Stacking Calibrated Odds Against Floods

Emmanuel Rassou

1. Introduction

Accurately predicting weather events has long been a scientific challenge, but modern advancements in weather technology have greatly improved the availability and structure of historical meteorological data. In this paper, we focus on Bangladesh—a country shaped by its extensive river systems and frequently affected by seasonal floods. Reliable forecasts of daily rainfall can significantly aid in flood preparedness and disaster mitigation.

Our goal is to develop a predictive model that satisfies two key criteria: (i) robustness, particularly for the positive class (i.e., when there is rain), and (ii) interpretability, where a predicted probability of, say, 0.8 should correspond to an 80% empirical frequency of rain. To achieve this, we propose a methodology that combines classical machine learning models using stacking, and improves interpretability through calibration.

2. Related work

Following the approach of Hussain et al. [4], we employ a combination of five statistical machine learning methods as base learners. Among these, Logistic Regression and Support Vector Classifier are parametric models, while K-Nearest Neighbors, Decision Tree, and Random Forest are non-parametric. To maintain consistency with their methodology, we retain the same set of base learners and conduct experiments on the original Kaggle dataset, which comprises ten years of weather data from Bangladesh [9]. One limitation with the dataset is that it is skewed with 74% of the dataset comprising negative examples. This means that with no special weighting employed, a naive classifier just predicting 'no' would receive an accuracy of 74 %. An indirect effect of this leads to precision (proportion of predicted positives that are true) being artificially high compared to the recall (proportion of true positives that are predicted true), even though we would ideally like to prioritize recall over precision in this setting. As seen by their results, replicated in Table 1, only the Decision Tree has a higher recall than precision, however this method turns out to have the lowest F1-score (harmonic mean of recall and precision). Their main contribution involves a hard-voting ensemble of these base learners such that each one receives an equal weighting in the final aggregation of outputs. The ensembling achieves a higher precision and F1-score confirming our intuition empirically, however, the recall sitting at 51.16% is still lower than three of the five base learners.

We aim to improve upon their approach by stacking, an extension of ensembling where a weighted combination of the base learners are instead learned by a *final estimator*. To the best of our knowledge, stacking applied to the setup described by Hussain et al.[4], has not been published before. We aim to improve upon the F1-score and recall by considering several final estimators in the second layer of the stack.

We also propose a different final estimator: the Passive-Aggressive Classifier(PAC)[1]. It has performed well before on timeseries forecasting for extreme weather effects [3], however, we propose using it for the final estimator in this setting. While canonical methods require more features to give better accuracy, PAC is an online learning method that performs particularly well when only a few covariates are given as they can adapt quickly based on new data. More details on the paper and intuition of the method is given in Appendix C and shown in Figure 3.

Finally, we simultaneously achieve our second goal of interpretability by surveying three calibration methods when applied to base learners as well as the final estimators of the stacking ensemble. Historically, the calibration of the output probabilities was first applied to the weather domain given its need for classification confidences [2]. The degree of calibration is usually visualized with a *Reliability Curve*[6], which compares the predicted average probability and the true fraction from labels across discrete bins. To the best of our knowledge, calibration has not been combined with stacking within this application domain. Platt Scaling[8] and Isotonic Regression[10] are two established methods that have implementations from sklearn. Angular Calibration [5] is a recent generalization with no implementation or experiments in any application domain at the time of writing. We open-source our implementation on GitHub (`https://github.com/emmanuel2406/StackingCalibratedOddsAgainstFloods`). Technical details for each method are given in Appendix D.

3. Proposal

3.1 Methodology. We first preprocess our dataset [9] in the same manner as [4]. This involves standardizing numerical features, as well as converting categorical features into dummy variables and dropping *RainToday* and *Rainfall* as covariates. For each of the base learners, we replicate their results by only using the default hyperparameters and the same random state of 101 for replicability. We split the dataset randomly into 65% training and 35% test set. This breaks the temporal dependency between consecutive data entries, and we leave time-series forecasting for future work. For angular calibration [5], we dedicate 20% of the training dataset to estimate the sign. In addition to the method in [4], we run soft-voting ensembling as a baseline, where the votes of each base learner are proportional to its output score as opposed to its binary decision as in hard-voting. Stacking involves a two layer scheme where the first layer comprises the same five base learners, but the second layer learns a dynamic weighting of the five output scores for the final output. This *Final Estimator* does not take into account the original input data(Passthrough=False), and we retain the same hyperparameters for simplicity. We compare the adoption of all five base learners , as well as PAC [1] as the final estimator. We then apply calibration to the final estimator of this stacking scheme.

