International Journal of Science, Strategic Management and Technology
Volume **02** Issue **08** August-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Reliability-Aware Screening with Subject-Independent Parkinson's Disease Classifiers: Uncertainty, Calibration, Selective Prediction, and Conformal Evaluation

Bhargavi Kapilavai[1]

[1] M.Tech Scholar, Department of Computer Science and Engineering,

JNTUK University College of Engineering Kakinada, Andhra Pradesh, India
Corresponding Author: Bhargavi Kapilavai

https://doi.org/10.55041/ijsmt.v2i8.091

Cite this Article: Kapilavai, B. (2026). Reliability-Aware Screening with Subject-Independent Parkinson's Disease Classifiers: Uncertainty, Calibration, Selective Prediction, and Conformal Evaluation. International Journal of Science, Strategic Management and Technology, *02*(8), 1-9. https://doi.org/10.55041/ijsmt.v2i8.091

License: This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

Abstract

Handwriting and drawing tasks such as spirals, circles, and meanders, together with pen-movement signals, are widely used to build machine learning models for Parkinson's disease detection. Most studies report classification accuracy but say little about whether the predicted probabilities can be trusted for screening decisions on new people. In this work we take previously trained, subject-independent classifiers - three MobileNetV2 CNNs for circle, spiral, and meander drawings and a random forest for pen signals - and treat them as frozen models. Rather than improving accuracy, we evaluate their reliability using four tools: Monte-Carlo dropout and tree-disagreement uncertainty, temperature-scaling calibration measured by the Expected Calibration Error, selective prediction with risk-coverage analysis, and split conformal prediction, all under a strict subject-independent protocol. Reliability varies clearly across modalities, with the signal and meander models showing stronger reliability characteristics than the circle model under the evaluated criteria. Because the held-out test set has only ten subjects, we present the results as a preliminary, screening-oriented reliability assessment rather than a clinical claim.

Keywords

Parkinson's disease; trustworthy machine learning; uncertainty estimation; Monte-Carlo dropout; probability calibration; Expected Calibration Error; selective prediction; conformal prediction; subject-independent evaluation.

I. INTRODUCTION

Parkinson's disease is a progressive neurological disorder that affects movement, and its early motor symptoms include tremor, rigidity, and reduced fine motor control. These symptoms leave measurable traces in handwriting and drawing, which is why tasks such as drawing spirals, circles, and meanders, as well as recording the pen trajectory during writing, have become popular low-cost inputs for computer-assisted screening. A large body of work has shown that convolutional neural networks and classical models

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

trained on these inputs can separate people with Parkinson's disease from healthy controls with encouraging accuracy.

However, accuracy alone does not tell a clinician how much to trust an individual prediction. A screening tool that is right most of the time can still be dangerously over-confident on the specific cases where it is wrong. For a decision-support setting, the more useful questions are whether the reported confidence matches the true chance of being correct, whether the model can flag its own uncertain cases so that they are referred to a specialist instead of being decided automatically, and whether we can attach an empirically supported coverage guarantee to the predictions under the conformal assumptions. These are questions of reliability and trustworthiness, and they are largely orthogonal to the raw classification score.

In this paper we deliberately separate the two concerns. We do not propose a new architecture and we do not retrain the classifiers to chase a higher accuracy. Instead, we take existing subject-independent Parkinson's classifiers as frozen models and ask a single question:

Can previously trained subject-independent Parkinson's disease classifiers provide reliable screening decisions when their uncertainty, calibration, and ability to defer uncertain cases are explicitly evaluated?

To answer this, our study makes the following contributions:

- We assemble a reliability evaluation pipeline around four frozen, subject-independent models - three MobileNetV2 drawing CNNs and a pen-signal random forest - without altering their weights.
- We quantify predictive uncertainty using Monte-Carlo dropout for the CNNs and tree-disagreement for the random forest, and test whether higher uncertainty is associated with wrong predictions.
- We measure probability calibration before and after temperature scaling using the Expected Calibration Error, fitting the temperature only on calibration subjects.
- We analyse selective prediction through risk-coverage behaviour and apply split conformal prediction to obtain prediction sets with empirical coverage.
- We report all results under a strict subject-independent protocol and openly discuss the limitations imposed by a small held-out test set.

