Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Mathematical and Statistical Framework For Uncertainty-Aware Machine Learning in Complex Data Environments

Sushitha S[1*], Dr Dinesh M[2], Dr Nandeesha R[3], Pragneshkumar L Thakkar[4], M Shanmugam[5], Dr Sowmyashree M S[6]

[*1]Asssistant Professor, Nagarjuna School of Computer Science, ORCID ID: https://orcid.org/0000-0003-2974-7290, Email I'd: bhatsushitha@gmail.com
[2]Department of Semiconductor Division, Product Manager, UST Global Technology, Bangalore 560066, Karnataka India. Email ID: dinesh.svdm@gmail.com
[3]Associate Professor, Departments of Electronics and Communication Engineering, Government Engineering College, Hassan-573201,Visvesvaraya Technological University, Belagavi, Karnataka, India, Email ID: rnandeesha@gmail.com, ORCID ID: https://orcid.org/0000-0001-7982-128X
[4]Associate Professor, Department of Science and Humanities, L D College of Engineering, Ahmedabad, Gujarat, ORCID ID: 0000-0001-5118-934X, Email ID: pragnesh83@ldce.ac.in
[5]Associate Professor, Departments of Electronics and Communication Engineering, Government Engineering College, Hassan-573201, Visvesvaraya Technological University, Belagavi, Karnataka, India. Email ID: mshanmugam168@gmail.com ORCID ID: https://orcid.org/0009-0003-6633-9728
[6]Associate Processor, Department of Electronics and Communication Engineering, Government Engineering College, Chamrajanagara 571313, Visvesvaraya Technological University, Belagavi, Karnataka, India, Email ID: sowmya.mtech@gmail.com, ORCID ID:https://orcid.org/0009-0002-8671-2422

ABSTRACT
High-dimensional data sets can be used to train machine-learning models that can show high discrimination abilities, but also can give probability estimates with varying calibrations and reliabilities. This study developed and evaluated a mathematical and statistical framework integrating predictive performance, calibration, uncertainty, selective prediction, and predictor stability in a complex high-dimensional classification setting. A total of 174 observations and 450 numerical predictors were included in the analysis. To assess model performance, repeated stratified five-fold cross-validation with 10 repetitions was applied to Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. The Random Forest model had the best ROC-AUC (0.961) and accuracy (0.880), while the Gradient Boosting model had the lowest Brier score (0.0925) and negative log-loss (0.3188). Support Vector Machine had the lowest value of ECE (0.0450). Predictive entropy was higher for misclassifications across all predictors and had the strongest correlation with participant-level misclassification for Gradient Boosting (rho = 0.7065). When using uncertainty-based selective prediction, the accuracy reached 1.000 for 50% coverage for Random Forest and Gradient Boosting, respectively. The predictor effects also showed good consistency between cross-validation folds. The results suggest a more complete picture of the reliability of machine learning models in high-dimensional data settings can be obtained by considering discrimination, probabilistic reliability, predictive uncertainty, selective coverage, and predictor stability together.

Keywords: uncertainty-aware machine learning, high-dimensional data, predictive entropy, probability calibration, selective prediction

INTRODUCTION

Structures that can predict the behavior of data sets with many variables, with very mixed and complex signal relationships, are often extracted using machine learning. The high dimensionality of environments is a special statistical challenge because candidate predictors can approach or even exceed the number of observations, making generalization difficult. Adaptive learning approaches are now being developed and applied to model this data, but their efficacy is reliant on the successful management of dimensional complexity and uncertainty (Wilson & Anwar, 2024). Previous studies on high-dimensional multimedia learning have also demonstrated that high-dimensional feature spaces bring along challenges in representation, model selection, and predictive robustness (Gao et al., 2017). With these problems, model performance alone is not enough to evaluate learning systems under complex data conditions. The level of

dimensionality may influence both the estimation of the statistical parameters as well as the stability of learned relationships. Reducing the number of predictors in the space makes it more tractable, but dimensionality reduction should not lose the signals of interest for the prediction (Terol et al., 2020). An added concern is the potential for strong dependence among predictors, as multicollinearity can cause model effects to be unstable and result in the effect of individual predictors being unclear (Kyriazos & Poga, 2023). The selection of the variables and the choice of regularisation parameters are thus important validation tasks, especially if sample size is limited compared to the number of predictors (Heinze et al., 2018). If the data is complex, a stringent framework for complex data should therefore not only examine the ability of a classifier to correctly classify an instance, but also the reliability of the probabilistic output and learned predictor contributions when they are repeated on new samples.

Uncertainty-aware machine learning adds to the traditional evaluation by providing metrics on the strength of the model's support for a specific prediction for a given individual. Uncertainty-aware attention has shown that it is useful to absorb uncertainty into the prediction and interpretation of a model, in lieu of using every prediction of the model as equally reliable (Heo et al., 2018). More recently, uncertainty has been introduced into classification along with ensemble learning, such as conformal deep forests, which make explicit use of uncertainty in classification decisions (Zhang et al., 2025). The same kind of approach has been used for reliable text classification: uncertainty information can be used to differentiate reliable from uncertain predictions (Hu & Khan, 2021). Multivariate and multi-view formulations also show that uncertainty could be explored in different views of the same classification problem (Kornaev et al., 2024).

