult samples, the impact of uncertainty on the larger system may be directly measured with a coverage-accuracy plot as in [16], [95]. These systems all depend on setting a threshold, which is usually arbitrary. Therefore, it is better to create a plot that shows the outcome for all possible thresholds by plotting the coverage against the accuracy. Showing the coverage and accuracy only for a single threshold makes it hard to compare models when the distribution of the uncertainty measure shifts.

However, these coverage-accuracy plots do not give direct insights into the Uncertainty Quantification performance per se. Gaining more insights into this may help improve the large system, rather than only evaluate it. For this, it may be worth casting the uncertainty as a classification task, so that regular classification metrics may be used. Be aware that this is typically an unbalanced task, where again the cost of false-positives and false-negatives is not well defined, so ROC curves may be a preferred approach. Since a perfect uncertainty measure is not able to provide perfect classification (as described in Section 4.1.1), it may be worth adjusting the metrics to give a more directly interpretable evaluation of the uncertainty.

For both of these cases, it is worthwhile to use a good baseline to assess whether the Uncertainty Quantification method actually provides an improvement. Setting a threshold against a standard Neural Network with Softmax as uncertainty gives a fair baseline.

5.3.2. Interpretation

While the rejection usecase does not demand a well-calibrated measure of uncertainty, this may be important

for interpretation by a person. In this case the best approximation that can be given is that a predicted probability should align with the true probability. This can be measured by the Expected Calibration Error (or ENCE for regression [118]), which is therefore an acceptable metric for evaluating an uncertainty that needs to be directly interpreted.

However, giving too many significant figures of a probability may give a false sense of precision, so it is possible that similar probabilities can be put in larger bins, which may even be mapped to natural language. In that case, the Expected Calibration Error is not ideal, as many small errors can have a substantial contribution to this metric, but may not actually affect the presented uncertainties. Instead, Maximum Calibration Error may be used, as this would ignore the small calibration errors and only focus on the large differences.

For a thorough understanding of what works best for interpretability, human evaluation and user studies are needed. Both for the general problem of using uncertainty quantifying ML models, as well as for specific user groups and specific tasks. For supporting interpretability in medical decision making user studies should focus on the specific medical discipline of the user.

5.3.3. Intermediary Features

When uncertainty is used as an intermediate, for example as a feature for a different model, or as an acquisition function for Active Learning, it can be hard to identify which properties are required for an optimal uncertainty measure.

ECE / ENCE may be used as a proxy for the quality of the uncertainty, but this is not specific to the usecase. Instead, the uncertainty method should be evaluated on the impact it has on the performance of the larger system.

For any case of using uncertainty, it may be good to perform some sanity checks to ensure the uncertainty is behaving as intended [123]. For systems that are expected to measure epistemic uncertainty, one may try to create out-of-distribution data, and validate whether the epistemic uncertainty increases. To observe the quality of aleatoric uncertainty, one may look at the samples in the training data that are classified with high aleatoric uncertainty, to assess whether they align with the intuitions for aleatoric uncertainty. Alternatively, aleatoric uncertainty may be evaluated with relevant and realistic induced noise in the training data.

6. Open Challenges

We close the review by highlighting several open challenges of using uncertainty quantification for biosignals that warrant attention. Overall, while uncertainty quantification has been gaining traction, there are still multiple obstacles for adoption and under-explored areas. This paper removed some obstacles by providing an outline of how

to add Uncertainty Quantification to a biosignal classifier in Section 5. We invite more researchers to incorporate Uncertainty Quantification methods into their models and the address remaining open questions, as discussed in this section.

6.1. Interpretability of Uncertainty

This review found 14 papers where the quantified uncertainty was explicitly or implicitly intended to be interpreted by a person, but none of them connected the uncertainty to thorough studies of how different representations affect uncertainty. Gill *et al.* [86] - for example - makes a visualization distinguishing predictive and epistemic uncertainty in FCD lesions detection, but it is not known how well such a visualization helps a clinician with identifying the true lesions and the false positives. Mendoza *et al.* [82] bins uncertainty estimates into natural language (including "Cannot rule out", "Consider" and "Possible") to be more intuitive, but the impact this has on interpretation is not yet known, and it may require different metrics for evaluating uncertainty.

Previous research about how well clinicians can interpret probabilistic tests exists [124], [125], but that is currently not tied to the way Uncertainty Quantification research is conducted. Research on what makes a well-interpretable (disentangled) uncertainty is needed, with an emphasis on designing visualisations.

6.2. Small Uncertainty Models for Biosignals

Bayesian Neural Networks cover the majority of uncertainty quantification methods encountered in this review. These methods have been popularized in Computer Vision, where Deep Neural Networks are dominating the state-of-the-art.

While Deep Learning has been gaining popularity and generating good results on large datasets [126], its infamy for requiring large amounts of training data means many Biosignal models prefer shallower Machine Learning systems such as Support Vector Machines [127] and Linear Discriminant Analysis [128]. This review did not find much uncertainty quantification for such models, although they do exist (see Prince [122]). More research implementing uncertainty quantification on shallow models is needed, preferably with the ability to disentangle aleatoric and epistemic uncertainty, but minimally with the ability to capture a mixture of aleatoric and epistemic uncertainty. Larsen *et al.* [11] provides a starting point with pseudo-bootstrap ensembles, but a more thorough analysis of uncertainty for such a model is needed.

