A Appendix

A.1 Source Code

We release *SmartCal* as an open-source package in addition to the scripts used in the knowledge base construction and benchmarking experiments to make the results reproducible.[3]

A.2 Collected Datasets Information

The following includes the datasets collected and used in our experiments. The largest portion was used to build the knowledge base, while a subset of the datasets was reserved to compare the performance of the proposed approach. These benchmark data sets were not included in the knowledge base and were unseen evaluation sets.

Tabular Datasets.

Dataset	Source	Classes	Instances	Used in
airlines V2	openml	2	26969	knowledgebase
analcatdata dmft	openml	6	797	knowledgebase
Apple Stock Price Trends	openml	3	2516	knowledgebase
Auction Verification	uci	2	2043	knowledgebase
bank marketing	uci	2	45211	knowledgebase
blood	openml	2	748	knowledgebase
breast cancer	kaggle	2	569	knowledgebase
bridges	openml	6	150	knowledgebase
cars	kaggle	4	1728	knowledgebase
cars1	openml	3	392	knowledgebase
cirrhosis	uci	3	418	knowledgebase
classification in asteroseismology	kaggle	2	1001	knowledgebase
cnae-9	openml	2	240	knowledgebase
compas-two-years	openml	2	5278	knowledgebase
Credit Approval Classification	openml	2	1000	knowledgebase
Credit Score	openml	3	50000	knowledgebase
credit-g	openml	2	1000	knowledgebase
crosswalk	kaggle	4	600	knowledgebase
CustomerSegmentation	kaggle	4	10695	knowledgebase
darwin	uci	2	174	knowledgebase
dataset china	openml	5	27522	knowledgebase
dataset credit risk file2	openml	4	51336	knowledgebase
dermatology database 1	kaggle	6	366	knowledgebase
diabetes risk prediction dataset	kaggle	2	520	knowledgebase
diagnosed cbc data v4	kaggle	5	1281	knowledgebase
Dry Bean Dataset	kaggle	5	13611	knowledgebase
EDA-Home-Mortgage-NY-2	openml	6	87930	knowledgebase
eeg-eye-state	openml	2	14980	knowledgebase
ETH-BTC-USD	kaggle	2	3654	knowledgebase
fetal health	kaggle	3	2126	knowledgebase
Financial Risk Assessment	openml	3	15000	knowledgebase

Continued on next page

[3] `https://anonymous.4open.science/r/SmartCal/README.md`

Dataset	Source	Classes	Instances	Used in
first-order-theorem-proving	openml	6	6118	knowledgebase
fitness class 2212	kaggle	2	1500	knowledgebase
Flare	openml	6	1066	knowledgebase
German-Credit-Data-Creditability-2	openml	2	1000	knowledgebase
gisette	openml	2	7000	knowledgebase
glass	kaggle	7	214	knowledgebase
happydata	kaggle	2	143	knowledgebase
hayes-roth clean	openml	3	160	knowledgebase
hill-valley	openml	2	1212	knowledgebase
ilpd-numeric	openml	2	583	knowledgebase
Indian pines	openml	8	9144	knowledgebase
Interest Rate	openml	3	32862	knowledgebase
Iris	kaggle	3	150	knowledgebase
Is fraud	openml	2	5227	knowledgebase
isolet	openml	26	7797	knowledgebase
King-rook-vs-King	openml	18	28056	knowledgebase
kits	openml	2	1000	knowledgebase
kr-vs-kp	openml	2	3196	knowledgebase
land mines	uci	5	338	knowledgebase
Lead Scoring	kaggle	2	9240	knowledgebase
letter	openml	26	20000	knowledgebase
liver cirrhosis	kaggle	3	25000	knowledgebase
MagicTelescope	openml	2	19020	knowledgebase
mfeat-pixel	openml	10	2000	knowledgebase
microaggregation2	openml	5	20000	knowledgebase
Midwest survey	openml	9	2494	knowledgebase
Mnist2D	kaggle	10	70000	knowledgebase
monks-problems-2	openml	2	601	knowledgebase
Corporate Credit	openml	10	5000	knowledgebase
mushrooms	kaggle	2	8124	knowledgebase
NATICUSdroid	uci	2	29333	knowledgebase
NHANES	uci	2	6287	knowledgebase
NPHA	uci	3	714	knowledgebase
open payments	openml	2	73558	knowledgebase
optdigits	openml	10	5620	knowledgebase
ozone-level-8hr	openml	2	2534	knowledgebase
pbcseq	openml	3	1945	knowledgebase
pendigits	openml	10	10992	knowledgebase
penguins	openml	3	344	knowledgebase
Phishing Websites	uci	2	11055	knowledgebase
phoneme	openml	2	5404	knowledgebase
students dropout & success	uci	3	4424	knowledgebase
qsar-biodeg	openml	2	1055	knowledgebase
Raisin Dataset	kaggle	2	900	knowledgebase
regensburg pediatric appendicitis	uci	2	782	knowledgebase
regime alimentaire	openml	2	202	knowledgebase