The value of uncertainty for statistics relies on whether it is a measure of the predictability that can be observed. To tackle this problem, there have been conformalized deep-learning approaches that have incorporated uncertainty awareness into the training of classifiers and the construction of prediction sets while formally considering coverage (Einbinder et al., 2022). Uncertainty can also be used as evidence from machine learning-assisted measurement and can be directly related to the prediction error probability, as confidence information can be used to predict classification errors (Shirmohammadi et al., 2024). This relationship is particularly important when the models are used to produce probabilities for decisions, as two classifiers with similar discrimination might have very different probability calibration and the ability to recognize challenging observations. The complexity of the model is not necessarily superior in terms of empirical performance either, underscoring the importance of considering a range of aspects when comparing competing algorithms, as more complex learning mechanisms are not necessarily better (Yao et al., 2017). Assessing Uncertainty is more challenging when the conditions under which data are available change from those used in the model development process. Uncertainty can be problematic when the datasets shift, and comparative studies have indicated that uncertainty estimates must be validated (Ovadia et al., 2019). This concern leads to a strategy of evaluation that involves prediction and calibration and to the idea of entropy and observed error as well as to a strategy for performance via repeated resampling. An operational test is possible by looking at the impact of excluding cases of high uncertainty on the remaining cases (selective prediction). This can be supplemented in high-dimensional spaces by examining the stability of influential predictors across the re-sampled training sets, using stability analysis.

Notwithstanding these advances, uncertainty quantification, calibration, predictive discrimination, selective prediction and feature stability are often considered as individual methodological concerns. Less attention has been paid to how they can be integrated into a single mathematical and statistical model for small-sample, high-dimensional classification, and in particular whether the uncertainty in the prediction is related to repeated empirical errors. There is a need for a unified evaluation of models to differentiate the models that merely provide good discrimination from those that also yield informative probability estimates, meaningful uncertainty signals, and predictable patterns of model performance indicators on repeated internal samples.

The present study thus proposes and tests a mathematical and statistical model of uncertainty-aware machine learning (ML) in a complex high-dimensional data space. It aims at comparing the predictive discrimination and the probabilistic calibration of various linear, kernel-based, bagging, and boosting classifiers; measuring the predictive uncertainty of classifiers and its relation with the misclassification at the participant level; analyze the predictive uncertainty-guided selective prediction at different coverage levels; and analyze the cross-validation stability of predictor contributions. These goals offer a coherent foundation for assessing the model's predictive success, its representation of uncertainty and the model stability in the context of the same resampling scheme.

2. METHODOLOGY

2.1 Research Design

To create an uncertainty-aware machine-learning framework for binary classification in a high-

dimensional setting, a quantitative computational design was used. The framework was a combination of repeated cross-validation and discrimination, probability calibration, entropy-based uncertainty estimation, selective prediction and coefficient-stability analysis. To ensure fair model comparison, the same validation scheme was used for Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting.

2.2 Data Source

The data were obtained from the publicly available resource contributed by Fontanella (2022). It includes quantitative measurements of handwriting from subjects with normal handwriting and PD without the binary diagnosis. The measurements characterize temporal, spatial, kinematic, as well as pressure-related handwriting features.

2.3 Data Preparation and High-Dimensional Structure

All but the participant identifier were placed in the predictor matrix, and the class was coded H = 0 and P = 1. The rest of the variables were considered to be numerical predictors. Data screening included checking for missing and infinite values, checking for duplicate features, checking for constant features, and checking for variance of predictors. The dimensionality of the feature space was written as:

$$p / n$$

in which p is the number of predictors and n is the number of observations. For each pair of features, the pairwise Pearson correlation was computed to measure the feature dependence, and the highly correlated feature pairs (with a high Pearson correlation greater than 0.95) were identified. Preprocessing was conducted in every training fold to avoid information leakage. No predictors matched the criteria for variance filtering, but it was added to each model pipeline. Random Forest and Gradient Boosting were fitted without scaling features.

2.4 Model Development and Validation

Four classifiers were tested: Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. For each model, the predicted classes and probabilities of the positive class were generated. Internal validation was performed via repeated stratified five-fold cross-validation with 10 repetitions, thus giving 50 training/validation partitions for each of the four models. The class proportions were maintained during stratification, and pre-processing was fitted separately on each training partition and applied to the respective validation partition. The accuracy, sensitivity, specificity, precision, F1-score and ROC-AUC were used for the evaluation of predictive performance. Means, standard deviations, medians and 95% resampling intervals were calculated using repeat-level estimates.

2.5 Probabilistic Performance and Calibration

The three measures of the probability quality were used to evaluate the results: Brier score, negative log-loss, and Expected Calibration Error (ECE). The Brier score was defined for two cases; observed binary outcome y and predicted probability p:

$$\text{Brier score} = (1/N) \Sigma (p_i - y_i)^2$$

where N is the number of out-of-fold predictions.

Negative log-loss was calculated as:

$$\text{Log-loss} = -(1/N) \Sigma [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

Ten probability bins were used to create calibration curves. Each bin's mean predicted probability was compared with the actual number of positive classes (frequency) in the bin.

