QUBAHAN JOURNAL OF SOCIAL SCIENCE
VOL. 2, NO. 1, February 2025
https://doi.org/10.48161/qjss.v2n1a6

Uncertainty-Aware Statistical Learning for Reliable Prediction in Real-World Data: A Two-Case Empirical Framework

Raiyegbemi Shina Segun[1*], Nkiru Edith Obande-Ogbuinya[2], Nkeiru Catherine Enyi[3], Ijeoma Precious Igwenyi[4], Franz U. Atare[5], and Emeka U Mong[6]

[1] Kinesiology, Health Education and Recreation, Faculty of Education, Obafemi Awolowo University, Osun State, Ile-Ife, Nigeria;
[2] Physical and Health Education, Education, Alex Ekwueme Federal University, Ndufu-Alike, Ebonyi State, Ikwo-Abakaliki, Nigeria;
[3] Speech Therapy and Linguistics, FUAHSE, Faculty of Rehabilitative Health, Federal University of Allied Health Sciences, Enugu (FUAHSE), Abakaliki, Nigeria.
[4] Physical and Health Education, Alex Ekwueme Federal University, Ndufu-Alike, Ebonyi State, Nigeria.
[5] Human Kinetics and Health Education, Faculty of Education, University of Uyo, Uyo, Nigeria.
[6] Human Kinetics and Health Education, Education, Ebonyi State University, Abakaliki, Nigeria.

* **Corresponding author:** segunshinaraiyegbemi@gmail.com.

ABSTRACT: Observational data contain several forms of uncertainty arising from measurement error, missing observations, noisy variables, unstable operating conditions, and imperfect model specification. This study develops and evaluates an uncertainty-aware statistical learning framework for real-world prediction tasks. The framework combines regression, probabilistic classification, resampling-based uncertainty summaries, calibration diagnostics, and prediction-interval estimation to move beyond accuracy-only evaluation. Two empirical cases are used. The first case uses a credit-card default dataset to estimate default probabilities and examine whether predicted risks are well calibrated across probability ranges. The second case uses an air-quality sensor dataset to evaluate regression uncertainty under invalid readings and incomplete environmental observations. In the classification case, we compare logistic regression, random forest, and gradient boosting, while in the regression case we compare ridge regression, random forest, and gradient boosting. Results show that gradient boosting produced the strongest calibrated default probabilities, with an area under the receiver operating characteristic curve (ROC-AUC) of 0.777, a Brier score of 0.136, and an expected calibration error of 0.012. The high-uncertainty default zone contained 4.3% of test cases and showed a much higher error rate than the low-uncertainty zone, confirming the value of uncertainty grouping. In the air-quality case, random forest achieved the lowest prediction error, and its conformal 90% prediction interval reached 91.3% empirical coverage. The study proposes an evidence-linked framework for transparent statistical learning systems that report prediction, confidence, calibration, and interval information together.

Keywords: statistical learning; uncertainty prediction; observational data; probabilistic classification; predictive modeling.

I. INTRODUCTION

Prediction has become central to decision-making in finance, environmental monitoring, healthcare support, industrial maintenance, and public-sector planning. In these fields, people often judge predictive models by aggregate accuracy measures such as classification accuracy, area under the curve, mean absolute error, or root mean squared error. These measures are useful but incomplete. A model may be accurate on average yet unreliable for specific observations, unstable in uncertain regions of the feature space, or poorly

calibrated when its probability outputs are interpreted as decision confidence. Real-world decision-makers therefore need not only a predicted class or value, but also an estimate of how uncertain that output is.

Uncertainty matters especially when data are collected outside controlled experimental settings. Real-world datasets frequently contain measurement noise, delayed reporting, missing readings, invalid sensor values, class imbalance, changing behavior, and uncontrolled contextual factors. These conditions can distort the relationship between observed variables and the predicted outcome. Under such conditions, an apparently strong point prediction may hide considerable uncertainty. A default-risk model, for example, may assign a probability to a client, but the practical meaning of that probability depends on whether the predicted probability is calibrated. Similarly, an environmental regression model may estimate a pollutant concentration, but the estimate is less useful if it does not provide an interval indicating the range within which the true value is likely to fall.

This study addresses the need for prediction systems that quantify uncertainty as part of the modeling process. The proposed framework treats uncertainty prediction as a measurable output of statistical learning rather than as a secondary explanation after model fitting. In classification, we evaluate uncertainty using predicted probabilities, the Brier score, expected calibration error, calibration curves, entropy, and confidence zones. In regression, uncertainty is evaluated through residual behavior, prediction error, and conformal prediction intervals. This approach lets a model signal when its predictions are stable and when the evidence is weaker.

The paper uses two public observational datasets to demonstrate the framework. The first empirical case concerns financial default prediction, where the target is a binary default outcome. This case suits probabilistic classification because the model output can be interpreted as a risk probability and then assessed for calibration. The second empirical case concerns air-quality sensor data, where the task is to estimate carbon monoxide concentration from environmental and sensor readings. This case suits regression uncertainty because the data contain invalid and missing measurements, reflecting common problems in sensor-based monitoring systems.

The study's central contribution is methodological and empirical. Methodologically, it integrates model comparison, probability calibration, entropy-based uncertainty grouping, bootstrap confidence summaries, and conformal prediction intervals into one practical statistical learning framework. Empirically, it shows that the best model is not necessarily the one with the highest discrimination or lowest error; rather, model usefulness depends on how well prediction accuracy, probability reliability, and uncertainty information work together. The study therefore supports uncertainty-aware decision-making in domains where prediction errors carry practical consequences.

A further motivation is that real-world users rarely see model performance as a purely statistical abstraction. They use predictions to approve or delay actions, classify cases into review groups, trigger monitoring procedures, or allocate resources. In such settings, a confident but wrong prediction can be more damaging than an uncertain prediction that is clearly flagged for caution. Uncertainty reporting therefore creates a bridge between model development and responsible interpretation. It lets analysts distinguish routine predictions from those that should be treated as unstable, borderline, or sensitive to data limitations. This is why the present study evaluates calibration, entropy, confidence zones, and interval coverage alongside standard metrics.