Continued on next page

Dataset	Source	Classes	Instances	Used in
riceClassification	kaggle	2	18185	knowledgebase
Satellite	openml	2	5100	knowledgebase
semeion	openml	2	319	knowledgebase
Sick numeric	openml	2	3772	knowledgebase
SIRTUIN6	uci	2	100	knowledgebase
sonar	kaggle	2	208	knowledgebase
spoken-arabic-digit	openml	10	263256	knowledgebase
Stars	kaggle	6	240	knowledgebase
steel-plates-fault	openml	2	1941	knowledgebase
TCGA InfoWithGrade	uci	2	839	knowledgebase
teachingAssistant	openml	3	151	knowledgebase
telco-customer-churn	openml	2	7043	knowledgebase
Telecust1	kaggle	4	1000	knowledgebase
Thyroid Diff	uci	2	383	knowledgebase
tic-tac-toe	openml	2	958	knowledgebase
titanic	openml	2	891	knowledgebase
total score	openml	4	719	knowledgebase
Traffic violations	openml	3	70340	knowledgebase
TUNADROMD	uci	2	4465	knowledgebase
variousCancers final	openml	9	383	knowledgebase
vehicle	kaggle	4	846	knowledgebase
waveform-5000	openml	3	5000	knowledgebase
weather classification data	kaggle	4	13200	knowledgebase
BankNote Authentication	kaggle	2	1372	knowledgebase
heart	kaggle	2	1026	knowledgebase
iris v2	kaggle	3	151	knowledgebase
wines SPA	kaggle	21	7500	knowledgebase
housing	kaggle	5	20640	knowledgebase
data	kaggle	2	569	knowledgebase
California-Housing-Classification	openml	2	20640	knowledgebase
auto-mpg	kaggle	3	398	knowledgebase
diabetes	kaggle	2	768	knowledgebase
student-por	kaggle	2	649	knowledgebase
bank	kaggle	2	11162	knowledgebase
insurance	kaggle	6	1338	knowledgebase
Social Network Ads	kaggle	2	400	knowledgebase
wine-quality-white-and-red	kaggle	2	6497	knowledgebase
Insurance claims data	kaggle	2	58592	knowledgebase
graduation dataset	kaggle	3	4424	knowledgebase
predictive maintenance	kaggle	6	10000	knowledgebase
retail store inventory	kaggle	4	73100	knowledgebase
all seasons	kaggle	27	12844	knowledgebase
Uncleaned employees final dataset	kaggle	9	17417	knowledgebase
Rice Cammeo Osmancik	uci	2	3810	knowledgebase
USPS	openml	2	1424	knowledgebase
ibm-employee-attrition	openml	2	1470	knowledgebase

Continued on next page

Dataset	Source	Classes	Instances	Used in
ibm-employee-performance	openml	2	1470	knowledgebase
amazon employee access	openml	2	32769	knowledgebase
kdd internet usage	openml	2	10108	knowledgebase
law-school-admission-bianry	openml	2	20800	knowledgebase
anneal	openml	2	898	knowledgebase
arcene	openml	2	100	knowledgebase
DiabeticMellitus	openml	2	281	knowledgebase
robert	openml	10	10000	knowledgebase
Meta Album SPT Micro	openml	20	800	knowledgebase
autoUniv-au6-750	openml	8	750	knowledgebase
Otto Group Product Challenge	openml	9	61878	knowledgebase
credit-score-classification-Hzl	openml	3	100000	knowledgebase
ASP-POTASSCO-classification	openml	11	1294	knowledgebase
MIP-2016-classification	openml	5	218	knowledgebase
CPMP-2015-runtime-classification	openml	4	527	knowledgebase
cjs	openml	6	2796	knowledgebase
analcatdata supreme	openml	10	4052	knowledgebase
Telco Customer Churn	kaggle	2	7043	knowledgebase
loan prediction	kaggle	2	614	knowledgebase
OVA Kidney	openml	2	1545	knowledgebase
eucalyptus	openml	2	736	knowledgebase
albert	openml	2	58252	knowledgebase
heloc	openml	2	10000	knowledgebase
riccardo	openml	2	20000	knowledgebase
madeline	openml	2	3140	knowledgebase
guillermo	openml	2	20000	knowledgebase
xd6	openml	2	973	knowledgebase
ada	openml	2	4147	knowledgebase
philippine	openml	2	5832	knowledgebase
road-safety	openml	2	111762	knowledgebase
rl	openml	2	4970	knowledgebase
Bioresponse	openml	2	3751	knowledgebase
cylinder-bands	openml	2	540	knowledgebase
user behavior dataset	kaggle	5	700	knowledgebase
okcupid-stem	openml	3	50789	knowledgebase
wine-quality-red	openml	6	1599	knowledgebase
football-player-position	openml	4	3611	knowledgebase
wall-robot-navigation	openml	4	5456	knowledgebase
dilbert	openml	5	10000	knowledgebase
satimage	openml	6	6430	knowledgebase
Advanced IoT Dataset	openml	6	30000	knowledgebase
fabert	openml	7	8237	knowledgebase
JapaneseVowels	openml	9	9961	knowledgebase
volkert	openml	10	58310	knowledgebase
ad click	kaggle	2	10000	benchmarking
aids clinical trials group study 175	uci	2	2139	benchmarking

