Confidence or Competence: A Calibration Study of Machine Learning Classifiers for Selective Prediction in High-Stakes Contexts

Saumyya Dalal
Unitedworld Institute of Technology, Karnavati University
Gandhinagar, Gujarat, India

ABSTRACT
The central claim of the paper is that, besides being capable of making correct predictions, a model must also have the ability to say "I do not know" in critical moments. To accomplish this, the indicator of model honesty should be given more significance than the performance measure modeled in terms of accuracy. Five classifiers have been trained on the COMPAS recidivism dataset, and their performance has been evaluated not only based on accuracy and F1 but also on their calibration quality and performance of selective predictions. Results indicate that the Multilayer Perceptron (MLP), despite its modest raw accuracy, achieves the best calibration. The Random Forest model, which is often the default recommendation, is the least well-calibrated regarding its uncertainty in the uncalibrated state but improves significantly under post-hoc calibration. Every single classifier improves accuracy when allowed to abstain on low-confidence cases. The conclusion reached in the paper is that selective prediction is a practical, model-independent approach that can lead to safer AI systems, along with accuracy evaluation not being enough for making deployment decisions.

Keywords
Calibration, Selective Prediction, Expected Calibration Error, COMPAS, Recidivism, Trustworthy AI, Temperature Scaling, Abstention.

1. INTRODUCTION
Most machine learning models follow a traditional approach where a model is built with data, then tested on a testing set, and the determining factor of the model's quality is its accuracy. In general, the model is considered successful if its accuracy reaches a certain acceptable level. Nevertheless, it overlooks a simpler question ‘Does the model have the ability to respond with confidence when it is merely guessing?’.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm serves in American courts to assess defendants according to their chances of committing future crimes. The algorithm was biased toward African-Americans and, generally speaking, did not correspond to the actual outcome of its predictions according to the findings of ProPublica investigations in 2016. Among other things, the analysis shows that African-American defendants had a higher chance of being classified wrongly as high-risk offenders than other groups (about 45% against 24%), as the analysis of the COMPAS recommender model states. Therefore, it proves that the model has some serious flaws in terms of its operational reliability [1].

The same gap is analyzed using a publicly accessible COMPAS database. The goal is not to develop a better recidivism prediction model, which is quite a complicated and ethically questionable task. The aim is to examine five types of classifiers according to two features neglected by traditional approaches:

1. Does calibration predict confidence and does it match actual accuracy? (calibration)
2. Does accuracy improve when the model abstains on cases it is unsure about? (selective prediction)

The five classifiers compared are Logistic Regression, k-Nearest Neighbors (k-NN), Decision Tree, Random Forest, and a multilayer perceptron (MLP). These cover the range of methods a practitioner would actually reach for: linear models, lazy learners, tree-based methods, ensembles, and simple neural networks. Throughout this paper, the use of terms such as 'honesty' is metaphorical to refer to a model's ability to produce well-calibrated confidence estimates.

2. BACKGROUND AND RELATED WORK
2.1. What Calibration Means
When the confidence scores of a classifier have validity, it is said to be calibrated. For instance, if the model produced a forecast with a probability value equal to 0.8, this means that 80 percent of those forecasts are accurate. Thus, the model can be proven as reliable [2].

A reliability diagram is one of the common ways of visualizing the calibration. Each predicted case is classified into confidence bins according to the confidence level (for example, 0-0.1, 0.1-0.2 and so on). The real percentage of the accurate predictions is shown against the predicted confidence in each bin. The diagram generated for a completely calibrated model has a diagonal line. The points above the diagonal line are designated as deviations indicating underconfidence. On the other hand, the points below the diagonal line indicate overconfidence.

The Expected Calibration Error (ECE) quantifies calibration. It is the weighted average of the absolute difference between the predicted confidence and the real accuracy for each bin. The Brier Score is another popular measure used to check the quality of the model [6]. Another major issue is that modern neural networks tend to be strongly overconfident [2]. To address this problem, post hoc correction methods such as Platt Scaling and Isotonic Regression can be used to adjust raw model outputs without retraining [4, 5].