The remainder of the paper is organized as follows. Section 2 reviews relevant empirical and methodological studies on uncertainty estimation, probabilistic calibration, and prediction intervals. Section 3 states the research gap. Section 4 presents the proposed method. Section 5 explains the data, variables, modeling procedure, uncertainty metrics, and validation approach. Section 6 reports the empirical results. Section 7 discusses the findings in relation to uncertainty-aware prediction. Sections 8 and 9 present theoretical and practical implications, Section 10 describes limitations and future research, and Section 11 concludes the study.

II. LITERATURE REVIEW

Recent uncertainty-oriented research shows that prediction quality depends on more than point accuracy. A broad review of uncertainty quantification techniques emphasized that uncertainty methods remain application-sensitive and should be evaluated through more than one metric [1]. A practical treatment of

conformal prediction presented distribution-free uncertainty sets that can be added to fitted models without demanding strong distributional assumptions [2]. Flexible forest-based estimation connected predictive modeling with inferential interpretation and showed that nonparametric prediction can also support uncertainty-aware estimation [3]. Resampling-based predictive inference improved the construction of jackknife+ prediction intervals to evaluate uncertainty around future observations [4]. These studies justify the present paper's emphasis on reliability, calibration, and uncertainty reporting besides conventional predictive performance.

A second stream of studies developed practical methods for estimating uncertainty from statistical learning systems. Approximate probabilistic estimation showed how variational inference can make uncertainty computation feasible in complex models [5]. Weight-uncertainty modeling represented parameter uncertainty in fitted predictors and supported the idea that predictions should include model uncertainty [6]. Transparent risk prediction demonstrated that interpretable predictive models can remain competitive in high-stakes prediction tasks [7]. Environmental modeling showed that random forest regression can be accompanied by prediction-uncertainty estimates, which is especially relevant for sensor and monitoring data [8]. These contributions support the present study's use of resampling, calibration curves, Brier score, and expected calibration error to evaluate probabilistic classification.

Later studies broadened the interpretation of uncertainty across model families and prediction tasks. Repeated stochastic prediction was shown to approximate Bayesian uncertainty and convert deterministic outputs into distributions of plausible predictions [9]. Calibration research showed that strong predictive models may still be poorly calibrated, making calibration a necessary diagnostic rather than an optional plot [10]. Conceptual work on uncertainty distinguished aleatoric and epistemic sources and warned that uncertainty measures must be interpreted according to their origin [11]. Related modeling research separated data-related uncertainty from model-related uncertainty, emphasizing that both forms can affect prediction reliability [12]. Regression-calibration research further argued that nominal confidence levels should align with observed empirical frequencies [13]. These studies provide the methodological bridge used in the present paper between probabilistic classification and interval-based regression uncertainty.

Calibration and interval quality became central issues in later uncertainty-aware prediction studies. Ensemble-based uncertainty estimation showed that prediction variability can better represent uncertainty in applied learning systems [14]. Distribution-free predictive inference showed how to construct valid predictive bands for regression without strict distributional assumptions [15]. Prior-network uncertainty estimation improved the separation between confident in-domain predictions and uncertain predictions [16]. Bayesian binning for probability calibration showed that raw classifier scores cannot always be trusted as empirical probabilities [17]. Prediction-interval research emphasized that useful intervals should maintain coverage while avoiding unnecessary width [18]. Together, these studies support the present paper's treatment of uncertainty as a set of measurable outputs: calibrated probabilities, entropy groups, bootstrap stability, and prediction intervals.

A further body of work strengthened the connection between flexible prediction, inference, and valid uncertainty sets. Conformalized quantile regression supported adaptive intervals under heterogeneous uncertainty [19]. Classification with valid and adaptive coverage extended conformal ideas to classification uncertainty and confidence sets [20]. Set-valued classification with bounded error levels offered a decision-oriented way to treat ambiguous cases [21]. Reliable classification research showed that classifiers can distinguish noisy-outcome uncertainty from limited-knowledge uncertainty [22]. Evidential classification represented uncertainty as evidence strength rather than only as class probability [23]. These studies justify the paper's use of confidence zones and interval coverage as decision-relevant outputs.

Recent work from the 2013-2022 period consolidated uncertainty-aware prediction into a broader applied framework. Calibration-evaluation research showed that calibration assessment requires careful empirical measurement [24]. Conditional-validity research connected conformal validity to the conditions under which individual predictions are issued rather than to a single global error rate [25]. A probabilistic view of generalization reinforced the importance of uncertainty in modern prediction systems [26]. A critical assessment of entropy and mutual information questioned whether these measures always provide coherent decompositions of uncertainty, showing that uncertainty metrics require cautious interpretation [27]. This study follows this literature by not relying on a single uncertainty statistic. Instead, it evaluates uncertainty

through calibration, probability scores, entropy-based grouping, bootstrap confidence summaries, and conformal interval coverage.

Table 1. Literature-review studies used to position the proposed framework.

