An Empirical Review of Uncertainty Estimation for Segmentation Quality Control in CAD with Point-Based Graph Neural Networks

Gerico Vidanes, David Toal, Daniel Xu Zhang, Andy Keane, Marco Nunez, Jon Gregory

Computational Engineering and Design Group
University of Southampton

September 11, 2024

Abstract

Deep neural networks are able to achieve high accuracy in automated feature recognition or semantic segmentation of geometries used in computational engineering. Being able to recognise abstract and sometimes hard to describe geometric features has applications for automated simulation, model simplification, structural failure analysis, meshing, and additive manufacturing. However, for these systems to be integrated into engineering workflows, they must provide some measures of predictive uncertainty such that engineers can reason about and trust their outputs. This work presents an empirical study of practical uncertainty estimation techniques that can be used with pre-trained neural networks for the task of boundary representation model segmentation. A point-based graph neural network is used as a base. Monte-Carlo (MC) Dropout, Deep Ensembles, test-time input augmentation, and post-processing calibration are evaluated for segmentation quality control. A point-based input augmentation is presented; which, when combined with MC Dropout, is top performing across 2 publicly available 3D datasets. It was found that the error of a human-in-the-loop system across a dataset can be reduced from 5% to 1% for MFCAD++ and from 34% to 23% for Fusion360 Gallery when 20% of the predictions are flagged for manual correction based on uncertainty estimations.

1 Introduction

Feature recognition (or semantic segmentation) of engineering geometry is a widely useful capability. One of the first applications of this was for the automated transition between computer-aided design (CAD) models and computer-aided manufacturing and process planning (Shah et al. (2000); Al-Wswasi et al. (2018)). Later, feature recognition was also used for automated analysis; where detected features are used to aid in downstream meshing, simulation, and post-processing (Zhang et al. (2014)). With the development and wider use of geometric deep learning within computational engineering (Lambourne et al. (2021); Colligan et al. (2022); Zhang et al. (2022); Vidanes et al. (2024); Cao et al. (2020); Jayaraman et al. (2021)), the recognised features could be more complex and abstract. This opens the door to future use cases like detecting structural failure or features which cause problems in meshing or additive manufacturing.

While the development of the underlying predictive models - neural networks (NN) - is proceeding rapidly in the literature, consideration for how these can be properly integrated into the engineering workflow is lacking. Despite being highly accurate and flexible, these models are not perfect since they are fundamentally statistical models (Goodfellow et al. (2016)). Coupled with their end-to-end nature, the basic system simply presents the engineer with a dense set of semantic segmentation predictions which may or may not be correct. Taking these at face value, errors in recognised features can, for example, lead to errors in the analysis models being built from these tags. In the best case this can cause simulations to fail, and in the worst case can be silent errors which give misleading simulation results. This is exacerbated when an input geometry is outside of their training distribution[1]. In contrast, traditional or algorithmic feature recognition approaches give engineers some confidence in their outputs. Unfamiliar inputs tend to produce runtime errors or simply produce blank labels which can be easily caught downstream.

To combat this and to move towards more robust and useable deep learning systems for engineering workflows, the current work studies NN uncertainty estimation techniques in so far as they can be used to make decisions about NN outputs. Essentially, an uncertainty (or confidence) value can be given to each NN prediction such that it is correlated to the likelihood of its correctness. Predictions that are very uncertain are then likely to be incorrect and thus can be discarded or flagged to the engineer for correction. Work on this area has been increasing, but as the survey from Gawlikowski et al. (2023) has identified, the literature is lacking on the validation of existing methods over real-world problems, especially for the 3D domain.

To the best of the authors' knowledge, this is the first application of uncertainty / confidence estimation techniques to NNs for 3D CAD segmentation or processing in general. Therefore, an empirical review of practical techniques is presented and the implications to engineering workflows is discussed in detail. This work is placed as an initial exploration of the space to be used as a starting point for further detailed research. Additionally, a novel test-time augmentation which involves repeated stochastic encoding of the 3D CAD model into a point cloud is presented as an uncertainty estimation technique. The paper is structured as follows. Section 2 discusses recent work that is directly related and the gaps which the current work fills. Section 3 provides necessary background on the data representation and NN inference. Section 4 discusses the uncertainty estimation techniques being reviewed and how they have been implemented in the current work. The experimental framework is discussed in Section 5, with the results being presented in Section 6. Discussions of the overall results and limitations are presented in Section 7 and conclusions are summarised in Section 8. Finally, some future work is suggested in Section 9.

The specific novel contributions of this work is as follows:

- Presenting an empirical review of uncertainty estimation techniques applied to quality control of semantic segmentation of 3D CAD geometries. Specifically using a point-based GNN.
- Point resampling is proposed as a test-time augmentation for uncertainty estimation.
- MC Dropout (Gal and Ghahramani (2016)) is combined with the point resampling test-time augmentation and is shown to be top performing across two large publicly available datasets.

[1] Whether an input is out-of-ditribution or not is also difficult to know a-priori and is a research area in itself (Yang et al. (2024)). Uncertainty estimation techniques can also be used for this.

2 Related Work

Much work has been done on uncertainty quantification for NNs in general (Gawlikowski et al. (2023)), with an important work in this space being the thesis by Gal (2016) which presents practical techniques and a rich review. Works which use convolutional / bottleneck like networks for computer vision are relevant to this work. One of the first was Kendall et al. (2015) which applied MC Dropout (Gal and Ghahramani (2016)) to a convolutional neural network (CNN) - however, only improvements in semantic segmentation accuracy were evaluated and per-pixel uncertainty masks were simply used for qualitative analysis. Filling this gap, Mukhoti and Gal (2018) present metrics for evaluating uncertainty estimation techniques in terms of how well they (inversely) correlate with semantic segmentation accuracy. They also compare MC Dropout and Concrete Dropout (Gal et al. (2017)) with standard NN inference and show improvement. Another important practical technique is the Deep Ensemble by Lakshminarayanan et al. (2017); they evaluate this technique on a regression task and an image classification task. They present the accuracy of predictions whose uncertainties pass a range of confidence thresholds, and show that the uncertainties estimated by the Ensemble technique is better than MC Dropout. However, they compare using 10 ensemble models and 10 stochastic forward passes for MC Dropout; these are arguably not comparable in terms of computation required (one can easily produce more MC Dropout samples) and more samples would produce better results as shown in Kendall et al. (2015). Gustafsson et al. (2020) also compares the Deep Ensemble technique with MC Dropout and includes semantic segmentation of a street scene as an evaluation task. They utilise a novel metric, *Area Under the Sparsification Error Curve*, which essentially measures the (reverse) correlation between uncertainty and predictive accuracy via prediction ranking. They also show that the Deep Ensemble is better, but they again compare an ensemble of 10 vs 10 MC Dropout samples. Somewhat contrary to the results just discussed is that from Hubschneider et al. (2019) which compares the MC Dropout, Deep Ensemble, and Gaussian Mixture techniques albeit only for a regression task. They show that MC Dropout is best here even with only 10 stochastic forward passes. The current work adds to this applied literature on a novel use case and input representation.

More recently, work has been done on uncertainty estimation for NNs with 3D unstructured inputs, namely point clouds. Vassilev et al. (2024) use the *KPConv* architecture (Thomas et al. (2019)) as a base and compares the standard probability output with MC Dropout and variational inference via parameter sampling. However, they only evaluate using segmentation accuracy and calibration error which is not directly relevant to this work. Petschnigg and Pilz (2021) instead use a *PointNet* architecture (Qi et al. (2017a)) and again compares standard probability with MC Dropout and variational inference. Relevant here is that they evaluate the accuracy of the predictions which pass an uncertainty threshold, similar to the filtering application in the current work. While these are important first steps into practical uncertainty estimation in the 3D domain, the range of techniques validated is lacking.

Worth mentioning at this stage is the work by Guo et al. (2017) which compares different post-processing calibration methods. These aim to transform the output of the NN such that it better reflects the confidence of the prediction. They propose the simple temperature scaling technique here and show that it is the best across many datasets, including 6 image classification and 4 document classification. Calibration and its evaluation is not directly relevant in this work as will be further discussed later but some of these methods will be tested in the current work for completeness. Another interesting work is by Laves et al. (2019) which applies temperature scaling on the individual outputs of MC Dropout inference before aggregating. They show a decrease in error rate when discarding uncertain predictions, across a range of uncertainty thresholds. The current work

performs a similar combined technique for estimating uncertainty - augmenting the input stochastically and performing dropout at test-time.

Finally, the most relevant work is the recent, large-scale, empirical review on a real use case by Ng et al. (2023). They compare Bayes by Backprop (Blundell et al. (2015)), MC Dropout, Deep Ensemble, and Stochastic Segmentation Networks (Monteiro et al. (2020)) as uncertainty estimation techniques for the quality control of NN medical image segmentation. They show that the Deep Ensemble is best. The current work aims to perform a similar empirical review on a range of techniques but for the 3D CAD application.

3 Background

3D feature recognition with deep learning is a wide field due to the different 3D representations available and the diverse applications. There exists approaches for 3D data encoded as voxels (Zhang et al. (2018); Wang et al. (2017)), triangular surface meshes (Hanocka et al. (2019)), and point clouds (Qi et al. (2017a); Thomas et al. (2019); Qi et al. (2017b); Wang et al. (2019c); Zhao et al. (2021)). On the other hand, this work is focused on CAD where geometry tends to be encoded as boundary representation (b-rep) models - a 3D shape is described by its bounding 2D surface. The surface is described by a parametric function $\mathbf{x} : \mathbf{U} \rightarrow \mathbb{R}^3$ - i.e. points on the 2D surface embedded in 3D space ($\mathbb{R}^3$) are given by the function $\mathbf{x}(u, v)$ defined on a rectangular parameter domain $\mathbf{U} \in \mathbb{R}^2$ (Piegl and Tiller (1996)). For non-trivial shapes, their bounding surfaces cannot be described by one parameterisation; therefore, they are composed of many patches or 'b-rep faces' (portions of the domain) with each face bounded by edges which are themselves parametric curves.