3.2 Assumptions. Angular calibration, as introduced by Li et al.[5], is specifically designed for linear classifiers, as it relies on the existence of a coefficient vector and assumes a Gaussian design. Consequently, we restrict its application to parametric models such as logistic regression, support vector classifiers, and PAC. However, in practice, the Gaussian assumption can often be relaxed while still yielding performance improvements.

3.3 Calibration Exploration. As a preliminary study, we compared the effectiveness of the calibration when applied to each of the base learners. When plotting the Reliability Curves of the trained models without calibration, we can reason about the tradeoff between recall and precision. We coin the *Reliability Curve Rule* as a qualitative rule of thumb for calibration. Below the perfect line means that model predictions are overconfident in predicting it will rain, and so correcting for this will gain some precision at the cost of recall. Above the perfect line represents underconfidence where precision is traded for an improvement in recall. In this setting a pre-calibration curve above the perfect line is more promising where recall and F1-score can be improved. One limitation with this rule is that when the curve is not uniformly below or above the line, then such a tradeoff becomes more complex in practice. At least for calibration methods such as Platt Scaling and Isotonic Regression that directly try to fit the perfect line, the rule gives a good sanity check for ensuring performance gains when applying calibration.

In our primary experiment involving calibration and stacking, calibration can be applied either to the base learners or to the final meta-estimator. Intuitively, aggregating the outputs of five individually calibrated base learners might be expected to yield a well-calibrated meta-learner. However, coordinating multiple calibration objectives introduces substantial complexity, making this approach more challenging than calibrating only the final output of the meta-estimator. We conduct an ablation study, applying calibration at each combination of both stages. Note: we do not apply angular calibration to the base learners because of the non-parametric models.

4. Experiments

From Table 1, we see that a hard-voting ensembling of the five base learners achieves a higher precision of 78.17% and F1-score of 61.85%, however, the recall of 51.16% is less than three of the base learners. We observe marginal improvements when using stacking with logistic regression and a support vector classifier as the final estimators, yielding F1-scores of 62.35% and 61.93%, respectively. These gains can be attributed to the more expressive weighting of base learners, which enables improved recall and is sufficient to shift the harmonic mean in favor of a higher F1-score. PAC [1] as a final estimator gives significant improvements over hard-voting ensembling with a F1-score of 66.88% and Recall of 69.10%. Although the precision is significantly lower, this balancing of recall and precision is preferable in a setting with flood crises.

For our calibration exploration, we test our Reliability Curve Rule with our five base learners. As shown in Figure 1, Logistic Regression and Decision Tree are below the perfect line, which reflects a loss in recall. On the other hand, KNN and Random Forest are above the perfect line, indicating the opposite tradeoff. Support Vector Classifier is an interesting case that seems to be both below and above the perfect line. We show that in Table 2, our reliability curve rule holds true for Platt Scaling and Isotonic Regression. The generality of Angular Calibration gives a gain in recall for both Logistic Regression and Support Vector Classifier, which can be leveraged for more complex miscalibrated patterns such as in stacking.

After applying all three calibration methods to the most promising Final Estimators as seen in Table 3, we observe that Platt Scaling and Isotonic Regression significantly decrease the F1-score of PAC, while Angular calibration retains the F1-score of 66.88 %. Similar effects are observed for the Log Reg and SVC but to a lesser extent. As shown in Figure 2, the post-calibration reliability curves indicate that both Platt Scaling and Isotonic Regression tend to overfit to the identity line, resulting in less robust calibration. In contrast, Angular Calibration produces a smoother and more consistent curve, albeit with a slight angular deviation. This suggests that the transformation learned by Angular Calibration is more generalizable. Notably, this outcome challenges our Reliability Curve Rule, where proximity to the identity line is often equated with better calibration—highlighting the importance of assessing calibration quality beyond visual closeness to the diagonal.