2.2. Selective Prediction
Selective prediction (also called learning with rejection or abstention) is the idea that a classifier should be able to decline a prediction when it lacks confidence. This concept was also formalized as a trade-off between coverage i.e. the fraction of inputs the model answers, and accuracy i.e. how correct it is on those inputs [8].

The basic mechanism is simple. A binary classifier's confidence is computed as the distance of the predicted probability from 0.5. This maps all predictions to a confidence value between 0.5 and 1.0. A threshold value is set, and the model only predicts on samples above the value. The rise in threshold may result in coverage drops, but the remaining predictions should be more accurate.

2.3. The COMPAS Dataset

The COMPAS dataset was released by ProPublica in 2016, as part of their investigation into algorithmic bias in criminal sentencing [1]. It has the records of defendants in Broward County, Florida, scored by the COMPAS tool between 2013 and 2014. The target variable is two-year recidivism, whether the defendant was arrested for a new crime within two years.

After applying ProPublica's standard filtering criteria, 6,172 samples remain. Out of these, 3,363 did not reoffend and 2,809 did. The base recidivism rate is 45.51%, making the task a near-balanced binary classification problem.

Features used are age, sex, juvenile felony count, juvenile misdemeanor count, juvenile other offense count, prior conviction count, and charge degree. Race is deliberately excluded. (Fig. 1)