A b-rep model is a complex data structure, but approaches have been proposed for processing these with NNs. Colligan et al. (2022), Cao et al. (2020), and Jayaraman et al. (2021) treat the b-rep face topology as a graph and process these with graph convolution. Lambourne et al. (2021) also exploits the topology of the b-rep faces but presents a novel way to perform kernel convolution on this. Alternatively, Zhang et al. (2022) and Vidanes et al. (2024) encode the surfaces as point clouds while still preserving information from the b-rep model.

Regardless of the specific approach and 3D encoding used, the overall process of feature recognition when using NNs is the same. In this work, similar to those above, feature recognition is formulated as semantic segmentation of the input (rather than object detection). This process involves dense prediction where each elementary entity - voxels, pixels, mesh faces, points, or b-rep faces - is classified into a category or otherwise given a label. This prediction takes the form of a score per category - $\mathbf{z} \in \mathbb{R}^K$, where K is the number of classes - often called *logits*. Therefore, the NN forward pass or inference can be formalised as

$$f_\theta : \mathbf{X} \in \mathbb{R}^N \times \mathbb{R}^D \rightarrow \mathbf{Z} \in \mathbb{R}^N \times \mathbb{R}^K$$

where f_θ is the NN parameterised with weights θ, $\mathbf{X}$ is a description of the input as a set of vectors, and $\mathbf{Z}$ is the output giving each of the N entities a logit vector. As an example, $\mathbf{X}$ for a b-rep model could be the set of N b-rep faces each described by D attributes. The argmax within each logit vector then gives the index corresponding to the predicted category.

The current work builds on the NN approach from Vidanes et al. (2024) as its flexibility allows the easy integration and evaluation of a wide range of uncertainty estimation techniques. This relatively simple approach is shown to be competitive with those which directly use the b-rep data. For this, the b-rep model is first encoded into an extended point cloud representation that retains its links with

the b-rep faces - $\mathcal{P} \in \mathbb{R}^N \times \mathbb{R}^{3+D}$, where D is the extra information other than the 3D coordinates and N becomes the number of points. The NN forward pass then becomes:

$$f_{\boldsymbol{\theta}} : \mathcal{P} \in \mathbb{R}^N \times \mathbb{R}^{3+D} \rightarrow \mathbf{Z} \in \mathbb{R}^F \times \mathbb{R}^K,$$

where F is the number of b-rep faces. Noting that the output shape is $(F \times K)$ since the network aggregates the point features to their associated b-rep faces to produce direct and differentiable b-rep face predictions.

While label predictions can be simply obtained from the logit vector outputs of the NN, it is often useful to transform this into a vector giving the probability that the entity belongs to each category. This can be done with the softmax function that normalises the input into a vector which sums to one:

$$\sigma_{SM}(\mathbf{z})_i = \frac{e^{\mathbf{z}_i}}{\sum_{j=1}^{K} e^{\mathbf{z}_j}}.$$

The 'probability' of the entity belonging to the predicted class is then $q = \max_k \sigma_{SM}(\mathbf{z})$.

However, literature suggests that this normalised vector should not be interpreted probabilistically as it tends to be uncalibrated and overconfident (Guo et al. (2017); Gal (2016)), especially for new inputs which are not within the training distribution. Therefore, much work has been done on proper NN uncertainty estimation - for a full review see Gawlikowski et al. (2023). Put simply, these techniques aim to provide an uncertainty score for each prediction which captures the uncertainty within the input data, or the model parameters, or both. With this, one can make decisions about the quality of the NN predictions for a given input.

This work chooses to review techniques which require little to no modification of the network architecture and no changes in the training scheme. These could be applied to pre-trained models that engineers already have. Four broad categories of approaches which implement different conceptual representations of uncertainty and represent a range of computational costs are reviewed. These are approximate Bayesian inference (MC Dropout from Gal and Ghahramani (2016)), test-time input augmentation, model ensembling, and post-processing calibration.

It is also worth noting that for semantic segmentation in image or point cloud uses cases, individual pixels or points do not have much significance in themselves. In other words, picking out the most uncertain pixels or points is often not useful for users of the NN and structural metrics are often used (Ng et al. (2023)). In the current application, it is assumed (at least in the first instance) that individual b-rep faces have much more self-contained meaning. It is possible that a feature could be represented by a single b-rep face; however, very complex shapes and features can require many b-rep faces.

4 Uncertainty Estimation

4.1 Post-Processing Calibration

There are methods which aim to transform the outputs of a trained model using a known calibration dataset such that the new probability vectors are well calibrated. Perfect calibration can be defined as

$$\mathbb{P}(\hat{Y} = Y | \hat{P} = p) = p, \forall p \in [0, 1]$$

where $\hat{Y}$ is the predicted class, Y is the true class, $\hat{P}$ is the predicted probability or predictive confidence, and p is the true (frequentist) probability (DeGroot and Fienberg (1983); Dawid (1982)). It is uncommon for works on classification/segmentation quality control to include these approaches, but they have been represented here with the reasoning that calibrated probability outputs are more useful in picking out incorrect predictions for quality control. Two methods are used here and explained in the following.

Temperature scaling (Guo et al. (2017)) is a simplified multi-class version of the Platt scaling method (Platt et al. (1999)) for calibrating NN probability predictions that only tunes one parameter, τ[2]:

$$\hat{q} = \max_k \sigma_{SM}(\mathbf{z}/\tau)^{(k)}.$$

The logits from the calibration set geometries are used to tune the temperature scaling parameter by minimising the negative log likelihood loss between $\hat{q}$ and the 'true' probability vector (which is just the one-hot encoded class index). The L-BFGS optimiser in *PyTorch* was used with a learning rate of 0.01 following the example implementation of Guo et al. (2017)[3]. After scaling, the maximum probability, $\hat{q}$, is used as the predictive confidence.

Histogram binning (Zadrozny and Elkan (2001)) is a frequentist approach which bins the 'predicted scores' from a calibration dataset. Given a new test example, it is placed into one of the bins according to its (raw) score. The 'corrected' or calibrated probability that this new test example belongs to the predicted class is the fraction of calibration examples in the same bin of the same predicted class which were correct. Here, the maximum probability, q, was used as the 'predicted scores'[4]. 20 equal-width bins for each calibration set was used.

4.2 Monte-Carlo Dropout

The dropout technique (Srivastava et al. (2014)) randomly 'drops' neurons in a layer (i.e. zeroes out elements in the weight matrices) during training for regularisation. This is nominally not done during 'test-time' or inference and thus the resulting network is effectively averaging the weights of slightly smaller subnetworks. This prevalent technique is one of the factors which allows modern neural networks to have so many layers. Gal and Ghahramani (2016) shows that using this at test-time produces stochastic forward passes which approximates the sampling of weights for the variational inference of Bayesian NNs.

A distribution of vector outputs is obtained for a given input - $f_{\boldsymbol{\theta}_t}(\mathcal{P}) = \mathbf{Z}_t$, where t is the t-th forward pass. This can then be collapsed to a prediction vector by simply obtaining the element-wise mean after applying softmax to each. This vector is treated similarly as that above - the argmax is the predicted class index and the corresponding value is the confidence. Early works show that this aggregated vector, with enough forward passes, is more accurate than basic inference. Kendall et al. (2015) applies this to an encoder-decoder architecture for image segmentation and presents an optimal configuration of dropout layers within the convolutional NN for maximum segmentation accuracy. This was found to be placing dropout layers within the deepest layer(s) of the encoder and decoder units. This was confirmed for the point-based NNs used here (results not shown). 50 stochastic forward passes were used and aggregated here which was found to be sufficient for converged uncertainty estimates and resultant classification accuracy.

[2] τ is used here instead of T from the original work to not confuse with the use of T later in this paper.

[3] See `http://github.com/gpleiss/temperature_scaling`.

[4] This seemed to produce similar but slightly better results than using the maximums within the raw logits

4.3 Test-Time Augmentation - Point Resampling

Another way to obtain a distribution of NN outputs given the same input is to do test-time augmentation (Wang et al. (2019b,a); Gawlikowski et al. (2023)). In the current work, the 'raw' input to the system is a b-rep CAD model but the NN's observation/input is a point cloud, $\mathcal{P}$, sampled from the surface. Therefore, a natural and effective way to do data augmentation is to repeatedly sample $\mathcal{P}$ with the stratified stochastic sampling method proposed by Vidanes et al. (2024). In other words, the network input is not simply transformed to look slightly different but is actually a different instantiation of the same fundamental geometry. For each forward pass, the points seen by the network are different and have no formal correspondence, thus the per-point predictions cannot be aggregated into distributions. However, because the network has a b-rep face prediction head which aggregates relevant points into the face space, this can form a distribution of outputs for each b-rep face - $f_{\boldsymbol{\theta}}(\mathcal{P}_t) = \mathbf{Z}_t$. Similarly to the above, the distribution can be aggregated into one prediction vector per face. 50 stochastic forward passes were also used and was sufficient for convergence.

4.4 Resampling & Dropout

This work also presents a combination method with only a small computation overhead when compared to the individual components. The point resampling test-time augmentation and MC Dropout inference can be used simultaneously to produce a wider variety in the distribution of output logits given a single input and trained NN. Each logit output is produced from a different point cloud (from the same geometry) and with a different sample of network nodes being dropped - $f_{\boldsymbol{\theta}_t}(\mathcal{P}_t) = \mathbf{Z}_t$. As above, 50 stochastic forward passes are used.