2.6 Predictive Uncertainty and Error Association

Predictive uncertainty was quantified using Shannon entropy of the binary class probabilities. For participant i, entropy was calculated as:

$$H_i = -[p_i \ln(p_i) + (1 - p_i) \ln(1 - p_i)]$$

where larger H_i values indicate greater predictive uncertainty. Comparisons of entropy between correct and incorrect predictions were performed. The mean predictive entropy and empirical misclassification rate were also computed by participant for repeated out-of-fold predictions. The participants were classified according to whether they were consistently correct or ever misclassified over the repetitions. The Mann-Whitney U test was used to test the difference between these groups, and the effect-size measure was rank-biserial correlation. Spearman rank correlation was employed to examine the correlation between the entropy of the participant and the empirical error rate. The model-wise multiple comparisons were corrected by Holm adjustment.

2.7 Selective Prediction

Using an entropy-based abstention procedure, it was determined if the uncertainty could be used to detect

predictions with lesser reliability. The participants' mean entropy was used to rank predictions, and the most uncertain were then eliminated. Coverage was considered as:
Coverage = Number of retained predictions / Total number of predictions
and selective accuracy as:
Selective accuracy = Correct retained predictions / Total retained predictions
Accuracy was evaluated over coverage levels from 50% to 100%, producing an accuracy-coverage curve for each classifier.

2.8 Predictor Stability

The stability of the predictors was evaluated by the standardized Logistic Regression coefficient values computed using each of 50 folds of the training set. The mean coefficient (M), the mean absolute coefficient (MAC), the standard deviation (SD), the sign consistency (SC), and the retention frequency (RF) were determined for every predictor. Sign consistency was regarded as the percentage of the cross-validation fits where the sign of the dominant coefficient did not change. The mean absolute standardized coefficient was used to rank the predictors, and the coefficient stability was quantified between folds.

2.9 Statistical Analysis

The performance of all predictive evaluations was estimated using out-of-fold estimates. The variability was represented by 95% resampling intervals, and the performance measures were summarised over the cross-validation repetitions. The differences between the groups of uncertainty were analysed non-parametrically, and uncertainty-error associations were calculated by Spearman correlation. The statistical tests were two-sided, and a p-value < 0.05 was regarded as significant after Holm adjustment if such adjustment was performed.

3. RESULTS

3.1 Dataset and Predictor-Space Characteristics

There were 174 observations and 450 numerical predictor variables for this analysis. Overall, there were 85 observations classified as H (48.85%) and 89 observations classified as P (51.15%), giving a binary outcome. The number of predictors (450) was greater than the number of observations (n = 174), and the predictor-to-observation ratio was 2.586. No predictor values were missing, and no duplicate observations or constant predictors were found. The mean absolute pairwise correlation was 0.1182, and the median absolute pairwise correlation was 0.0893 for all the predictor matrices. The highest correlation was 1.0000. 81 predictor pairs had an absolute correlation of 0.95 or higher. The composition and structural features of the data set of the predictor space are presented in Table 1.

Table 1. Dataset and structural characteristics

Characteristic	Value
Number of observations	174
Number of predictor variables	450
Predictor-to-observation ratio	2.586
Outcome classes	2
Class distribution	H: 85 (48.85%); P: 89 (51.15%)
Missing predictor values	0
Duplicate observations	0
Exact constant predictors	0
Mean absolute pairwise correlation	0.1182
Median absolute pairwise correlation	0.0893

Maximum absolute pairwise correlation	1.0000
Feature pairs with absolute r >= 0.95	81

3.2 Cross-Validated Predictive Performance

The mean accuracy values of the four classifiers obtained through the repeated stratified cross-validation were between 0.818 and 0.880. Random Forest had the highest mean accuracy at 0.880 (95% resampling interval: 0.863-0.898), followed by Gradient Boosting at 0.873 (0.853-0.890), Support Vector Machine at 0.854 (0.846-0.867), and Logistic Regression at 0.818 (0.790-0.844). Support Vector Machine (SVC) had the highest sensitivity of 0.909 (0.890-0.930). Random Forest and Gradient Boosting produced sensitivities of 0.900 (0.868-0.921) and 0.890 (0.865-0.928), respectively, while Logistic Regression produced a sensitivity of 0.799 (0.753-0.829). Specificity was 0.860 (0.829-0.880) for Random Forest, 0.855 (0.826-0.880) for Gradient Boosting, 0.838 (0.814-0.859) for Logistic Regression, and 0.796 (0.770-0.812) for Support Vector Machine. The highest F1-score was obtained by Random Forest at 0.885 (0.870-0.903), followed by Gradient Boosting at 0.878 (0.858-0.894), Support Vector Machine at 0.864 (0.856-0.876), and Logistic Regression at 0.818 (0.787-0.844). The classification scores and 95% resampling intervals for the cross-validated results are presented in Table 2.