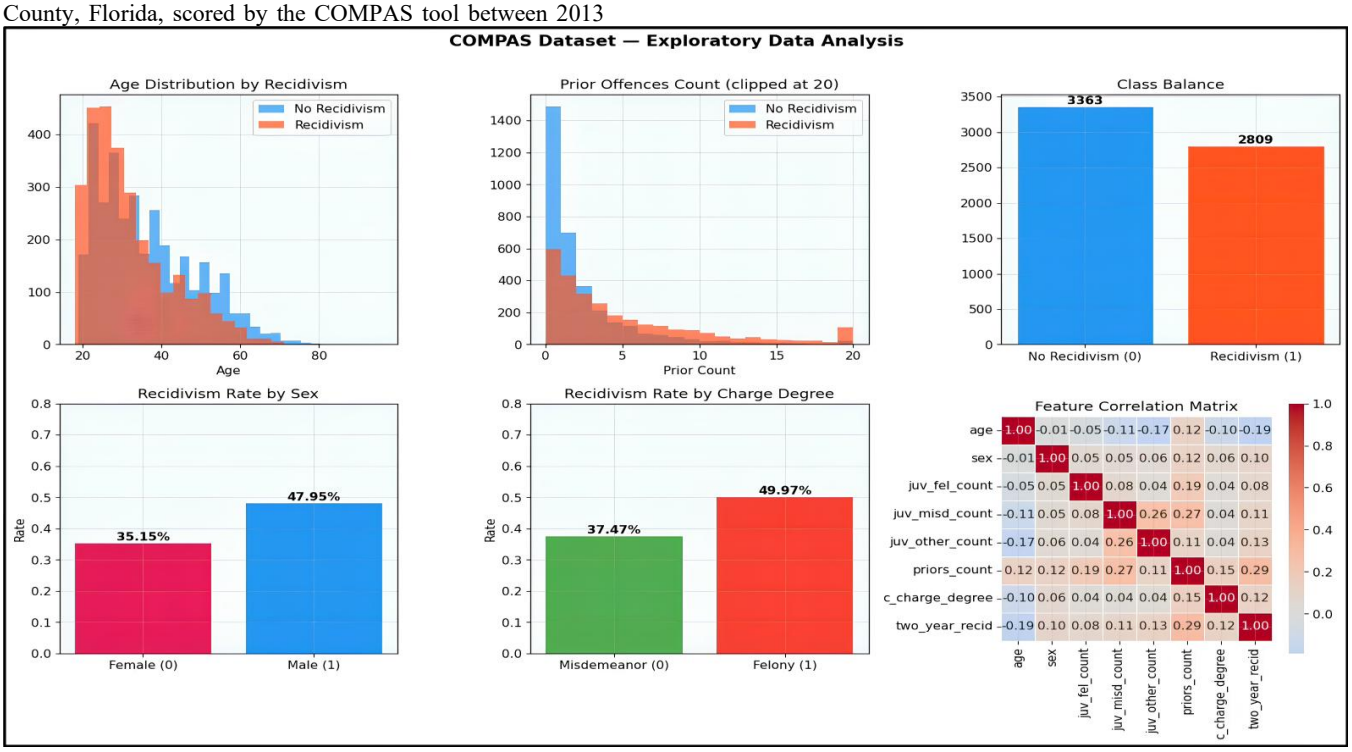

Fig. 1. Exploratory analysis of feature distributions and recidivism outcomes in the COMPAS dataset

3. METHODOLOGY

3.1. Experimental Setup

All experiments used Python 3.12 with scikit-learn 1.7.0 [9]. The dataset was split 80/20 with stratification. Features were standardized using StandardScaler.

Five classifiers were trained:

- Logistic Regression calibration baseline.
- The k-Nearest Neighbors probability is calculated from the class proportions of the neighbors.
- Decision Tree (max. depth 5) constrained to avoid overfitting.
- Random Forest (150 trees) [10]
- A single-layer MLP network (100 hidden units with early stopping).

3.2. Calibration Evaluation

Each model's predicted probabilities were evaluated using reliability diagrams (10 uniform bins), expected calibration error (ECE), and Brier score. Post-hoc calibration was applied using Temperature Scaling [2] from scikit-learn in prefit mode. Because models like Decision Trees and k-Nearest Neighbors do not natively output logits, the predicted probabilities were extracted and pseudo-logits reconstructed using the inverse sigmoid transformation. A single temperature parameter T was then optimized for each model on the held-out validation set. The model with the best resulting ECE was reported.

3.3. Selective Prediction Evaluation

Confidence was computed as $|P(y=1|x) - 0.5| + 0.5$, mapping predictions between 0.5 and 1.0. A sweep of 40 thresholds from 0.50 to 0.95 was applied with a minimum 5% coverage floor. Abstention gain at 80% coverage was defined as Acc 80% - Acc 100%.

4. RESULTS AND DISCUSSION

4.1. Standard Performance Metrics

Table 1 shows Accuracy, F1 score, ROC-AUC, and Brier score for all five classifiers on the held-out test set (1,234 samples). The models don't show drastic differences in the metrics. Accuracy wise, it is seen that four of the five models are within one percentage point of each other. Logistic Regression leads at 67.6% (0.6759), with MLP at the second highest position with 67.4% (0.6742).

Random Forest has the least accuracy at 63.7% (0.6370). If this were the only table to be considered, the recommendation would be to use Logistic Regression or MLP, ignoring other significant factors.

Fig. 2 depicts the robustness of rankings across 1,000 bootstrap resamples. It shows 95% confidence intervals showing that differences between top-performing models (MLP, Logistic Regression, Decision Tree) are within the margin of sampling variance. In contrast, Random Forest and k-NN consistently underperform across these resamples. The results mean that sample variance has little impact on the relative performance of the best models. A detailed calibration analysis is therefore warranted. It is essential to point out that the impact of overfitting may have further exacerbated the poor calibration of the Random Forest, which was run using the untuned baseline (150 trees, no limit in depth).

The presence of proxy bias (e.g. systematic enforcement patterns that are correlated with prior convictions) is still a problem, even when race is excluded from the models. Thus, rather than directly addressing the fairness concerns previously discovered by ProPublica, the calibration and selective prediction methods proposed here function as a form of intervention designed to improve model reliability.