4.5 Deep Ensemble

An ensemble of neural networks (Hansen and Salamon (1990)) can also be used to obtain a distribution of outputs given the same input. In this work, an ensemble of models with the same architecture and trained with the same dataset is used following Lakshminarayanan et al. (2017). The models are trained using different random initialisations (and different mini-batch sampling of the dataset) and thus take a different trajectory through weight space[5]. The outputs of each separate neural network for a given input geometry can be treated as samples from a distribution - $f_{\boldsymbol{\theta}_m}(\mathcal{P}) = \mathbf{Z}_m$ for model m - and aggregated as above. 10 models were used.

4.6 Predictive Confidence

As noted above, a predictive confidence in $[0, 1]$ can be obtained as the maximum element in the predicted vector (corresponding to the predicted class index one would obtain with argmax). To formalise this for the methods which produce a distribution of outputs, the predictive confidence is given by

$$\hat{q} = \max_k \left(\frac{1}{T} \sum_t \sigma_{SM}(\mathbf{z}_t) \right),$$

[5] Fort et al. (2019) show that this is sufficient for the models to take different modes in function space. In comparison, while MC Dropout might provide diversity in weight space, they stay in a relatively small subset of function space (i.e. $f_{\theta_1}(\mathbf{x}) = f_{\theta_2}(\mathbf{x})$ even though $\theta_1 \neq \theta_2$).

where T is the number of stochastic forward passes or the number of models in the ensemble.

Alternatively, Gal (2016) and Mukhoti and Gal (2018) use information theoretic (Shannon (1948)) scalars which summarise the uncertainty within the distributions more formally. The predictive entropy can be approximated from this distribution as:

$$\hat{\mathbb{H}} = -\sum_k \left(\frac{1}{T} \sum_t \sigma_{SM}(\mathbf{z}_t) \right) \log \left(\frac{1}{T} \sum_t \sigma_{SM}(\mathbf{z}_t) \right),$$

and represents the combined epistemic and aleatoric uncertainty. Alternatively, mutual information can be approximated as:

$$\hat{\mathbb{I}} = \hat{\mathbb{H}} + \frac{1}{T} \sum_{k,t} \sigma_{SM}(\mathbf{z}_t) \log \sigma_{SM}(\mathbf{z}_t),$$

and represents just the epistemic or model uncertainty. In other words, predictive entropy also captures the uncertainty or spread in the individual predicted vectors whereas mutual information only captures the spread across the predicted vectors. See the thesis by Gal (2016) for a deeper explanation.

The scalars obtained from these are unbounded and need to be normalised to the interval $[0, 1]$ such that they can be interpreted as a probability or confidence. This can be done with a separate calibration set to obtain the range for normalisation; alternatively, a different range can be obtained for each class since they can have separate distributions. A high uncertainty scalar value corresponds to a low predictive confidence therefore a reversed scale min-max normalisation was used.

5 Method

5.1 Base Neural Network

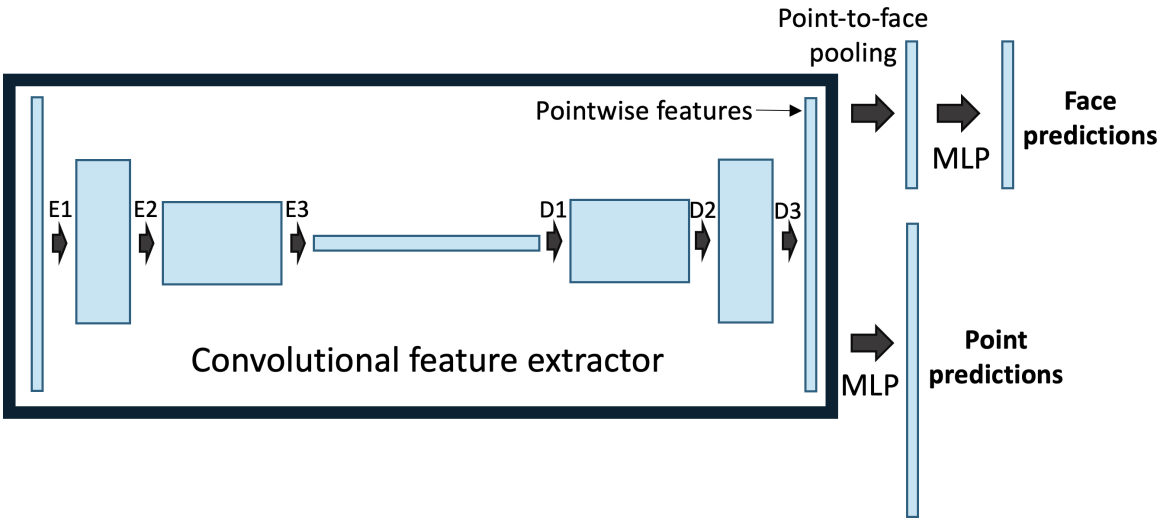


FIGURE 1: Block diagram of neural network architecture. The convolutional-type point-based feature extractor is followed by a multi-head structure. *MLP* = Multi-Layer Perceptron. *EX* = Encoder X. *DX* = Decoder X.

The underlying models used in this work are point-based graph neural networks based on the work from Vidanes et al. (2024), which are an extension on *PointNet++* (Qi et al. (2017b)). The architecture

is illustrated as a block diagram in Figure 1. The unscaled network described in their work was used for computational efficiency - i.e. default depth and width resulting in 1.4M learnable parameters. The 'facewise' prediction branch was included and the average of the point and face loss was used for training. The multi-head structure shown in Figure 1 is only to aid training as discussed in the original work; only the face predictions from the facewise branch are used in the following. To take full advantage of MC Dropout inference, extra dropout layers with a dropout probability of 0.5 were added in the final encoder layer (E3 in Figure 1) and the first decoder layer (D1 in Figure 1) - the optimal configuration suggested by Kendall et al. (2015).

The *ADAM* optimiser (Kingma and Ba (2014)) was used with the same standard training hyperparameters as Vidanes et al. (2024), except that the network weights corresponding to the minimum cross-entropy loss in the validation set were extracted, instead of those corresponding to the maximum b-rep face classification accuracy on the validation set. This was expected to give better calibrated probability outputs on the base model as suggested by Guo et al. (2017).

Following the original work, the b-rep geometries were encoded into 7D point clouds - encoding 3D coordinates, 3D surface normals, and a b-rep face index - using b-rep stratified sampling with at least 4096 points. The suggested b-rep face type feature used in Vidanes et al. (2024) and Colligan et al. (2022) was not used here since it can make the trained networks tied to a specific CAD kernel.

5.2 Experimental Framework

This empirical review includes methods with many layers of stochasticity. It is well-known that NN training is stochastic due to the random mini-batching, weight initialisation and dropout layers. Moreover, this work is interested in estimating the 'real-world' performance of the above methods. From this perspective, the evaluation of trained NNs is also stochastic since the training, validation, and testing datasets are samples from the underlying distribution or pattern being learned. On top of this, some uncertainty estimation methods being reviewed here are inherently stochastic. To maximise the reliability of the results, many repetitions and cross-validations are needed to capture the stochasticity in the performance metrics. The following details the individual layers of evaluation repetitions and what aspects of randomness they are trying to capture.

Multiple models are necessary for the Deep Ensemble inference approach, therefore 10 separate models were trained on the same training and validation data. For the other approaches, the following was also repeated for each model with results being aggregated to capture the variation caused by the random weight initialisation and the random mini-batch sampling. Resampling cross-validation (CV) was used with a separate set of 3000 unseen geometries to estimate the unbiased performance of the methods - i.e. not tied to the specific geometries in the dataset splits. For each CV run, 1000 geometries were randomly sampled from this pool and used to compute the performance metrics, with the remaining 2000 geometries being used as a calibration set where required. As an example, for the histogram binning approach, during each CV run, the bins were computed independently using the samples within the corresponding calibration set. 20 CV runs were done. Finally, the sampling of a set of stochastic forward passes and aggregation was repeated 100 times. While the estimates are converged with $T = 50$, in the sense that it remains stable with increasing T, there is still some variation in the result when sampling a different set of 50 stochastic NN outputs. As will become apparent in the next section, many repetitions here do not incur a high computational cost.

In summary, for the Deep Ensemble method there were 20 separate evaluations being aggregated (due to CV runs). For the baseline and post-processing calibration methods, 200 evaluations were

being aggregated across the different models. For the MC Dropout, resampling, and combined methods, 20000 evaluations were aggregated since each set of $T = 50$ forward passes is treated as a separate sample.

Finally, unless otherwise stated, metrics are calculated for each individual class separately first then averaged to obtain the evaluation metric for that model and CV run. This is often called 'macro' averaging, instead of 'micro' averaging where metrics are calculated globally regardless of class. 'Macro' averaging is useful in cases where there is significant class imbalance so that poor performance in less frequent classes is not hidden by the dominating effect of good performance on very frequent classes.

5.3 Computational Implementation

The experimental framework detailed above together with the size of the datasets being used here results in this being a significant computational task with compute, memory, and time trade-offs. This section details the data generation approach to efficiently obtain the performance metrics.

Studying the inference and uncertainty estimation techniques, one can see that there are overlaps in the required computations. For instance, a single standard NN forward pass produces a logit vector which is directly used for the baseline case, used for transformation in the post-processing calibration methods, and can be aggregated with other forward passes to obtain a prediction for the Deep Ensemble and point resampling methods. In addition, because resampling CV is being used, the same geometry could be present in multiple CV runs either in the testing or calibration set. With this in mind, the logit vectors produced from each geometry from different models could be stored and re-used to save on computation and time at the cost of memory and/or disk space.

