Towards a Probabilistic Fusion Approach for Robust Battery Prognostics

Jokin Alcibar[1], Jose I. Aizpurua[2,3], and Ekhi Zugasti[4]

[1,2,4] *Electronics & Computer Science Department, Mondragon University, Spain*
jalcibar@mondragon.edu
ezugasti@mondragon.edu

[3] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
jiaizpurua@mondragon.edu

ABSTRACT

Batteries are a key enabling technology for the decarbonization of transport and energy sectors. The safe and reliable operation of batteries is crucial for battery-powered systems. In this direction, the development of accurate and robust battery state-of-health prognostics models can unlock the potential of autonomous systems for complex, remote and reliable operations. The combination of Neural Networks, Bayesian modelling concepts and ensemble learning strategies, form a valuable prognostics framework to combine uncertainty in a robust and accurate manner. Accordingly, this paper introduces a Bayesian ensemble learning approach to predict the capacity depletion of lithium-ion batteries. The approach accurately predicts the capacity fade and quantifies the uncertainty associated with battery design and degradation processes. The proposed Bayesian ensemble methodology employs a stacking technique, integrating multiple Bayesian neural networks (BNNs) as base learners, which have been trained on data diversity. The proposed method has been validated using a battery aging dataset collected by the NASA Ames Prognostics Center of Excellence. Obtained results demonstrate the improved accuracy and robustness of the proposed probabilistic fusion approach with respect to (i) a single BNN model and (ii) a classical stacking strategy based on different BNNs.

1. INTRODUCTION

Batteries are key components in the transition towards a sustainable carbon-free economy. In this transition, the development of remaining useful life (RUL) prediction of batteries is a crucial activity. The accuracy and reliability of the RUL prediction models is essential to build trust in the predictions (Liu et al., 2023). In this context, robust and reliable battery prognostics models support the development of accurate monitoring strategies and cost-effective solutions.

The estimation of the state-of-health (SOH) of batteries is a key activity for the design of RUL prognostics models. SOH-based prognostics models focus on capturing the run-to-failure ageing dynamics and battery health state estimation (Toughzaoui et al., 2022). It is frequently used to determine age-related degradation that reduces energy capacity and rises safety risks, including overheating and explosions (Wang et al., 2022). Therefore, accurate SOH monitoring and forecasting are key activities to design and operate safe, reliable and effective battery-powered systems (H. Zhao et al., 2023).

SOH estimation is an ongoing area of research (Yang, Chen, Chen, & Huang, 2023). SOH refers to the ratio of the current maximum capacity relative to its original specified capacity (X. Zhao, Wang, Li, & Miao, 2024). SOH can be quantified through different factors, including resistance and maximum power. However, discharge capacity is the most common definition (Vanem, Salucci, Bakdi, & Alnes, 2021), and this is adopted in this research.

Recent data-driven approaches have focused on modeling the capacity degradation of lithium-ion (Li-ion) batteries. (Lee, Kwon, & Lee, 2023) used convolutional neural network (CNN) to estimate the future SOH value of Li-ion batteries, transforming the capacity degradation data into two-dimensional images. Estimates of the SOH and RUL are commonly found together in the literature. For example, (Toughzaoui et al., 2022) developed a CNN-LSTM architecture, and (Wei & Wu, 2023) presented a graph CNN complemented by dual attention mechanisms for the estimation of SOH and RUL of batteries. However, due to the variability inherent in battery manufacturing process, it is essential to quantify this uncertainty to ensure robust and reliable prognostics predictions

(Abdar et al., 2021; Nemani et al., 2023).

There are different sources of uncertainty present in the design, operation and maintenance of batteries (Hadigol, Maute, & Doostan, 2015). (Y. Zhang, Zhang, Liu, Feng, & Xu, 2024) introduced a SOH assessment method that estimates uncertainty through the quantile distribution of deep features, which are inferred from a Residual Neural Network (ResNet) architecture. This approach generates SOH values accompanied by confidence intervals. However, the proposed ResNet architecture lacks probabilistic layers, overlooking the uncertainty inherent in the model parameters. (Che et al., 2024) developed a prognostic framework to assess battery aging, using a CNN-LSTM Bayesian neural network. However, this approach limits the uncertainty to the final dense layers, which are the only components modeled probabilistically.

With the aim of capturing uncertainty associated with complex processes, recent studies in the broader machine learning (ML) community have focused on ensembles of probabilistic models. (Fan, Olson, & Evans, 2017) introduced a Bayesian posterior predictive framework for weighting ensemble climate models. (Cobb et al., 2019) present a new ML retrieval method based on an ensemble of Bayesian Neural Networks (BNNs). In this scenario, the overall output from the ensemble is treated as a Gaussian mixture model. However, models are equally weighted with no adaptation to the observed data. (S. Zhang, Liu, & Su, 2022) present a Bayesian Mixture Neural Network (BMNN) for Li-ion battery RUL prediction. The BMNN framework incorporates a Bayesian Convolutional Neural Network as feature extractor and a Bayesian Long Short-Term Memory to learn degradation patterns over time. However, the absence of a weighted model combination limits the analysis of individual model contributions.

Alternatively, (Bai & Chandra, 2023) described a Bayesian ensemble learning framework that uses gradient boosting by combining multiple Neural Networks trained by Markov Chain Monte Carlo (MCMC) sampling. Finally, (Dai, Pollock, & Roberts, 2023) demonstrate the robustness of Bayesian fusion by embedding the Monte Carlo fusion framework within a sequential Monte Carlo algorithm.

In this context, inspired by the use of probabilistic ensemble models to capture model uncertainty, the main contribution of this research is the development of a novel probabilistic model fusion approach for battery SOH predictions. Bayesian convolutional neural networks (BCNNs) are used as base models for SOH prediction, and the fusion approach integrates individual BCNN probabilistic predictions. The fusion strategy balances between precision and reliability of individual predictions, adopting an optimal tradeoff between accuracy and uncertainty of predictions through the proposed stacking approach.

The proposed approach has been compared with (i) individual BCNN models and (ii) fusion strategies focused on stacking of BCNN models using point prediction information. Obtained results confirm that the proposed framework infers accurate, well-calibrated, and reliable probabilistic predictions, which improve predictive performance and contribute to estimate uncertainty in a robust and reliable manner in complex data-driven tasks. The proposed approach has been tested and validated with the publicly available NASA's battery dataset (Saha & Goebel, 2007).

The remainder of this article is organized as follows. Section 2 outlines our probabilistic fusion approach for robust battery prognostics. Section 3 describes a case study to demonstrate the application of our methodology. Section 4 presents and analyzes the results obtained from the case study. Section 5 discusses the implications of these findings. The article concludes with Section 6, summarizing our main conclusions and suggesting avenues for future work.

2. Probabilistic Fusion Approach for Robust Battery Prognostics

The proposed probabilistic fusion framework integrates BCNNs with probabilistic ensemble strategies. The main objective of the integration is to generate accurate predictions with robust uncertainty quantification, thanks to the uncertainty quantification of Bayesian modelling (Blundell, Cornebise, Kavukcuoglu, & Wierstra, 2015) and the robustness and accuracy of ensemble strategies (S. Zhang et al., 2022).

The approach is divided into offline and online stages. Starting from a set of battery datasets, in the offline process, data pre-processing and model training steps are completed. In the online process, trained models are stacked in an ensemble model according to computed weight and stacking criteria. The outcome of the approach is a one-step-ahead probabilistic capacity estimate. Figure 1 shows the high-level block diagram of the proposed approach.