The remainder of the paper is organised as follows. Section 2 reviews related work on Parkinson's detection and on trustworthy machine learning. Section 3 describes the frozen models, the data protocol, and the four reliability methods. Section 4 gives the experimental setup. Section 5 presents and discusses the results. Section 6 concludes and outlines future work.

II. RELATED WORK

Handwriting- and drawing-based Parkinson's detection has been studied extensively. Public datasets of spiral, circle, and meander drawings, together with pen-signal recordings, have enabled models ranging from hand-crafted kinematic features fed to classical classifiers to end-to-end convolutional neural networks [1]. Transfer learning with lightweight backbones such as MobileNetV2 [2] is attractive here because the datasets are small, and random forests [3] remain a strong baseline for the tabular pen-signal features. The dominant evaluation metric in this literature is classification performance - accuracy, precision, recall, F1-score, and ROC-AUC - often supported by interpretability tools such as Grad-CAM. Our work reuses this kind of trained model but shifts the question from how accurate to how reliable.

The trustworthy machine learning literature provides the tools for that shift. Uncertainty in deep networks can be approximated cheaply with Monte-Carlo dropout, which interprets dropout at inference time as approximate Bayesian sampling [4]. For ensembles such as random forests, disagreement between trees is a natural, model-specific source of uncertainty. Calibration - the agreement between predicted confidence and observed correctness - is commonly summarised with the Expected Calibration Error [5] and the Brier score [6], and modern neural networks are known to be miscalibrated and over-confident, a problem that simple temperature scaling can often reduce [7]. Selective prediction, in which a model may abstain on low-confidence inputs, is analysed through the risk-coverage trade-off [8], [9]. Conformal

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

prediction offers distribution-free prediction sets with a user-chosen coverage level under an exchangeability assumption [10], [11]. Broad studies of predictive uncertainty under dataset shift have shown that these properties do not come for free and must be measured directly [12].

What is comparatively rare in the Parkinson's drawing literature is a systematic, subject-independent reliability evaluation that brings uncertainty, calibration, selective prediction, and conformal prediction together on frozen classifiers. Our study fills that gap: rather than proposing another classifier, it characterises the trustworthiness of the predictions that existing classifiers already produce.

III. PROPOSED METHODOLOGY

The pipeline has five stages: (i) organising the data under a strict subject-independent protocol, (ii) loading the frozen classifiers, (iii) estimating predictive uncertainty, (iv) calibrating and analysing the probabilities, and (v) evaluating selective and conformal prediction. Figure 1 shows the overall workflow of the proposed evaluation pipeline.

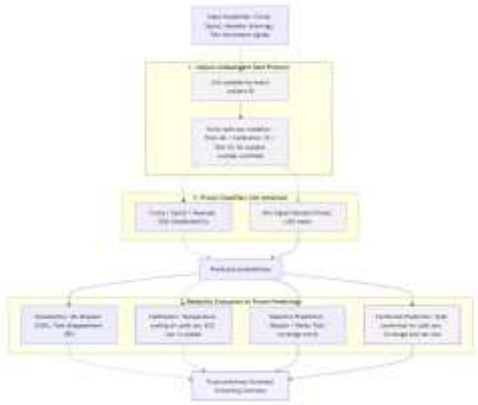

Fig. 1. Overall workflow of the proposed reliability-aware screening framework, from subject-independent data organisation through frozen classifiers to the four reliability analyses and the final trustworthy-screening summary.

A. Data and Subject-Independent Protocol

Four modalities are used: circle, spiral, and meander drawings, and pen-movement signals. Every exported sample carries an explicit subject identifier, and the multimodal rows are joined by exact subject ID rather than by row position. Each modality is divided into training, validation (used as the calibration set), and test partitions of 46, 10, and 10 subjects respectively for the strict splits. The central rule is that a subject never appears in more than one partition. We verified this independently from the stored subject identifiers: across all four modalities and both the strict and standard split types, the intersection between train, validation, and test subjects is empty. The validation partition is reused as the calibration set for temperature scaling and for the conformal quantile, so no calibration parameter is ever fitted on the test subjects. All reported metrics are computed at the sample level over the held-out test subjects. Because a subject may contribute more than one drawing or signal window, the number of test samples per modality is ten for circle, forty for spiral, forty for meander, and one hundred and twenty for pen signals, while every sample belonging to a given subject remains confined to a single partition.