Table 1. Standard performance metrics on the test set

Model	Accuracy	F1	AUC	Brier Score
Logistic Reg.	0.676 ± 0.013	0.609 ± 0.018	0.726 ± 0.014	0.210 ± 0.005
k-NN	0.646 ± 0.014	0.620 ± 0.017	0.682 ± 0.015	0.242 ± 0.008
Decision Tree	0.666 ± 0.014	0.629 ± 0.018	0.712 ± 0.015	0.216 ± 0.006
Random Forest	0.638 ± 0.014	0.595 ± 0.018	0.674 ± 0.015	0.244 ± 0.007
MLP	0.675 ± 0.014	0.632 ± 0.017	0.733 ± 0.015	0.207 ± 0.006

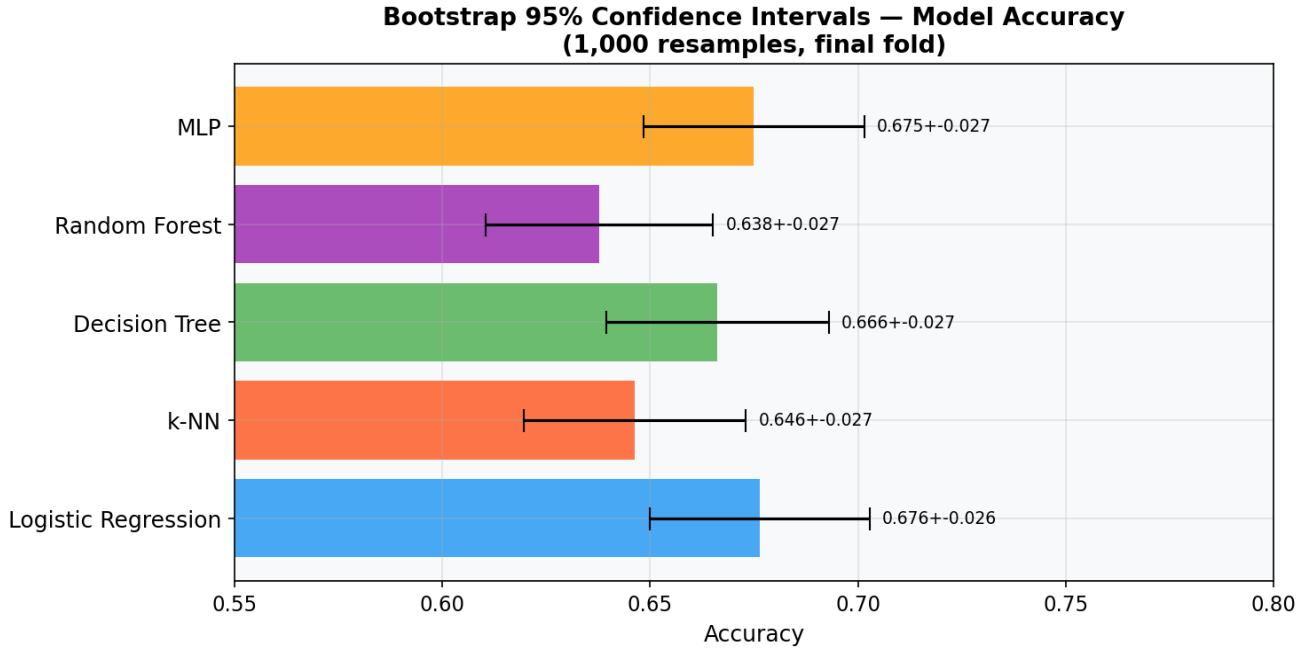

Fig 2. Bootstrap confidence intervals for model accuracy across 1,000 resamples.

4.2. Calibration Analysis

Table 2 shows the ECE before and after the post-hoc correction of each model. The ECE of the MLP is 0.042 +/- 0.011 which is superior to the next best performing model, Logistic Regression with an ECE of 0.064 +/- 0.014. k-NN is the worst-calibrated of the five (ECE = 0.143 +- 0.014), with Random Forest close behind (ECE = 0.128 +- 0.013). Post-hoc calibration (Temperature Scaling) works well for some models, but is ineffective for others. Random Forest ECE drops significantly (0.128 +- 0.013 to 0.092 +- 0.022), and k-NN also gains significantly (0.143 +- 0.014 to 0.083 +- 0.015). However, the Decision Tree ECE worsens under Temperature Scaling (0.088 +- 0.016 to 0.155 +- 0.045), which provides a practical caution against treating post-hoc calibration as a universal fix.