6.3. Appropriate Benchmarks for Uncertainty

Xia *et al.* [29] offers some benchmark data. They do this by introducing noise to existing biosignal datasets with the intention that uncertainty should go up as dataset shift makes the accuracy go down. While this is a good starting point, the type of introduced noise may not be reflective

of real dataset shifts that may be observed when UQ models are implemented in practice. Instead, there is a need for datasets that realistically capture the aleatoric and epistemic uncertainty they may encounter when biosignal models are deployed in practice.

Epistemic uncertainty presents most realistically in cross-subject generalisability, rare comorbidities, or unusual erroneous recordings. By tailoring a dataset with these sources of epistemic uncertainty, we can improve the construct validity of UQ research. For designing such datasets we encourage looking at out-of-distribution detection datasets in Computer Vision as a starting point [112], but with clear attention to what is realistic in biosignals.

6.4. Vacuity-Dissonance and Aleatoric-Epistemic

Two frameworks for understanding uncertainty were encountered. The most common is the distinction between aleatoric (data) and epistemic (knowledge) uncertainty. However, the vacuity (absence of class features) and dissonance (contradicting class features) distinction could provide a more directly interpretable disentangling of uncertainty. It is not clear how these frameworks interact, and clarifying this may provide a more complete understanding of the uncertainty a model encounters.

Future research may explore their interactions, their differences, and other interpretations of uncertainty that may be useful for biosignal classification tasks.

6.5. Uncertainty in Regression

Most of the reviewed literature focused on classification tasks, with only a few papers focused on uncertainty in regression. Methods for predicting, evaluating and communicating uncertainty in regression do exist, but since they are less prevalent, less is known about possible unique properties. There have been several extensive comparisons of UQ methods specifically using biosignals, but only for classification problems.

Thorough comparison of regression methods with uncertainty, as well as a critical look at how these methods are evaluated, is still needed. As discussed, biosignals can suffer from high dimensionality, noise, and low sample size, which may have a specific impact on the quality of different regression methods and how they can be evaluated.

Additionally, regression problems may come with unique challenges for communicating uncertainty in a medical setting. Real-time monitoring of vitals typically does not include a representation of heteroscedastic uncertainty. Research is needed on whether quantiles, variance, or histograms would make for usable and interpretable methods for uncertainty estimation in regression. The only work on interpretable uncertainty in regression we found was Martinez *et al.* [63], which looks at generating interpretable ECG with uncertainty estimates based on bio-impedance. However, they do not evaluate their method with users.

6.6. The Needs of Clinicians

Elul *et al.* [52] discusses the needs of clinicians in three concepts: estimating uncertainty, handling unknown classes, and detecting a failure to generalise.

Under the aleatoric-epistemic uncertainty framework, the *estimating uncertainty* corresponds to aleatoric uncertainty, while both out-of-distribution unknown and known classes fall under epistemic uncertainty. In order to better address the clinical concerns, each of these problems may be addressed separately. While the path towards this is not known, the unification of aleatoric-epistemic and vacuity-dissonance uncertainties may provide a starting point.

6.7. Using Uncertainty for Biosignal Applications

56.6% of the reviewed papers use uncertainty either for presenting a confidence with a prediction, or for rejecting difficult samples. However, there is an unknown number of other possible things that uncertainty quantification may be used for that need exploring.

A promising purpose is to use uncertainty in an online setting while recording a biosignal. An increase in uncertainty may correspond with artefacts in the data, making uncertainty an artefact detector with possibly better properties than normal artefact classifiers. One advantage is that it may only detect artefacts that are obstructing a good classification, allowing it to tolerate artefacts in channels or at timepoints where they do not pose a problem for the specific task.

There may be many more unexplored opportunities to use estimated uncertainties when these uncertainty-enabled models are integrated in a task environment. Perhaps in a neurorehabilitation BCI the uncertainty may be used to support the patient in improving their movement attempts, or in situations where the labels may be erroneous an uncertainty measure is able to detect mislabeled training samples [129].

6.8. Informative Priors

Variational Inference gives a modeler the option to specify a prior $p(\theta)$. This prior may be very helpful in training good Bayesian Neural Networks when data is limited. Efforts to cast domain knowledge into a probability distribution for $p(\theta)$ may be non-trivial, but this has the potential to improve these models.

Alternatively, the prior $p(\theta)$ may also be learned on datasets similar to the task at hand [37].

6.9. Rejected Samples

We see that several works reject difficult samples to improve accuracy. In medical diagnosis systems the assumption is that these difficult samples may be offered to a diagnostician, so that their quality of diagnosis may not be compromised by mistakes in the Neural Network. However, it is unclear what the resulting diagnostic performance of the whole clinical system would be when combining the assessment of the doctor with the prediction of

the Neural Network. It may be that they find the same samples difficult,

In Breast Cancer and Tuberculosis screening some theoretical work with historical data has been done [130]. Similar research may be done within the biosignal domain as a step towards implementing models with Uncertainty Quantification in the medical biosignal domain.

6.10. Label Ambiguity

Supervised Machine Learning considers the labels as *ground truth*. However, in reality these ground truths may not be entirely accurate. This is often due to ambiguity or annotator error. To achieve appropriate estimates of uncertainty, the uncertainty of the ground truth should also be considered, but we found that this does not yet get enough attention.