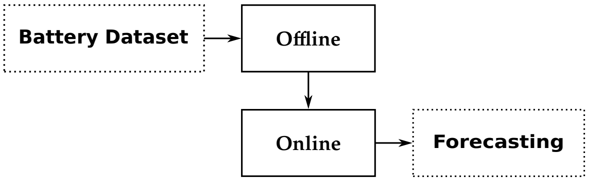


Figure 1. High-level block diagram of the proposed approach.

The high-level concepts in Figure 1 are implemented through the detailed model architecture shown in Figure 2.

The base models are BCNN models, which are trained (offline) through a leave-one-out cross validation (LOOCV) process. The probabilistic results of individual BCNN models are aggregated through a stacking process that includes accuracy and uncertainty metrics. In the testing (online) phase, each BCNN model weights are computed using learned mod-

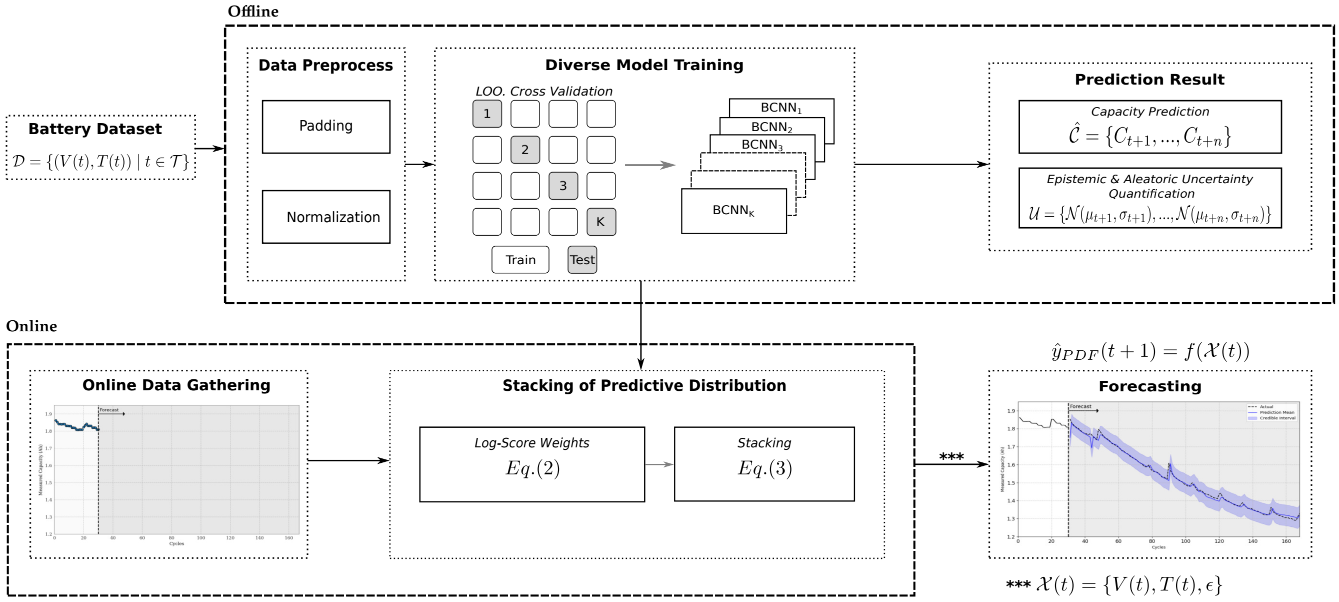

Offline
Battery Dataset
$\mathcal{D} = \{(V(t), T(t)) \mid t \in \mathcal{T}\}$
Data Preprocess
Padding
Normalization
Diverse Model Training
LOO. Cross Validation
1
2
3
K
Train
Test
BCNN$_1$
BCNN$_2$
BCNN$_3$
BCNN$_K$
Prediction Result
Capacity Prediction
$\hat{\mathcal{C}} = \{C_{t+1}, ..., C_{t+n}\}$
Epistemic & Aleatoric Uncertainty Quantification
$\mathcal{U} = \{\mathcal{N}(\mu_{t+1}, \sigma_{t+1}), ..., \mathcal{N}(\mu_{t+n}, \sigma_{t+n})\}$
Online
Online Data Gathering
Forecast
Stacking of Predictive Distribution
Log-Score Weights
$Eq.(2)$
Stacking
$Eq.(3)$
***
$\hat{y}_{PDF}(t+1) = f(\mathcal{X}(t))$
Forecasting
Actual
Prediction Mean
Credible Interval
Cycles
Measured Capacity (Ah)
*** $\mathcal{X}(t) = \{V(t), T(t), \epsilon\}$

Figure 2. Block diagram of the proposed approach.

els (log-score weights) and the stacking model is designed to combine them and generate a distribution from a mixture model. The following subsections explain in detail the main parts of the approach.

2.1. Offline Phase

During the offline phase, starting from a battery dataset with different run-to-failure trajectories on the same type of batteries, different base models are designed through a training strategy which seeks diversity in the training set to develop complementary predictive models.

2.1.1. Ensemble Base Models: BCNNs

BCNN models are a Bayesian extension of the classical CNN models to include uncertainty associated with parameter estimation. This requires modification of the classical backpropagation algorithm through Bayesian techniques that involves incorporating uncertainty into the model by treating weights as random variables, and applying variational inference to approximate posterior distributions. This results in a more robust model that predicts the complete probability density function (PDF).

Consequently, BCNN models have been selected to improve the robustness and accuracy of model prediction. To this end, BCNNs make use of probabilistic distributions to model parameters and the uncertainty related to their training process, and prior distributions to incorporate previous knowledge, generate uncertainty estimations and mitigate over-fitting (Blundell et al., 2015). In contrast, the classical learning models, e.g. non-Bayesian CNN models, focus on maximum likelihood estimation (MLE) and they overlook prior and posterior distributions. This leads to increasing error and decreasing model robustness in high uncertainty contexts, e.g. out-of-distribution data or manufacturing drifts.

The proposed approach utilizes data pre-processing techniques to standardize the length of discharge cycles through padding. This technique involves repeating the last discharge value until the desired cycle length is reached, ensuring consistent input dimensions for all models. Additionally, normalization is carried out scaling the discharge values between 0 and 1.

The architecture of the BCNN models is shown in Figure 3 defined as follows:

- Input data: the input data for the BCNN is structured in a tensor format. The rows represent data samples of discharge cycles, and columns that correspond to features, such as the voltage and temperature over time. Notably, the input does not include the current discharge as it remains constant in this scenario.
- Convolutional 1D Reparametrization: this layer creates a convolution kernel that is applied to the input data. During the forward pass, kernel and bias parameters are drawn from a Gaussian distribution. It uses the reparameterization estimator to approximate distributions through Monte Carlo trials, integrating over the kernel and bias.
- Global Average Pooling 1D: this layer performs average pooling specifically for temporal data. It reduces the spatial dimensions of the input data to a single value per channel by calculating the average over the temporal dimension.
- Flatten: this layer reshapes input data into a one dimensional array, enabling compatibility between Bayesian

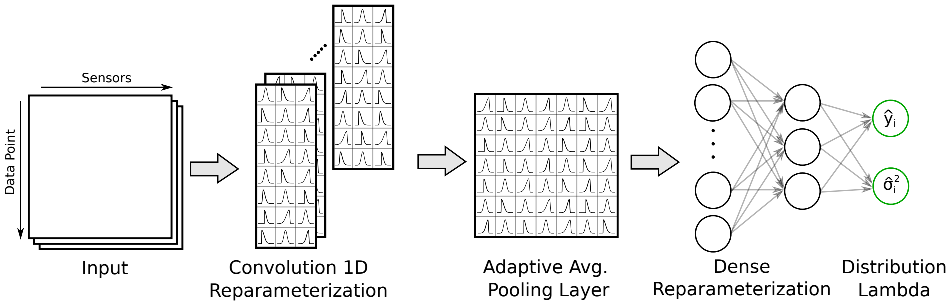


Figure 3. Schematic of the Bayesian convolution neural network.

convolutional layers and Bayesian dense layers.