As per Fig. 3, the MLP is the best calibrated model (ECE = 0.042), its curve running extremely close to the diagonal across the probability range where predictions concentrate. Logistic Regression lands at ECE = 0.037, though it shows slightly more deviation. Decision Tree (ECE = 0.038) is also reasonably well-calibrated. k-NN (0.143 +- 0.014) and Random Forest (0.128 +- 0.013) both sit above the diagonal at low probabilities: majority voting and bagging each produce confident-looking outputs that the observed rates do not support.

Table 2. ECE before and after post-hoc correction

Model	ECE (Original)	ECE (Corrected)	Method
Decision Tree	0.088 ± 0.016	0.155 ± 0.045	Temp. Scaling (T = 1.688)
Logistic Reg.	0.064 ± 0.014	0.062 ± 0.017	Temp. Scaling (T = 1.065)
MLP	0.042 ± 0.011	0.039 ± 0.010	Temp. Scaling (T = 1.125)
k-NN	0.143 ± 0.014	0.083 ± 0.015	Temp. Scaling (T = 10.000)
Random Forest	0.128 ± 0.013	0.092 ± 0.022	Temp. Scaling (T = 4.606)

As seen in Fig. 4, the MLP calibration remains relatively stable under Temperature Scaling, with its ECE changing from 0.042 to 0.039. Random Forest sees a clean improvement where its ECE falls from 0.128 to 0.092, correcting much of the overconfidence. k-NN gains slightly, moving to 0.083 ECE, while Decision Tree worsens to 0.155 +- 0.045, showing that simple temperature scaling can over-correct models that are already well calibrated. MLP produces the tightest match between predicted risk and observed outcomes of any model tested here.

4.3. Selective Prediction Results

Table 3 shows accuracy at fixed coverage levels. The gain column represents the accuracy improvement achieved by excluding the least confident 20% of predictions. Most models are seen to improve under selective prediction. It is noticed that MLP showed the largest gains (4.4 percentage points at 80% coverage, reaching 71.9% accuracy). The Random Forest accuracy increased by 2.5 points to yield an accuracy of 66.2%, while Logistic Regression achieved 70.5% accuracy at 80% coverage. This demonstrates that selective prediction is one of the more effective techniques for improving accuracy without the need for retraining.