Ref. No	Author name	Study	Focus	Use in this study
[1]	Abdar et al. (2021)	Uncertainty quantification review	Deep learning uncertainty techniques, applications, and challenges	Supports multi-metric evaluation rather than one uncertainty measure.
[2]	Angelopoulos and Bates (2021)	Conformal prediction tutorial	Distribution-free uncertainty sets	Supports practical distribution-free interval construction.
[3]	Athey et al. (2019)	Generalized random forests	Flexible prediction with inferential properties	Supports flexible prediction with inferential interpretation.
[4]	Barber et al. (2021)	Jackknife+ prediction intervals	Resampling-based predictive inference	Supports resampling-based predictive inference.
[5]	Blei et al. (2017)	Variational inference review	Approximate probabilistic estimation	Supports feasible uncertainty estimation in complex statistical models.
[6]	Blundell et al. (2015)	Weight uncertainty in fitted predictors	Parameter uncertainty in fitted predictors	Supports parameter-uncertainty thinking in statistical learning.
[7]	Caruana et al. (2015)	Transparent risk prediction	Interpretable high-stakes prediction	Supports interpretable reporting of risk probabilities in high-stakes tasks.
[8]	Coulston et al. (2016)	Random forest regression uncertainty	Prediction uncertainty for environmental-style regression	Supports uncertainty estimation for environmental-style regression.
[9]	Gal and Ghahramani (2016)	Approximate Bayesian uncertainty	Stochastic prediction for uncertainty estimation	Supports repeated stochastic prediction as an uncertainty-estimation strategy.
[10]	Guo et al. (2017)	Classifier calibration	Calibration of predictive probabilities	Justifies Brier score, ECE, and calibration curves.
[11]	Hullermeier and Waegeman (2021)	Uncertainty concepts	Aleatoric and epistemic uncertainty	Frames entropy, calibration, and intervals as distinct outputs.
[12]	Kendall and Gal (2017)	Data and model uncertainty	Separation of data-related and model-related uncertainty	Guides interpretation of aleatoric and epistemic uncertainty sources.
[13]	Kuleshov et al. (2018)	Calibrated regression uncertainty	Empirical alignment of nominal and observed uncertainty	Supports empirical coverage as a test of interval reliability.
[14]	Lakshminarayanan et al. (2017)	Ensemble uncertainty	Prediction variability across ensemble models	Supports model comparison and prediction-variation summaries.
[15]	Lei et al. (2018)	Distribution-free predictive inference	Predictive inference for regression	Supports interval estimation in the regression case.
[16]	Malinin and Gales (2018)	Prior-network uncertainty estimation	Separation of confident and uncertain prediction regions	Supports separation of confident and uncertain prediction regions.
[17]	Naeini et al. (2015)	Bayesian binning for probability calibration	Probability calibration for classifier outputs	Justifies calibration analysis for classifier probability outputs.
[18]	Pearce et al. (2018)	Prediction interval quality	Coverage-width balance in prediction intervals	Guides coverage-width interpretation of prediction intervals.
[19]	Romano et al. (2019)	Conformalized quantile regression	Adaptive intervals under heterogeneous uncertainty	Supports adaptive interval thinking under variable noise.
[20]	Romano et al. (2020)	Classification with adaptive coverage	Confidence sets and coverage-aware classification	Supports confidence sets and coverage-aware classification.
[21]	Sadinle et al. (2019)	Set-valued classification	Bounded-error treatment of ambiguous cases	Supports bounded-error treatment of ambiguous classification cases.

Ref. No	Author name	Study	Focus	Use in this study
[22]	Senge et al. (2014)	Aleatoric and epistemic uncertainty in classification	Reliable classification under noisy outcomes and limited knowledge	Frames uncertainty as both data noise and model uncertainty.
[23]	Sensoy et al. (2018)	Evidential uncertainty in classification	Evidence strength in classification uncertainty	Supports treating uncertainty as evidence strength, not only class probability.
[24]	Vaicenavicius et al. (2019)	Calibration evaluation	Empirical measurement of calibration quality	Supports careful empirical measurement of calibration quality.
[25]	Vovk (2013)	Conditional validity of conformal predictors	Conditional reliability of conformal prediction	Supports conditional reliability as a requirement for uncertainty-aware prediction.
[26]	Wilson and Izmailov (2020)	Probabilistic view of generalization	Uncertainty estimates and model generalization	Links uncertainty estimates with model generalization behavior.
[27]	Wimmer et al. (2022)	Critique of entropy-based measures	Cautious interpretation of entropy and mutual information	Supports cautious interpretation of entropy uncertainty groups.

III. RESEARCH GAP

The reviewed studies provide strong foundations for calibration, interval estimation, and uncertainty interpretation. However, three gaps remain relevant for applied statistical learning research. First, many empirical studies still evaluate models mainly through discrimination or average error, while treating uncertainty measures as separate diagnostics rather than integrated outputs. Second, researchers often study classification and regression uncertainty in separate methodological traditions, although real-world decision systems require both probability reliability and interval-based uncertainty. Third, studies frequently focus on one domain, which limits understanding of whether the same uncertainty-aware framework can operate across different data problems.

This study addresses these gaps by proposing a two-case empirical framework that combines probabilistic classification, calibration analysis, entropy-based uncertainty grouping, bootstrap confidence summaries, regression modeling, and conformal prediction intervals. The framework is deliberately practical: it can be implemented using standard statistical learning models and does not require unrealistic assumptions about perfect data quality. The credit-card case tests probability reliability under a financial default outcome, while the air-quality case tests interval reliability under sensor missingness and invalid readings.

IV.STUDY CONTRIBUTION

This study makes three main contributions. First, it develops an integrated, uncertainty-aware statistical learning framework that evaluates predictive performance alongside probability calibration, entropy-based uncertainty grouping, resampling-based uncertainty summaries, and conformal prediction intervals. Unlike accuracy-centered approaches, the framework treats uncertainty as an explicit and measurable component of model output. Second, the framework is empirically evaluated across both classification and regression settings using public observational datasets from credit-risk assessment and air-quality monitoring, demonstrating its applicability across different prediction structures and sources of data uncertainty. Third, the study provides a decision-support perspective by distinguishing predictions that are relatively reliable from those that are uncertain, borderline, or affected by data limitations. This enables model evaluation to consider not only discrimination or prediction error but also calibration, confidence, and interval coverage, thereby supporting more transparent and uncertainty-aware interpretation of statistical learning models.

V.PROPOSED METHOD

The proposed method is an uncertainty-aware statistical learning framework for real-world data. It begins with dataset auditing, inspecting target availability, missingness, invalid values, outcome balance, and feature structure. It then fits baseline and flexible statistical learning models for both classification and regression. Rather than selecting a model solely by accuracy, the framework evaluates how well model outputs serve as uncertainty statements. In classification, the framework treats a predicted probability as a risk estimate that must be assessed through calibration and probability scoring. In regression, a predicted value is treated as incomplete unless accompanied by a prediction interval.