- Dense Reparameterization: this layer implements a reparameterization estimator for Bayesian variational inference. It implements a stochastic forward pass via sampling from the kernel and bias distributions. This approach improves the robustness of the model, allowing uncertainty estimation in parameter values and supporting probabilistic modeling in deep learning.
- Distribution Lambda: this layer is responsible for producing the final results given the inputs and the learned weights from the previous layers. The output layer consists of two neurons representing the mean, $\hat{y}$ and variance, $\hat{\sigma}^2$, in order to quantify the expected value and its associated uncertainty. To ensure a positive variance, the neuron is activated using an exponential function.

BCNN combines feature extraction capabilities of classical CNN models with the uncertainty quantification of Bayesian theory. The proposed architecture is built using the Bayesian layers of `TensorFlow Probability` in Python (Dillon et al., 2017).

2.1.2. Training for Diversity

Model diversity is a key concept for effective ensemble models (Nam, Yoon, Lee, & Lee, 2021). Accordingly, in this case, the training set for each battery model is modified to learn different battery aging properties. Historical capacity fading data are used to build aging models for each battery in the dataset.

Namely, using the LOOCV strategy, if K run-to-failure trajectories are available, K diverse BCNN models are built changing the training set in each iteration (cf. Figure 2). That is, the model is trained on all batteries except one, which is held as a test set. This process is repeated so that each battery serves as a test set exactly once. Thus, all available data are used for training, maximizing the diversity of training scenarios.

Training the BCNN models through LOOCV strategy, enhances the ability of individual models to generalize across different battery types and manufacturing conditions.

This stage completes the offline training process, which results in a set of BCNN models:

$$\mathcal{M} = \{BCNN_1, BCNN_2, \ldots, BCNN_K\}, \qquad (1)$$

which are used in the subsequent online inference process to build ensemble models.

2.2. Online: Stacking of Predictive Distribution

During the online phase, the proposed stacking of predictive distribution strategy is designed and tested. The proposed approach takes as input individual base models [cf. Eq. (1)] and monitored data up to the prediction instant t, which is used to forecast the probability density function (PDF) of the capacity at $t+1$, $\hat{y}_{PDF}(t+1)$. The objective of the stacking process is to integrate the predictive distributions of different base models and propagate all the information end-to-end.

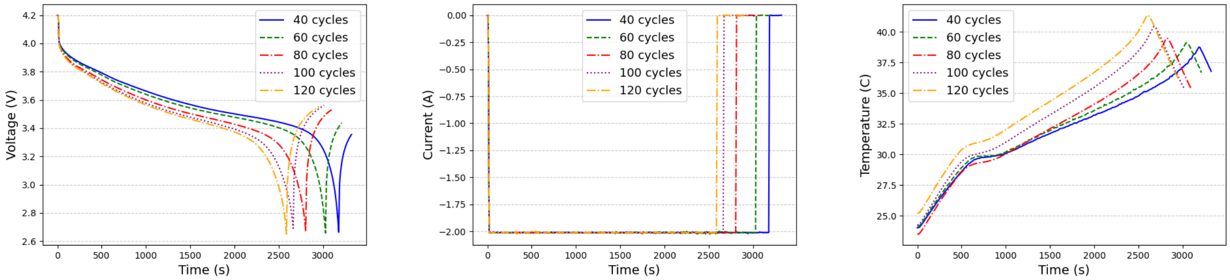
For comparison and benchmarking purposes, an alternative stacking approach is also implemented named stacking of point prediction (cf. Subsection 3.3).

Log-Score Weights

The optimal way to combine a set of Bayesian posterior predictive distributions is by using the logarithmic score (Yao, Vehtari, Simpson, & Gelman, 2018). This method maximizes the average log-likelihood of the observed data, which is a proper scoring rule used to evaluate the accuracy of probabilistic forecasts. It measures the accuracy of a forecast and penalizes overconfidence and underconfidence in the predicted probability. The logarithmic score is defined as follows:

$$\hat{w} = \arg\max_{w} \frac{1}{N} \sum_{i=1}^{N} log \sum_{k=1}^{K} w_k p(y_i \mid y_{-i}, M_k) + \lambda_{reg} \sum_{k=1}^{K} w_k^2 \quad (2)$$

where N denotes the total number of data points and K denotes the total number of base models. The leave-one-out predictive distribution for each model, *i.e.* $p(y_i \mid y_{-i}, M_k)$, is used to compute the model's prediction for the data point i. To avoid overfitting, a regularization term λ_{reg} is added to the likelihood function, penalizing large weights.



(a) Voltage variation

(b) Current variation

(c) Temperature variation

Figure 4. Feature variations due to an increasing number of discharge cycles in battery #5.

Stacking

Stacking is a method to average point estimates from several models (LeBlanc & Tibshirani, 1996). In its simplest form, it can be seen as a weighted average method. Through the weighted average, it facilitates the construction of ensembles that incorporate predictions from multiple models. In the proposed framework, the goal of weighted average ensemble is to leverage the predictive capabilities of K pre-trained BCNN models [cf. Eq. (1)]. It seeks to mitigate forecasting errors by assigning weights to the linear combination of these models, thereby enhancing the accuracy of predictions.

In the Bayesian framework, stacking extends beyond the limitations of averaging point predictions by combining multiple Bayesian posterior predictive distributions. This approach develops a *stacking model* that leverages the strengths of various predictive models, enhancing overall predictive accuracy. The stacking of the predictive distribution enables the fusion of uncertainties from various models into a unified predictive framework. This approach improves the accuracy of forecasts and offers a comprehensive evaluation of the uncertainty associated with these forecasts, providing advantages across diverse decision-making scenarios. The fundamental equation governing this process is defined as follows:

$$\hat{p}(\tilde{y}|y) = \sum_{k=1}^{K} \hat{w}_k p(\tilde{y}|y, M_k) \tag{3}$$

where $\hat{p}(\tilde{y}|y)$ represents the aggregate probability estimation based on the ensemble model, ω_k denotes the weight assigned to the k-th component within the ensemble, and $p(\tilde{y}|y, M_k)$ refers to the probabilistic forecast generated by each base model, denoted as $\mathrm{BCNN_k}$, given the observed data y.

This probabilistic prediction indicates the likelihood of observing the predicted outcome $\tilde{y}$, dependent on the specific base model employed.

2.2.1. Forecasting

Online forecasting is computed for one-step-ahead predictions. In order to forecast battery capacity at instant $t+1$, previous data until the instant t is used, plus an uncertainty factor expressed as noise:

$$\mathcal{X}(t) = \{V(t), T(t), \epsilon\} \tag{4}$$

where $\{V(t), T(t)\}$ denote the values of voltage and temperature at instant t, and ϵ denotes the Gaussian noise term, $N(0, \sigma)$ with $\sigma = 0.1$, that introduces variability in the progression of X over time.

The one-step-ahead capacity distribution prediction is thus defined as follows:

$$\hat{y}_{PDF}(t+1) = f(\mathcal{X}(t)) \tag{5}$$

where $f(.)$, denotes the designed ensemble model, $\hat{y}_{PDF}(t+1)$ is the distribution of the capacity estimate at $t+1$.