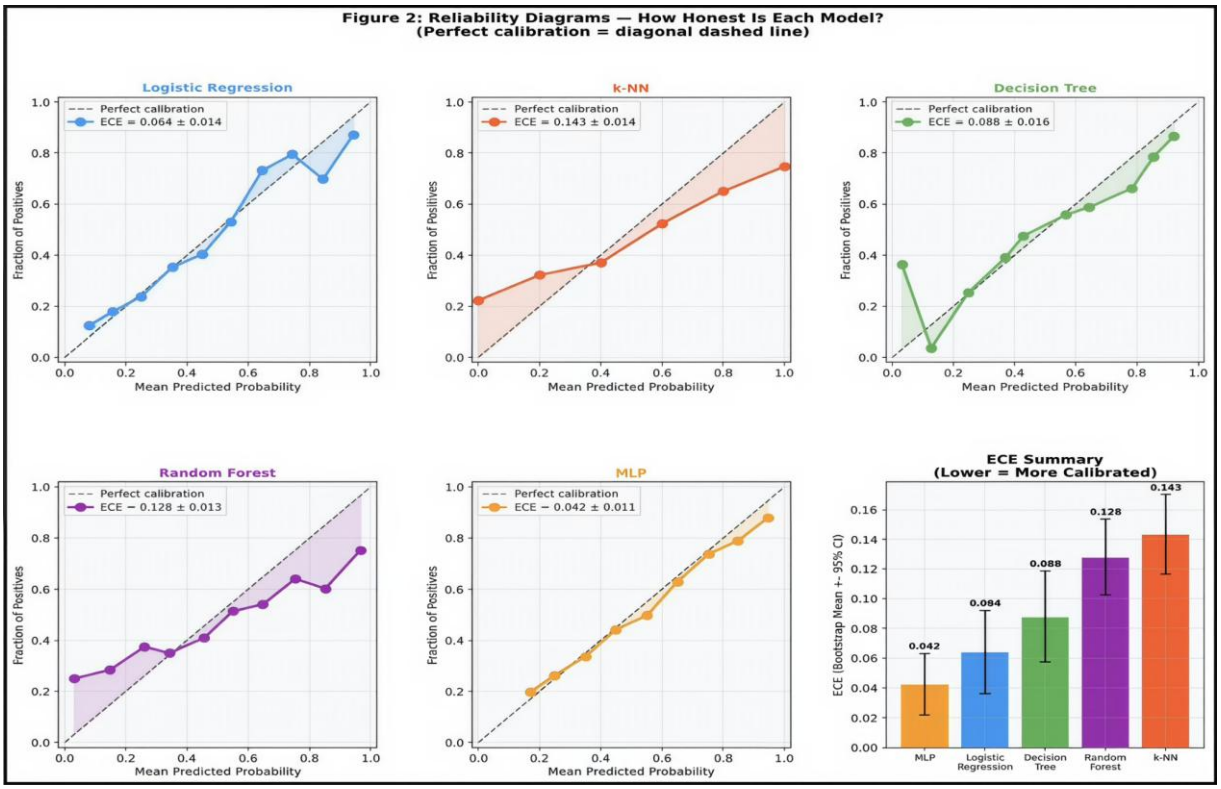

Fig. 3. Reliability diagrams for the evaluated classifiers, showing predicted confidence against observed accuracy.

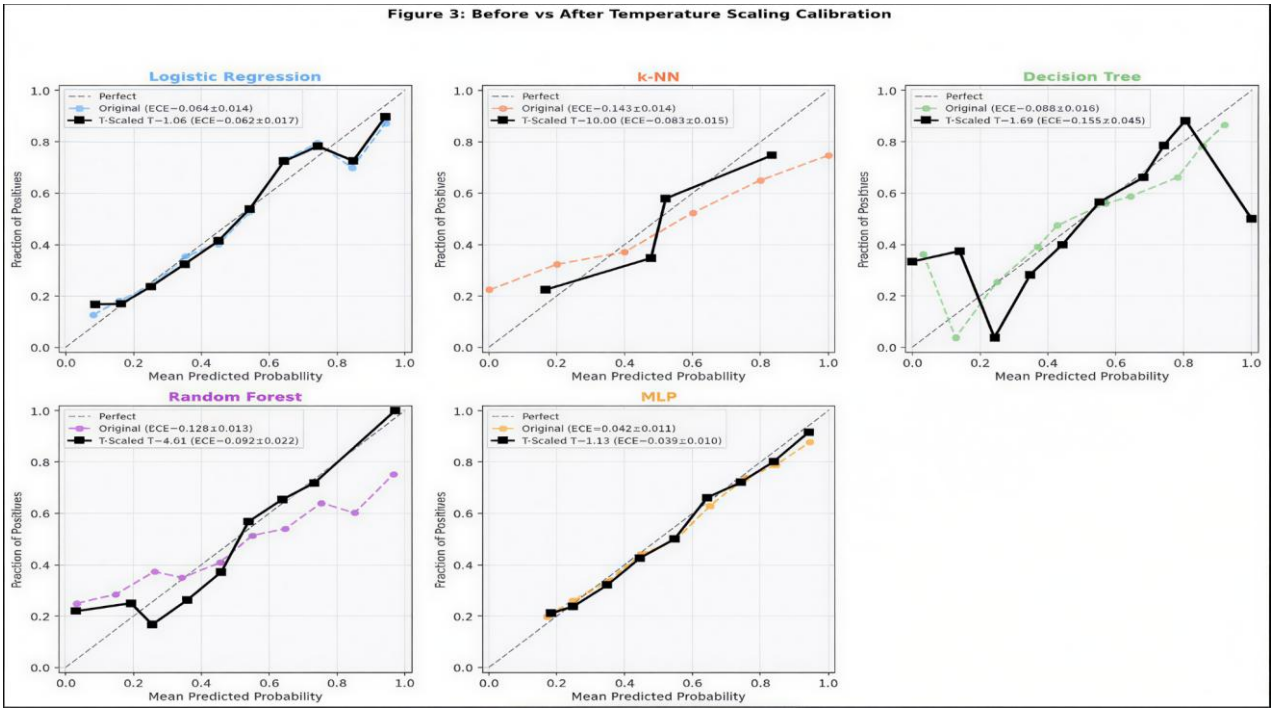

Fig. 4. Comparison of calibration performance before and after temperature scaling