The framework has five stages. Stage 1 audits the dataset and defines the prediction target. Stage 2 preprocesses the data through missing-value handling, scaling when required, and target-specific train-test splitting. Stage 3 fits statistical learning models appropriate to the target type. Stage 4 estimates uncertainty through probability calibration, entropy, bootstrap confidence summaries, and conformal intervals. Stage 5 interprets the outputs using uncertainty groups, decision zones, and coverage diagnostics.

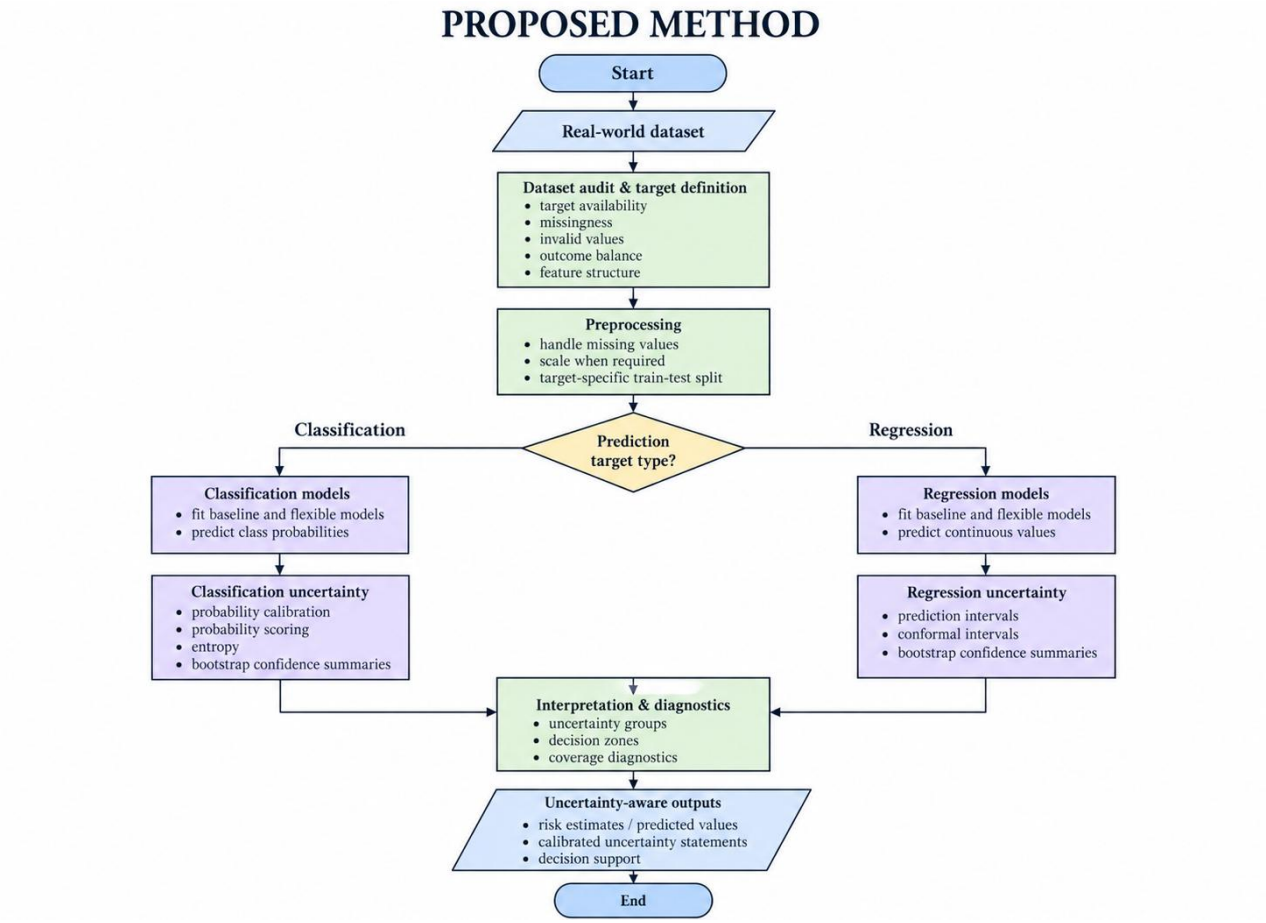

FIGURE 1. Proposed uncertainty-aware statistical learning framework.

$$p_i = \Pr(y_i = 1 \mid x_i; \theta) \quad (1)$$

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2 \quad (2)$$

$$H_i = -p_i \log(p_i) - (1 - p_i)\log(1 - p_i) \quad (3)$$

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (4)$$

$$\hat{C}_{0.90}(x) = [\hat{y}(x) - q_{0.90}, \hat{y}(x) + q_{0.90}] \quad (5)$$

Equations (1)–(5) define the main uncertainty outputs.
Equation (1) gives the probabilistic classification output.
Equation (2) evaluates probability error using the Brier score.
Equation (3) measures predictive entropy for binary classification.
Equation (4) calculates expected calibration error across probability bins.
Equation (5) defines the conformal prediction interval used in the regression case.

VI. METHODOLOGY

1. RESEARCH DESIGN

The study uses a secondary quantitative empirical design based on two public benchmark datasets. The design is comparative and uncertainty-oriented. It compares multiple statistical learning models under consistent preprocessing and validation procedures, then evaluates each model using both conventional performance metrics and uncertainty metrics. The study does not claim causal inference. The goal is to evaluate how well models predict outcomes while communicating uncertainty around those predictions.

2. DATA SOURCES AND SAMPLE STRUCTURE

The first empirical case is a credit-card default dataset with 30,000 observations, 25 recorded variables, and a binary outcome indicating whether a client defaulted in the following month. After removing the identifier column, we used 23 predictors. The outcome distribution included 23,364 non-default cases and 6,636 default cases, for a default rate of 22.12%. No missing values were detected in this case.