Ju *et al.* [131] demonstrates various methods for dealing with annotator disagreement on medical image classification. They demonstrate that the usual approach of establishing the ground-truth annotation by majority-vote is insufficient, and proposes a method that achieves better accuracy.

Label ambiguity is especially common in biosignal analysis, as the ground truth often cannot be reliably established. Zhang *et al.* [83] found that models with highly confident (incorrect) predictions corresponded with error or ambiguity in ECG labels. They found a subject with two arrhythmia, but the original label only included one of them. In sleep stage classification there is often disagreement about the exact onset of a different stage [95], and class definitions might even change with differences in the gold standards [132]. Generally it can be assumed that there is at least some label ambiguity in these biosignal datasets, but quantifying how much and for which samples is also important. Knowledge of label ambiguity can improve uncertainty estimation, and is valuable for further investigation.

6.11. Large Language Models with Uncertainty

Large Language Models (LLMs) have gained popularity due to their easy adaptability and minimal data requirements. LLMs have even been tested in their ability to analyse EEG[133], ECG and PPG [134], but no specific attention has been given to using them with uncertainty estimation for biosignals. We consider LLMs to be a promising method for prototyping with small datasets.

Uncertainty estimation for LLMs comes with some special properties compared to regular Machine Learning methods. The uncertainty of the predictions from the model indicates the uncertainty of how likely the token is, not necessarily how correct that token is. Additional steps are needed to ensure the token probabilities are predictive of correctness [135]. Alternatively, LLMs can also predict an uncertainty estimate as part of their answer, which is called verbalized uncertainty [136]. These verbalized uncertainties are typically overconfident [117], but provide a unique challenge for uncertainty estimation.

6.12. Detecting Distribution Shift with Uncertainty

Possibly the biggest risk in deploying Machine Learning systems in clinical settings is distribution shift. When a model is trained and evaluated on data from one setting, but is ultimately applied in a different setting the quality of the predictions will degrade. This may come from differences in the operators, recording hardware, patient populations or even bugs in the data handling. Such changes degrade the quality of predictions in ways that are typically not predictable.

Standard Machine Learning methods that only consider aleatoric uncertainty will be overconfident in these settings, while with epistemic uncertainty it may be possible to detect when the data distribution at deployment becomes different from the training distribution.

Some effort on detecting distribution shift in biosignals exists, but no effective methods have been established [29]. Additionally, current work is limited to synthetic shifts that might not represent shifts that would occur in reality. Datasets that combine recordings across clinics, contexts (inpatient vs outpatient), patient populations and time can show which distribution shifts occur and allow Machine Learning models to be trained in one context and thoroughly evaluated in another. Good epistemic uncertainty estimation should be able to estimate the lower quality of predictions under distribution shift, and research to establish this is duly needed.

7. Conclusion

This review finds that Uncertainty Quantification methods for Neural Networks have been gaining increasing attention in the biosignal domain for the last five years, but that there are some hurdles to overcome.

By providing clarification about how uncertainty measures relate to aleatoric and epistemic uncertainty, and by providing an end-to-end guideline on how to add uncertainty quantification to a biosignal classifying Neural Network we make uncertainty quantification more accessible to researchers working with EEG, ECG, EMG and EOG.

Many areas still remain to be explored. Uncertainty Quantification methods should be further studied in situ, where clinicians may perform specific actions based on predicted uncertainty. To this end, studies that investigate the performance of a (clinical) environment containing an uncertainty-estimating model are needed.

Acknowledgment

The authors received no financial support for the research, authorship, and/or publication of this article.

References

- [1] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature medicine*, vol. 25, no. 1, pp. 30–36, 2019.
- [2] M. Abdar, F. Pourpanah, S. Hussain, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.
- [3] A. Malhotra, M. Younes, S. T. Kuna, *et al.*, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, 2013.
- [4] R. Chai, G. R. Naik, S. H. Ling, Y. Tran, A. Craig, and H. T. Nguyen, "Channels selection using independent component analysis and scalp map projection for EEG-based driver fatigue classification," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, ISSN: 1558-4615, Jul. 2017, pp. 1808–1811. DOI: 10.1109/EMBC.2017.8037196.
- [5] Rifai Chai, Y. Tran, G. R. Naik, *et al.*, "Classification of EEG based-mental fatigue using principal component analysis and Bayesian neural network," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2016, pp. 4654–4657, Aug. 2016, Place: United States Publisher: [IEEE], ISSN: 2694-0604. DOI: 10.1109/EMBC.2016.7591765.
- [6] Rifai Chai, M. R. Smith, T. N. Nguyen, Sai Ho Ling, A. J. Coutts, and H. T. Nguyen, "Comparing features extractors in EEG-based cognitive fatigue detection of demanding computer tasks," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2015, pp. 7594–7597, 2015, Place: United States Publisher: [IEEE], ISSN: 2694-0604. DOI: 10.1109/EMBC.2015.7320150.
- [7] F. Burden and D. Winkler, "Bayesian regularization of neural networks," *Artificial neural networks: methods and applications*, pp. 23–42, 2009.
- [8] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.
- [9] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.
- [10] H. van Gorp, I. A. M. Huijben, P. Fonseca, R. J. G. van Sloun, S. Overeem, and M. M. van Gilst, "Certainty about uncertainty in sleep staging: A theoretical framework," *English, Sleep*, vol. 45, no. 8, Aug. 2022, Place: Cary Publisher: Oxford Univ Press Inc WOS:000814108500001, ISSN: 0161-8105. DOI: 10.1093/sleep/zsac134. (visited on 01/16/2023).
- [11] K. Larsen, C. Zhao, J. Keyak, *et al.*, "A new method of modeling the multi-stage decision-making process of crt using machine learning with uncertainty quantification," *arXiv preprint arXiv:2309.08415*, 2023.
- [12] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier, "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?" In *Uncertainty in Artificial Intelligence*, PMLR, 2023, pp. 2282–2292.
- [13] B. Mucsányi, M. Kirchhof, and S. J. Oh, "Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks," *arXiv preprint arXiv:2402.19460*, 2024.
- [14] I. P. de Jong, A. I. Sburlea, and M. Valdenegro-Toro, "How disentangled are your classification uncertainties?" *arXiv preprint arXiv:2408.12175*, 2024.
- [15] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