Table 2. Comparative predictive performance with 95% resampling intervals

Model	Accuracy	Sensitivity	Specificity	F1-score	ROC-AUC
Gradient Boosting	0.873 (0.853-0.890)	0.890 (0.865-0.928)	0.855 (0.826-0.880)	0.878 (0.858-0.894)	0.954 (0.942-0.964)
Logistic Regression	0.818 (0.790-0.844)	0.799 (0.753-0.829)	0.838 (0.814-0.859)	0.818 (0.787-0.844)	0.896 (0.875-0.914)
Random Forest	0.880 (0.863-0.898)	0.900 (0.868-0.921)	0.860 (0.829-0.880)	0.885 (0.870-0.903)	0.961 (0.957-0.965)
Support Vector Machine	0.854 (0.846-0.867)	0.909 (0.890-0.930)	0.796 (0.770-0.812)	0.864 (0.856-0.876)	0.929 (0.920-0.938)

ROC-AUC was 0.961 (0.957-0.965) for Random Forest, 0.954 (0.942-0.964) for Gradient Boosting, 0.929 (0.920-0.938) for Support Vector Machine, and 0.896 (0.875-0.914) for Logistic Regression. The estimates of ROC-AUC and the 95% resampling intervals for the four classifiers are presented in Figure 1A. Each plot of the probability calibration curves (Figure 1B) displays the observed frequencies in the positive class against the mean (average) probabilities predicted by each classifier over 10 probability bins.

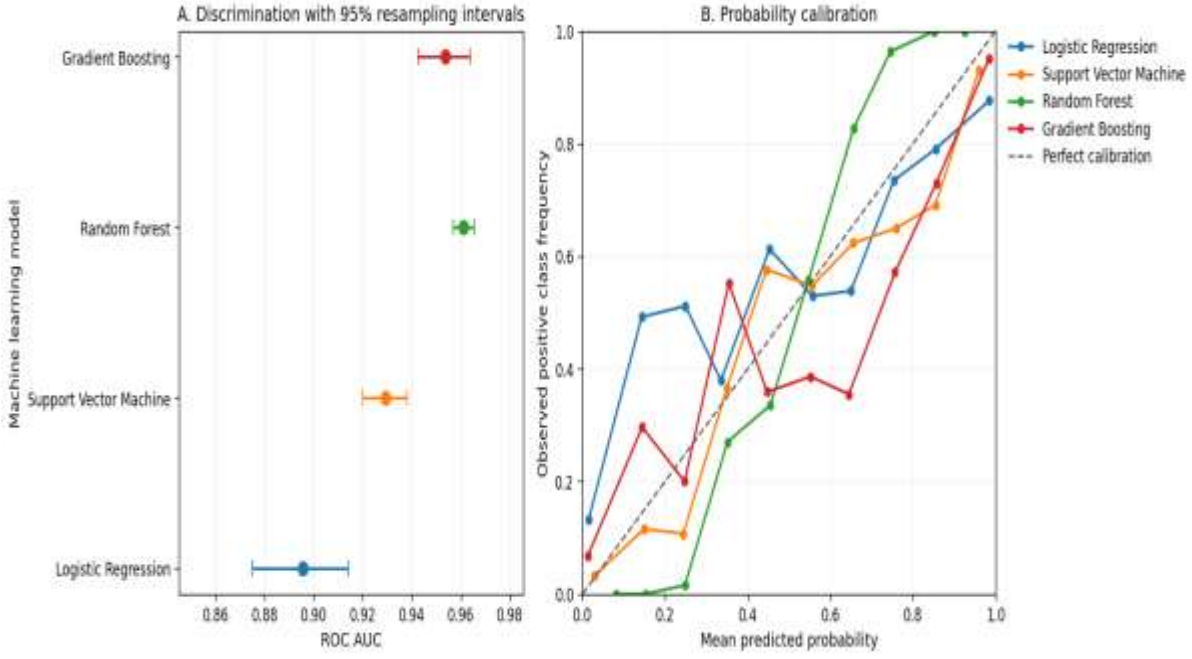

Figure 1. Predictive discrimination and calibration. (A) ROC-AUC with 95% resampling intervals. (B) Calibration curves for the four classifiers.

3.3 Probabilistic Performance and Uncertainty Estimates

The Brier score and log-loss were 0.0925 and 0.3188 for Gradient Boosting, 0.1036 and 0.3572 for Random Forest, 0.1075 and 0.3429 for Support Vector Machine, and 0.1478 and 0.7123 for Logistic Regression, respectively. The calibration error of respective algorithms was Support Vector Machine (0.0450), Gradient Boosting (0.0657), Logistic Regression (0.1188) and Random Forest (0.1391). Predictive entropy was different for correct and incorrect predictions by each of the classifiers. For Gradient Boosting, mean entropy was 0.1217 for correct predictions and 0.4273 for incorrect predictions, giving a difference of 0.3057. For Support Vector Machine, mean entropy was 0.2513 for correct predictions and 0.4853 for incorrect predictions, giving a difference of 0.2340. The values were 0.1211 and 0.3137 for Logistic Regression, with a difference of 0.1926. For Random Forest, mean entropy was 0.5085 for correct predictions and 0.6696 for incorrect predictions, giving a difference of 0.1612. The rank-biserial correlation of Gradient Boosting was 0.9577, the Random Forest was 0.9243, the Logistic Regression was 0.8272, and the Support Vector Machine was 0.7677. The Holm-adjusted p-values were 5.003×10^{-21}, 1.537×10^{-16}, 4.307×10^{-18}, and 1.198×10^{-12}, respectively. The Spearman's correlation coefficients between mean entropy at the participant level and empirical misclassification rate were 0.5125 to 0.7065. The probabilistic, calibration, entropy and uncertainty-error statistics of the models are shown in Table 3.