It is possible to perform SOH predictions for longer prediction horizons through a recursive forecasting strategy. However, due to the accumulation of individual forecasting errors, this approach may lead to decrease long-term forecasting performance. Long-term SOH forecasting activities are left open for future work.

This approach allows the model to learn continuously and adapt to changing conditions. Online forecasting is particularly beneficial in environments that require immediate decision making based on the latest available data.

3. Case study

3.1. Dataset description

The effectiveness of the proposed method has been tested using a battery dataset from the NASA Ames Prognostics Center of Excellence (Saha & Goebel, 2007).

A subset of available battery data has been selected, focusing on batteries #5, #6, #7 and #18. Each battery is operated under various conditions including charging, discharging, and impedance analysis. Throughout the charge and discharge cycles, temperature, current, and voltage were meticulously recorded. During charging, a constant current mode at 1.5 A was maintained until the voltage reached 4.2 V, followed by a switch to constant voltage mode until the current dropped to 20 mA. Discharge cycles involved a constant load mode at 2 A until the voltage levels reached 2.7 V, 2.5 V, 2.2 V and 2.5 V for batteries #5, #6, #7 and #18, respectively. The experiment ended once the battery capacity decreased by 30%. These batteries had a maximum capacity of 2Ah with an end-of-life capacity set at 1.4Ah.

Figures 4(a), 4(b) and 4(c) show the evolution of voltage, current (constant), and temperature measurements with the increment of discharge cycles for the battery #5. Figure 5 shows variations in capacity degradation rates for identical batteries. This is an indicator of uncertainty inherent in the manufacturing process, which affects SOH estimates.

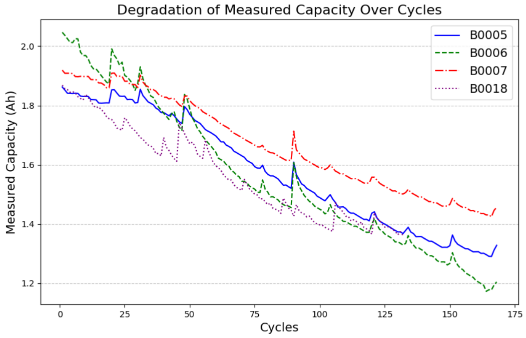


Figure 5. Capacity degradation data of Li-ion batteries.

3.2. BCNN structure and hyperparameters

The design of the base BCNN model structure is developed through experimentation. The BCNN architecture for SOH forecasting is detailed in Table 1, where 'None' is indicative of the batch size. The input for the model comprises 371 data points per discharge cycle, with each point aggregating 3 features: voltage, temperature, and time.

The proposed structure encompasses a total of 1300 trainable parameters, designed to extract features from battery discharge cycle data for forecasting purposes. Figure 3 details the convolutional layer hyperparameters, which includes 16 kernels, each with a dimension of 3, adopting a Laplace distribution for the prior and employing a ReLU activation function. In addition, the model incorporates Bayesian dense layers with 16 units, Adam optimizer, a learning rate of 0.01, and Evidence Lower Bound (ELBO) as its loss function (S. Zhang et al., 2022).

Table 1. BCNN model architecture

Layer	Description	Output Shape	# Param.
-	Input	(None, 371, 4)	0
1	Conv.1D Reparameter.	(None, 369, 16)	416
2	Conv.1D Reparameter.	(None, 368, 8)	528
3	Global Average Pooling	(None, 8)	0
4	Flatten	(None, 8)	0
5	Dense Reparameter.	(None, 16)	288
6	Dense Reparameter.	(None, 2)	68
7	Distribution Lambda	(None,1),(None,1)	0
Total params: 1300 (5.08 KB)			

3.3. Benchmarking

In order to compare the designed stacking approach with alternative stacking strategies, another stacking approach has been designed using point prediction information instead of the full distribution.

Stacking of Point Prediction

An effective method for determining the weight of each model in the stacking process is by minimizing the leave-one-out mean squared error with a L_2 regularization term, λ_{reg}. The purpose of this term is to penalize large weights, thus preventing overfitting and balancing individual model contributions. The weights are obtained through the following optimization problem:

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} w_k \hat{f}_K^{(-i)}(x_i) \right)^2 + \lambda_{reg} \sum_{k=1}^{K} w_k^2 \quad (6)$$

where $\hat{f}_K^{(-i)}(x_i)$ represents the predicted value of the k-th model, when the i-th observation is left out of the training set. The regularization parameter, λ_{reg}, controls the strength of the regularization applied. To ensure a feasible solution, the weights are restricted to $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$.

Accordingly, the stacking of point prediction approach is defined as follows:

$$\hat{y} = \sum_{k=1}^{K} \hat{w}_k f_k(x|\theta_k) \quad (7)$$

where $\hat{y}$ represents the prediction of the ensemble for the test battery capacity, $\hat{w}_k$ denotes the weight assigned to the k-th battery base model, and $f_k(x|\theta_k)$ is the prediction made by the corresponding base model (BCNN$_k$).

3.4. Evaluation criteria

The accuracy of the regression is measured by Mean Squared Error, while Negative Log Likelihood assesses model performance by quantifying prediction probabilities. Finally, The correctness of probability predictions is assessed through the CRPS.

Mean Square Error (MSE) is a metric for measuring the quality of an estimator. It is a measure of the average squared differences between the estimated values and what is estimated. MSE is calculated by taking the average of the square of the differences between the predicted values and the actual values (Hodson, 2022).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{8}$$

where, n represents the number of observations, Y_i denotes the actual value for the ith observation, and $\hat{Y}_i$ signifies the predicted value for the ith observation.

Coefficient of Determination (R^2) is a metric used to assess the goodness of fit of the model. It provides a measure of how well the observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model (Barrett, 1974).

$$R^2 = 1 - \frac{\sum\limits_{i=1}^{n}(Y_i - \hat{Y}i)^2}{\sum\limits_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{9}$$

where, n is the number of observations, Y_i is the actual value, $\hat{Y}_i$ the predicted value for the i-th observation and $\bar{Y}$ the mean of Y. R^2 of 1 implies perfect model predictions, while 0 means no explained variability.

Continuous Ranked Probability Score (CRPS) can be formally expressed as a quadratic measure of discrepancy between the predicted Cumulative Distribution Function (CDF), $F(\cdot)$, and the observed empirical CDF for a given scalar observation y (Zamo & Naveau, 2018):

$$CRPS(F, y) = \int (F(x) - \mathbb{I}(x \geq y_i))^2 dx, \tag{10}$$

where $\mathbb{I}(x \geq y_i)$ is the indicator function, which models the empirical CDF.

To obtain a single score value from Eq. (10), a weighted average is calculated for each individual observation of the test set (Gneiting, Raftery, Westveld, & Goldman, 2005):

$$CRPS = \frac{1}{N}\sum_{i=1}^{N} CRPS(F_i, y_i) \tag{11}$$

where N denotes the total number of predictions.

