Summary

This thesis is framed at the intersection between modern Machine Learning techniques, such as Deep Neural Networks, and reliable probabilistic modeling. In many machine learning applications, we do not only care about the prediction made by a model (e.g. *this lung image presents cancer*) but also in how confident is the model in making this prediction (e.g. *this lung image presents cancer with* 67% *probability*). In such applications, the model assists the decision-maker (in this case a doctor) towards making the final decision. As a consequence, one needs that the probabilities provided by a model reflects the true underlying set of outcomes, otherwise the model is useless in practice. When this happens, we say that a model is perfectly calibrated.

Bayes Decision Rule provides a principled framework for decision making under uncertainty and guarantees optimal performance (i.e. minimum error probabilities). For Bayes Decision Rule to work, one needs to use a calibrated model since that implies that the model has (better) recovered the data generating distribution. Calibration is not the only thing that matters, but also the refinement which is the ability of the classifier to recover how the data of the different classes are separated.

However, modern machine learning techniques, such as Deep Neural Networks, are uncalibrated, which compromises their deployment in high-risk applications. Many works have attempted to solve the miscalibration of modern Deep Neural Networks, and this is one of the main objectives of this thesis.

This thesis starts by reviewing the elements involved in Bayes Decision Rule, through the lens of Proper Scoring Rules. This let us introduce one of the key concepts, at least in my personal opinion, that one should take into consideration in order to train a machine learning model. This is the data uncertainty in the target distribution, i.e. how the different samples from the different classes overlap. This observation is used in the first contribution of this thesis to justify why any Data Augmentation techniques will not guarantee calibrated distributions, even though empirical evidence has been provided in the opposite direction. We show how a proposed loss function that takes into account data uncertainty solves the miscalibration introduced by Mixup training, a state-of-the-art Data Augmentation technique.

However, since Deep Neural Networks are expensive models to train, techniques that aim at implicitly calibrate these models are costly to be deployed in practice since one has to deal with model selection techniques. To this end, the second contribution proposes to recalibrate the output of a Deep Neural Network using a Bayesian Neural Network. With this, we show that one can use expressive models as long as uncertainty is incorporated, which contrasts with many of the recent contributions that hypothesize that the calibration space is inherently simple because simpler techniques work better than their complex counterparts. We also show that the main criticism of Bayesian techniques, when applied to modern Neural Networks, is overcome by combining the capabilities of Deep Neural Networks with the proposed decoupled Bayesian Neural Network.

One of the problems of Bayesian Neural Networks is the specification of a meaningful prior in the parameter space that induces a useful prior in the function space. A bad specified prior will wrongly bias the way we quantify uncertainty with the posterior, leading to suboptimal Bayesian predictions. In the final contribution of this thesis, we introduce a new prior that directly applies to the function space, named the Transformed Gaussian Process. This new prior over functions is constructed by warping samples from a Gaussian Process using an invertible transformation. These warping functions are parameterized by Bayesian Neural Networks, which allow us to model non-stationary processes accounting for parameter uncertainty, clearly improving the performance over the point estimate counterpart. We introduce a sparse variational inference algorithm that allows us to lighten the computational burden that we would inherit from standard Gaussian Processes, to target the intractable posterior, to train the model using Stochastic variational inference and to use any observation model; among other nice properties.