Of course extra computation, and large memory requirements in some cases, is required for performing the 'simulation' of the different inference methods from this pool of data. But an extra advantage of this approach is that the logit pool generation step is trivially parallelisable - the generation of each logit is independent. For this work, the *IRIDIS 5* compute cluster at the University of Southampton was utilised; logit generation was parallelised on the multicore machines as well as being run across multiple nodes. In practice, inference of multiple models on one geometry was done in series within a process such that the b-rep geometry was only transformed into a point cloud once (for that process). This saves on computation and ensures that one set of logits from the 10 models for one geometry was from the same point cloud, without needing to repeatedly re-seed the random number generator.

Figure 2 illustrates the pool of logit vectors created by passing the 3000 geometries through each of the 10 models multiple times with dropout either on or off. To obtain a single prediction with accompanying confidence estimate, the appropriate subset of the data pool is sampled. For instance, to simulate MC Dropout inference, T logit vectors for a given point cloud are sampled from the subset produced by a model with dropout on. The prediction is obtained by averaging across T and the confidence can be obtained from either the mean vector or the predictive entropy or mutual information estimates. A similar process is done for simulating the combined (point resampling and MC Dropout) case but logit vectors from models with dropout on across different point encodings of a single b-rep geometry are used. To simulate inference with an ensemble of 5 models, the logit vectors produced by a random 5 models (with dropout off) for the same geometry are sampled. This can then be aggregated as before to obtain a mean prediction and confidence estimate. Not only does this allow for efficient evaluation of the different methods, the recombining of stored logit vectors allows convergence studies to be easily performed with increasing T or number of models. Many

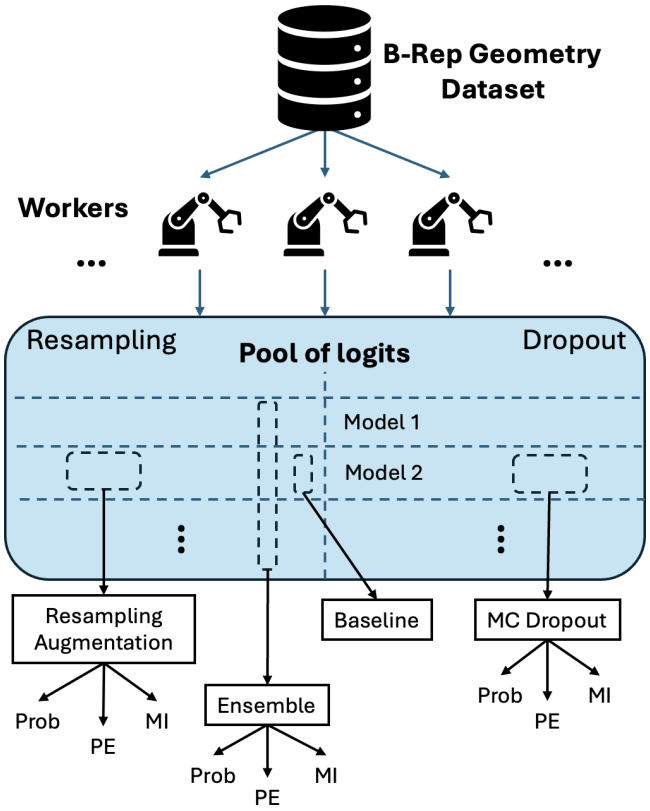


FIGURE 2: Simplified illustration of data flow from the dataset of b-rep geometries to re-usable logit vectors for simulating different inference methods.

CV runs can also be trivially performed by sampling the logits of appropriate subsets of geometries. For instance, for temperature scaling, one can obtain the logit vectors (produced by a model with dropout off) from a sample of 2000 geometries and tune the τ parameter. This can then be used to transform the logits from the remaining 1000 geometries from the pool and a set of metrics can be computed. Figure 3 shows a sample of the data being stored by the parallel workers.

	model	sampling_seed	dropout	geom_id	face_idx	gt	logit_0	logit_1	logit_2	...
0	2	674885	False	45559	0	24	-11.395023	-9.771173	-6.572227	...
1	2	656221	False	45559	1	3	-5.850572	-6.814668	-3.885224	...
2	2	833777	False	45559	2	0	20.654837	-3.994620	3.113753	...
3	2	747463	False	45559	3	3	-3.582532	-5.484285	-3.095505	...
4	2	464341	False	45559	4	3	-4.179116	-5.199742	-2.358987	...

FIGURE 3: Sample of logit vector dataset with metadata annotations.

5.4 3D CAD Datasets

This work uses two large and publicly available datasets of 3D CAD geometries with semantic segmentation labels. MFCAD++ from Colligan et al. (2022) is an algorithmically generated dataset where each b-rep face in the model is labelled with the manufacturing operation which created it. The geometries start as a 'stock' cuboid and manufacturing operations are simulated by applying various (random) cuts in series. There are a total of 25 classes - 24 machining features plus any remaining 'stock' faces. They provide lists of geometries for the training, validation, and test splits

with 41766, 8950, and 8949 geometries respectively. The entire training and validation split was used for NN parameter tuning and early stopping. While, as mentioned previously, 3000 geometries from the test set are used for evaluation of the uncertainty estimation methods. The number of faces per geometry is approximately normally distributed with a mean of 30, ranging between 6 and 86.

The Fusion360 Gallery Segmentation Dataset from Lambourne et al. (2021) is a collection of user submitted geometries with faces labelled according to the CAD modelling operation which created it. There are 8 possible classes - 'Extrude Side', 'Extrude End', 'Cut Side', 'Cut End', 'Fillet', 'Chamfer', 'Revolve Side', and 'Revolve End'. This labelling is inherently ambiguous since the same 3D shape can be obtained with different sequences of modelling operations; this is noted in their work. The public release only provides a list of geometries for the train and test split with 30314 and 5366 geometries respectively. In the current work, the provided 'training' geometries were randomly split with a 85/15 ratio for use as a training and validation set. As before, 3000 geometries from the unseen test split are used for the evaluations in the next section. The mean number of faces per geometry is around 15, but the distribution is dramatically skewed with a range from 1 to 421.

6 Results

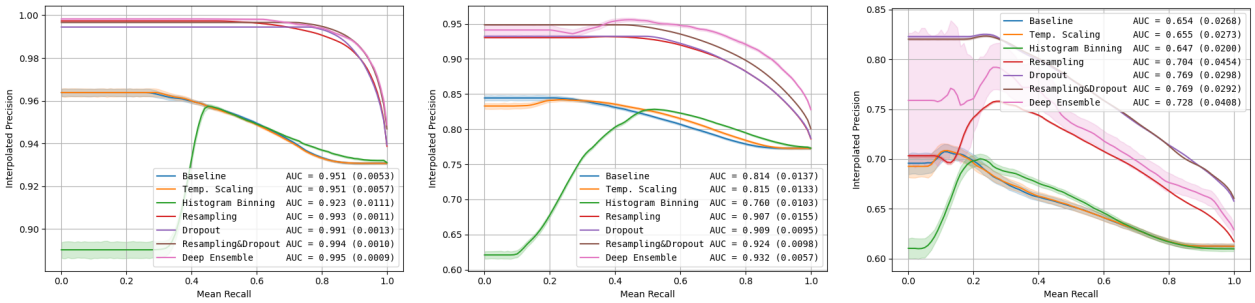
6.1 Filtering Performance

The first way to evaluate the uncertainty estimation methods is how well they can be used to filter for predictions which are likely to be correct. This task can be viewed as a binary classification or detection task allowing the use of the familiar Precision-Recall (PR) curve for comparison, with area-under-the-PR-curve (AUPRC) as a summary metric. Predictions passing the confidence threshold and having a correct label are 'true positive' (TP) samples and those passing the threshold but having an incorrect label are 'false positive' (FP) samples. It also follows that correctly labelled faces which did not pass the confidence threshold are 'false negative' (FN) samples. This allows the calculation of the standard precision and recall metrics for a given threshold:

$$\text{precision} = \frac{TP}{TP + FP}, \text{recall} = \frac{TP}{TP + FN}.$$

It is worth noting that the commonly used $P(\text{accurate}|\text{certain})$ from Mukhoti and Gal (2018) maps directly to the precision metric described above. A similar probabilistic interpretation to recall would be $P(\text{certain}|\text{accurate})$. This work chooses to use the ubiquitous PR and AUPRC metrics over the conditional probabilities and *Patch Accuracy vs Patch Uncertainty* metrics from Mukhoti and Gal (2018).

Figure 4 summarises the performance of the methods across different datasets. The traditional class probability q was used as the baseline. For all inference methods, the maximum element of the (mean) probability vector has been used as the predictive confidence. For each model, CV run, and class (and sample of stochastic forward passes) the precision and recall metrics were calculated for different threshold values and macro averaged across classes. Here, an increment of 5 from 0 to 95 was used, changing to an increment of 1 up to 100. The model predictions are already quite accurate overall, so it's expected that there will be more predictions with high confidences. This gives a set of curves (for each model, CV run and sample of T) on the PR axis that are misaligned for each method - i.e. there is no simple correspondence of x-values. To collapse them into one mean line with error bounds, the raw PR values were interpolated onto a 'mean recall' axis from 0 to 100 with increment 1. The AUPRC metric was also computed using the interpolated values; the value shown in the plot



(A) MFCAD++ (B) MFCAD++ (5%) (C) Fusion360 Gallery

FIGURE 4: Precision-Recall mean curves and 95% confidence bounds for different uncertainty estimation methods, evaluated on different datasets. Standard max probability is used as predictive confidence. Average area-under-the-curve (AUC) metric and standard deviation is shown for each method - higher is better. (B) shows results when using models trained on 5% of the MFCAD++ training set.