Negative Log Likelihood (NLL) metric assesses probabilistic models by using the likelihood concept, which indicates how likely the observed data is given model parameters (Bosman & Thierens, 2000). Likelihood ($\mathcal{L}$) is the product of each observation's probability density function (PDF), expressed mathematically as

$$\mathcal{L}(\theta \mid X) = \prod_{i=1}^{N} f(x_i|\theta) \tag{12}$$

where θ denotes model parameters and X includes N data points. NLL is preferred for optimization since minimizing NLL is equivalent to maximizing the log-likelihood, facilitating the discovery of model parameters that best explain the observed data, represented by

$$-\log \mathcal{L}(\theta \mid X) = -\sum_{i=1}^{n} \log f(x_i \mid \theta) \tag{13}$$

Calibration refers to the statistical consistency between the predictive distributions and the actual observations. It represents a joint property of forecasts and empirical data (Jung, Jo, Choo, & Lee, 2022). Namely, it is stated that the model is calibrated if (Kuleshov, Fenner, & Ermon, 2018):

$$\frac{\sum_{t=1}^{T} \mathbb{I}\{y_t \leq F_t^{-1}(p)\}}{T} \to p \text{ for all } p \in [0, 1] \tag{14}$$

In this expression, T refers to the total number of data points, while the indicator function $\mathbb{I}\{y_t \leq F_t^{-1}(p)\}$ takes a value of 1 when the condition $y_t \leq F_t^{-1}(p)$ is true, and 0 otherwise. Given this condition, y_t express the observed outcome at time t, and $F_t^{-1}(p)$ is the inverse of the CDF for the forecast, evaluated at probability p. Therefore, the condition represents the threshold below which a random sample from the distribution would occur with a probability p.

Sharpness means that the confidence intervals should be optimized for minimal width around a singular value. That is, the goal is to reduce the variance, denoted as $var(F_n)$, of the random variable characterized by the cumulative distribution function F_n (Kuleshov et al., 2018; Tran et al., 2020):

$$sha = \sqrt{\frac{1}{N}\sum_{n=1}^{N} var(F_n)} \tag{15}$$

Table 2. Comparison of different ensemble strategies for different batteries used as test.

	Baseline Model				Benchmarking Ensemble				Proposed Ensemble			
	$MSE(\downarrow)$	$R^2(\uparrow)$	$NLL(\downarrow)$	$CRPS(\downarrow)$	$MSE(\downarrow)$	$R^2(\uparrow)$	$NLL(\downarrow)$	$CRPS(\downarrow)$	$MSE(\downarrow)$	$R^2(\uparrow)$	$NLL(\downarrow)$	$CRPS(\downarrow)$
B0005	0.0007	0.9732	2.3397	0.0183	**0.0002**	**0.9901**	-1.9523	0.0145	0.0003	0.9886	**-2.1001**	**0.0131**
B0006	0.0013	0.9636	8.0947	0.0213	**0.0009**	**0.9753**	-1.8222	0.0183	0.0009	0.9741	**-1.9358**	**0.0178**
B0007	0.0005	0.9696	-0.0409	0.0149	**0.0003**	**0.9814**	**-1.9755**	**0.0145**	**0.0004**	**0.9763**	**-1.9769**	**0.0145**
B0018	0.0013	0.8943	9.0342	0.0223	**0.0010**	**0.9183**	**-1.9478**	**0.0174**	**0.0010**	**0.9141**	**-1.9312**	**0.0178**

4. Results

To evaluate the proposed approach, firstly, different ensemble strategies are compared to evaluate their strengths and identify the most suitable approach. Subsequently, a sensitivity analysis is developed with respect to the contribution of individual base-models to the overall ensemble.

4.1. Probabilistic Ensemble Strategies

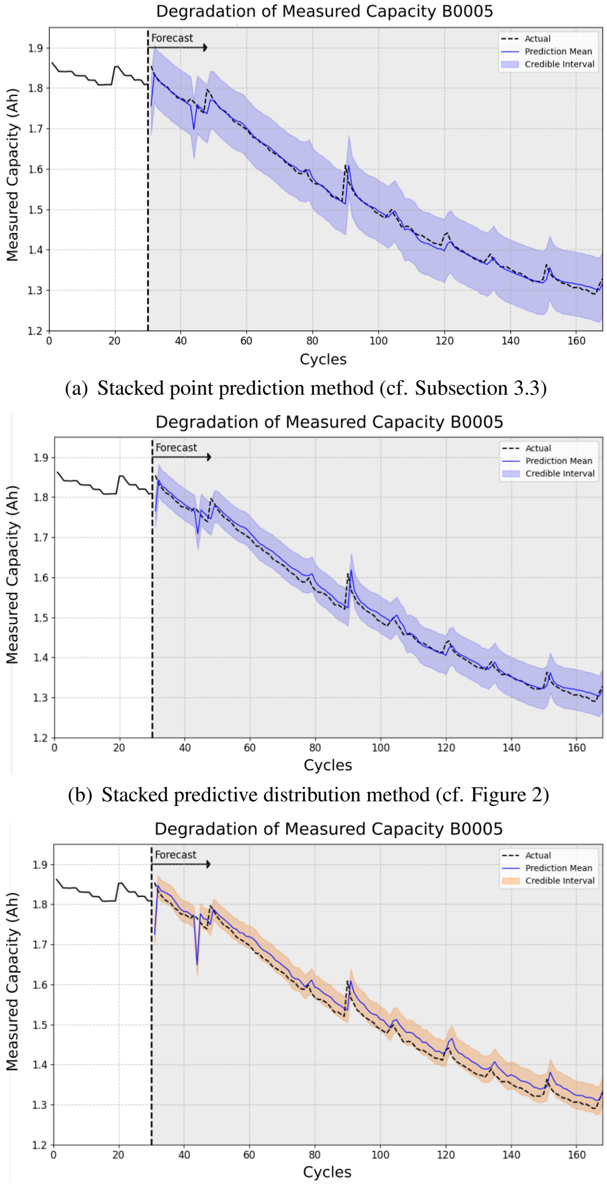
This section focuses on the comparison between (i) the baseline model, *i.e.* BCNN model trained with all available data, (ii) ensemble of point prediction and (iii) proposed ensemble method (cf. Figure 2) to further evaluate the improvement of ensemble strategies over baseline model.

Table 2 presents a comparative analysis in terms of accuracy and probabilistic metrics. This comparison highlights that, for different test scenarios, the ensemble methodologies enhance the performance of the baseline model.

A notable observation from the results in Table 2 is the variance between the proposed ensemble approach (cf. Figure 2) and the benchmarking ensemble model (cf. Subsection 3.3) in specific scenarios. For batteries #5 and #6, the proposed approach exhibited superior outcomes, particularly in probabilistic metrics (NLL and CRPS). This suggests that within a Bayesian framework, prioritizing likelihood maximization, leads to accurately modelling uncertainty, and therefore, it is more advantageous than focusing on MSE minimization (as in Subsection 3.3).

The model optimization criterion has a direct impact on the performance of the tested methods and on the effectiveness of the ensemble approach. However, for batteries #7 and #18, no significant differences were observed between the tested ensemble approaches, which indicates that the results are associated to the prior models. That is, it is possible that the same prior model minimizes the MSE and maximizes the likelihood at the same time.

Figure 6(a) shows the comparison between the ensemble model generated by stacking point predictions (cf. Subsection 3.3), Figure 6(b) shows the ensemble model generated through stacking of predictive distributions (cf. Figure 2), and Figure 6(c) shows the individual BCNN trained with the entire dataset, e.g. for the battery #5, train with batteries #6, #7, and #18, and test with #5.