Table 3. Probabilistic calibration and uncertainty-aware model assessment

Model	Brier Score	Negative Log-Loss	ECE	Entropy, Correct	Entropy, Incorrect	Entropy Difference	Rank-Biserial r	Spearman rho	Holm-adjusted p
Gradient Boosting	0.0925	0.3188	0.0657	0.1217	0.4273	0.3057	0.9577	0.7065	5.003 x 10^-21
Random Forest	0.1036	0.3572	0.1391	0.5085	0.6696	0.1612	0.9243	0.6129	1.537 x 10^-16
Support Vector Machine	0.1075	0.3429	0.0450	0.2513	0.4853	0.2340	0.7677	0.5125	1.198 x 10^-12
Logistic Regression	0.1478	0.7123	0.1188	0.1211	0.3137	0.1926	0.8272	0.6171	4.307 x 10^-18

3.4 Participant-Level Uncertainty and Prediction Error

Ten out-of-fold predictions were made per model by each participant. There were 120 cases correctly classified every time by Logistic Regression and 54 cases misclassified at least one time. For Support Vector Machine, 139 participants were always right, and 35 were wrong at least one time. For Random Forest, these numbers were 141 and 33, and for Gradient Boosting, 130 and 44. Predictive entropy of the participants who were consistent in their predictions was 0.0724, 0.2277, 0.4930 and 0.0712 for Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting, respectively. The mean entropy value for each category of participants misclassified at least one time were 0.3423, 0.5149, 0.6761, and 0.4244, respectively. Figure 2A displays the average participant-level predictive entropy for each of the classifiers. For Logistic Regression, the participant-level uncertainty-error relationship is illustrated in Figure 2B, which has a Spearman correlation of rho = 0.6171 and a p-value of 1.23×10^{-19}. The same relationship is seen in the Support Vector Machine with rho = 0.5125 and $p = 4.86 \times 10^{-13}$ as seen in Figure 2C. The correlation for the Random Forest results is shown in Figure 2D, with a value of rho = 0.6129 and $p = 2.51 \times 10^{-19}$. The Gradient Boosting results are displayed in Figure 2E, and the correlation is rho = 0.7065, $p = 1.29 \times 10^{-27}$.