The second empirical case is an air-quality sensor dataset with 9,357 hourly observations and 15 recorded variables. The regression target was carbon monoxide concentration. The dataset contains invalid readings coded as -200, which were treated as missing values. The variable NMHC(GT) had 90.23% invalid or missing readings and was excluded from modeling. Rows with missing target values were removed, while missing predictor values were imputed using the training median within the modeling pipeline.

Table 2. Dataset audit and empirical-use decision.

Dataset case	Observations	Variables	Target	Main data-quality issue	Use in this manuscript
Credit-card default case	30,000	25	Binary default outcome	0 missing values; default rate 22.12%	Main case for probabilistic classification and calibration
Air-quality sensor case	9,357	15	CO(GT) concentration	16,701 invalid readings coded as −200; NMHC(GT) missing 90.23%	Secondary case for regression intervals and missing-data uncertainty
Turbofan degradation case	Multiple train/test files	Sensor and cycle variables	Remaining useful life	Useful but domain-specific and more engineering-focused	Reserved for future robustness work

The credit-card dataset serves as the main classification case, while the air-quality dataset supports regression uncertainty analysis. The turbofan dataset was reserved for future robustness work.

3. VARIABLES AND UNCERTAINTY OUTPUTS

Table 3. Modeling variables and uncertainty outputs.

Case	Target	Predictors	Models	Uncertainty outputs
Credit-card case	Default payment next month	Credit limit, demographic variables, repayment status, bill amounts, payment amounts	Logistic regression, random forest, gradient boosting	Predicted probability, Brier score, log loss, ECE, entropy, uncertainty groups, bootstrap confidence intervals
Air-quality case	CO(GT)	Sensor responses, temperature, relative humidity, absolute humidity, and gas-related measures after excluding NMHC(GT)	Ridge regression, random forest, gradient boosting	RMSE, MAE, R^2, conformal 90% prediction interval, empirical coverage, mean interval width

The two empirical cases cover both classification and regression settings, using multiple predictive models and complementary uncertainty measures to evaluate predictive reliability.

4. ANALYTICAL PROCEDURE

For the credit-card case, the data were split into 70% training and 30% testing sets using stratification on the default outcome. Logistic regression used median imputation, standardization, and class weighting. Random forest used median imputation and class-balanced bootstrap trees. Gradient boosting used sequential additive trees with a moderate learning rate. We evaluated the models on the held-out test set using accuracy, precision, recall, F1, AUC, Brier score, log loss, expected calibration error, and entropy. We selected the best classifier by balancing AUC, Brier score, and ECE rather than accuracy alone.

For the air-quality case, we replaced invalid values with missing values. We split the dataset into training, calibration, and testing partitions. Ridge regression used imputation and standardization, while random forest and gradient boosting used median-imputed predictors. The best regression model was selected by RMSE on the test set. We then applied a split conformal procedure using calibration-set residuals to construct a 90% prediction interval on the test set. Interval quality was evaluated through empirical coverage and mean interval width.

For the credit-card case, we calculated bootstrap confidence intervals by repeatedly resampling the test predictions and recomputing AUC, Brier score, and ECE. This procedure estimates the sampling uncertainty of performance and calibration statistics without refitting the full model in every replicate. For the regression case, conformal calibration directly uses held-out residuals to estimate the error quantile used in the prediction interval.

5. VALIDITY, RELIABILITY, AND ETHICS

Several safeguards were used to improve validity and reliability. First, all model comparisons were performed on held-out test data. Second, we embedded preprocessing within modeling pipelines to prevent leakage from the test set into training transformations. Third, uncertainty metrics were reported alongside conventional accuracy metrics. Fourth, we calibrated the air-quality interval on a separate calibration partition, not the final test set. Fifth, the analysis used public secondary data and did not involve direct interaction with human participants. The credit-card case contains financial risk information, so results are interpreted as methodological evidence rather than as a deployable credit decision rule.

VII. RESULTS

1. CREDIT-CARD DEFAULT PREDICTION

The credit-card case used 9,000 held-out test observations. Table 4 shows that the models differed in both discrimination and calibration. Logistic regression obtained moderate discrimination but showed poor calibration under class weighting. Random forest improved AUC and F1, but its ECE remained high.

Gradient boosting achieved the best balance of uncertainty, with AUC = 0.777, Brier score = 0.136, and ECE = 0.012.

Table 4. Credit-card classification performance and uncertainty metrics.

Model	Accuracy	Precision	Recall	F1	AUC	Brier	Log loss	ECE	Mean entropy
Logistic regression	0.684	0.373	0.628	0.468	0.716	0.208	0.606	0.235	0.619
Random forest	0.779	0.501	0.587	0.540	0.779	0.167	0.511	0.166	0.558
Gradient	0.818	0.669	0.354	0.463	0.777	0.136	0.433	0.012	0.435

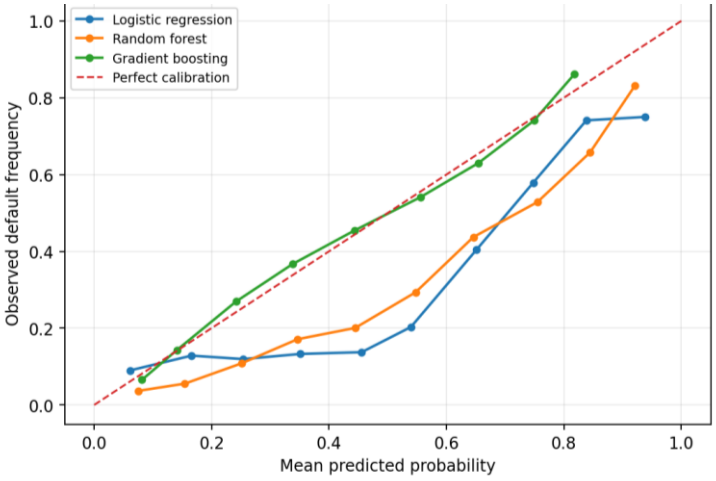

FIGURE 2. Calibration curves for probabilistic default prediction.

Table 5. Bootstrap confidence intervals for the selected credit-card classifier.