Lastly, Table 4 presents results from experiments involving calibration at both the base learner and final estimator levels. Performance declines when PAC is used as the final estimator, with the highest F1-score reaching only 64.61% under Platt Scaling of the base learners and no final estimator calibration. In contrast, for both Logistic Regression and Support Vector Classifier as final estimators, incorporating calibration at both stages yields modest improvements: the F1-score increases from 62.35% to 62.43% , and from 61.93% to 62.57% respectively.

5. Future directions

While our contributions represent a meaningful step forward, as evidenced by an improvement in F1-score from 61.85% to 66.88% and recall from 51.16% to 69.10% over the baseline in [4], as well as the preservation of F1-score performance after calibration, several avenues remain for further investigation. In particular, exploring calibration strategies tailored to individual base learners may enable a more fine-grained ablation study and uncover nuanced performance effects. Another promising direction involves a deeper analysis of the weights learned by the final estimator, which could provide insights into how stacking balances precision and recall. Furthermore, a quantitative approach to evaluating calibration—beyond visual reliability curves—could offer a more robust foundation for interpreting calibration behavior. We also hypothesize that performance discrepancies across calibration methods may be partially explained by a failure mode referred to as collapse, where the meta-learner underutilizes the diversity among base learners. Understanding this phenomenon may clarify the limitations of certain calibration schemes when used in ensemble contexts. Overall, the interplay between calibration and stacking, particularly in relation to classification robustness and interpretability, remains a complex yet compelling area for future research.

References

[1] Koby Crammer et al. "Online Passive-Aggressive Algorithms". In: *Journal of Machine Learning Research* 7.19 (2006), pp. 551–585. URL: http://jmlr.org/papers/v7/crammer06a.html.

[2] Renate Hagedorn, Thomas M. Hamill, and Jeffrey S. Whitaker. "Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures". In: *Monthly Weather Review* 136.7 (2008), pp. 2608–2619. DOI: 10.1175/2007MWR2410.1. URL: https://journals.ametsoc.org/view/journals/mwre/136/7/2007mwr2410.1.xml.

[3] Michael William Hopwood, Hector Mendoza, and Thushara Gunda. "Generating actionable information through the fusion of text and timeseries data: A case study of extreme weather effects at Photovoltaic plants." In: Sandia National Lab. (SNL-NM), Albuquerque, NM (United States). July 2020. URL: https://www.osti.gov/biblio/1811807.

[4] Adil Hussain et al. *Weather Forecasting Using Machine Learning Techniques: Rainfall and Temperature Analysis.* Sept. 2024. DOI: 10.20944/preprints202402.1566.v2.

[5] Yufan Li and Pragya Sur. "Optimal and Provable Calibration in High-Dimensional Binary Classification: Angular Calibration and Platt Scaling". In: *arXiv preprint arXiv:2502.15131* (2025).

[6] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd International Conference on Machine Learning.* ICML '05. Bonn, Germany: Association for Computing Machinery, 2005, pp. 625–632. ISBN: 1595931805. DOI: 10.1145/1102351.1102430. URL: https://doi.org/10.1145/1102351.1102430.

[7] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd International Conference on Machine Learning.* ICML '05. Bonn, Germany: Association for Computing Machinery, 2005, pp. 625–632. ISBN: 1595931805. DOI: 10.1145/1102351.1102430. URL: https://doi.org/10.1145/1102351.1102430.

[8] John Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Adv. Large Margin Classif.* 10 (June 2000).

[9] Apurbo Shahid Shawon. *Weather Data Bangladesh: Rain and Temperature Prediction Based on Weather Data Using Machine Learning.* https://www.kaggle.com/datasets/apurboshahidshawon/weatherdatabangladesh. Accessed: 2025-04-25. Sept. 2023.

[10] Bianca Zadrozny and Charles Elkan. "Transforming classifier scores into accurate multiclass probability estimates". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '02. Edmonton, Alberta, Canada: Association for Computing Machinery, 2002, pp. 694–699. ISBN: 158113567X. DOI: 10.1145/775047.775151. URL: https://doi.org/10.1145/775047.775151.