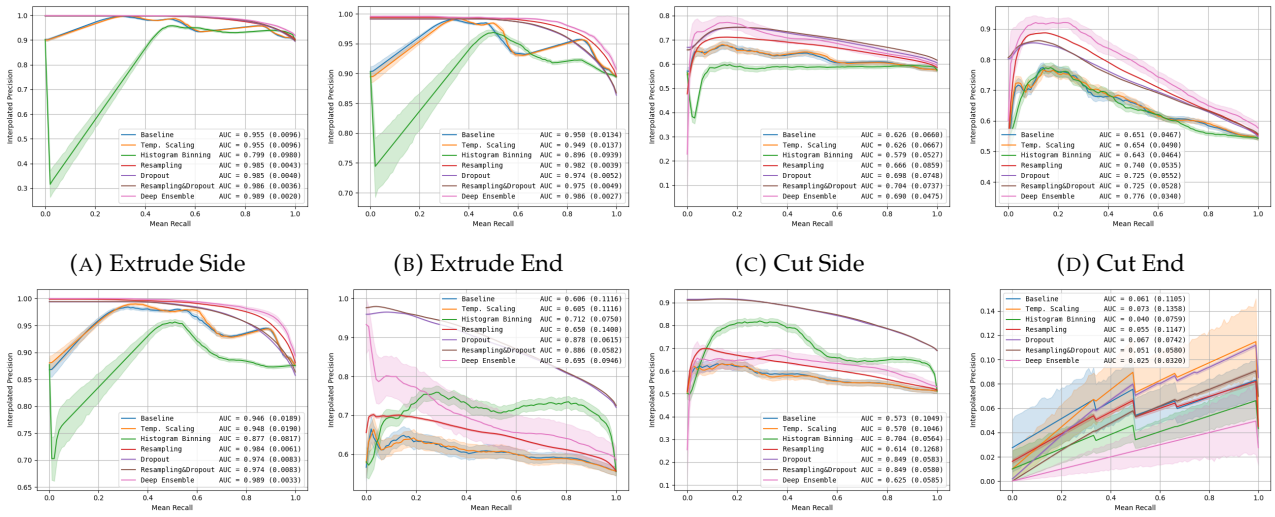
is averaged across models and CV runs. Figure 4b shows results from models which were trained on a small subset of the MFCAD++ training set - 5% of the original. The full validation set was still used for early stopping.

The first thing to notice is that the precision values of different methods at 100% recall (i.e. all predictions pass the threshold) are different. This corresponds to the base accuracy of the predictions. As expected, the post-processing calibration methods have the same value as the baseline here since these do not change the prediction - the probability vector is scaled but the argmax remains the same. On the other hand, the methods which aggregate different predictions can have a different argmax value than a specific individual forward pass. Literature shows that MC Dropout inference and Deep Ensembles increase predictive accuracy and these results reflect that.

From Figure 4a, the aggregation based methods all perform similarly well, with the Deep Ensemble being slightly better. This trend broadly remains in Figure 4b, but the separation between methods grows. Looking at Figure 4c, the Deep Ensemble method is now significantly worse than the combination method or even MC Dropout by itself, with a very large variance at low recall values or high confidence thresholds (i.e. the accuracy of the most confident predictions). This variance is partly driven by the differences in performance across the CV runs for less frequent classes like 'Chamfer' and 'Revolve End' as shown in Figure 5. However, interpolation effects are also playing a role here which will be discussed in Section 7. The ambiguity of the classes together with the massive diversity of geometries in the dataset adds confounding factors when analysing results for this dataset, thus further analysis is left for future work.

6.2 Quality Control

Following Ng et al. (2023), the performance of these methods for the application of segmentation quality control is evaluated with slightly different but related metrics. Put simply, the application involves flagging the most uncertain outputs (based on some threshold) of the system for manual correction such that the overall output of the human-in-the-loop system is more accurate. This is a semi-automation case where the manual effort of a human expert is ideally concentrated into the most difficult feature recognition cases whilst the system automates those which (in theory) the NN is confident in. In this section, more practical and intuitive metrics are used instead of the familiar but more conceptual precision and recall from the previous section. Instead of specifying an arbitrary



(A) Extrude Side (B) Extrude End (C) Cut Side (D) Cut End

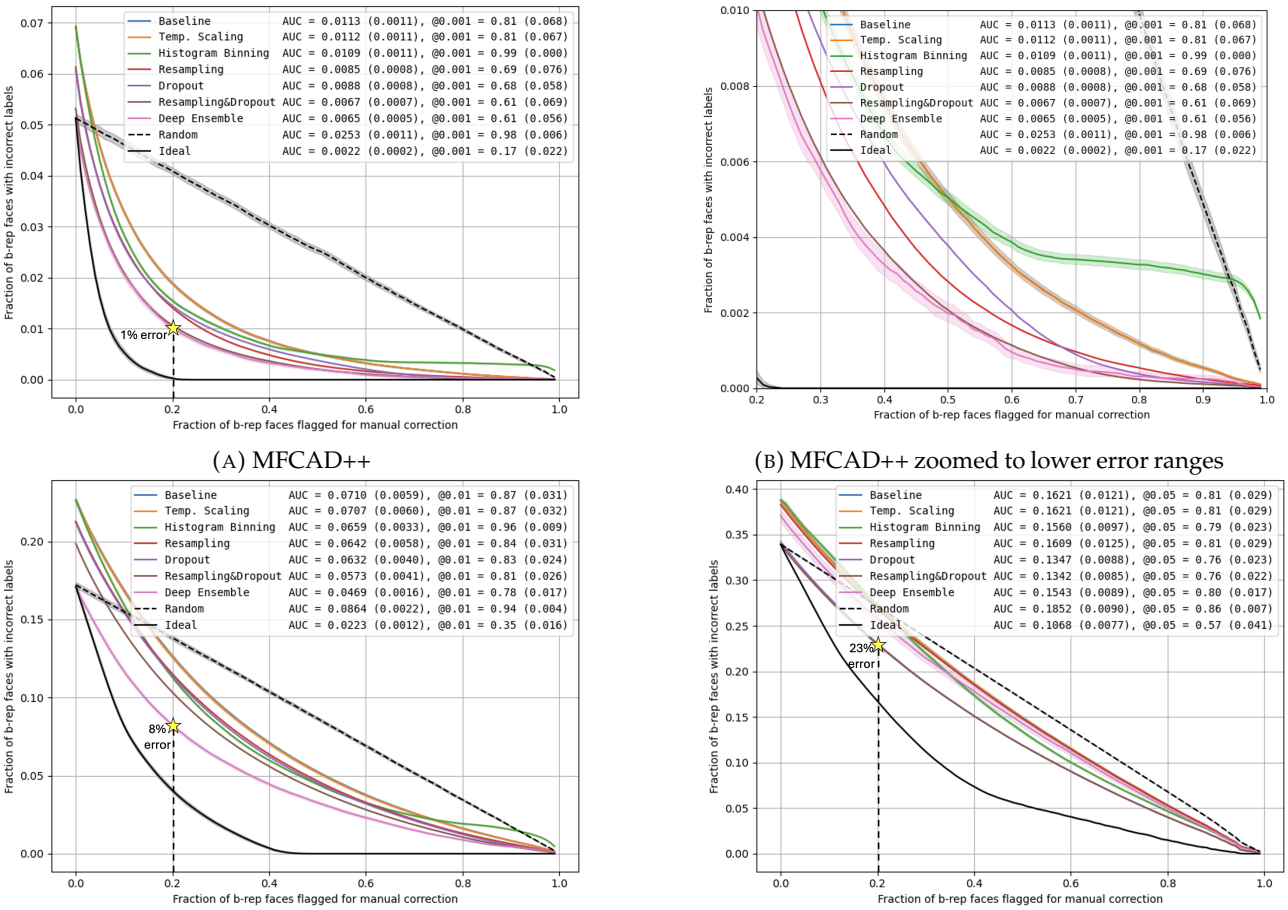
(E) Fillet (F) Chamfer (G) Revolve Side (H) Revolve End

FIGURE 5: Precision-Recall curves for individual Fusion360 Gallery classes using different uncertainty estimation methods. Average area-under-the-curve (AUC) metric is shown for each method.

range of thresholds, the predictions for the faces in the sample of 1000 test geometries are ordered in increasing confidence. A range of fractions (1 to 100 with an increment of 1 in this case) are then specified such that the least confident (or most uncertain) predictions are flagged for 'manual correction'. 'Manual correction' in this case simply means that the predictions become correct regardless of their value - emulating a perfect oracle. The error rate remaining, after correction, for the faces in the test geometries can then be calculated.

Figure 6 summarises the results of this experiment across the different datasets. Again, the maximum value within the (mean) probability vector has been used here as the confidence for all methods. Two extra cases are also shown for context. The dotted black line represents the case where predictions are flagged for manual correction randomly, regardless of their estimated confidence. The solid black line represents the ideal case where incorrect predictions are always flagged first. One would expect the ideal case to be a straight line and intersect the horizontal axis at the same value as the vertical axis intersection - i.e. the fraction of b-rep faces corrected is the exact fraction which have incorrect labels originally. This is not the case in these plots since the error rate for each class is computed then averaged (macro averaging). Results for single classes, not shown, do show a straight line for the ideal case as expected. In addition, because the correctness of the predictions are different across some methods, these cases depend on which predictions are used - i.e. which curve they intersect with on the vertical axis. Here, the method with the best base prediction accuracy for each dataset is used to visualise the random and ideal cases, similar to Ng et al. (2023). Similar to the previous section, metrics are computed for each model, CV run, class, and sample of Ts and then averaged across classes. However, because constant fractions (or approximately constant recalls) are being used here instead of thresholds, the set of metrics for each method can then simply be aggregated without interpolation to obtain the mean curves and confidence thresholds shown in Figure 6. The average AUC is computed as before. Finally, a specific interpolated point is shown for each method as a summary metric following Ng et al. (2023). In other words, this is the fraction of b-rep face predictions which needs to be flagged for manual correction to achieve a given error rate overall.