- [16] Y. Lin, R. Palaniappan, P. De Wilde, and L. Li, "Reliability Analysis for Finger Movement Recognition With Raw Electromyographic Signal by Evidential Convolutional Networks," English, *Ieee Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 96–107, 2022, Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc WOS:000748370800010, ISSN: 1534-4320. DOI: 10.1109/TNSRE.2022.3141593. (visited on 01/31/2023).
- [17] A. Jøsang, *Subjective logic*. Springer, 2016, vol. 4.
- [18] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.
- [19] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.
- [20] F. S. Costabal, K. Matsuno, J. Yao, P. Perdikaris, and E. Kuhl, "Machine learning in drug development: Characterizing the effect of 30 drugs on the QT interval using Gaussian process regression, sensitivity analysis, and uncertainty quantification," English, *Computer Methods in Applied Mechanics and Engineering*, vol. 348, pp. 313–333, May 2019, Place: Lausanne Publisher: Elsevier Science Sa WOS:000462472400012, ISSN: 0045-7825. DOI: 10.1016/j.cma.2019.01.033. (visited on 01/31/2023).
- [21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, PMLR, 2017, pp. 1321–1330.
- [22] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Lower upper bound estimation method for construction of neural network-based prediction intervals," *IEEE transactions on neural networks*, vol. 22, no. 3, pp. 337–346, 2010.
- [23] D. Betancourt and R. Muhanna, "Interval deep learning for uncertainty quantification in safety applications," *arXiv preprint arXiv:2105.06438*, 2021.
- [24] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.
- [25] S. R. Jantre, S. Bhattacharya, and T. Maiti, "Quantile regression neural networks: A bayesian approach," *Journal of Statistical Theory and Practice*, vol. 15, no. 3, p. 68, 2021.
- [26] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, PMLR, 2016, pp. 1050–1059.
- [27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.
- [28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [29] T. Xia, J. Han, and C. Mascolo, "Benchmarking uncertainty quantification on biosignal classification tasks under dataset shift," in *Multimodal AI in healthcare: A paradigm shift in health intelligence*, Springer, 2022, pp. 347–359.
- [30] R. Harper and J. Southern, "A Bayesian Deep Learning Framework for End-To-End Prediction of Emotion From Heartbeat," English, *Ieee Transactions on Affective Computing*, vol. 13, no. 2, pp. 985–991, Jun. 2022, Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc WOS:000804643000033, ISSN: 1949-3045. DOI: 10.1109/TAFFC.2020.2981610. (visited on 01/16/2023).
- [31] T. Pearce, F. Leibfried, and A. Brintrup, "Uncertainty in neural networks: Approximately bayesian ensembling," in *International conference on artificial intelligence and statistics*, PMLR, 2020, pp. 234–244.
- [32] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *Advances in neural information processing systems*, vol. 33, pp. 4697–4708, 2020.
- [33] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Benamoun, "Hands-on bayesian neural networks - a tutorial for deep learning users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, 2022.
- [34] T. Fawden, L. Qendro, and C. Mascolo, "Uncertainty-informed on-device personalisation using early exit networks on sensor signals," in *2023 31st European Signal Processing Conference (EUSIPCO)*, IEEE, 2023, pp. 1305–1309.
- [35] Y. Ovadia, E. Fertig, J. Ren, *et al.*, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.
- [36] T. Heskes, "Practical confidence and prediction intervals," *Advances in neural information processing systems*, vol. 9, 1996.
- [37] R. Shwartz-Ziv, M. Goldblum, H. Souri, *et al.*, "Pre-train your loss: Easy bayesian transfer learning with informative priors," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 706–27 715, 2022.
- [38] J. Belen, S. Mousavi, A. Shamsoshoara, and F. Afghah, "An Uncertainty Estimation Framework for Risk Assessment in Deep Learning-based AFib Classification," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, ISSN: 2576-2303, Nov. 2020, pp. 960–964. DOI: 10.1109/IEEECONF51394.2020.9443466.
- [39] R. R. Van De Leur, B. Hompot, R. J. Hassink, P. A. Doevenmans, and R. Van Es, "Interpretable uncertainty estimation for automated triage of 12-lead electrocardiogram using deep convolutional neural networks," English, *European Heart Journal*, vol. 42, pp. 3163–3163, Oct. 2021, Place: Oxford Publisher: Oxford Univ Press WOS:000720456903464, ISSN: 0195-668X. (visited on 01/31/2023).
- [40] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," *arXiv preprint arXiv:2203.09168*, 2022.
- [41] J. F. Vranken, R. R. van de Leur, D. K. Gupta, *et al.*, "Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms," *European Heart Journal-Digital Health*, vol. 2, no. 3, pp. 401–415, 2021.
- [42] X. Jin, J. Wang, L. Liu, and Y. Lin, "Uncertainty-aware denoising network for artifact removal in eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4470–4480, 2023.
- [43] Y. Lin, R. Palaniappan, P. De Wilde, and L. Li, "Robust Long-Term Hand Grasp Recognition With Raw Electromyographic Signals Using Multidimensional Uncertainty-Aware Models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 1–1, 2023, Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering, ISSN: 1558-0210. DOI: 10.1109/TNSRE.2023.3236982.
- [44] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems*, vol. 31, 2018.
- [45] M. Jürgens, N. Meinert, V. Bengs, E. Hüllermeier, and W. Waegeman, "Is epistemic uncertainty faithfully represented by evidential deep learning methods?" *arXiv preprint arXiv:2402.09056*, 2024.
- [46] H. Li, J. Wang, S. Zhao, F. Tian, J. Yang, and M. Sawan, "Real-time Biosignal Recording and Machine-Learning Analysis System," 2022, pp. 427–430. DOI: 10.1109/AICAS54282.2022.9869982.