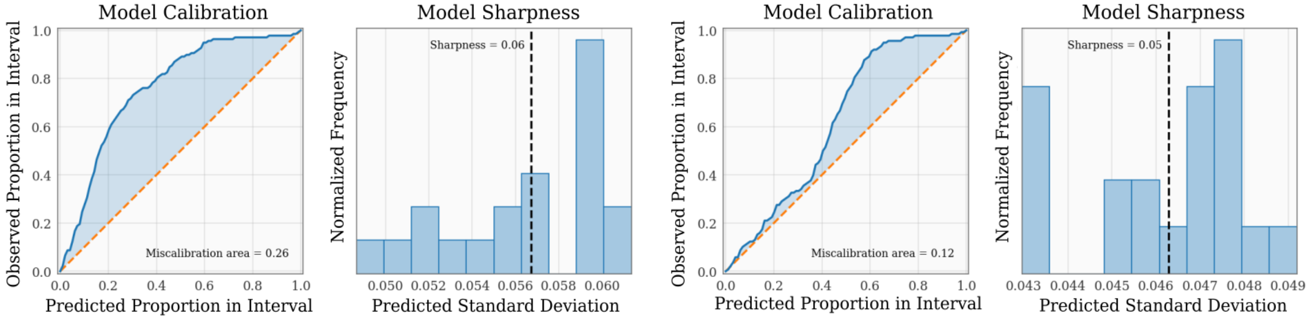
(a) Stacked point prediction method (cf. Subsection 3.3)

(b) Stacked predictive distribution method (cf. Figure 2)

(c) Baseline model

Figure 6. Battery capacity degradation forecasting results.

It is observed that the ensemble models enhance the performance of baseline model in terms of accuracy and uncertainty



(a) calibration and sharpness for the benchmarking ensemble model

(b) calibration and sharpness for the proposed ensemble model

Figure 7. Evaluation of calibration and sharpness for battery #5.

quantification. This is indicated by the positioning of the ground truth (dashed lines) at the limit of the lower boundary in Figure 6(c), which means that the uncertainty does not accurately cover the observed values. That is, the uncertainty bounds are not well-calibrated, compromising the model's ability to accurately represent the underlying variability in the data and in the model compared to ensemble strategies.

Figure 6(a) shows an improvement in the prediction accuracy. However, it simultaneously introduces a higher level of uncertainty compared to the proposed ensemble method in Figure 6(b). This is reflected in the NLL and CRPS metrics, where the stacking of the predictive distribution demonstrates superior performance (cf. Table 2). Such probabilistic metrics indicate that the model parameters make the observed data more probable, indicating a good fit to the observed data.

The evaluation of the shape of the PDF is a crucial aspect of uncertainty quantification. Accordingly, the calibration and the sharpness assessment of PDFs is performed through a python toolbox for predictive uncertainty quantification (Chung, Char, Guo, Schneider, & Neiswanger, 2021). Figure 7 shows the calibration and sharpness of the analysed ensemble methods designed for probabilistic forecasting for the battery #5.

The calibration plot for the point-prediction ensemble model [cf. Figure 7(a)] reveals a miscalibration area of 0.26, indicating a gap between predicted probabilities and actual outcomes, generally overestimating event probabilities. On the contrary, the proposed ensemble model [cf. Figure 7(b)] shows better calibration with a miscalibration area of 0.12, aligning closer to the ideal, especially in midrange probabilities.

In terms of sharpness, the predictions of the point-prediction based ensemble model have a mean sharpness value of 0.06 and are right-skewed, reflecting higher uncertainty. However, the proposed ensemble model has a mean sharpness value of 0.05, with a slightly left-skewed distribution, indicating more predictions with lower uncertainty and greater confidence.

4.2. Sensitivity of the Ensemble Strategy with Base-Models

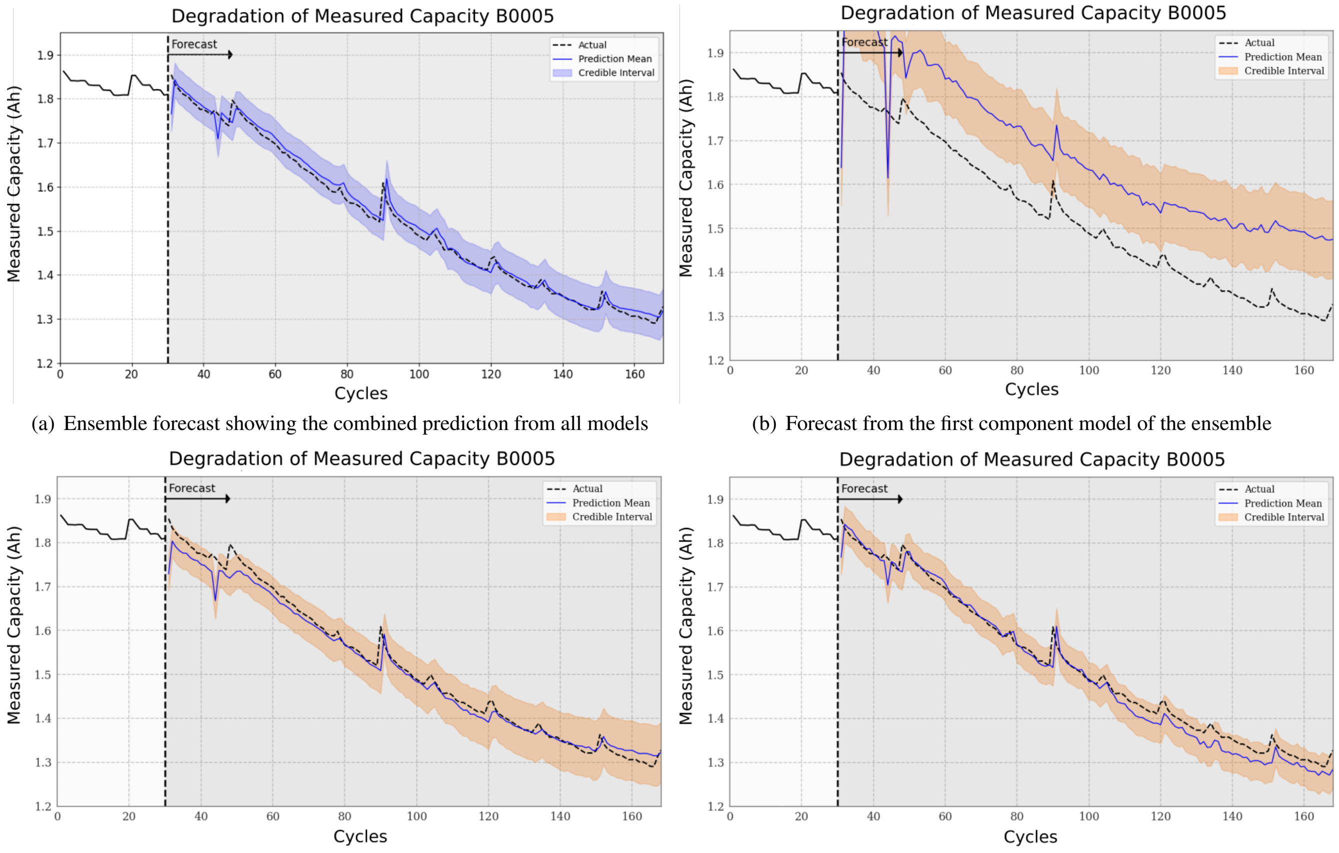
To evaluate the contribution of each individual BCNN model to the ensemble approach, a sensitivity assessment has been performed. Namely, the performance of the different leave-one-out iterations has been evaluated, sequentially training with different battery datasets and testing with the leave-out battery dataset. This has been compared with the proposed ensemble approach results to identify individual contributions from different models. Table 3 displays the obtained results.

Table 3. Performance evaluation of BCNN models and the ensemble approach.