B. Frozen Classifiers

The classifiers are inherited from a prior classification framework and are treated as fixed. The circle, spiral, and meander drawings are each modelled by a MobileNetV2-based binary CNN, and the pen signals are modelled by a random forest with 100 trees. We confirmed that all three CNNs load and that each contains usable dropout layers, which is a precondition for Monte-Carlo dropout, and that the random forest loads and exposes probability outputs. No model is retrained for the main experiments; the study evaluates the reliability of the predictions these frozen models already make.

C. Uncertainty Estimation

For the CNNs we apply Monte-Carlo dropout: dropout remains active at inference time and multiple stochastic forward passes are performed for each input. The mean of these passes is used as the predictive probability, and the spread across passes provides an uncertainty measure. For the random forest we do not pretend that Monte-Carlo dropout applies; instead we use the disagreement among the individual trees, which is the natural model-specific analogue. In both cases the key test is not the absolute uncertainty value but whether uncertainty is higher on incorrect predictions than on correct ones, which is what would make it useful for flagging unreliable cases.

D. Calibration

We assess calibration on the raw probabilities and on temperature-scaled probabilities. The temperature is a single scalar fitted only on the calibration subjects,

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

never on the test subjects. Calibration quality is summarised with the Expected Calibration Error, which compares average confidence with average accuracy within confidence bins, and with reliability diagrams. Because temperature scaling is fitted on a small calibration set, it is not guaranteed to help; where it makes calibration worse, we retain and report this result rather than omitting it.

E. Selective Prediction

Selective prediction allows the system to abstain on the least confident inputs, producing three outcomes: Parkinson's, Healthy, or Refer. Test samples are ranked from most to least confident using the model's uncertainty measure, and coverage is defined as the fraction of samples that are retained rather than deferred. Sweeping this fraction traces the full risk-coverage relationship, and the selective accuracy at a representative operating point of eighty percent coverage is read directly from this curve rather than tuning a threshold to obtain a target accuracy, so that the trade-off is reported transparently.

F. Conformal Prediction

We use split conformal prediction with the validation partition as the calibration set, which is disjoint from the test subjects. We set the target error level to $\alpha = 0.10$, corresponding to a 90% target coverage. For each test sample the method produces a prediction set that may contain a single class, {Parkinson's} or {Healthy}, or both classes when the model cannot separate them confidently. We report the empirical coverage and the average prediction-set size. Given the small calibration set, we treat the coverage as coarse and describe it under the assumptions of the conformal method rather than as an unconditional clinical guarantee.

IV. EXPERIMENTAL SETUP

The pipeline is implemented in Python. The drawing models use a Keras/TensorFlow MobileNetV2 backbone and the signal model uses a scikit-learn random forest, loaded from saved artifacts. Monte-Carlo dropout uses repeated stochastic forward passes per test image; the random forest uncertainty uses per-tree probability spread. Temperature scaling and the conformal quantile are fitted on the ten calibration subjects of each modality, and all reported metrics are computed at the sample level using the samples from the ten subject-disjoint test subjects, balanced as five healthy and five patient. Table 1 lists the data configuration and the evaluation settings.

Item	Setting
Modalities	Circle, Spiral, Meander drawings; Pen signals
Frozen models	3 x MobileNetV2 CNN; 1 x Random Forest (100 trees)
Split (strict, per modality)	Train 46 / Calibration 10 / Test 10 subjects
Subject overlap	None (verified from subject IDs)
CNN uncertainty	Monte-Carlo dropout (stochastic forward passes)
RF uncertainty	Per-tree disagreement / probability spread
Calibration	Temperature scaling fitted on calibration set only
Calibration metric	Expected Calibration Error (raw and scaled)
Selective metric	Risk-coverage curve; selective accuracy at 0.8 coverage
Conformal	Split conformal; empirical coverage and set size
Test set	10 subjects (5 healthy / 5 patient) per modality

Table 1. Data configuration and reliability-evaluation settings.