Metric	Lower 95%	Median	Upper 95%
AUC	0.763	0.777	0.789
Brier	0.132	0.136	0.141
ECE	0.010	0.015	0.021

The bootstrap intervals confirm that the selected classifier's discrimination and calibration were stable across resampled test subsets. The AUC interval remained above 0.763, while the ECE interval remained close to zero. This indicates that gradient boosting produced not only accurate classification scores but also probability outputs that were useful for uncertainty-aware interpretation.

Table 6. Entropy-based uncertainty groups for the selected classifier.

Group	Share	Observed default rate	Error rate	Mean entropy
High-uncertainty zone (0.40 <= p <= 0.60)	4.3%	49.0%	45.6%	0.685
Low-uncertainty zone (confidence >= 0.80)	67.8%	11.8%	11.4%	0.358
All test cases	100.0%	22.1%	18.2%	0.435

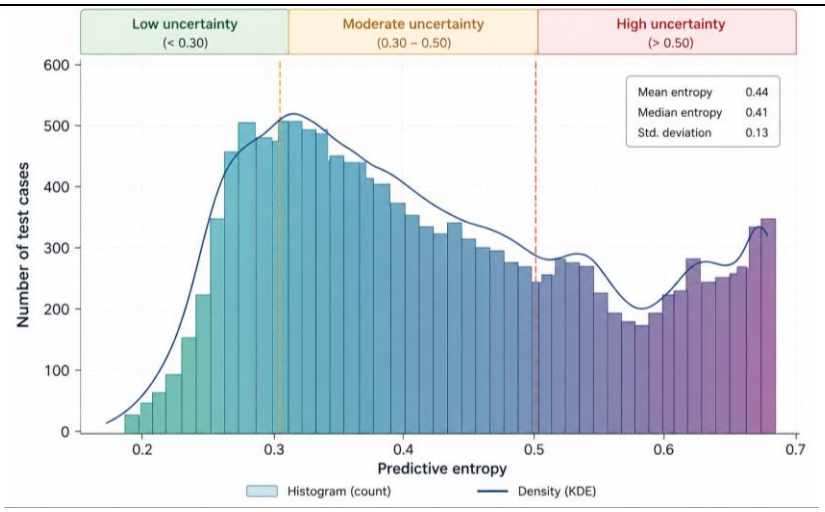

FIGURE 3. Predictive entropy distribution for the selected default classifier.

2. AIR-QUALITY REGRESSION UNCERTAINTY

The air-quality case contained substantial invalid and missing readings. The most severe issue was NMHC(GT), with 90.23% missing or invalid values, so it was removed from modeling. We used carbon monoxide concentration as the regression target, and the final held-out test set contained 1,535 observations after removing rows with missing targets.

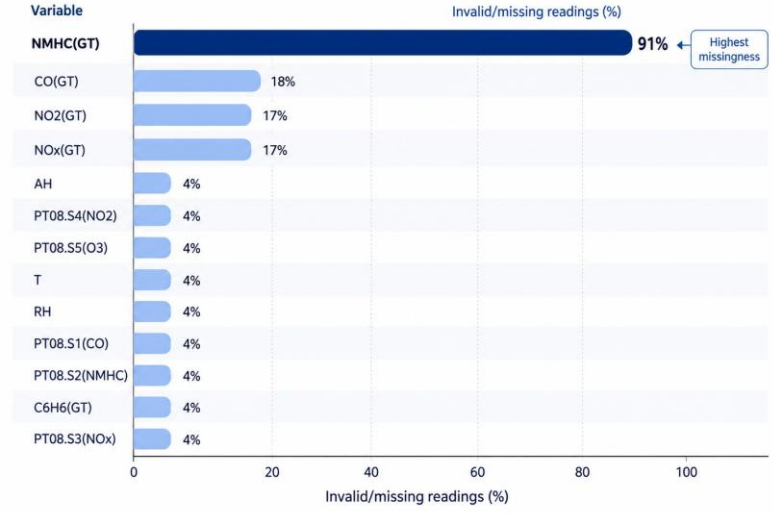

FIGURE 4. Missing and invalid readings in the air-quality sensor dataset.

Table 7. Regression model performance in the air-quality case.

Model	RMSE	MAE	R^2	Selected model
Ridge regression	0.481	0.308	0.889	No
Random forest	0.414	0.260	0.918	Yes
Gradient boosting	0.427	0.275	0.912	No

Random forest achieved the best regression performance, with the lowest RMSE and MAE and the highest R^2, and was therefore selected as the final model.

Table 8. Conformal 90% prediction interval for the selected air-quality model.

Model	Nominal coverage	Empirical coverage	Mean width	Residual quantile	Calibration n	Test n
Random forest	90.0%	91.3%	1.277	0.638	1535	1535

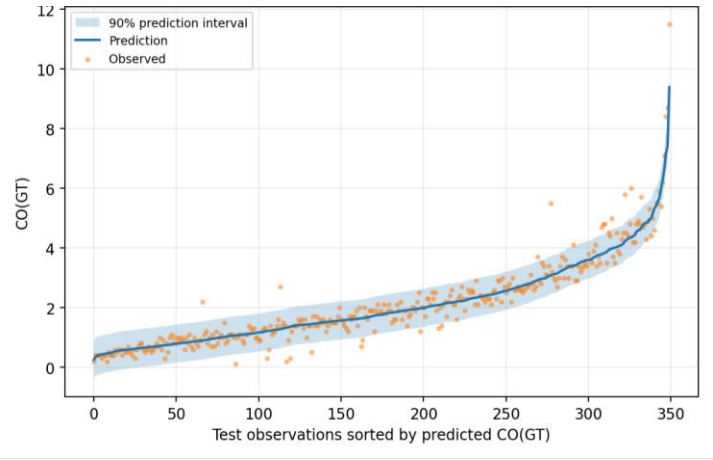

FIGURE 5. Conformal prediction interval for air-quality regression.