It can be seen that the overall results here correlate to those shown when using the PR metrics. The combination method is consistently well performing across classes. Deep Ensemble is best for the full MFCAD++ case, albeit within a standard deviation to the combination method as measured



(A) MFCAD++

(B) MFCAD++ zoomed to lower error ranges

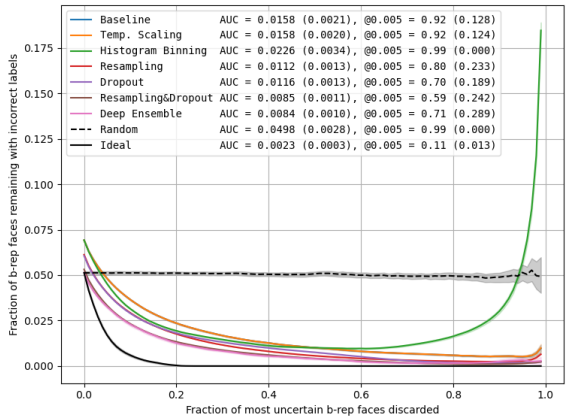
(C) MFCAD++ (5%)

(D) Fusion360 Gallery

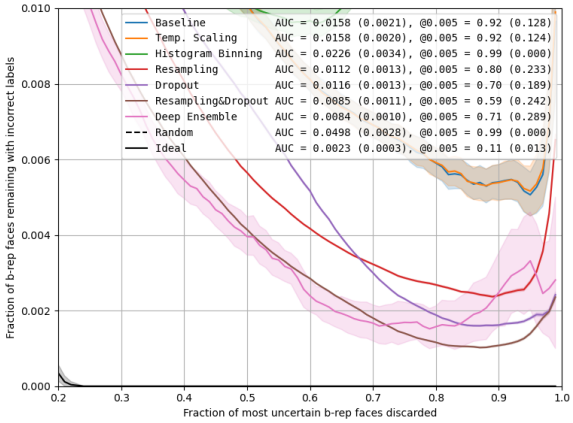
FIGURE 6: Fraction of b-rep faces with incorrect labels after flagging a fraction of predictions for 'manual correction'. Average area-under-the-curve (AUC) metric and fraction of predictions required to be flagged for a given error is shown for each method - lower is better. (C) shows results when using models trained on 5% of the MFCAD++ training set. Best improved error rate given 20% of predictions are flagged is also shown.

by the mean AUC. For the predictions from models trained on 5% of the MFCAD++ training set, Deep Ensemble is best with a significant margin. In contrast, for the Fusion360 Gallery case, the combination method is top performing by a significant margin. The point metrics shown in Figure 6 are for very low error rates since this is hard to distinguish in the plots, but this is likely not of most interest in practice as it corresponds to a large amount of manual intervention. More reasonable is perhaps for a human expert to check 20% of the most uncertain predictions. In this case, both the Deep Ensemble and the combination methods reduce the error rate for predictions on the MFCAD++ dataset from around 5% to 1%. For the models trained on the small MFCAD++ training set, the Deep Ensemble uncertainty estimation improves the error rate from around 17% to around 8% with a human in the loop. Finally, for the Fusion360 Gallery case, MC Dropout and the combination method improves error rate from around 34% to 23%.

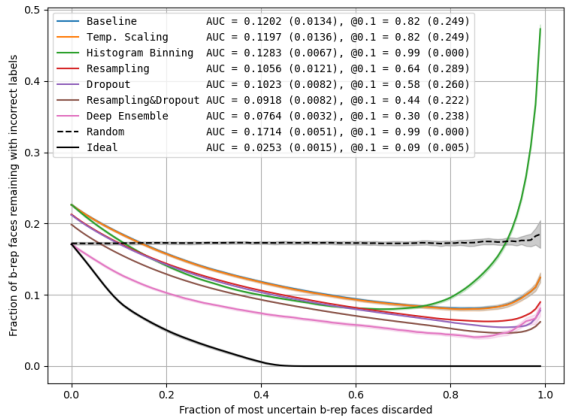
Figure 7 presents the same results in a slightly different way. Instead of considering the error rate on the entire test set after manual correction, the error rate on the filtered or 'certain' predictions is shown. Put simply, a fraction of the most uncertain b-rep face predictions are discarded and the error rate of the remaining predictions is plotted against this fraction. This better shows the performance at the strictest confidence thresholds, whereas this is diluted in the previous metric since all 'uncertain'



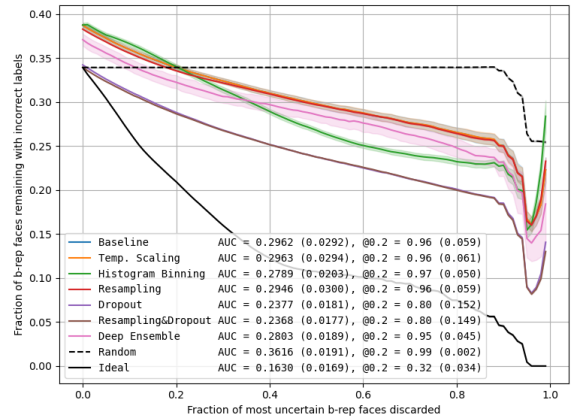
(A) MFCAD++



(B) MFCAD++ zoomed to lower error ranges



(C) MFCAD++ (5%)



(D) Fusion360 Gallery

FIGURE 7: Error rate in the remaining predictions after a fraction of uncertain b-rep faces are discarded. Average area-under-the-curve (AUC) metric is shown for each method - lower is better. The fraction of predictions required to be flagged for a given error is also shown - lower is better. (C) shows results when using models trained on 5% of the MFCAD++ training set.

predictions become correct after manual intervention. This is relevant to a more feature detection type application where dense prediction (a label for every b-rep face) is not needed. As mentioned before, this includes use cases like identifying portions of the geometry that might have structural failure, or be difficult to mesh, or be difficult to additively manufacture.

Essentially, this is the inverse of the PR metrics in subsection 6.1 but with a more practical interpretation. In this case, the random baseline is (approximately) a straight line since removing predictions uniformly randomly should preserve the overall distribution of correct and incorrect labels in the remaining predictions. The main trends here and the performances of methods relative to each other are of course the same as before, but the curves no longer go towards (1,0) since the perfect oracle is not correcting an increasing fraction of predictions. The increase in error rate at higher discarded fractions, most dramatic in the histogram binning approach, suggests that overconfidence in some incorrect predictions is still present. This is also shown, but more subtly, in Figure 6 by the decrease in the magnitude of the curve's gradient. The histogram approach shows the most dramatic case likely because of the equal width bins used. The top bin tends to occur from 0.95 confidence and up; therefore, predictions with a raw 'score' of $>= 0.95$ will all have the same calibrated predictive confidence (for each class).

6.3 Uncertainty/Confidence Estimation

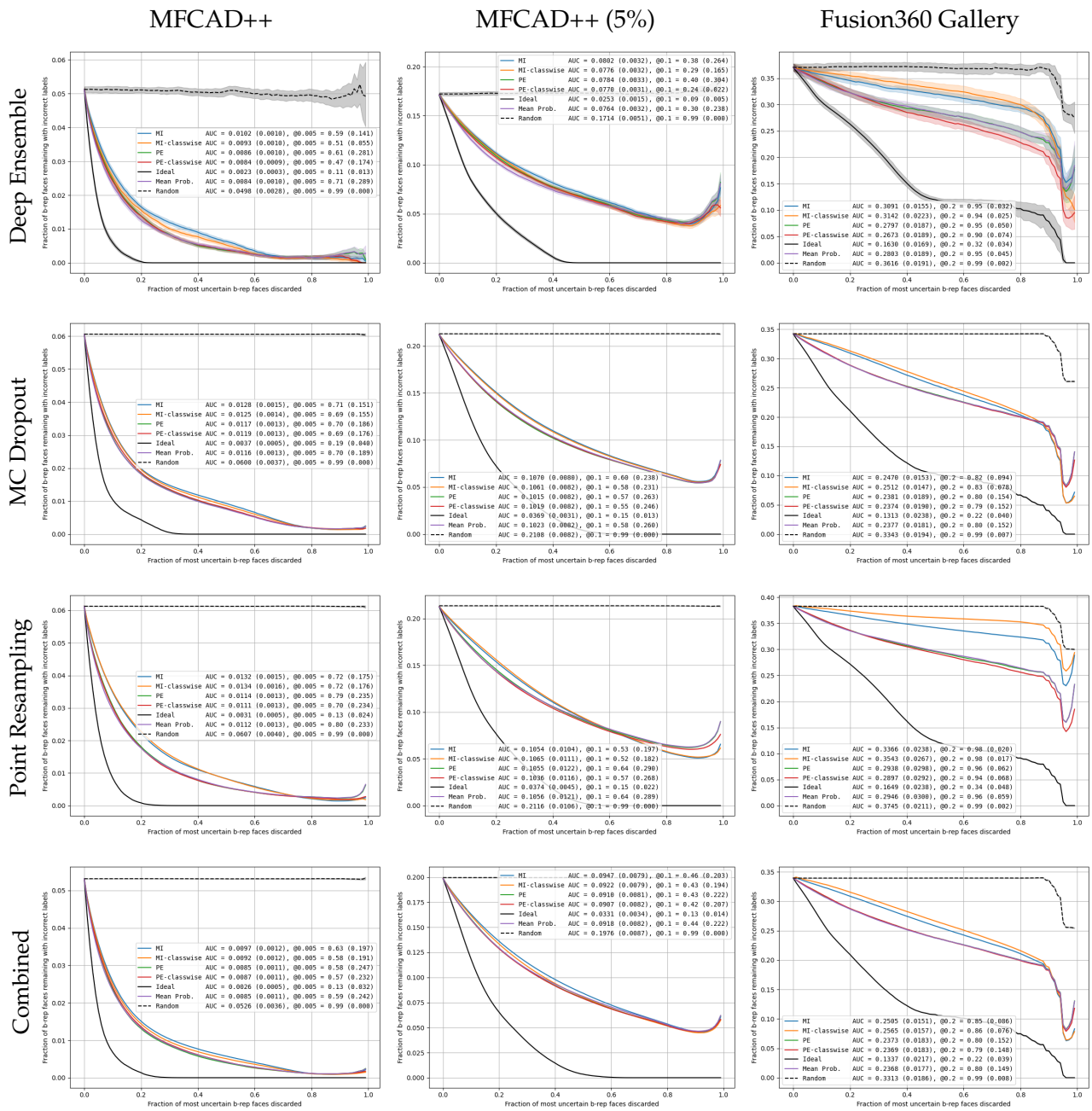


FIGURE 8: Quality control curves for different uncertainty scalar estimation methods, for each inference method and dataset. The accuracy of predictions after discarding uncertain ones is shown. AUC and x-axis value at given error rates are also shown - lower is better.