Appendix A Tables

ML Method	F1 Score(%)	Recall(%)	Precision(%)
Logistic Regression (Log Reg)	**61.65**	54.49	71.00
KNN	42.72	31.23	67.63
Decision Tree	53.99	**56.15**	52.00
Support Vector Classifier(SVC)	61.63	52.82	73.95
Random Forest	61.45	50.83	**77.66**
Passive-Aggressive Classifier(PAC)	54.25	**56.15**	52.84
Hard-Voting Ensembling	61.85	51.16	**78.17**
Soft-Voting Ensembling	60.32	50.50	74.88
Stacking [Log Reg]	62.35	52.82	76.08
Stacking [KNN]	59.54	51.83	69.96
Stacking [Decision Tree]	53.57	54.82	52.38
Stacking [SVC]	61.93	52.16	76.21
Stacking [Random Forest]	60.31	52.49	70.85
Stacking [PAC]	**66.88**	**69.10**	64.80

Table 1: Stacking experiments using the same 5 base learners(Log Reg, KNN, Decision Tree, SVC, Random Forest). The final estimator method for stacking is provided in square brackets

Method	Log Reg	KNN	Decision Tree	SVC	Random Forest
Uncalibrated	61.65	42.72	**53.99**	61.63	61.45
Platt Scaling	60.31	46.41	38.66	60.20	**62.18**
Isotonic	61.83	**48.40**	38.66	60.63	61.54
Angular	**63.87**	–	–	**63.28**	–

Table 2: Comparing F1 Scores of different calibration methods. **Red** indicates a tradeoff favoring precision over recall; **Green** indicates a tradeoff favoring recall over precision. Recall that Angular Calibration only applies to parametric models.

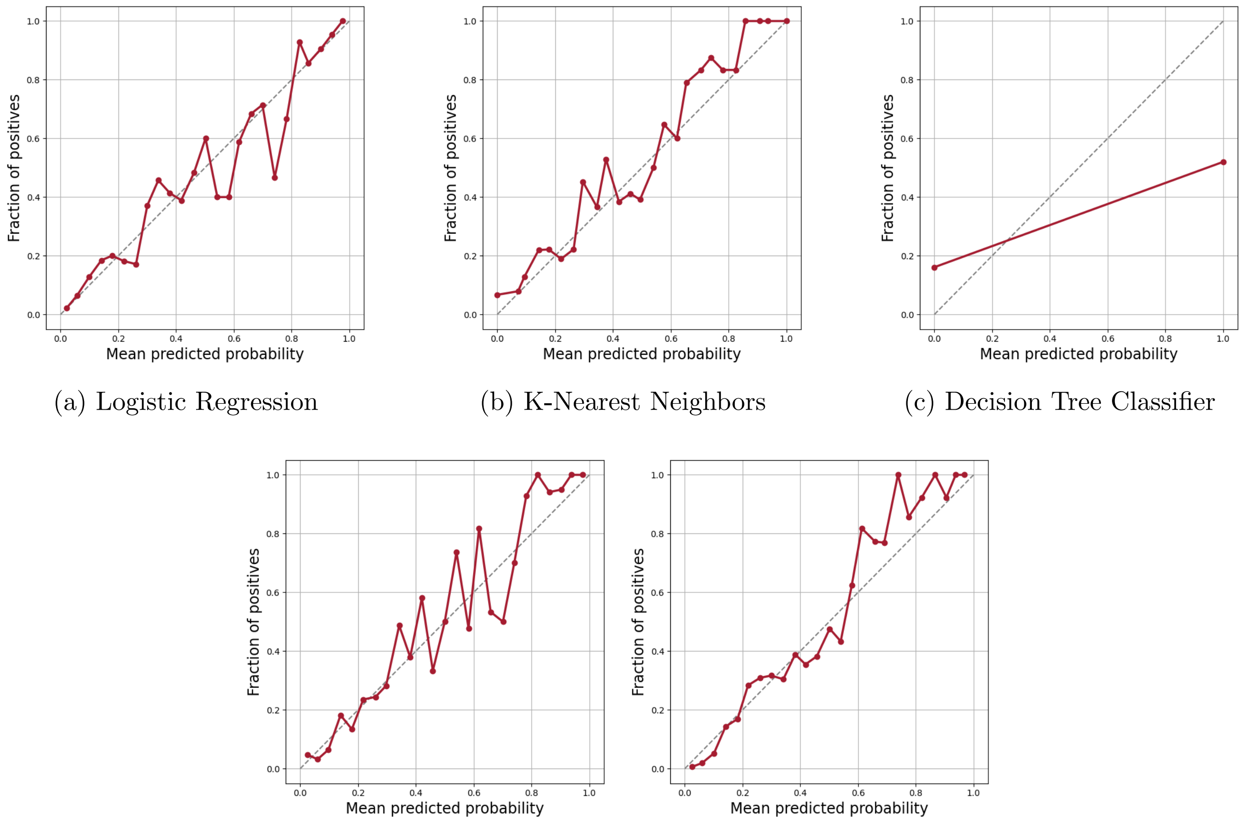
Calibration Method	Log Reg	SVC	PAC
Uncalibrated	62.35	61.93	**66.88**
Platt Scaling(Δ)	+0	-0.37	-6.43
Isotonic(Δ)	-0.33	-1.60	-5.99
Angular(Δ)	+0	+0	+0