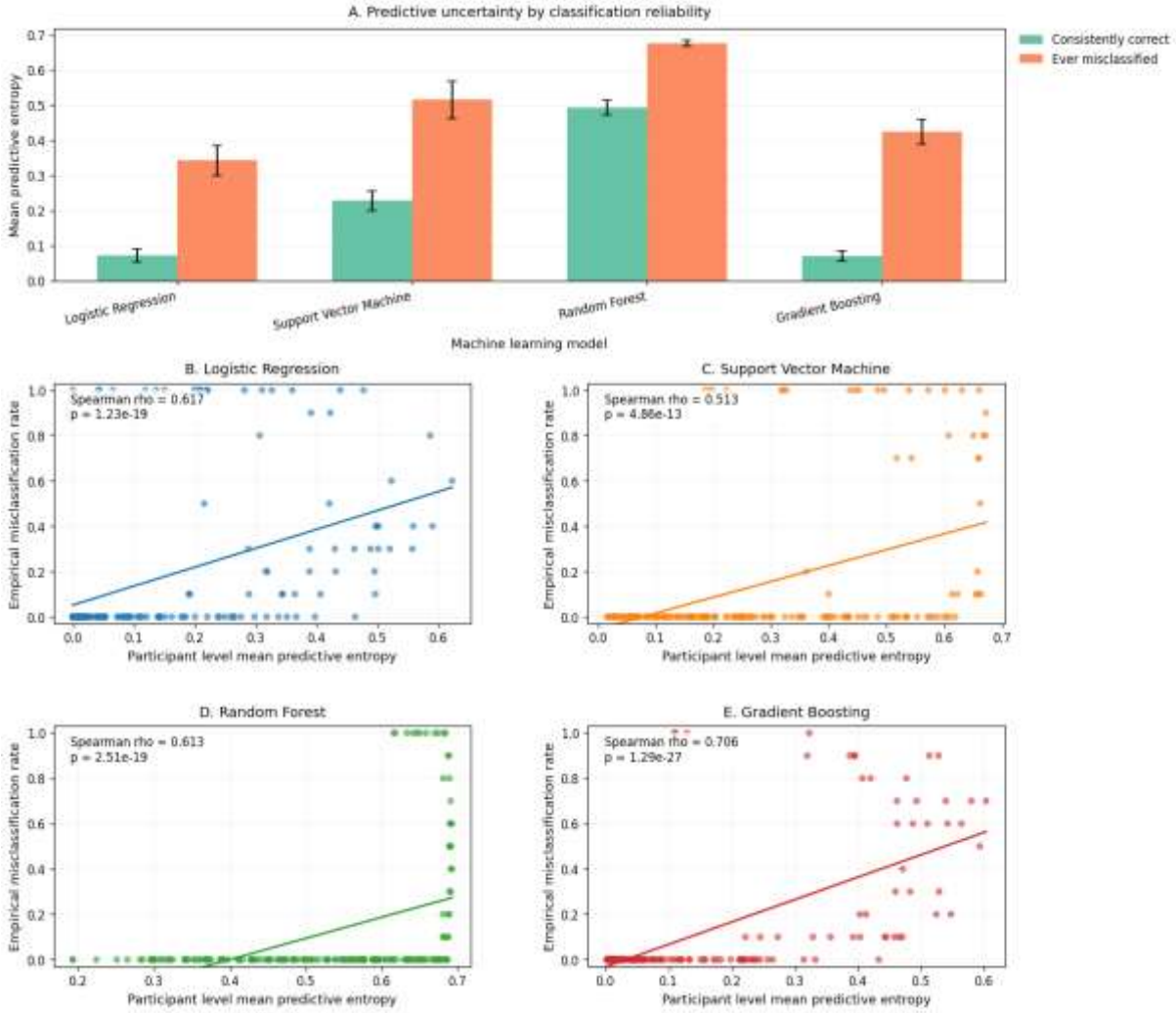

Figure 2. Predictive uncertainty and misclassification. (A) Entropy by classification status. (B–E) Participant-level entropy–error relationships across the four classifiers.

3.5 Selective Prediction and Predictor Stability

The predictive accuracy of the selective method has been determined for coverage rates between 50% and 100%. For Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting, the accuracy at full coverage was 0.8391, 0.8506, 0.8851 and 0.8621, respectively. At 90% coverage, the corresponding values were 0.8535, 0.8917, 0.9236, and 0.9236. The accuracy for 80% coverage was Logistic Regression's 0.8714, Support Vector Machine's 0.9143, Random Forest's 0.9429, and Gradient Boosting's 0.9571. At 70% coverage, the respective values were 0.8934, 0.9426, 0.9754, and 0.9754. At 50% coverage, selective accuracy was 0.9540 for Logistic Regression and Support Vector Machine and 1.0000 for Random Forest and Gradient Boosting. Selective accuracy is plotted against the prediction coverage for the four models in Figure 3A. Logistic Regression coefficient stability was evaluated across the 50 cross-validation training partitions. The results are obtained from all 450 predictors used in each training fold. There were 255 predictors with coefficient-sign consistency ≥ 0.90 and 210 predictors with sign consistency ≥ 0.95. The median sign consistency was 0.9400, and the median coefficient of variation was 0.5471. Pressure var19 and num of pendown19 had the two highest mean absolute standardised coefficients, 0.3691 (SD 0.0540) and 0.3620 (SD 0.0467), respectively. These were followed by air time24 (0.2727; SD 0.0400), max y extension25 (0.2553; SD 0.0627), mean jerk on paper24 (0.2505; SD 0.0468), total time24 (0.2495; SD 0.0378), gmrt on paper2 (0.2385; SD 0.0443), max x extension25 (0.2371; SD 0.0382), air time15 (0.2304; SD 0.0267), and total time15 (0.2273; SD 0.0273). The 15 predictors having the highest mean absolute standardized coefficient as well as their interfold variability are presented (Figure 3B).

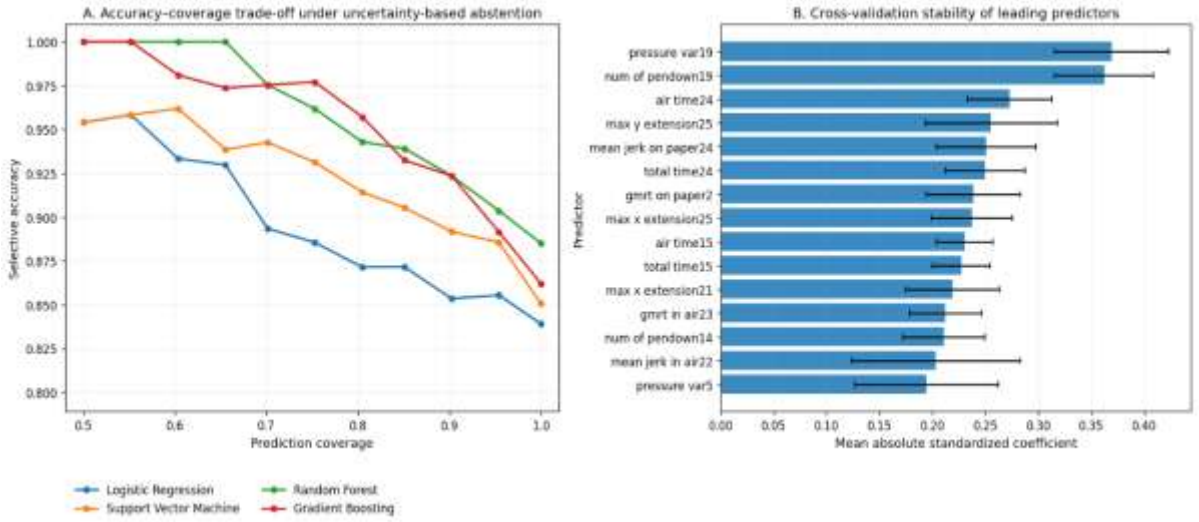

Figure 3. Selective prediction and predictor stability. (A) Accuracy-coverage trade-off across classifiers. (B) Stability of leading Logistic Regression predictors across cross-validation folds.