Test[1]	Model	*MSE* (↓)	R^2 (↑)	*NLL* (↓)	*CRPS* (↓)
#5	BCNN [#6,#7][2]	0.0005	0.9802	-1.0707	0.0135
	BCNN [#6,#18]	0.0244	0.1016	19.4417	0.1411
	BCNN [#7,#18]	0.0006	0.9795	-2.0774	0.0132
	Ensemble	**0.0003**	**0.9886**	**-2.1001**	**0.0131**
#6	BCNN [#5,#7]	0.0011	0.9695	3.7012	0.0197
	BCNN [#5,#18]	0.0147	0.5861	0.5852	0.0849
	BCNN [#7,#18]	0.0018	0.9491	-0.7498	0.0252
	Ensemble	**0.0009**	**0.9741**	**-1.9358**	**0.0178**
#7	BCNN [#5,#6]	0.0008	0.9543	-1.5462	0.0166
	BCNN [#5,#18]	0.004	0.7704	2.1996	0.0326
	BCNN [#6,#18]	0.0026	0.854	-1.5735	0.0286
	Ensemble	**0.0004**	**0.9763**	**-1.9769**	**0.0145**
#18	BCNN [#5,#6]	0.0091	0.2534	14.708	0.0833
	BCNN [#5,#7]	0.0041	0.6663	1.5441	0.0459
	BCNN [#6,#7]	0.0013	0.8929	1.8299	0.0213
	Ensemble	**0.0010**	**0.9141**	**-1.9312**	**0.0178**

[1] Battery identifier used for testing.
[2] BCNN [#A,#B]: BCNN trained with batteries #A and #B.

The ensemble BCNN model demonstrates significantly higher accuracy and predictive power than individual BCNN models, as evidenced by its superior performance across multiple metrics. It achieves the lowest MSE in every testing battery, indicating more precise predictions, and the highest R^2 score, showing its ability to explain a greater proportion of variance.



(a) Ensemble forecast showing the combined prediction from all models

(b) Forecast from the first component model of the ensemble

(c) Forecast from the second component model of the ensemble

(d) Forecast from the third component model of the ensemble

Figure 8. Capacity fade forecasting for battery #5 employing an ensemble of BCNN models.

The ensemble model also shows a notable improvement in the NLL metric, suggesting a more reliable uncertainty estimation. Additionally, by achieving the lowest CRPS, it emphasizes its proficiency in probabilistic forecasting and precise uncertainty quantification. Overall, the ensemble method outperforms individual models, highlighting its effectiveness in contexts that require high accuracy and reliability.

Figure 8 presents the forecasts generated by individual models for battery #5 (cf. Table 3). Figures 8(b)-8(d), show individual models and Figure 8(a) shows the combined forecast of the ensemble model.

It can be seen that the ensemble effectively combines the characteristics of models 2 and 3, thereby improving the overall performance of the final forecast of the ensemble.

5. Discussion

The proposed research work demonstrates that the stacking of predictive distributions based on a Bayesian framework improves the accuracy and robustness of predictions compared with stacking of point predictions. Furthermore, it has been observed that the use of an ensemble of BCNN models improves the modeling of uncertainty when compared to relying on a single BCNN model (baseline). However, before drawing definitive conclusions about the application of the proposed solution in real-world applications, further work is necessary testing the robustness, scalability, and sensitivity with respect to noise.

Robustness

Credible intervals reflect the uncertainty associated with the data and the model (cf. Figure 6). The robustness of the proposed approach is therefore directly dependent on model and data uncertainty. The reduction of credible intervals align with the objective of increasing robustness. To this end, increasing the number of observations would reduce the uncertainty attributed to the model, which results in more precise credible intervals. Additionally, employing priors like maximum entropy priors or weakly informative priors may further tighten credible intervals, thereby improving the reliability of the model predictions.

Scalability

To analyze larger fleets of batteries, instead of using leave-one-out methodologies, it may be more appropriate to develop generalized training methodologies. In this direction, one approach would be to cluster batteries that exhibit similar operation and degradation conditions. This strategy would enable capturing data diversity, which is a key property for ensemble strategies. Alternatively, a hierarchical modelling strategy may be adopted. This method involves a global model for overall battery behavior, supplemented by smaller models for specific groups, enabling precise adaptations without the need for separate models per battery. This strategy ensures scalability and flexibility in handling various battery operation and degradation conditions efficiently.

Noise Sensitivity

The proposed approach assumes a Gaussian noise to model the variability of the modeled process and measurements [cf. Eq. (4)]. To analyze the impact of Gaussian noise levels on prediction results, a sensitivity analysis has been performed. Figure 9 shows the obtained results.

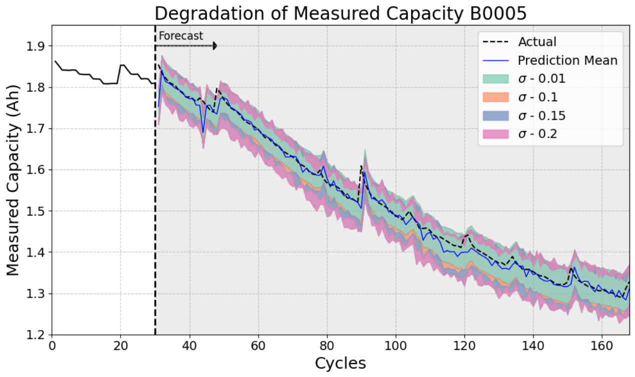


Figure 9. Impact of Gaussian Noise on Predictive Modeling of Battery Capacity Degradation.

Obtained results indicate that, when testing data diverges from training data, the epistemic uncertainty increases. The increase in Gaussian noise causes a greater deviation, and therefore, there is a significant rise in epistemic uncertainty. Analysing the model's behaviour in the presence of different types of uncertainty is crucial to evaluate the robustness of the model and determine if additional training stages are needed to enhance its reliability. Consequently, this research adopts a noise level of 0.1 as a trade-off decision between prediction accuracy and uncertainty.

Application Limits

Some of the adopted practices may limit the applicability of the proposed framework in real-world applications. The experimental setup, conducted in a controlled environment with specified load conditions, may not entirely replicate the diverse sources of uncertainty present in real-world applications. Such controlled conditions could potentially skew the understanding of uncertainty due to environmental and operational variabilities. Consequently, the predictive performance observed in this study may differ under less predictable conditions. In this direction, for controlled operation environments, the complexity of the proposed approach may be reduced. However, the proposed methodology complexity is designed to capture a wide range of uncertainties found in real operating systems.

6. Conclusion and future work

Batteries are key components in power and energy systems and ensuring a robust and reliable remaining useful life (RUL) prediction of batteries is crucial to develop accurate monitoring strategies, and build cost-effective solutions.

In this context, battery RUL prediction models generally focus on individual prediction models. They may be able to capture uncertainty associated with the battery ageing process, but the uncertainty modelling and capturing ability is also limited to the individual model. This research presents a probabilistic ensemble prognostics approach which combines Bayesian Convolutional Neural Network (BCNN) models in a probabilistic stacking strategy. The proposed framework leverages the probabilistic predictive information of individual BCNN models, which are integrated through a probabilistic stacking approach that calibrates between accuracy and robustness of probabilistic predictions.

The proposed approach has been tested on NASA's battery dataset. Obtained results show that the proposed probabilistic stacking approach improves accuracy and uncertainty of predictions with respect to other ensemble strategies and individual BCNN models.

This research study contributes towards understanding and predicting the capacity fade in Li-ion batteries. Namely, it highlights the role of probabilistic approaches and ensemble methods in modelling the uncertainties inherent in battery manufacturing and operation.

Looking forward, there are different opportunities to expand the scope and applicability of this work. On the one hand, the use of a larger battery dataset, which includes diverse environmental and operational conditions, would allow for a more comprehensive understanding of capacity fade across various scenarios. On the other hand, it may be possible to perform a more exhaustive comparative analysis of different fusion strategies, including Bayesian Model Averaging, Pseudo Bayesian Model Averaging, or Mixture Models. This comparative will provide further insights into the optimal approaches for integrating predictive models in the context of battery life prediction, enhancing both the accuracy and reliability of capacity fade forecasts.