V. RESULTS AND DISCUSSION

A. Per-Modality Reliability Summary

Table 2 collects the main reliability metrics for the four frozen models on the held-out test subjects. The columns report test accuracy as context, the Expected Calibration Error before and after temperature scaling, the selective accuracy at eighty percent coverage, the conformal coverage and average prediction-set size, and the mean uncertainty on correct versus incorrect predictions.

Modality	Test Acc	ECE raw	ECE cal	Sel. Acc @0.8	Conf. Cov	Avg set size	Unc (correct)	Unc (wrong)
Circle	0.60	0.428	0.413	0.50	0.50	0.90	0.015	0.026
Spiral	0.80	0.135	0.191	0.812	0.80	1.10	0.074	0.119
Meander	0.80	0.123	0.136	0.844	0.80	1.00	0.063	0.135
Signals	0.708	0.067	0.061	0.76	0.867	1.333	0.326	0.438

Table 2. Trustworthy-screening summary across the four frozen modalities (using samples from the held-out test subjects).

B. Calibration

Figure 2 shows the Expected Calibration Error before and after temperature scaling. The signal model is the best calibrated, with a low raw error that improves slightly after scaling, and the meander model is also reasonably calibrated. The circle model is clearly the worst, with a large calibration error that temperature scaling only marginally reduces. For the spiral model,

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **08** August-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

temperature scaling actually increases the calibration error on this small test set. We keep this result as reported rather than omitting it: with only ten calibration subjects, a single scalar temperature is not guaranteed to generalise, and this reflects the variability expected with such a small calibration set.

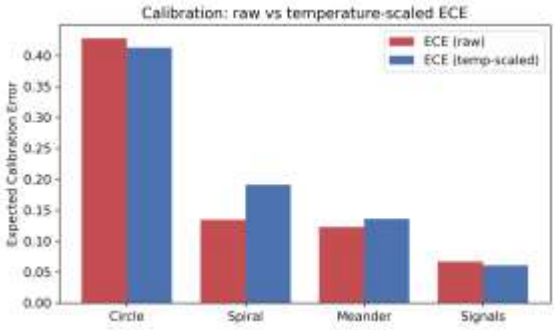

Fig. 2. Expected Calibration Error before and after temperature scaling for each modality. Lower is better; note that scaling does not help every modality.

C. Uncertainty and Error

Figure 3 compares the mean uncertainty on correct and incorrect predictions. For the spiral, meander, and signal models the uncertainty is consistently higher on wrong predictions than on correct ones, which is the desired behaviour: uncertainty carries usable information about when the model is likely to be mistaken. For the circle model the gap is very small and both values are low, meaning its uncertainty is not a reliable warning signal - consistent with its poor calibration in Figure 2.

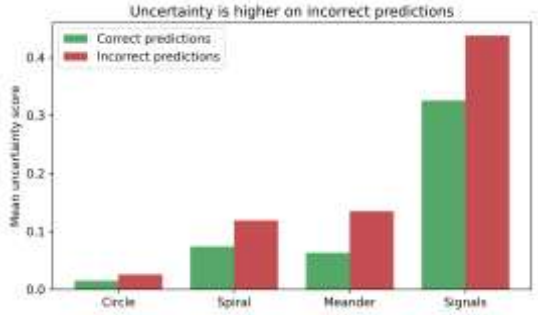

Fig. 3. Mean predictive uncertainty on correct versus incorrect predictions. A larger gap indicates that uncertainty is more useful for flagging errors.

D. Selective Prediction and Conformal Coverage

Figure 4 summarises selective prediction and conformal coverage. Allowing the model to defer uncertain cases raises the accuracy on the retained cases for the stronger modalities: at eighty percent coverage the meander and spiral models reach selective accuracies of roughly 0.84 and 0.81, and the signal model about 0.76, all above their full-coverage accuracy. The circle model does not benefit, again reflecting its uninformative uncertainty. Against the 90% target coverage, the observed conformal coverage varies substantially across modalities, with the drawing models reaching 0.50 to 0.80 and the signal model reaching 0.867; given the small calibration and test sets, these empirical values should be interpreted cautiously. Figure 5 shows the corresponding average prediction-set sizes: values above one indicate that some test samples receive the ambiguous two-class set {Parkinson's, Healthy}, which is the uncertainty-aware 'refer' behaviour we want from a screening tool that is unsure.