Table 3: Retention of F1 score after calibration is applied to stacking final estimators

Base, Final Calibration	**Log Reg**	**SVC**	**PAC**
None, None	62.35	61.93	**66.88**
None, Platt Scaling	62.35	61.56	60.45
None, Isotonic	62.02	60.23	60.89
None, Angular	62.35	61.93	**66.88**
Platt Scaling, None	62.14	62.15	64.61
Platt Scaling, Platt Scaling	62.14	**62.57**	62.75
Platt Scaling, Isotonic	62.33	61.32	63.36
Platt Scaling, Angular	62.14	62.15	64.61
Isotonic, None	**62.43**	61.60	64.03
Isotonic, Platt Scaling	**62.43**	62.35	62.04
Isotonic, Isotonic	**62.43**	61.20	62.60
Isotonic, Angular	**62.43**	61.60	64.03

Table 4: Comparing F1-scores of different combinations of base learner and final estimator calibration. X,Y implies that X is the calibration method for each of the base learners, and Y is the calibration method of the final estimator

Appendix B Reliability Curves



(a) Logistic Regression

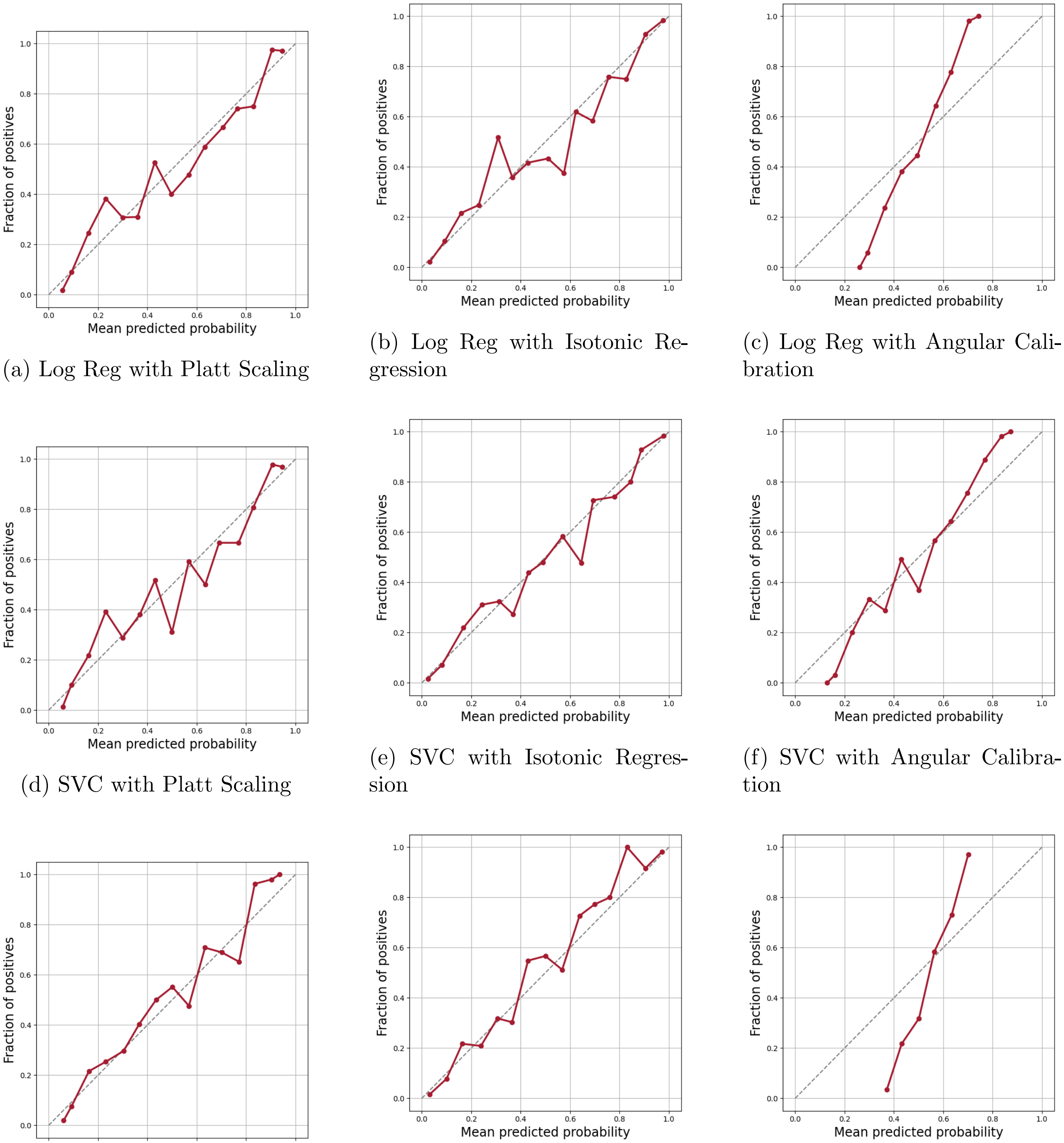
(b) K-Nearest Neighbors

(c) Decision Tree Classifier

(d) Support Vector Classifier

(e) Random Forest

Figure 1: Pre-calibration reliability curves of each base learner



(a) Log Reg with Platt Scaling

(b) Log Reg with Isotonic Regression

(c) Log Reg with Angular Calibration

(d) SVC with Platt Scaling

(e) SVC with Isotonic Regression

(f) SVC with Angular Calibration

(g) PAC with Platt Scaling

(h) PAC with Isotonic Regression

(i) PAC with Angular Calibration

Figure 2: Post-calibration reliability curves for different final estimators and calibration methods. Although Angular Calibration seems to have the highest calibration error, the generalizability of its shape retains the F1-score better

Appendix C Details on the Passive-Aggressive Classifier(PAC)

The Passive-Aggressive Classifier (PAC) is an online learning algorithm designed for binary classification tasks. Unlike traditional batch learning methods such as Logistic Regression and Support Vector Machines, which process the entire training dataset simultaneously, PAC updates its model incrementally as new data instances arrive. This characteristic makes it particularly suitable for scenarios involving streaming data or situations where data is too large to fit into memory.

The algorithm operates under a simple yet effective principle: it remains passive when the current prediction is correct, making no changes to the model, and becomes aggressive by updating the model only when a prediction error occurs. This approach ensures that the model focuses on correcting mistakes without altering correct predictions, thereby maintaining stability and preventing overfitting.

In the context of ensemble methods like stacking, PAC serves as an effective meta-learner. Its ability to adjust weights dynamically in response to misclassifications allows it to balance the contributions of diverse base learners effectively. This dynamic adjustment is crucial for optimizing performance metrics such as the F1-score, as it helps in managing the trade-off between precision and recall.