As mentioned in Subsection 4.6, there are a number of ways in the literature to summarise the uncertainty from the distribution of (stochastic) predictions. Figure 8 shows the behaviour when using the different uncertainty scalar estimation methods on the set of outputs produced from the different inference methods for each dataset case. It is observed that they perform more or less similarly, with the simple maximum mean probability used thus far being one of the best. Predictive entropy is similarly well performing. Interestingly, using the mutual information scalar tends to perform well when a large fraction of uncertain faces are discarded (i.e. high confidence threshold). In other

words, it has a high accuracy when only the most certain faces are considered. There is a larger spread of performance when using the Fusion360 Gallery dataset, but the general trends remain the same.

It seems that the uncertainty information is sufficiently captured by looking at the maximum element of the mean vector across forward passes - without explicitly incorporating the rest of the vector. However, it is noted that the probability vectors being averaged have already been normalised through the softmax function which does incorporate information from the whole logit vector.

6.4 Convergence

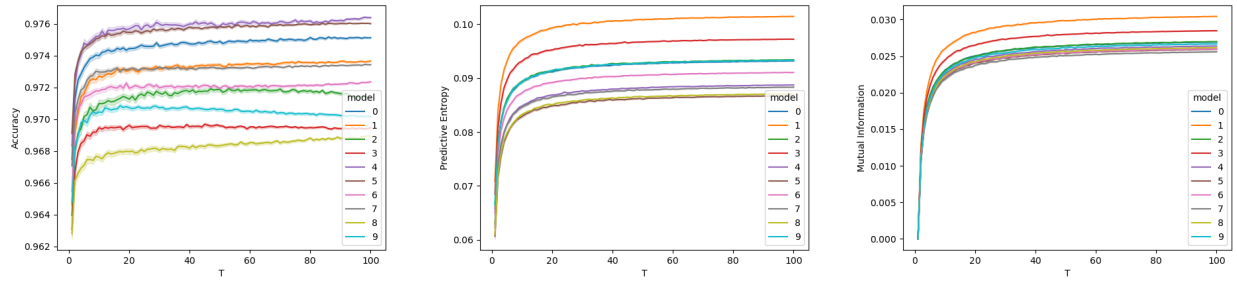
Figure 9 shows that the predictive accuracy and uncertainty scalar estimates are more or less stable after $T = 50$, for all three tested datasets. This is also the case for the MC Dropout and point resampling individual inference methods, not shown. Apparent from these plots is also the range of predictive performance across the trained models.

Conversely, at $M = 10$ ensemble models, the metrics have not yet stabilised for any of the three cases, as shown in Figure 10. This technique may benefit from an increased number of ensemble models however this has significant computational cost and is left as future work. The authors conjecture that while the numerical results may change, the overall conclusions will remain the same.

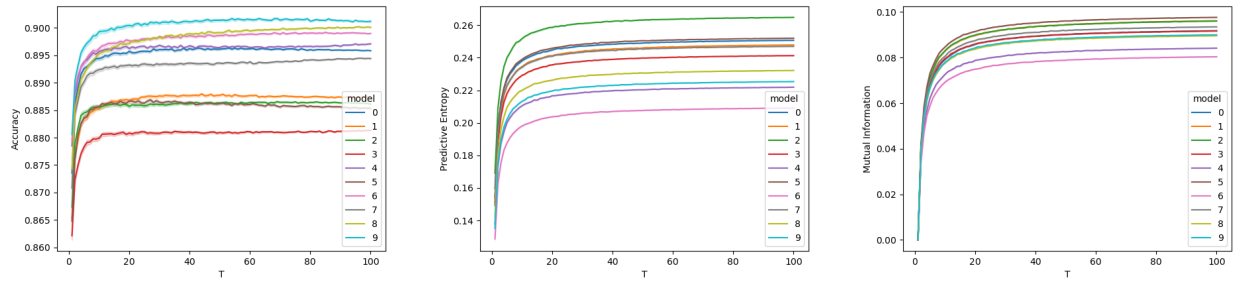
7 Discussion

Results from literature tend to suggest either MC Dropout or Deep Ensembles are best. The results in this work suggest that it is dataset (and possibly trained NN) dependent. For both the MFCAD++ cases, Deep Ensemble is significantly better than MC Dropout when measuring using the quality control metrics in Subsection 6.2. However, for the Fusion360 Gallery trained models, Deep Ensemble was significantly worse than MC Dropout. The point resampling augmentation introduced in this work has mixed results. It performs similarly to MC Dropout for the MFCAD++ cases, but worse in the Fusion360 Gallery case. However, when combined with MC Dropout, this method becomes consistently good across all cases. The combined method is equally top performing for the MFCAD++ case, although lags behind when considering models trained on the small subset of MFCAD++. For the Fusion360 Gallery case, it is a clear winner together with simple MC Dropout. This method also has the added bonus that it is more computationally efficient than a Deep Ensemble - only one model needs to be trained. However, with the not insignificant variance in predictive accuracy across the trained models shown in Figure 9, there is an argument to be made that multiple models should be trained in practice anyway to find a 'good one'.

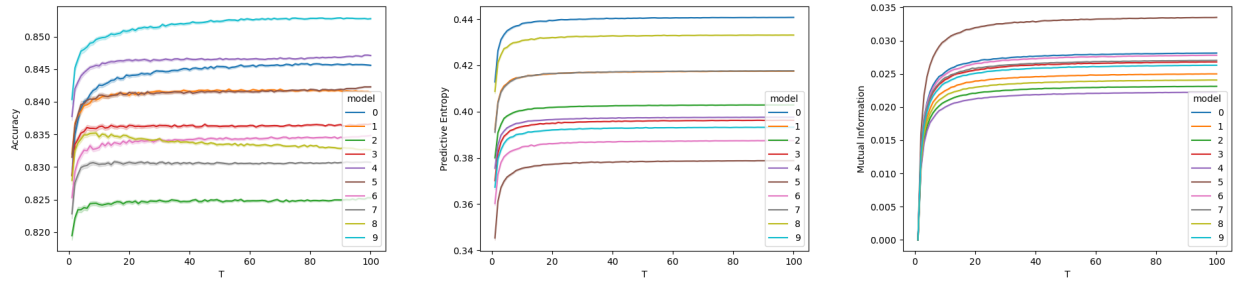
An interesting observation from these results is that the standard predictive confidence obtained from the softmax of the basic NN forward pass is not as miscalibrated as is often suggested by the literature. In this work, it is not significantly worse than the extra uncertainty estimation methods. Looking at some of the factors that Guo et al. (2017) propose that cause 'modern neural networks' to be uncalibrated and overconfident - some of these do not apply to the networks used here. For instance, they observe that NNs can overfit to negative log likelihood (NLL) loss without overfitting to the 0/1 predictive accuracy loss therefore NNs with weights extracted at the minimum of the latter can have miscalibrated probability outputs. As stated in Subsection 5.1, cross entropy loss (directly correlated to NLL loss) is used as the early stopping criteria here instead of predictive accuracy. They



(A) MFCAD++



(B) MFCAD++ (5%)

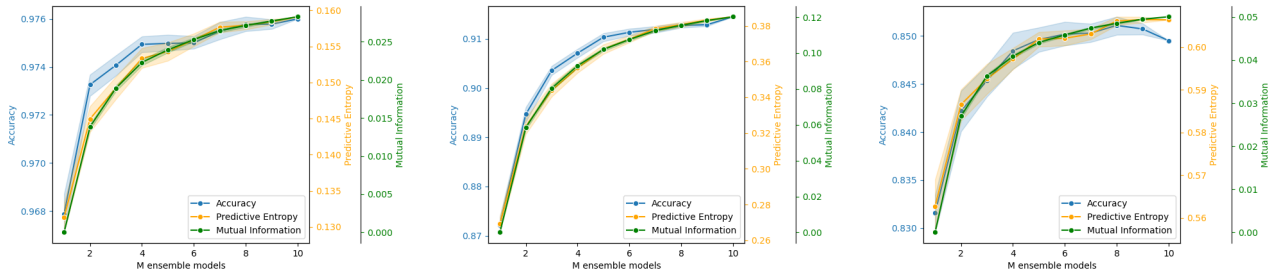


(C) Fusion360

FIGURE 9: Aggregate metrics computed from each trained model performing combined stochastic inference, with increasing number of stochastic forward passes T.

also state that miscalibration grows substantially with model capacity (i.e. number of parameters); the NNs here are small compared to most used in the state-of-the-art.