Table 3. Accuracy at fixed coverage levels. Gain = Acc 80% - Acc 100%

Model	100%	90%	80%	70%	Gain
Logistic Regression	0.676	0.692	0.705	0.732	+0.029
k-NN	0.647	0.647	0.647	0.647	+0.000
Decision Tree	0.666	0.683	0.690	0.706	+0.024
Random Forest	0.637	0.649	0.662	0.678	+0.025
MLP	0.674	0.701	0.718	0.733	+0.044

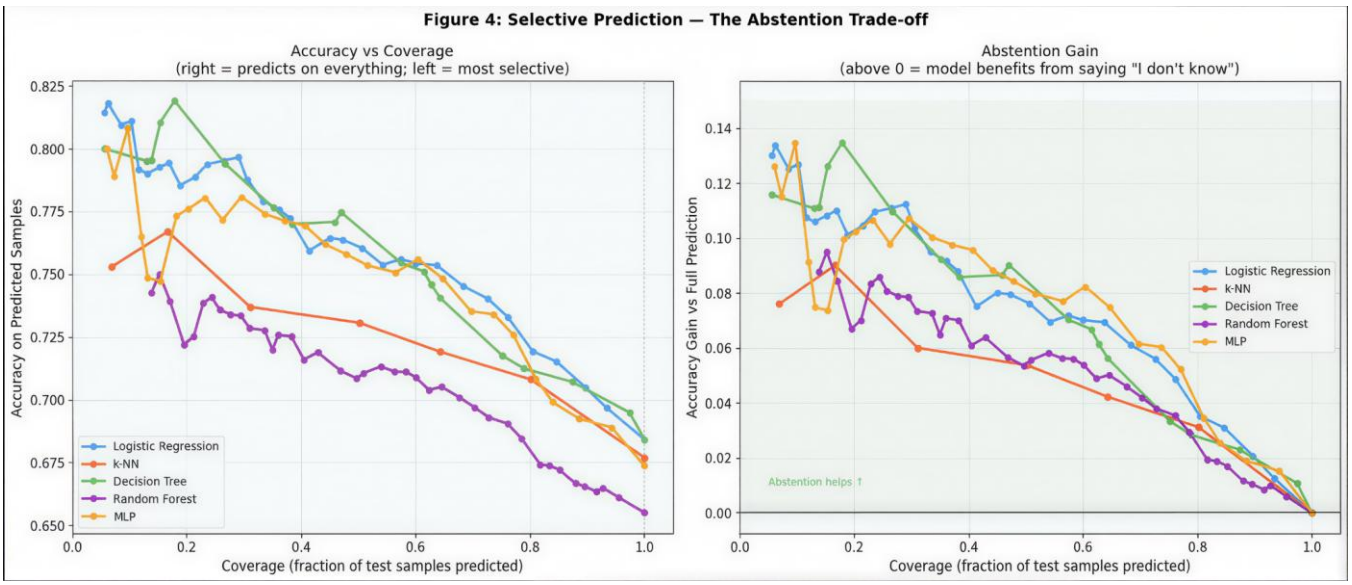

Fig. 5. Selective prediction performance showing accuracy–coverage trade-offs and abstention gains across classifiers.

In the left figure of Fig. 5, both the MLP and the Logistic Regression start with accuracy greater than 0.75 with low coverage and slowly deteriorate as coverage increases. In contrast, Random Forest is predicted to remain below 0.68 for quite a while before achieving full coverage, suggesting that the respective model does not have sufficient confidence level to separate between correct and incorrect forecasts in a consistent manner. Additionally, when low-confidence instances are

incorporated into the prediction procedure, the accuracy of Random Forest drops markedly.