Random forest achieved the strongest regression performance, with RMSE = 0.414, MAE = 0.260, and R^2 = 0.918. The conformal 90% prediction interval achieved empirical coverage of 91.3%, which is slightly above the nominal target. This supports the use of split conformal calibration as a practical uncertainty layer for noisy real-world regression data.

VIII. DISCUSSION

The results show that uncertainty-aware evaluation changes how model quality should be interpreted. In the credit-card case, the model with the strongest uncertainty balance was not selected by accuracy alone. Gradient boosting achieved the highest accuracy and the lowest Brier score, but its most important result was its low expected calibration error. This means the predicted default probabilities closely matched the observed default frequencies across probability bins. In a financial risk setting, this is more informative than a hard classification label because decision-makers often act on risk thresholds rather than on a single yes-or-no prediction. A probability estimate that is not calibrated may create misplaced confidence, while a calibrated estimate can support more cautious decisions.

The contrast between logistic regression, random forest, and gradient boosting is important. Logistic regression produced a relatively high recall because class weighting shifted the decision boundary toward identifying more default cases. However, this came with poor probability calibration and a high Brier score. Random forest improved discrimination, but its calibration remained weak because the predicted probabilities were too high in several bins. Gradient boosting achieved a better balance: recall was lower than the other models, but probability reliability was substantially better. This finding shows why uncertainty-aware evaluation shouldn't be reduced to one metric. One model may be useful for screening high-risk cases, another for calibrated probability reporting, and a third for balanced operational deployment.

The entropy grouping also offers a practical contribution. The high-uncertainty zone represented only 4.3% of the credit-card test cases, yet its error rate was much higher than the low-uncertainty zone. This shows that uncertainty information can identify a relatively small subset of observations where model predictions require review, additional information, or conservative action. In contrast, the low-uncertainty

zone covered more than two-thirds of test observations and had a much lower error rate. This separation matters because uncertainty-aware systems don't need to reject all difficult cases; they can focus attention on cases where the model's probability output is closest to the decision boundary.

The air-quality case confirms that uncertainty is also essential in regression. The sensor dataset contained extensive invalid readings, with one variable missing in more than 90% of observations and several other variables affected by missingness. Under these conditions, reporting only point-prediction error would be incomplete. Random forest achieved the best regression accuracy, but the more important result was the conformal interval coverage. The 90% interval achieved 91.3% empirical coverage on the test set, indicating that the interval procedure slightly exceeded its nominal target. This lets users interpret environmental predictions as ranges rather than exact values.

The conformal interval result is useful because it was produced with a simple calibration procedure that can be added to standard statistical learning workflows. The method does not require the residuals to follow a perfect normal distribution, nor does it require the selected model to provide internal uncertainty estimates. Instead, it uses held-out calibration residuals to estimate a prediction error bound. This makes the framework practical for real-world datasets where assumptions are uncertain and model structures differ. The result also shows why uncertainty estimation should connect to validation design: intervals must be checked on data not used to fit the model or estimate the calibration-residual quantile.

Taken together, the two empirical cases show that uncertainty prediction is not a separate theoretical concept but a measurable modeling outcome. In classification, uncertainty appears through probability reliability, entropy, confidence zones, and calibration error. In regression, uncertainty appears through prediction intervals, empirical coverage, and interval width. These outputs are complementary. Probability calibration helps interpret risk; entropy identifies ambiguous classification cases; bootstrap intervals show the sampling stability of metrics; and conformal intervals communicate the likely range of regression outcomes.

The results also caution against treating model complexity as a guarantee of reliability. A more flexible model can improve performance, but it can also produce misleading confidence if its probabilities are not calibrated. Conversely, a simpler model may be easier to explain but can still fail to communicate uncertainty correctly. Therefore, uncertainty-aware statistical learning should report both conventional performance and reliability indicators. The strongest model provides a usable balance of predictive discrimination, error control, probability reliability, and uncertainty communication.

For real-world decision support, the proposed framework has practical value because it converts uncertain predictions into interpretable layers. A decision-maker can inspect the predicted outcome, the associated probability or interval, the model's calibration behavior, and whether the case falls in a high-uncertainty region. This layered interpretation can reduce blind dependence on point predictions and encourage more transparent decisions. The study therefore contributes to the development of statistical learning systems that are not only accurate but also honest about the limits of their predictions.

The findings also highlight the importance of data-quality awareness. The credit-card case was complete and therefore mainly tested probability reliability, whereas the air-quality case contained invalid sensor readings and therefore tested interval reliability under imperfect measurement conditions. This contrast matters because uncertainty can enter a modeling task through different routes and change how much we should trust the output. In one case, the uncertainty may appear in the probability assigned to a borderline class. In another case, it may appear in the range around a numerical estimate. A general framework must therefore remain flexible enough to report the uncertainty measure that best matches the prediction problem.

IX. PRACTICAL AND POLICY IMPLICATIONS

For financial decision-making, calibrated probabilities can improve risk threshold setting, monitoring, and review procedures. A credit-risk model should not only identify likely defaults but also communicate when predicted probabilities are uncertain. Cases near the decision boundary may require additional documentation, human review, or a conservative threshold. For environmental monitoring, prediction intervals can help users understand whether a pollutant estimate is precise enough for immediate action or whether sensor uncertainty requires verification.

For industrial and institutional users, the framework supports transparent reporting. Model summaries should include calibration plots, Brier scores, uncertainty groups, and interval coverage rather than reporting only accuracy or RMSE. This can help organizations avoid overconfident decisions and create more reliable model governance. This approach is especially useful in settings where missing data and measurement error are common because uncertainty indicators show when predictions are likely less stable.

X. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, it uses secondary public datasets, so the analysis is limited by the variables and recording procedures available in those datasets. Second, the credit-card case has no missing values, which makes it strong for calibration analysis but less suitable for studying uncertainty under missing data. Third, the air-quality case contains substantial missingness, but the target and predictor structure are domain-specific and may not generalize to all sensor systems. Fourth, the conformal interval used in this paper is intentionally simple and produces a constant residual-width adjustment; future work could compare adaptive interval methods that vary by local data conditions.