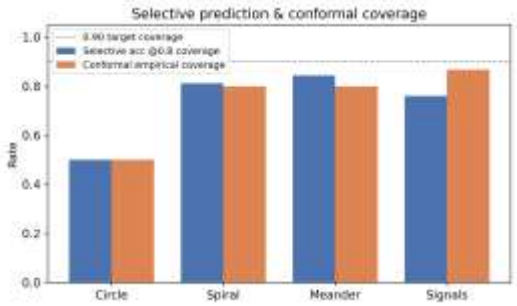

Fig. 4. Selective accuracy at eighty percent coverage and empirical conformal coverage per modality.

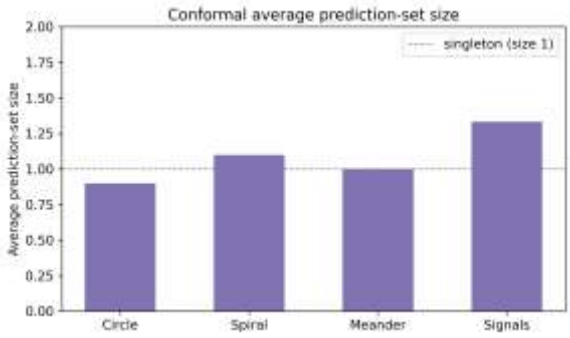

Fig. 5. Average conformal prediction-set size per modality. A size above one reflects ambiguous two-class sets, i.e. cases the model refers rather than decides.

E. Fusion as a Reference Baseline

For completeness, and only as an inherited reference rather than a new contribution, we include the behaviour of the existing fusion baselines under the same subject-independent protocol. Table 3 reports the late-fusion soft-voting ensemble and its component models, evaluated at the subject level by aligning one prediction per subject across the four modalities. The late-fusion ensemble reaches an accuracy of 0.80 and a

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

ROC-AUC of 0.88, comparable to the better single modalities. Because this inherited baseline operates on one aligned prediction per subject rather than on individual samples, its per-modality values are reported over the ten test subjects; the signal model, for example, reads 0.70 here (seven of ten subjects) as against 0.708 at the sample level in Table 2. We do not present this as an accuracy improvement; it is included so that the reliability discussion is anchored to the same fixed models used elsewhere in the project.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Circle CNN	0.60	0.667	0.40	0.500	0.76
Spiral CNN	0.80	0.714	1.00	0.833	0.68
Meander CNN	0.80	0.714	1.00	0.833	0.92
Signal RF	0.70	0.625	1.00	0.769	0.76
Late Fusion Ensemble	0.80	0.714	1.00	0.833	0.88

Table 3. Inherited late-fusion baseline evaluated at the subject level on the ten held-out test subjects (reference only, not a contribution).

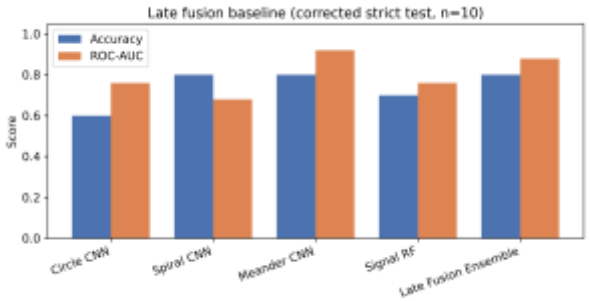

Fig. 6. Late-fusion baseline compared with its component models (reference).

The early-fusion baseline provides a cautionary illustration that motivates the whole study. Under subject-grouped five-fold cross-validation the early-fusion random forest reports a perfect mean accuracy of 1.00, yet on the independent held-out test subjects its accuracy falls to 0.50 with a ROC-AUC of 0.68, as shown in Figure 7. This large gap between development-time cross-validation and held-out performance is precisely the kind of over-optimism that accuracy-only reporting can hide, and it is a strong argument for the reliability-focused evaluation proposed here. We deliberately retain this result as a finding and do not 'correct' it by retraining or reselecting subjects.