- [47] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Uncertainty quantification in DenseNet model using myocardial infarction ECG signals," English, *Computer Methods and Programs in Biomedicine*, vol. 229, p. 107308, Feb. 2023, Place: Clare Publisher: Elsevier Ireland Ltd WOS:000905967100001, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2022.107308. (visited on 01/31/2023).
- [48] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions," *Journal of Mathematical Psychology*, vol. 85, pp. 1–16, 2018.
- [49] L. Zhang, X. Zhang, X. Zhu, R. Wang, and E. M. Gutierrez-Farewik, "Knee joint torque prediction with uncertainties by a neuromusculoskeletal solver-informed gaussian process model," in *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*, IEEE, 2023, pp. 1035–1040.
- [50] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, 2013.
- [51] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2.
- [52] Y. Elul, A. A. Rosenberg, A. Schuster, A. M. Bronstein, and Y. Yaniv, "Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning-based ECG analysis," English, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 24, e2020620118, Jun. 2021, Place: Washington Publisher: Natl Acad Sciences WOS:000668851100003, ISSN: 0027-8424. DOI: 10.1073/pnas.2020620118. (visited on 01/31/2023).
- [53] V. Schetinin and C. Maple, "A Bayesian model averaging methodology for detecting EEG artifacts," 2007, pp. 499–502. DOI: 10.1109/ICDSP.2007.4288628.
- [54] R. Hagan, C. Gillan, and F. Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease," English, *Informatics in Medicine Unlocked*, vol. 24, 2021, ISSN: 2352-9148. DOI: 10.1016/j.imu.2021.100606.
- [55] Y. Bekhti, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: Application to M/EEG source imaging," English, *Inverse Problems*, vol. 34, no. 8, p. 085010, Aug. 2018, Place: Bristol Publisher: Iop Publishing Ltd WOS:000436948900001, ISSN: 0266-5611. DOI: 10.1088/1361-6420/aac9b3. (visited on 01/16/2023).
- [56] D. J. C. Mackay, "Bayesian methods for adaptive models," in *California Institute of Technology*, 1992, ch. 5.2.
- [57] N. Mohamed, D. Rubin, and T. Marwala, "Detection of epileptiform activity in human EEG signals using Bayesian neural networks," in *IEEE 3rd International Conference on Computational Cybernetics, 2005. ICC3 2005.*, Apr. 2005, pp. 231–237. DOI: 10.1109/ICCCYB.2005.1511578.
- [58] R. S. Wabina and C. Silpasuwanchai, "Neural stochastic differential equations network as uncertainty quantification method for EEG source localization," *Biomedical physics & engineering express*, Nov. 2022, Place: England Publisher: IOP Publishing Ltd, ISSN: 2057-1976. DOI: 10.1088/2057-1976/aca20b.
- [59] L. Kong, J. Sun, and C. Zhang, "Sde-net: Equipping deep neural networks with uncertainty estimates," *arXiv preprint arXiv:2008.10546*, 2020.
- [60] A. Campbell, L. Qendro, P. Lió, and C. Mascolo, "Robust and Efficient Uncertainty Aware Biosignal Classification via Early Exit Ensembles," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, May 2022, pp. 3998–4002. DOI: 10.1109/ICASSP43922.2022.9746330.
- [61] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, IEEE, 2016, pp. 2464–2469.
- [62] A. Montanari, M. Sharma, D. Jenkus, M. Alloulah, L. Qendro, and F. Kawsar, "Eperceptive: Energy reactive embedded intelligence for batteryless sensors," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 382–394.
- [63] J. Martinez, A. Akbari, K. Sel, and R. Jafari, "Strategic Attention Learning for Modality Translation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, May 2020, pp. 1030–1034. DOI: 10.1109/ICASSP40776.2020.9053515.
- [64] G. Sovatzidi, M. Vasilakakis, and D. Iakovidis, "Constructive Fuzzy Cognitive Map for Depression Severity Estimation," *Studies in Health Technology and Informatics*, vol. 294, pp. 485–489, 2022. DOI: 10.3233/SHTI220506.
- [65] Y.-T. Liu, N. R. Pal, A. R. Marathe, and C.-T. Lin, "Weighted fuzzy dempster-shafer framework for multimodal information integration," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 1, pp. 338–352, 2017.
- [66] J. Mishra and M. Tiwari, "Cardiolabelnet: An uncertainty estimation using fuzzy for abnormalities detection in ecg," *Health Care Science*, vol. 2, no. 1, pp. 60–74, 2023.
- [67] A. Amirkhani, E. I. Papageorgiou, A. Mohseni, and M. R. Mosavi, "A review of fuzzy cognitive maps in medicine: Taxonomy, methods, and applications," *Computer methods and programs in biomedicine*, vol. 142, pp. 129–145, 2017.
- [68] T. Duan, Z. Wang, S. Liu, Y. Yin, and S. N. Srihari, "UNCER: A framework for uncertainty estimation and reduction in neural decoding of EEG signals," en, *Neurocomputing*, vol. 538, p. 126210, Jun. 2023, ISSN: 09252312. DOI: 10.1016/j.neucom.2023.03.071. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0925231223003156> (visited on 09/20/2023).
- [69] N. V. Rodrigues, L. R. Abramo, and N. S. Hirata, "The information of attribute uncertainties: What convolutional neural networks can learn about errors in input data," *Machine Learning: Science and Technology*, vol. 4, no. 4, p. 045019, 2023.
- [70] M. Valdenegro-Toro, I. P. de Jong, and M. Zullich, "Unified uncertainties: Combining input, data and model uncertainty into a single formulation," *arXiv preprint arXiv:2406.18787*, 2024.
- [71] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International conference on machine learning*, PMLR, 2020, pp. 9690–9700.
- [72] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5710–5719.
- [73] Z. Deng, C. Li, R. Song, X. Liu, R. Qian, and X. Chen, "EEG-based seizure prediction via hybrid vision transformer and data uncertainty learning," en, *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106401, Aug. 2023, ISSN: 09521976. DOI: 10.1016/j.engappai.2023.106401. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0952197623005857> (visited on 09/20/2023).
- [74] A. Borovac, T. P. Runarsson, G. Thorvardsson, and S. Gudmundsson, "Calibration of Automatic Seizure Detection Algorithms," in *2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, ISSN: 2473-716X, Dec. 2022, pp. 1–6. DOI: 10.1109/SPMB55497.2022.10014868.