Acknowledgenemts

This publication is part of the research projects KK-2023-00041, IT1451-22 and IT1676-22 funded by the Basque Government. J. I. Aizpurua is funded by Juan de la Cierva Incorporacion Fellowship, Spanish State Research Agency (grant No. IJC2019-039183-I).

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., … Nahavandi, S. (2021, December). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297. doi: 10.1016/j.inffus.2021.05.008

Bai, G., & Chandra, R. (2023, November). Gradient boosting Bayesian neural networks via Langevin MCMC. *Neurocomputing*, *558*, 126726. doi: 10.1016/j.neucom.2023.126726

Barrett, J. P. (1974). The coefficient of determination—some limitations. *The American Statistician*, *28*(1), 19–20.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In F. Bach & D. Blei (Eds.), *International conference on machine learning* (Vol. 37, pp. 1613–1622). Lille, France: PMLR.

Bosman, P. A., & Thierens, D. (2000). Negative log-likelihood and statistical hypothesis testing as the basis of model selection in ideas. In *Proceedings of the tenth dutch–netherlands conference on machine learning. tilburg university.*

Che, Y., Zheng, Y., Forest, F. E., Sui, X., Hu, X., & Teodorescu, R. (2024, January). Predictive health assessment for lithium-ion batteries with probabilistic degradation prediction and accelerating aging detection. *Reliability Engineering & System Safety*, *241*, 109603. doi: 10.1016/j.ress.2023.109603

Chung, Y., Char, I., Guo, H., Schneider, J., & Neiswanger, W. (2021). Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*.

Cobb, A. D., Himes, M. D., Soboczenski, F., Zorzan, S., O'Beirne, M. D., Baydin, A. G., … Angerhausen, D. (2019, June). An Ensemble of Bayesian Neural Networks for Exoplanetary Atmospheric Retrieval. *The Astronomical Journal*, *158*(1), 33. doi: 10.3847/1538-3881/ab2390

Dai, H., Pollock, M., & Roberts, G. O. (2023, February). Bayesian fusion: Scalable unification of distributed statistical analyses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(1), 84–107. doi: 10.1093/jrsssb/qkac007

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., … Saurous, R. A. (2017, November). *TensorFlow Distributions* (No. arXiv:1711.10604). arXiv.

Fan, Y., Olson, R., & Evans, J. P. (2017). A bayesian posterior predictive framework for weighting ensemble regional climate models. *Geoscientific Model Development*, *10*(6), 2321–2332.

Gneiting, T., Raftery, A. E., Westveld, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098–1118. doi: 10.1175/MWR2904.1

Hadigol, M., Maute, K., & Doostan, A. (2015). On uncertainty quantification of lithium-ion batteries: Application to an lic6/licoo2 cell. *Journal of Power Sources*, *300*, 507–524.

Hodson, T. O. (2022). Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, *2022*, 1–10.

Jung, Y., Jo, H., Choo, J., & Lee, I. (2022, June). Statistical model calibration and design optimization under aleatory and epistemic uncertainty. *Reliability Engineering & System Safety*, *222*, 108428. doi: 10.1016/j.ress.2022.108428

Kuleshov, V., Fenner, N., & Ermon, S. (2018, July). Accurate uncertainties for deep learning using calibrated regression. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2796–2804). PMLR.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641–1650.

Lee, G., Kwon, D., & Lee, C. (2023). A convolutional neural network model for SOH estimation of Li-ion batteries with physical interpretability. *Mechanical Systems and Signal Processing*, *188*, 110004. doi: 10.1016/j.ymssp.2022.110004

Liu, Y., Sun, J., Shang, Y., Zhang, X., Ren, S., & Wang, D. (2023, May). A novel remaining useful life prediction method for lithium-ion battery based on long short-term memory network optimized by improved sparrow search algorithm. *Journal of Energy Storage*, *61*, 106645. doi: 10.1016/j.est.2023.106645

Nam, G., Yoon, J., Lee, Y., & Lee, J. (2021). Diversity matters when learning from ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 8367–8377). Curran Associates, Inc.

Nemani, V., Biggio, L., Huan, X., Hu, Z., Fink, O., Tran, A., … Hu, C. (2023). Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, *205*, 110796. doi: 10.1016/j.ymssp.2023.110796

Saha, B., & Goebel, K. (2007). Nasa ames prognostics data repository. *NASA Ames, Moffett Field, CA, USA*.

Toughzaoui, Y., Toosi, S. B., Chaoui, H., Louahlia, H., Petrone, R., Le Masson, S., & Gualous, H. (2022). State of health estimation and remaining useful life assessment of lithium-ion batteries: A comparative study. *Journal of Energy Storage*, *51*, 104520. doi: 10.1016/j.est.2022.104520

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., & Ulissi, Z. W. (2020, May). Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, *1*(2), 025006. doi: 10.1088/2632-2153/ab7e1a

Vanem, E., Salucci, C. B., Bakdi, A., & Alnes, Ø. Å. S. (2021, November). Data-driven state of health modelling—A review of state of the art and reflections on applications for maritime battery systems. *Journal of Energy Storage*, *43*, 103158. doi: 10.1016/j.est.2021.103158

Wang, C.-j., Zhu, Y.-l., Gao, F., Bu, X.-y., Chen, H.-s., Quan, T., … Jiao, Q.-j. (2022). Internal short circuit and thermal runaway evolution mechanism of fresh and retired lithium-ion batteries with lifepo4 cathode during overcharge. *Applied Energy*, *328*, 120224.

Wei, Y., & Wu, D. (2023, February). Prediction of state of health and remaining useful life of lithium-ion battery using graph convolutional network with dual attention mechanisms. *Reliability Engineering & System Safety*, *230*, 108947. doi: 10.1016/j.ress.2022.108947

Yang, Y., Chen, S., Chen, T., & Huang, L. (2023). State of health assessment of lithium-ion batteries based on deep gaussian process regression considering heterogeneous features. *Journal of Energy Storage*, *61*, 106797.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018, September). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3). doi: 10.1214/17-BA1091

Zamo, M., & Naveau, P. (2018, February). Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences*, *50*(2), 209–234. doi: 10.1007/s11004-017-9709-7

Zhang, S., Liu, Z., & Su, H. (2022). A bayesian mixture neural network for remaining useful life prediction of lithium-ion batteries. *IEEE Transactions on Transportation Electrification*, *8*(4), 4708–4721. doi: 10.1109/TTE.2022.3161140

Zhang, Y., Zhang, M., Liu, C., Feng, Z., & Xu, Y. (2024, February). Reliability enhancement of state of health assessment model of lithium-ion battery considering the uncertainty with quantile distribution of deep features. *Reliability Engineering & System Safety*, 110002. doi: 10.1016/j.ress.2024.110002

Zhao, H., Chen, Z., Shu, X., Shen, J., Lei, Z., & Zhang, Y. (2023). State of health estimation for lithium-ion batteries based on hybrid attention and deep learning. *Reliability Engineering & System Safety*, *232*, 109066. doi: 10.1016/j.ress.2022.109066

Zhao, X., Wang, Z., Li, E., & Miao, H. (2024, January). Investigation into Impedance Measurements for Rapid Capacity Estimation of Lithium-ion Batteries in Electric Vehicles. *Journal of Dynamics, Monitoring and Diagnostics*. doi: 10.37965/jdmd.2024.475