Continued on next page

Dataset	Source	Classes	Instances	Used in
wdbc	openml	2	569	benchmarking
wheat	kaggle	3	210	benchmarking
wilt	openml	2	4839	benchmarking
wine data	kaggle	7	21000	benchmarking
Zombies-Apocalypse	openml	2	200	benchmarking
postoperative-patient-data	openml	2	88	benchmarking
kc2	openml	2	522	benchmarking
gender	kaggle	2	66	benchmarking
desharnais	openml	3	81	benchmarking
GCM	openml	14	190	benchmarking
detect dataset	kaggle	2	12001	benchmarking
income evaluation	kaggle	2	32561	benchmarking
preterm	kaggle	2	58	benchmarking
BraidFlow	openml	3	72	benchmarking
Obesity Classification	kaggle	3	108	benchmarking
Dataset-Mental-Disorders	kaggle	4	120	benchmarking
drug200	kaggle	5	200	benchmarking
heart-long-beach	openml	5	200	benchmarking
Period Changer	uci	2	90	benchmarking
SomervilleHappinessSurvey2015	uci	2	143	benchmarking
Toxicity	uci	2	171	benchmarking
monks-problems-1	openml	2	556	benchmarking
Asteroid Dataset	openml	2	126131	benchmarking
gina prior2	openml	10	3468	benchmarking

Image Datasets. We have used 5 language datasets with different task types. The datasets were divided between our knowledge base and benchmarking *SmartCal* performance. The datasets are listed in Table 6

Table 6: Image Classification Datasets

Dataset	Classes	Instances	Library	Mean	STD	Used in
SVHN	10	99289	torchvision	(0.4377, 0.4438, 0.4728)	(0.1980, 0.2010, 0.1970)	knowledgebase
CIFAR10	10	60000	torchvision	(0.4914, 0.4822, 0.4465)	(0.2470, 0.2435, 0.2616)	knowledgebase
CIFAR100	100	60000	torchvision	(0.5071, 0.4867, 0.4408)	(0.2675, 0.2565, 0.2761)	knowledgebase
MNIST	10	70000	torchvision	0.1307	0.3081	benchmarking
USPS	10	9298	torchvision	0.2179	0.3394	benchmarking

Language Datasets. We have used 4 language datasets with different task types. The datasets were divided between our knowledge base and benchmarking *SmartCal* performance. The datasets are listed in Table 7

A.3 Classification Models

This appendix outlines all classification models employed in our experiments across both the knowledge base and benchmarking datasets. These models were subjected to calibration techniques

Table 7: Language Classification Datasets

Dataset	Source	Classes	Instances	Task Type	Used in
IMDB	Kaggle	2	50000	Sentiment Analysis	knowledgebase
AGNews	Kaggle	4	127600	Categorization	knowledgebase
Language Detection	Kaggle	17	10338	Language Detection	benchmarking
Hate Speech	Kaggle	2	726119	Categorization	benchmarking

to ensure consistent confidence estimates and performance generalization across diverse data domains.

Tabular Classification Models. The tabular datasets were modeled using a variety of supervised classifiers, including `RandomForestClassifier`, `XGBClassifier`, `ProbabilisticSVC`, `GaussianNB`, `DecisionTreeClassifier`, `GradientBoostingClassifier`, and `AdaBoostClassifier`. All models were trained using the default hyper-parameters and using a fixed random seed of 42 to ensure reproducibility across multiple runs.