- [75] S. Wong, A. Simmons, J. R. Villicana, and S. Barnett, "Estimating patient-level uncertainty in seizure detection using group-specific out-of-distribution detection technique," *Sensors*, vol. 23, no. 20, p. 8375, 2023.
- [76] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-Lite: A Simplified Automatic Sleep Stage Scoring Model With Uncertainty Estimates," English, *Ieee Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2076–2085, 2021, Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc WOS:000709073200001, ISSN: 1534-4320. DOI: 10.1109/TNSRE.2021.3117970. (visited on 01/16/2023).
- [77] D. Milanes-Hermosilla, R. T. Codorniu, R. Lopez-Baracaldo, *et al.*, "Monte Carlo Dropout for Uncertainty Estimation and Motor Imagery Classification," English, *Sensors*, vol. 21, no. 21, p. 7241, Nov. 2021, Place: Basel Publisher: Mdpi WOS:000719045200001. DOI: 10.3390/s21217241. (visited on 01/16/2023).
- [78] K. S. Rathore, S. Vijayarangan, P. Sp, and M. Sivaprakasam, "A Multifunctional Network with Uncertainty Estimation and Attention-Based Knowledge Distillation to Address Practical Challenges in Respiration Rate Estimation," en, *Sensors*, vol. 23, no. 3, p. 1599, Feb. 2023, ISSN: 1424-8220. DOI: 10.3390/s23031599. [Online]. Available: <https://www.mdpi.com/1424-8220/23/3/1599> (visited on 09/20/2023).
- [79] M. Barandas, L. Famiglini, A. Campagner, *et al.*, "Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram," en, *Information Fusion*, vol. 101, p. 101978, Jan. 2024, ISSN: 15662535. DOI: 10.1016/j.inffus.2023.101978. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1566253523002944> (visited on 09/20/2023).
- [80] A. O. Aseeri, "Uncertainty-Aware Deep Learning-Based Cardiac Arrhythmias Classification Model of Electrocardiogram Signals," English, *Computers*, vol. 10, no. 6, p. 82, Jun. 2021, Place: Basel Publisher: Mdpi WOS:000665560000001, ISSN: 2073-431X. DOI: 10.3390/computers10060082. (visited on 01/31/2023).
- [81] M. F. Islam, S. Zabeen, M. H. K. Mehedi, S. Iqbal, and A. A. Rasel, "Monte carlo dropout for uncertainty analysis and ecg trace image classification," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Springer, 2022, pp. 173–182.
- [82] A. Mendoza, M. Razavi, and J. R. Cavallaro, "Deep learning system for left ventricular assist device candidate assessment from electrocardiograms," in *2023 Computing in Cardiology (CinC)*, IEEE, vol. 50, 2023, pp. 1–4.
- [83] W. Zhang, X. Di, G. Wei, S. Geng, Z. Fu, and S. Hong, "Cardiac arrhythmia classification with rejection of ecg recordings based on uncertainty estimation from deep neural networks," *Neural Computing and Applications*, vol. 36, no. 8, pp. 4047–4058, 2024.
- [84] K. Zanna, K. Sridhar, H. Yu, and A. Sano, "Bias Reducing Multitask Learning on Mental Health Prediction," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, ISSN: 2156-8111, Oct. 2022, pp. 1–8. DOI: 10.1109/ACII55700.2022.9953850.
- [85] C. Stoean, R. Stoean, M. Atencia, *et al.*, "Automated Detection of Presymptomatic Conditions in Spinocerebellar Ataxia Type 2 Using Monte Carlo Dropout and Deep Neural Network Techniques with Electrooculogram Signals," English, *Sensors*, vol. 20, no. 11, p. 3032, Jun. 2020, Place: Basel Publisher: Mdpi WOS:000552737900025. DOI: 10.3390/s20113032. (visited on 01/31/2023).