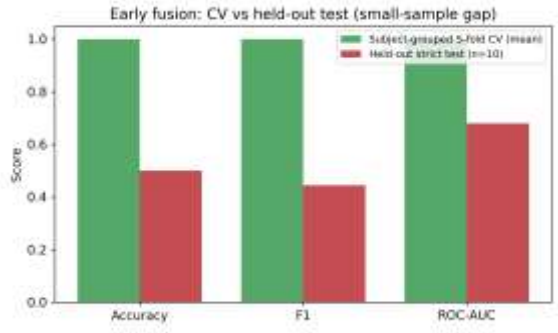

Fig. 7. Early-fusion cross-validation performance versus held-out test performance, illustrating development-to-test over-optimism.

F. Limitations

Several limitations must be stated plainly. First, the strict held-out test set contains only about ten subjects per modality, so every metric is preliminary and high variance. Second, temperature scaling and the conformal quantile are fitted on a similarly small calibration set, which is why calibration does not always improve and why conformal coverage is coarse. Third, the study is screening-oriented and makes no claim of clinical validation, diagnosis, or deployment. Fourth, the early-fusion cross-validation-to-test gap shows that results on this dataset size should be read as indicative rather than definitive. These constraints are inherent to the available data and are reported rather than engineered away.

VI. CONCLUSION AND FUTURE WORK

This paper presented a reliability-aware evaluation of subject-independent Parkinson's disease classifiers. Rather than proposing a new architecture or chasing higher accuracy, we treated three MobileNetV2 drawing CNNs and a pen-signal random forest as frozen models and characterised the trustworthiness of their predictions using Monte-Carlo dropout and tree-disagreement uncertainty, temperature-scaling calibration with the Expected Calibration Error, selective prediction with risk-coverage analysis, and split conformal prediction, all under a strictly subject-independent protocol. The results show that reliability varies markedly across modalities: the signal and meander models are comparatively well calibrated with informative uncertainty, the spiral model is usable but does not benefit from temperature scaling on this small set, and the circle model is poorly calibrated with uninformative uncertainty. Selective prediction

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

improved accuracy on the retained cases for the stronger modalities, and conformal prediction produced empirical coverage together with ambiguous prediction sets that can be interpreted as candidate referral cases.

The practical message is that classification accuracy is not sufficient evidence that a Parkinson's screening model can be trusted, and that uncertainty, calibration, selective prediction, and conformal prediction give a more complete and more transparent picture. Future work should validate these findings on larger, external, subject-independent cohorts, explore calibration methods better suited to very small calibration sets, and study reliability-aware fusion of the modalities. We emphasise again that all findings here are preliminary and screening-oriented, and that external validation is required before any clinical use.

Data and Code Availability

The trained models, split definitions, result tables, and figure-generation scripts used in this study are maintained within the project repository and can be made available for reproducibility on reasonable request.

Ethics Statement

This study used previously collected, de-identified handwriting, drawing, and pen-signal data. No new data collection involving human participants was conducted as part of this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

The authors received no specific funding for this research from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Bhargavi Kapilavai: conceived and implemented the reliability-evaluation pipeline, conducted the experiments and analysis, prepared the figures and tables, and wrote the manuscript.

REFERENCES

[1] C. R. Pereira et al., "A step towards the automated diagnosis of Parkinson's disease: Analyzing handwriting movements," in Proc. IEEE Int. Symp. Computer-Based Medical Systems (CBMS), 2015, pp. 171-176.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520.

[3] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[4] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. Machine Learning (ICML), 2016, pp. 1050-1059.

[5] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in Proc. AAAI Conf. Artificial Intelligence, 2015, pp. 2901-2907.

[6] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1-3, 1950.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2017, pp. 1321-1330.

[8] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," Journal of Machine Learning Research, vol. 11, pp. 1605-1641, 2010.

[9] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4878-4887.

[10] G. Shafer and V. Vovk, "A tutorial on conformal prediction," Journal of Machine Learning Research, vol. 9, pp. 371-421, 2008.

[11] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv:2107.07511, 2021.

[12] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 13991-14002.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.