The gains of abstention in terms of total accuracy versus full coverage baseline are presented in the right plot of the figure. Apart from k-NN, all the other models are above 0 level, meaning that selective prediction is valid in most cases. Being based on the neighbors' votes, k-NN receives the least gains, while MLP and Logistic Regression have significant gains with a decrease in coverage. Thus, it could be argued that abstention is not appropriate for the model that has confidence scores reflecting true uncertainty.

4.4. Comprehensive Comparison

Not one key metric gives a comprehensive image of how models perform, which is demonstrated in Table 4 that combines all evaluation methods. The conclusion can be drawn that the calibration ranking and the accuracy ranking are not the same, meaning that the most dependable model is not always chosen by practitioners based on its accuracy results only.

Furthermore, it can be said that the efficiency of all five models can be enhanced significantly without further improvement made to the machine if the machine is able to defer grey areas to a human examiner. Under these evaluation criteria, the Multilayer Perceptron emerges as the strongest overall choice, as it has the lowest original ECE (0.042) and one of the highest accuracy values (67.4%).

Additionally, the Decision Tree is still a great choice with regard to the need for explicit decision rules (accuracy of 66.6%, ECE = 0.038), as it can provide a visible structure of decision-making that allows its predictions to be followed. Thus, explicit decision rules are very important when a model can affect a person's life.

Table 4 makes this trade-off explicit. Ranking the five classifiers by accuracy alone places Logistic Regression first and MLP a close second; ranking them by ECE reverses this, placing MLP first by a wide margin (0.042 against 0.064). Because ECE captures how much a model's stated confidence can be trusted on a case-by-case basis and because this trust is what selective prediction depends upon the two other rankings answering different questions rather than contradicting one another. A practitioner interested only in aggregate correctness could reasonably choose Logictic Regression; a practitioner who needs to know which individual predictions to trust, defer, or escalate would be better served by MLP.

In conclusion, it can be stated that the results of the study demonstrate that taking other parameters into account besides accuracy when selecting a model for high-risk scenarios is of great importance.

Table 4. Final comprehensive results across all evaluation dimensions

Model	Accuracy	AUC	ECE	Acc 80%	Gain
Logistic Regression	0.676 ± 0.013	0.726 ± 0.014	0.064 ± 0.014	0.705 +- 0.015	+0.029
k-NN	0.646 ± 0.014	0.682 ± 0.015	0.143 ± 0.014	0.646 +- 0.014	+0.000
Decision Tree	0.666 ± 0.014	0.712 ± 0.015	0.088 ± 0.016	0.690 +- 0.015	+0.024
Random Forest	0.638 ± 0.014	0.674 ± 0.015	0.128 ± 0.013	0.662 +- 0.015	+0.025
MLP	0.675 ± 0.014	0.733 ± 0.015	0.042 ± 0.011	0.719 +- 0.015	+0.044

5. CONCLUSION

The findings reveal that using other metrics beyond raw accuracy provides a better informative assessment for model reliability. In particular, k-NN exhibited the weakest calibration, while Random Forest, despite being the most commonly recommended ensemble classifier, also showed poor calibration. Among all models, MLP achieves the best overall calibration and gains the most from selective prediction.

Most models improved when allowed to abstain from their least confident predictions. This suggests that selective prediction can provide a practical mechanism for handling cases in which the model is uncertain. Future work could explore whether selective prediction reduces the frequency of necessary model retraining.

Calibration and selective prediction are measurable, interpretable, and practically useful, and they should be included alongside accuracy in any evaluation of a model intended for deployment in a high-stakes context. Aggregate accuracy and prediction reliability are not equivalent measures, as a model may perform well across an entire test set while providing very little indication about which individual prediction is reliable. Thus, evaluating these factors is key to determining not just about model accuracy but also about the model's consistent performance owing to its confidence.

6. REFERENCES

[1] J. Angwin, J. Larson, L. Kirchner, S. Mattu, and D. Phiffer, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks," ProPublica, May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. 34th Int. Conf. Machine Learning (ICML), 2017, pp. 1321–1330.

[3] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4878–4887.

[4] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[5] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in Proc. 22nd Int. Conf. Machine Learning (ICML), 2005, pp. 625–632.

[6] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, Jan. 1950.

[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[9] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[10] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.