- [86] R. S. Gill, H.-M. Lee, B. Caldairou, *et al.*, "Multicenter Validation of a Deep Learning Detection Algorithm for Focal Cortical Dysplasia," English, *Neurology*, vol. 97, no. 16, E1571–E1582, Oct. 2021, Place: Philadelphia Publisher: Lippincott Williams & Wilkins WOS:000708601400019, ISSN: 0028-3878. DOI: 10.1212/WNL.0000000000012698. (visited on 01/16/2023).
- [87] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL," English, *Ieee Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, May 2021, Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc WOS:000649625200018, ISSN: 2168-2194. DOI: 10.1109/JBHI.2020.3022989. (visited on 01/31/2023).
- [88] J. Park, K. Lee, N. Park, S. C. You, and J. Ko, "Self-Attention LSTM-FCN model for arrhythmia classification and uncertainty assessment," en, *Artificial Intelligence in Medicine*, vol. 142, p. 102570, Aug. 2023, ISSN: 09333657. DOI: 10.1016/j.artmed.2023.102570. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0933365723000842> (visited on 09/20/2023).
- [89] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.
- [90] R. Ma, H. Zhang, J. Zhang, *et al.*, "Bayesian Uncertainty Modeling for P300-Based Brain-Computer Interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2789–2799, 2023, ISSN: 1534-4320, 1558-0210. DOI: 10.1109/TNSRE.2023.3286688. [Online]. Available: <https://ieeexplore.ieee.org/document/10153625/> (visited on 09/20/2023).
- [91] D. Milanes-Hermosilla, R. Trujillo-Codorniu, S. Lamar-Carbonell, *et al.*, "Robust Motor Imagery Tasks Classification Approach Using Bayesian Neural Network," en, *Sensors*, vol. 23, no. 2, p. 703, Jan. 2023, ISSN: 1424-8220. DOI: 10.3390/s23020703. [Online]. Available: <https://www.mdpi.com/1424-8220/23/2/703> (visited on 09/20/2023).
- [92] M. M. Rahman, M. W. Rivolta, F. Badilini, and R. Sassi, "Quantifying uncertainty of a deep learning model for atrial fibrillation detection from ecg signals," in *2023 Computing in Cardiology (CinC)*, IEEE, vol. 50, 2023, pp. 1–4.
- [93] T. Siddique and M. S. Mahmud, "Classification of fnirs data under uncertainty: A bayesian neural network approach," in *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*, IEEE, 2021, pp. 1–4.
- [94] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, architectures and applications*, Springer, 1990, pp. 227–236.
- [95] H. Phan, K. Mikkelsen, O. Y. Chen, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic Sleep Staging With Interpretability and Uncertainty Quantification," English, *Ieee Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, Aug. 2022, Place: Piscataway Publisher: Ieee-Inst Electrical Electronics Engineers Inc WOS:000838529800011, ISSN: 0018-9294. DOI: 10.1109/TBME.2022.3147187. (visited on 01/16/2023).
- [96] S. Vavaroutas, L. Qendro, and C. Mascolo, "Uncertainty estimation with data augmentation for active learning tasks on health data," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2023, pp. 1–4.
- [97] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient hamiltonian monte carlo," in *International conference on machine learning*, PMLR, 2014, pp. 1683–1691.