4. DISCUSSION

The results demonstrate the added value of uncertainty-aware evaluation in providing information not captured by discrimination metrics. Random Forest showed the highest ROC-AUC and overall accuracy, while Gradient Boosting showed the lowest Brier score and log-loss, and Support Vector Machine had the lowest Expected Calibration Error. This separation implies that the goodness of the models is measured differently when performance is measured by class separation as opposed to probabilistic accuracy and/or calibration. Predictive entropy was consistent and higher for incorrect predictions for all four classifiers, thus providing another piece of information. The participant-level analysis confirmed this trend, showing positive correlations between mean entropy and the empirical misclassification rate for Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting. Gradient Boosting had the highest uncertainty-error association, and Random Forest had the highest absolute entropy despite good discrimination. These results were further supported by the selective-prediction results, where all the accuracy scores improved with the increase of the percentage of the withheld high-uncertainty cases, with the best accuracies obtained at 50% coverage for the models of the type Random Forest and Gradient Boosting. Predictor-stability analysis was also performed, and it was discovered that a large number of coefficients had the same direction of effect over repeated training folds even in the high-dimensional case.

Observed performance characteristics agree with recent research that has demonstrated the advantages of learning strategies that can cope with redundant and highly correlated information in high-dimensional classification. In high dimensionality, Salhi et al. (2025) showed a benefit of feature-oriented optimization for classification, which reinforces the importance of dimensionality control in case predictive information is spread over a number of variables. Overall, the current findings support the idea that mathematical optimization and statistical learning should be viewed together when modelling complex systems instead of independently judging their merits based on raw predictive accuracy (Gopinath, 2025). The separation between the uncertainty, discrimination, and calibration measures is consistent with uncertainty-aware representation learning methods that quantify the quality of the decision by incorporating predictive confidence (Abdullah, 2025). The same issue exists in handwriting recognition: uncertainty-aware evaluation under domain shift showed that classification performance can be strong at times and weak at others despite the predictive reliability on the hidden domain (Klaß et al., 2022).

This correlation between entropy and misclassifications also aligns with uncertainty-driven recognition strategies involving the treatment of uncertain observations differently from high-confidence observations. For sequential uncertainty-based pseudo-label selection, uncertainty has been proven to help identify more reliable pseudo-labels or predictions from less reliable pseudo-labels or predictions, especially when pseudo-labels or predictions vary across observations (Patel et al., 2023). The research on multiview learning has also been extended to address the issue of uncertainty in prediction to make the learning more robust from multiple information sources (Xu et al., 2022). A related empirical pattern emerges from the selective-prediction findings: excluding high-entropy cases from a model's coverage actually resulted in higher classification accuracy, which suggests that the uncertainty estimates were still useful in providing practical information about the reliability of predictions. This interpretation also aligns with machine-learning systems and frameworks that use uncertainty as a component of model evaluation, instead of a secondary diagnostic (Dorst et al., 2023). Uncertainty-aware predictive modeling has also

been suggested as a way to enhance the reliability and transparency of data-driven decisions, as there can be different degrees of evidential support for the predictions (Kaiser et al., 2022).

The findings have some methodological implications. When a model is to be used for the prediction of probabilities to make decisions, model selection should not only be based on ROC-AUC or accuracy in high-dimensional spaces. This comparison of Random Forest, Gradient Boosting and Support Vector Machine shows that good discrimination, low probabilistic loss and a good calibration can be achieved with different classifiers. Predictive entropy then proves to be a useful ancillary measure of observations with high error-risk. Selective prediction is the process of turning this information into a usable decision rule by allowing uncertain cases to remain unclassified instead of being classified. The stability analysis is a second dimension of reliability as it demonstrates the stability of the contributions of the predictors when the training set is repeatedly perturbed. These components provide a foundation for a framework with predictive performance, probability quality, uncertainty and stability of the model as distinct but related properties.

There are a few restrictions to keep in mind. The size of the uncertainty-performance relationships observed here is based on a single, high-dimensional dataset with a relatively small sample, and thus may be different in other settings. Logistic Regression coefficients were used for the predictor-stability assessment, and this does not indicate feature stability for all classifiers. Possible extensions include extending the framework to larger and more diverse datasets, extending the use of entropy to conformal or Bayesian uncertainty measures, and assessing selective prediction under external validation or controlled distribution shift. The extensions would provide clarity to the generalizability of the framework while maintaining its core purpose: assessing the accuracy of the machine-learning predictions and their associated uncertainty and qualification.