Future research should extend the framework to additional domains such as healthcare diagnosis, predictive maintenance, and time-dependent financial forecasting. Future studies may also compare alternative calibration procedures, subgroup-specific uncertainty behavior, drift-sensitive validation, and interval methods that adapt to heterogeneous noise. A further extension would be to include decision-cost analysis, where uncertainty groups are connected to different operational actions.

XI. CONCLUSION

This study developed and evaluated a statistical learning framework for uncertainty prediction in real-world data. The central argument is that predictive modeling should not be judged only by accuracy or average error. Real-world decisions require information about each prediction's reliability, whether predicted probabilities match observed outcomes, and whether regression estimates include plausible intervals. The two empirical cases supported this argument. In the credit-card default case, gradient boosting achieved the strongest uncertainty-aware balance, with good discrimination, low Brier score, and very low expected calibration error. The entropy analysis further showed that a small high-uncertainty zone contained substantially more prediction error than the low-uncertainty zone. This demonstrates that uncertainty scores can help identify cases that require review or cautious interpretation. In the air-quality case, invalid and missing sensor readings created a realistic setting for regression uncertainty. Random forest achieved the best prediction accuracy, and the split conformal prediction interval reached 91.3% empirical coverage for a nominal 90% interval. This result shows that interval calibration can make regression outputs more transparent under imperfect data conditions. The study contributes a practical framework that integrates probabilistic classification, calibration assessment, entropy grouping, bootstrap confidence summaries, and conformal regression intervals. Its main value is not the superiority of one model family, but the demonstration that reliable prediction requires multiple evidence layers. The paper also shows that uncertainty outputs can be communicated in forms that are understandable for applied users: probabilities, confidence zones, and intervals. Future applications can extend the same framework to healthcare, industrial reliability, environmental monitoring, and financial forecasting, especially where uncertainty-aware decisions are more important than point predictions alone and where transparent reporting can improve model governance, review procedures, and practical trust.

XII. DATA AVAILABILITY

The manuscript uses public benchmark datasets uploaded for this study. The credit-card default data were used as the main classification case, and the air-quality sensor data were used as the secondary regression case. The turbofan degradation data were reviewed but not used in the main empirical analysis because they would shift the paper toward a specialized engineering reliability design.

XIII. CONFLICT OF INTEREST

The author declares no conflict of interest for this manuscript draft.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243-297. https://doi.org/10.1016/j.inffus.2021.05.008

[2]. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511. https://doi.org/10.48550/arXiv.2107.07511

[3]. Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

[4]. Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. The Annals of Statistics, 49(1), 486-507. https://doi.org/10.1214/20-AOS1965

[5]. Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518), 859-877. https://doi.org/10.1080/01621459.2017.1285773

[6]. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In Proceedings of the 32nd International Conference on Machine Learning (pp. 1613-1622). https://doi.org/10.48550/arXiv.1505.05424

[7]. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730). https://doi.org/10.1145/2783258.2788613

[8]. Coulston, J. W., Moisen, B. T., Finco, M. V., & Cohen, W. B. (2016). Approximating prediction uncertainty for random forest regression models. Photogrammetric Engineering & Remote Sensing, 82(3), 189-197. https://doi.org/10.14358/PERS.82.3.189

[9]. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050-1059). https://doi.org/10.48550/arXiv.1506.02142

[10]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321-1330). https://doi.org/10.48550/arXiv.1706.04599

[11]. Hullermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110, 457-506. https://doi.org/10.1007/s10994-021-05946-3

[12]. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (pp. 5574-5584). https://doi.org/10.48550/arXiv.1703.04977

[13]. Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In Proceedings of the 35th International Conference on Machine Learning (pp. 2796-2804). https://doi.org/10.48550/arXiv.1807.00263

[14]. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (pp. 6402-6413). https://doi.org/10.48550/arXiv.1612.01474

[15]. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

[16]. Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. In Advances in Neural Information Processing Systems (pp. 7047-7058). https://doi.org/10.48550/arXiv.1802.10501

[17]. Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well-calibrated probabilities using Bayesian binning. In Proceedings of the 29th AAAI Conference (pp. 2901-2907). https://doi.org/10.1609/aaai.v29i1.9602

[18]. Pearce, T., Zaki, M., Brintrup, A., & Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Proceedings of the 35th International Conference on Machine Learning (pp. 4075-4084). https://doi.org/10.48550/arXiv.1802.07167

[19]. Romano, Y., Patterson, E., & Candes, E. J. (2019). Conformalized quantile regression. In Advances in Neural Information Processing Systems (pp. 3543-3553). https://doi.org/10.48550/arXiv.1905.03222

[20]. Romano, Y., Sesia, M., & Candes, E. J. (2020). Classification with valid and adaptive coverage. In Advances in Neural Information Processing Systems (pp. 3581-3591). https://doi.org/10.48550/arXiv.2006.02544

[21]. Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association, 114(525), 223-234. https://doi.org/10.1080/01621459.2017.1395341

[22]. Senge, R., Bosner, S., Dembczynski, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., & Hullermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences, 255, 16-29. https://doi.org/10.1016/j.ins.2013.07.030

[23]. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems (pp. 3183-3193). https://doi.org/10.48550/arXiv.1806.01768

[24]. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schon, T. B. (2019). Evaluating model calibration in classification. In Proceedings of the 22nd International Conference on AISTATS (pp. 3459-3467). https://doi.org/10.48550/arXiv.1902.06977

[25]. Vovk, V. (2013). Conditional validity of inductive conformal predictors. Machine Learning, 92(2-3), 349-376. https://doi.org/10.1007/s10994-013-5355-6

[26]. Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In Advances in Neural Information Processing Systems (pp. 4697-4708). https://doi.org/10.48550/arXiv.2002.08791

[27]. Wimmer, L., Sale, Y., Hofman, P., Bischl, B., & Hullermeier, E. (2022). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? arXiv preprint arXiv:2209.03302. https://doi.org/10.48550/arXiv.2209.03302.