- [98] E. I. Chetkin, S. L. Shishkin, and B. L. Kozyrskiy, "Bayesian opportunities for brain-computer interfaces: Enhancement of the existing classification algorithms and out-of-domain detection," *Algorithms*, vol. 16, no. 9, p. 429, 2023.
- [99] L. Zhang, X. Zhang, X. Zhu, R. Wang, and E. M. Gutierrez-Farewik, "Neuromusculoskeletal model-informed machine learning-based control of a knee exoskeleton with uncertainties quantification," *Frontiers in neuroscience*, vol. 17, p. 1254088, 2023.
- [100] T. M. Frago, W. Bertoli, and F. Louzada, "Bayesian model averaging: A systematic review and conceptual classification," *International Statistical Review*, vol. 86, no. 1, pp. 1–28, 2018.
- [101] D. P. Mandic, S. Kanna, and A. G. Constantinides, "On the intrinsic relationship between the least mean square and kalman filters [lecture notes]," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 117–122, 2015.
- [102] S. De Rooij, K. Batselier, and B. Hunyadi, "Enabling large-scale probabilistic seizure detection with a tensor-network kalman filter for ls-svm," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP)*, IEEE, 2023, pp. 1–5.
- [103] L. Ruff, R. Vandermeulen, N. Goernitz, *et al.*, "Deep one-class classification," in *International conference on machine learning*, PMLR, 2018, pp. 4393–4402.
- [104] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3369–3378.
- [105] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," *Advances in neural information processing systems*, vol. 31, 2018.
- [106] Y.-C. Li and J. Zhan, "Effect of dimensionality reduction on uncertainty quantification in trustworthy machine learning," in *2023 International Conference on Machine Learning and Cybernetics (ICMLC)*, IEEE, 2023, pp. 326–332.
- [107] R. T. Schirmer, J. T. Springenberg, L. D. J. Fiederer, *et al.*, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
- [108] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.
- [109] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," *Frontiers in neuroscience*, vol. 10, p. 196, 2016.
- [110] M. Valdenegro-Toro and D. S. Mori, "A deeper look into aleatoric and epistemic uncertainty disentanglement," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2022, pp. 1508–1516.
- [111] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018.
- [112] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty: A simple baseline," *arXiv e-prints*, arXiv:2102, 2021.
- [113] X. Huang, J. Yang, L. Li, H. Deng, B. Ni, and Y. Xu, "Evaluating and boosting uncertainty quantification in classification," *arXiv preprint arXiv:1909.06030*, 2019.
- [114] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, *et al.*, "Overview and state-of-the-art of uncertainty visualization," *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, pp. 3–27, 2014.
- [115] K. Potter, P. Rosen, and C. R. Johnson, "From quantification to visualization: A taxonomy of uncertainty visualization approaches," *IFIP advances in information and communication technology*, vol. 377, p. 226, 2012.
- [116] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632.
- [117] T. Groot and M. Valdenegro-Toro, "Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models," in *Proceedings of TrustNLP Workshop@ NAACL 2024*, 2024.
- [118] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, "Evaluating and calibrating uncertainty prediction in regression tasks," *Sensors*, vol. 22, no. 15, p. 5540, 2022.
- [119] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International conference on machine learning*, PMLR, 2017, pp. 1183–1192.
- [120] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1249, 2018.
- [121] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.
- [122] S. Prince, *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.
- [123] M. Valdenegro-Toro, "Exploring the limits of epistemic uncertainty quantification in low-shot settings," *arXiv preprint arXiv:2111.09808*, 2021.
- [124] O. Kostopoulou, K. Arora, and B. Pálfi, "Using cancer risk algorithms to improve risk estimates and referral decisions," *Communications Medicine*, vol. 2, no. 1, p. 2, 2022.
- [125] B. Pálfi, K. Arora, and O. Kostopoulou, "Algorithm-based advice taking and clinical judgement: Impact of advice distance and algorithm information," *Cognitive research: principles and implications*, vol. 7, no. 1, p. 70, 2022.
- [126] S. Somani, A. J. Russak, F. Richter, *et al.*, "Deep learning and the electrocardiogram: Review of the current state-of-the-art," *EP Europace*, vol. 23, no. 8, pp. 1179–1191, 2021.
- [127] I. Kawashima and H. Kumano, "Prediction of mind-wandering with electroencephalogram and non-linear regression modeling," *Frontiers in Human Neuroscience*, vol. 11, p. 365, 2017.
- [128] Y.-C. Yeh, W.-J. Wang, and C. W. Chiou, "Cardiac arrhythmia diagnosis method using linear discriminant analysis on eeg signals," *Measurement*, vol. 42, no. 5, pp. 778–789, 2009.
- [129] O. Arriaga, S. Palacio, and M. Valdenegro-Toro, "Difficulty estimation with action scores for computer vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 245–253.
- [130] K. Dvijotham, J. Winkens, M. Barsbey, *et al.*, "Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians," *Nature Medicine*, pp. 1–7, 2023.
- [131] L. Ju, X. Wang, L. Wang, *et al.*, "Improving medical images classification with label noise using dual-uncertainty estimation," *IEEE transactions on medical imaging*, vol. 41, no. 6, pp. 1533–1546, 2022.
- [132] H. Danker-Hopfe, P. Anderer, J. Zeithofer, *et al.*, "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard," *Journal of sleep research*, vol. 18, no. 1, pp. 74–84, 2009.
- [133] J. W. Kim, A. Alaa, and D. Bernardo, "Eeg-gpt: Exploring capabilities of large language models for eeg classification and interpretation," *arXiv preprint arXiv:2401.18006*, 2024.

- [134] Z. Liu, C. Chen, J. Cao, *et al.*, “Large language models for cuffless blood pressure measurement from wearable biosignals,” *arXiv preprint arXiv:2406.18069*, 2024.
- [135] Z. Jiang, J. Araki, H. Ding, and G. Neubig, “How can we know when language models know? on the calibration of language models for question answering,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 962–977, 2021.
- [136] S. Lin, J. Hilton, and O. Evans, “Teaching models to express their uncertainty in words,” *arXiv preprint arXiv:2205.14334*, 2022.