5. CONCLUSION

Uncertainty-aware evaluation gave a more comprehensive evaluation of the reliability of machine learning systems than discrimination measures. Random Forest had the highest overall discrimination, Gradient Boosting had the highest probabilistic performance, and Support Vector Machine had the lowest calibration error. Predictive entropy was always found to be higher when it was wrong and was also found to have a positive correlation with the empirical error rates of all classifiers when the cases were repeatedly misclassified. Predictive entropy was always higher in the cases that were wrong and was seen to be positively correlated with the empirical error rates of all classifiers at the participant level when cases were repeatedly misclassified. Selective prediction also showed that classification accuracy increased when high-uncertainty cases were excluded, with the strongest selective performance observed for the ensemble models. Predictor-stability analysis also revealed that a significant portion of the variables maintained their signs of association across a series of training sets, even in the high-dimensional setting. These results justify a mathematical and statistical approach that considers together predictive accuracy, calibration, uncertainty, selective coverage and model stability. This can help to make machine learning systems more informative for evaluation in real-world data settings, especially when getting a good probability estimate and knowing which predictions are uncertain are crucial.

REFERENCES

1. Abdullah, D. (2025). Uncertainty-aware representation learning with physical consistency for robust decision processes in large-scale data systems. Journal of Scalable Data Engineering and Intelligent Computing, 9-17.
2. Dorst, T., Schneider, T., Eichstädt, S., & Schütze, A. (2023). Uncertainty-aware automated machine learning toolbox. tm-Technisches Messen, 90(3), 141-153.
3. Einbinder, B. S., Romano, Y., Sesia, M., & Zhou, Y. (2022). Training uncertainty-aware classifiers with conformalized deep learning. Advances in neural information processing systems, 35, 22380-22395.
4. Fontanella, F. (2022). DARWIN [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55D0K.
5. Gao, L., Song, J., Liu, X., Shao, J., Liu, J., & Shao, J. (2017). Learning in high-dimensional multimedia data: the state of the art. Multimedia Systems, 23(3), 303-313.
6. Gopinath, M. (2025). ADVANCED MACHINE LEARNING ALGORITHMS FOR PREDICTIVE MODELLING IN COMPLEX SYSTEMS: INTEGRATING MATHEMATICAL OPTIMIZATION AND STATISTICAL METHODS FOR ENHANCED DECISION-MAKING. International Journal of Applied Mathematics, 38(11s), 771-798.
7. Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. Biometrical journal, 60(3), 431-449.
8. Heo, J., Lee, H. B., Kim, S., Lee, J., Kim, K. J., Yang, E., & Hwang, S. J. (2018). Uncertainty-aware attention for reliable interpretation and prediction. Advances in neural information processing systems, 31.
9. Hu, Y., & Khan, L. (2021, August). Uncertainty-aware reliable text classification. In Proceedings of the

27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 628-636).
10. Kaiser, P., Kern, C., & Rügamer, D. (2022). Uncertainty-aware predictive modeling for fair data-driven decisions. arXiv preprint arXiv:2211.02730.
11. Klaß, A., Lorenz, S. M., Lauer-Schmaltz, M. W., Rügamer, D., Bischl, B., Mutschler, C., & Ott, F. (2022). Uncertainty-aware evaluation of time-series classification for online handwriting recognition with domain shift. arXiv preprint arXiv:2206.08640.
12. Kornaev, A., Kornaeva, E., Ivanov, O., Pershin, I., & Alukaev, D. (2024). Awareness of uncertainty in classification using a multivariate model and multi-views. arXiv preprint arXiv:2404.10314.
13. Kyriazos, T., & Poga, M. (2023). Dealing with multicollinearity in factor analysis: the problem, detections, and solutions. Open Journal of Statistics, 13(3), 404-424.
14. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems, 32.
15. Patel, G., Allebach, J. P., & Qiu, Q. (2023). Seq-ups: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 6180-6190).
16. Salhi, A., Alshamrani, R., Althbiti, A., Ismail, A., Abd-ElRahman, M., & Hassan, B. M. (2025). Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema. Scientific Reports, 15(1), 35038.
17. Shirmohammadi, S., Amiri, M. H., & Al Osman, H. (2024). Uncertainty as a predictor of classification accuracy in machine learning-assisted measurements [measurement methodology]. IEEE Instrumentation & Measurement Magazine, 27(7), 37-45.
18. Terol, R. M., Reina, A. R., Ziaei, S., & Gil, D. (2020). A machine learning approach to reduce dimensional space in large datasets. IEEE Access, 8, 148181-148192.
19. Wilson, A., & Anwar, M. R. (2024). The future of adaptive machine learning algorithms in high-dimensional data processing. International Transactions on Artificial Intelligence, 3(1), 97-107.
20. Xu, C., Zhao, W., Zhao, J., Guan, Z., Song, X., & Li, J. (2022). Uncertainty-aware multiview deep learning for internet of things applications. IEEE Transactions on Industrial Informatics, 19(2), 1456-1466.
21. Yao, Y., Xiao, Z., Wang, B., Viswanath, B., Zheng, H., & Zhao, B. Y. (2017, November). Complexity vs. performance: empirical analysis of machine learning as a service. In Proceedings of the 2017 internet measurement conference (pp. 384-397).
22. Zhang, J., Qiu, Y., & Dong, L. (2025). Conformal deep forest for uncertainty-aware classification. Journal of King Saud University Computer and Information Sciences, 37(6), 155.