Related to the above, it is observed that the results for the temperature scaling method are very similar to the baseline, often having overlapping performance curves. During the tuning of the temprature scaling parameter, τ, the optimisation was often not able to improve on the already low NLL. It is shown in Table 1 that only a small change in NLL is observed for the MFCAD++ and Fusion360 Gallery case, with their optimised τ remaining close to 1. This is unsurprising since the base NN training is already trying to minimise NLL over the large training sets; tuning an extra parameter on slightly more unseen data is unlikely to make a significant difference. Even for the Fusion360 Gallery case where the base accuracy of the trained models is not very high, the limiting factor does not seem to be data. In contrast, for the small MFCAD++ training set case, a 6.8% average decrease in NLL loss is observed after temperature scaling. In this case, the limited training data seems to be the bottleneck and the supplement of 2000 calibration geometries is helpful. However,



(A) MFCAD++ (B) MFCAD++ (5%) (C) Fusion360

FIGURE 10: Aggregate metrics computed from increasing number of models within a Deep Ensemble.

Dataset	NLL before	Optimised Temperature	NLL after	NLL % change
MFCAD++	0.1350 (0.0040)	0.959 (0.022)	0.1346 (0.0040)	0.301 (0.366)
MFCAD++ (5%)	0.5004 (0.0272)	1.368 (0.045)	0.4664 (0.0223)	6.764 (1.250)
Fusion360	0.5000 (0.0151)	0.977 (0.035)	0.4994 (0.0153)	0.127 (0.162)

TABLE 1: Behaviour of temperature scaling optimisation across 10 models and 20 CV runs for each dataset. Format is *mean* (*std.*).

this still only makes a very small change in the PR and quality control metrics, suggesting that probabilistic calibration is perhaps not a useful metric for the application being studied in this work.

Lastly, the results for the PR curves tends to have larger variance than that of the quality control metrics - most evident in the Deep Ensemble result for Fusion360 Gallery. Both of these metrics are being computed from the same set of results and therefore the variance from the differences in trained models and CV runs should be the same. However, the PR curve computation introduces an extra source of 'variance' in that misaligned curves are being interpolated into the same intervals. This is because the 'raw' PR values are obtained by specifying a range of thresholds; the fraction of predictions which pass these thresholds is different across trained models and CV runs and therefore the obtained sets of PR points are not on the same vertical lines. For instance, the fractions of predictions from the Deep Ensemble which pass a threshold of 0.5 for the Fusion360 dataset across CV runs ranges from 0.71 to 0.92. Interpolating this onto the 'mean recall' interval introduces extra noise.

8 Conclusions

The authors present this work as a first of its kind exploration into the application of uncertainty estimation techniques to feature recognition in CAD, specifically using point-based graph neural networks. A number of techniques were applied and compared to two 3D CAD geometry datasets with different semantics. It was found that combining MC Dropout with a point resampling test-time augmentation method outperformed all others when considering models trained on large datasets, in the context of segmentation quality control. In other words, it is good at estimating predictive uncertainties such that if a prediction is less uncertain than another one it is more likely to be correct. Therefore, the uncertainty estimates could be used in a human-in-the-loop approach to dramatically decrease error rates given moderate manual effort. However, Deep Ensemble seems to be better for both base accuracy and uncertainty estimation when considering small training dataset cases. It was

also found that the mean probability vector from the distribution of forward passes was sufficient in capturing uncertainty for the purposes of segmentation quality control.

The results suggest that practical and relatively simple techniques for uncertainty estimation are effective; with some techniques being able to be applied to already trained networks. Therefore, it is hoped that this work can be used as a base for tackling real case studies and helps the adoption of predictive deep learning methods into the engineering workflow.

9 Future Work

As an initial exploration into the space, there is naturally much future work to be done in this area. Some suggestions for further research are presented in the following. It was suggested in Subsection 4.6 that aleatoric and epistemic uncertainty can be decomposed and estimated from the distribution of predictions obtained from some inference methods given a single b-rep geometry input. A natural extension of the current study is to investigate how decomposed uncertainty could be useful, just as Petschnigg and Pilz (2021) and Kendall and Gal (2017) do for other application areas. Going further than filtering or flagging predictions, active learning and transfer learning in the continuously changing area of CAD could benefit from accurate uncertainty estimation techniques.

The results of this study also suggest further avenues of research involving the Deep Ensemble technique. Firstly, as shown in Subsection 6.4, the accuracy and uncertainty estimates are not quite converged yet with 10 models. A large scale study with more trained models may yield different results. In addition, it was observed that a Deep Ensemble has a much higher base prediction accuracy than single trained networks in the case where only a small amount of training data is available. There could be interesting further work here for developing data-efficient deep learning systems at the cost of increased compute and training.

Finally, this study treats the b-rep faces in the testing dataset completely independently when doing uncertainty estimation (after the NN inference step). In reality, they are embedded within geometries and have spatial and contextual relationships with the other faces within the geometry. Leveraging information from the whole geometry during uncertainty estimation could be useful. Alternatively, structural quality metrics like those used in image segmentation could be useful here for deciding whether a geometry as a whole has poor predictions rather than individual b-rep faces.

Acknowledgements

The authors acknowledge funding from Rolls-Royce plc. The authors also acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

References

Mazin Al-Wswasi, Atanas Ivanov, and Harris Makatsoris. A survey on smart automated computer-aided process planning (acapp) techniques. *The International Journal of Advanced Manufacturing Technology*, 97(1):809–832, 2018.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Weijuan Cao, Trevor Robinson, Yang Hua, Flavien Boussuge, Andrew R Colligan, and Wanbin Pan. Graph representation of 3d cad models for machining feature recognition with deep learning. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 84003, page V11AT11A003. American Society of Mechanical Engineers, 2020.

Andrew R. Colligan, Trevor T Robinson, Declan C. Nolan, Yang Hua, and Weijuan Cao. Hierarchical cadnet: Learning from b-reps for machining feature recognition. *Computer-Aided Design*, 147, February 2022. ISSN 0010-4485. doi: 10.1016/j.cad.2022.103226.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379): 605–610, 1982.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/2987588.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/84ddfb34126fc3a48ee38d7044e87276-Paper.pdf.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319, 2020.

Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

Christian Hubschneider, Robin Hutmacher, and J. Marius Zöllner. Calibrating uncertainty models for steering angle estimation. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1511–1518, 2019. doi: 10.1109/ITSC.2019.8917207.

Pradeep Kumar Jayaraman, Aditya Sanghi, Joseph G Lambourne, Karl DD Willis, Thomas Davies, Hooman Shayani, and Nigel Morris. Uv-net: Learning from boundary representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11703–11712, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Joseph G. Lambourne, Karl D.D. Willis, Pradeep Kumar Jayaraman, Aditya Sanghi, Peter Meltzer, and Hooman Shayani. Brepnet: A topological message passing system for solid models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12773–12782, 06 2021.

Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*, 2019.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33:12756–12767, 2020.

Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.

Matthew Ng, Fumin Guo, Labonny Biswas, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, and Graham Wright. Estimating uncertainty in neural networks for cardiac mri segmentation: A benchmark study. *IEEE Transactions on Biomedical Engineering*, 70(6):1955–1966, 2023. doi: 10.1109/TBME.2022.3232730.

Christina Petschnigg and Jürgen Pilz. Uncertainty estimation in deep neural networks for point cloud segmentation in factory planning. *Modelling*, 2(1):1–17, 2021.

L. Piegl and W. Tiller. *The NURBS Book*. Monographs in Visual Communication. Springer Berlin Heidelberg, 1996. ISBN 9783540615453. URL https://books.google.co.uk/books?id=7dqY5dyAwWkC.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Jami J. Shah, David Anderson, Yong Se Kim, and Sanjay Joshi. A Discourse on Geometric Feature Recognition From CAD Models . *Journal of Computing and Information Science in Engineering*, 1 (1):41–51, 11 2000. ISSN 1530-9827. doi: 10.1115/1.1345522. URL https://doi.org/10.1115/1.1345522.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

Hristo Vassilev, Marius Laska, and Jörg Blankenbach. Uncertainty-aware point cloud segmentation for infrastructure projects using bayesian deep learning. *Automation in Construction*, 164:105419, 2024.

Gerico Vidanes, David Toal, Xu Zhang, Andy Keane, Jon Gregory, and Marco Nunez. Extending point-based deep learning approaches for better semantic segmentation in cad. *Computer-Aided Design*, 166:103629, 2024.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019a.

Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 61–72. Springer, 2019b.

Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019c.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 2024. doi: 10.1007/s11263-024-02117-4. URL https://doi.org/10.1007/s11263-024-02117-4.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616, 2001.

Hang Zhang, Shusheng Zhang, Yajun Zhang, Jiachen Liang, and Zhen Wang. Machining feature recognition based on a novel multi-task deep learning network. *Robotics and Computer-Integrated Manufacturing*, 77:102369, 2022. ISSN 0736-5845. doi: https://doi.org/10.1016/j.rcim.2022.102369.

X. Zhang, David J.J. Toal, N.W. Bressloff, A.J. Keane, F. Witham, J. Gregory, S. Stow, C. Goddard, M. Zedda, and M. Rodgers. Prometheus: a geometry-centric optimisation system for combustor design. In *ASME Turbo Expo 2014: Turbine Technical Conference and Exposition (15/06/14 - 19/06/14)*, 06 2014. URL https://eprints.soton.ac.uk/363186/.

Zhibo Zhang, Prakhar Jaiswal, and Rahul Rai. Featurenet: Machining feature recognition based on 3d convolution neural network. *Computer-Aided Design*, 101:12–22, 2018.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.