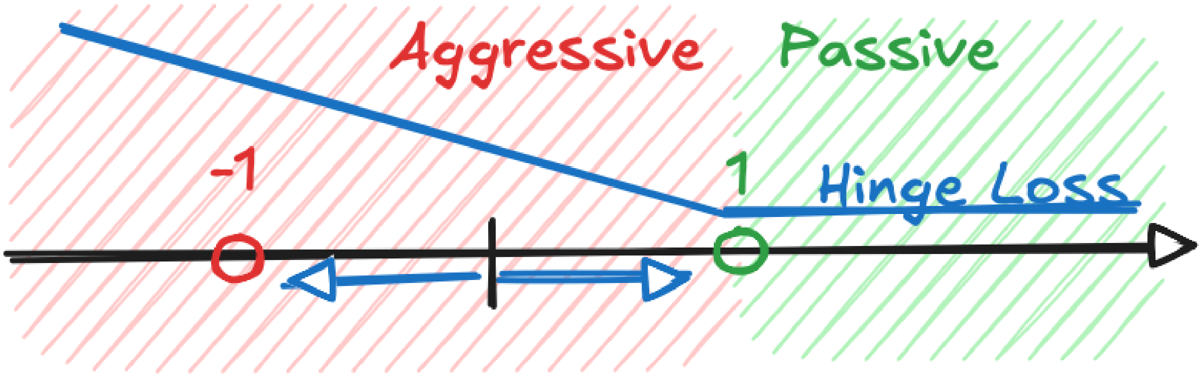


Figure 3: Our summary visualization of how the Passive-Aggressive Classifier works

As stated in the original paper [1], the classifier $\mathbf{w}$ predicts $y \in \{-1, +1\}$ and uses the hinge loss to mimic the passive-aggressive behavior discussed above. Then we have across time t, w_{t+1} is set to the projection of w_t onto the half-space of vectors which attain a hinge-loss of zero on the current example. This maximizes the margin between the classes in aggressive mode, and leaves the parameters alone in passive mode.

Appendix D Theoretical Underpinnings of Calibration

D.1 Platt Scaling

Published in 2000 by Platt [8], it applies a sigmoid function to raw uncalibrated predictions f_i. Two real numbers A and B are then fitted via maximum likelihood.

$$P(y_i = 1|f_i) = \sigma(Af_i + B) = \frac{1}{1 + \exp(Af_i + B)}$$

D.2 Isotonic Regression

This calibration method, as first described by Zadrozny et al. [10], learns a piecewise non-decreasing mapping from $f_i \to \hat{f}_i$ such that the order is preserved, i.e. $\hat{f}_i \geq \hat{f}_j \iff f_i \geq f_j$, and minimizes the following mean squared error:

$$\sum_{i}^{n}(y_i - \hat{f}_i)^2$$

While Isotonic Regression is more general than Platt Scaling, it has been shown that this method is also more prone to overfitting of smaller datasets [7].

D.3 Angular Calibration

Li and Sur [5] use sin and cos to interpolate between the informative projection $\hat{w}^T x_{\text{new}}$ and non-informative Gaussian noise $Z \sim \mathcal{N}(0, 1)$:

$$\hat{f}(\hat{w}^T x_{\text{new}}; \hat{\theta}) = \mathbb{E}_Z \left[\sigma \left(\cos(\hat{\theta}) \frac{\hat{w}^T x_{\text{new}}}{\|\hat{w}\|_\Sigma} + \sin(\hat{\theta}) Z \right) \right]$$

This formulation allows for a smooth transition between fully confident predictions (when $\hat{\theta} \approx 0$, so the model relies entirely on the projection $\hat{w}^T x_{\text{new}}$) and complete uncertainty (when $\hat{\theta} \approx \pi/2$, so the prediction is dominated by noise). The angle $\hat{\theta}$ thus controls the degree of epistemic uncertainty: it represents the model's confidence in its linear prediction relative to random noise. When $\hat{\theta}$ is small, the model believes its learned weights are highly informative; when large, it acknowledges potential overfitting or ambiguity in the prediction.

Angular calibration therefore provides a principled way to incorporate confidence (or lack thereof in terms of uncertainty) in post-hoc predictive distributions, particularly in high-dimensional or overparameterized settings, where classical calibration techniques may be unreliable due to overfitting.