Language Classification Models. For language-based datasets, we employed two model families: `TinyBERT` and `FastText`. The training process was governed by the following general hyperparameters: a batch size of 8, 50 training epochs, an early stopping patience of 5 epochs, a minimum delta of 0.1 for early stopping, and the classification accuracy metric for monitoring model performance. A fixed random seed of 42 was used throughout.

For learning rate tuning, we used a learning rate finder strategy. For `TinyBERT`, the learning rate was explored between $1e^{-5}$ and $5e^{-3}$ across 5 epochs. For `FastText`, the range extended from 0.0001 to 1.0, also over 5 epochs.

Image Classification Models. The image classification tasks relied on four deep learning architectures: `MobileNetV2`, `VGG16`, `VGG19`, and `ResNet18`. The training setup involved 50 epochs, a base learning rate of 0.001, a batch size of 32, and an image input size of 224 pixels. A random seed of 42 was applied to standardize the results. Model performance was monitored using validation loss.

To improve generalization and prevent overfitting, we incorporated early stopping with a patience of 5 epochs and a minimum delta of 0.0001. Additionally, a learning rate finder was used to sweep learning rates between $1e^{-7}$ and 1.0 over 50 iterations.

A.4 Calibration Methods and Hyperparameter Search Space

This section outlines the detailed calibration techniques considered in our study along with their corresponding hyperparameter search spaces. These methods were applied to improve the probabilistic outputs of various classifiers, ensuring reliable confidence estimations across datasets.

Empirical Binning Calibrator.. This method discretizes predicted probabilities into bins and computes calibrated probabilities for each bin. The default number of bins is set to 10. During hyperparameter search, we evaluate the number of bins from the set $\{5, 10, 15, 20\}$.

Isotonic Calibrator.. A non-parametric method that fits a free-form line to the predicted probabilities while preserving order. It does not require hyperparameter tuning.

Beta Calibrator.. This method applies a Beta-transformation to the probability scores. The model type is set to `'abm'` by default, while the search space includes model types `'abm'`, `'am'`, and `'ab'`.

Temperature Scaling Calibrator.. A parametric method that learns a single scalar parameter to scale logits. Default settings include `initial_T = 1.0`, `lr = 0.01`, and `max_iter = 100`. The grid search explores initial temperature values in $\{0.5, 1.0, 1.5, 2.0\}$, learning rates in $\{0.001, 0.01, 0.1, 1.0\}$, and iterations in $\{50, 100, 300, 500, 700, 1000, 1500, 2000\}$.

Vector, Matrix, and Dirichlet Calibrators.. These advanced parametric calibrators optimize weight matrices or distributions over logits. All three use the same default settings of `lr = 0.01` and `max_iter = 100`, and their hyperparameter grid search includes the same value sets as temperature scaling.

Meta Calibrator.. This method incorporates constraints based on calibration error or accuracy. The default settings are `alpha = 0.1`, `acc = 0.85`, and `default_constraint = 'ALPHA'`. The search space spans constraint types `'ALPHA'` and `'ACC'`, with alpha values in $\{0.01, 0.05, 0.1, 0.15, 0.2\}$ and accuracy targets in $\{0.8, 0.85, 0.9, 0.95\}$.

Platt Calibrator.. A logistic regression-based calibration method. The default calibrator type is `'PLATT'`. We also explore types `'PLATTBINNER'` and `'PLATT_SCALING'`, with `num_bins` in $\{5, 10, 15, 20\}$.

Histogram Calibrator.. This non-parametric method assigns calibrated scores based on histogram binning. The default is `'HISTOGRAM'`, and search space includes `'HISTOGRAM_TOP'` and `'HISTOGRAM_BINNING'` with `num_bins` in $\{5, 10, 15, 20\}$.

Adaptive Temperature Scaling Calibrator.. An advanced hybrid technique that adapts temperature scaling based on entropy and confidence levels. Defaults are `lr = 0.1`, `max_iter = 100`, `confidence_bins = 10`, `entropy_bins = 10`, and `initial_T = 1.0` with mode set to `'hybrid'`. The search space includes mode options `'linear'`, `'entropy'`, and `'hybrid'`, and explores the same learning rate and iteration values used in other parametric calibrators.

Mix-And-Match Calibrator.. A composite calibration method combining parametric and non-parametric approaches. By default, it uses `TemperatureScalingCalibrator` and `IsotonicCalibrator`. The search space includes a variety of parametric calibrators such as Temperature, Platt, Vector, Matrix, Beta, Meta, Dirichlet, and Adaptive Temperature Scaling, combined with non-parametric options including Isotonic, Empirical Binning, and Histogram calibrators.