Cal-Net: Jointly Learning Classification and Calibration On Imbalanced Binary Classification Tasks

Arghya Datta[1], Noah R. Flynn[2], S. Joshua Swamidass[2,†]
[1]Department of Computer Science and Engineering, Washington University in Saint Louis
[2]Department of Pathology and Immunology, Washington University in Saint Louis
[†]swamidass@wustl.edu

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534411

***Abstract*—Datasets in critical domains are often class imbalanced, with a minority class far rarer than the majority class, and classification models face challenges to produce calibrated predictions on these datasets. A common approach to address this issue is to train classification models in the first step and subsequently use post-processing parametric or non-parametric calibration techniques to re-scale the model's outputs in the second step without tuning any underlying parameters in the model to improve calibration. In this study, we have shown that these common approaches are vulnerable to class imbalanced data, often producing unstable results that do not jointly optimize classification or calibration performance. We have introduced Cal-Net, a "self-calibrating" neural network architecture that simultaneously optimizes classification and calibration performances for class imbalanced datasets in a single training phase, thereby eliminating the need for any post-processing procedure for confidence calibration. Empirical results have shown that Cal-Net outperforms far more complex neural networks and post-processing calibration techniques in both classification and calibration performances on four synthetic and four benchmark class imbalanced binary classification datasets. Furthermore, Cal-Net can readily be extended to more complicated learning tasks, online learning and can be incorporated in more complex architectures as the final state.**

I. Introduction

Advances in deep learning [1] encouraged the use of neural networks in several domains, including medicine and healthcare [2]. They are increasingly playing a critical role in decision-making processes. In these settings, neural networks must not only be accurate in their predictions, but should also be calibrated to output well-scaled probabilities. Predictions from a binary classifier are said to be well-calibrated if the outcomes predicted to occur with a probability p occur p fraction of the time. Figure 1 exhibits a hypothetical example of the calibration performance of a classifier on an imbalanced dataset (solid line) using reliability plots alongside the ideal calibration curve (dotted line) [3]. The closer the calibration curve of a classifier corresponds to the ideal calibration curve, the better is its calibration performance. However, datasets in critical domains can be highly imbalanced, with one class far less common than the other. Little work has been done to develop well-calibrated models on imbalanced datasets [4].

A classifier minimizes error during training, but most error functions assign equal weights to all instances which lead to the total error function being dominated by the performance on the majority class. To combat this, error functions assign higher costs on misclassifications for the minority class than the majority class, thereby aiming to maximize classification performance but often overlooking calibration performance. Furthermore, common parametric and non-parametric calibration approaches are often unstable on class imbalanced datasets [5]. Calibrated predictions are crucial in establishing trust and driving adoption of neural network-based classifiers on class imbalanced datasets in critical systems.

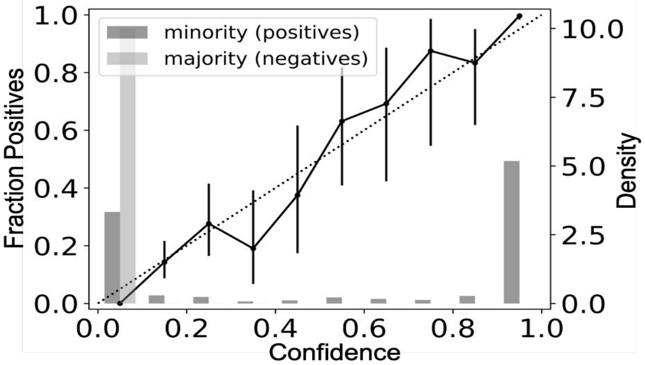

Fig. 1. Reliability plot (bins=10) showing a calibration curve of predicting positive events in an imbalanced dataset for a classifier. The solid line represents the performance of a classifier whereas the dotted line represents the ideal calibration curve. The x-axis or "confidence" is the average prediction of each bin, the primary y-axis or "fraction positives" is the fraction of minority (positive) samples in each bin and the secondary y-axis or "density" is the histogram density that shows the distribution of predictions for the classes.

Very little work has been done in exploring the possibility of addressing confidence calibration and class imbalance using neural network architectures [6]. In this work, we introduce Cal-Net, a neural network architecture and associated loss functions that simultaneously optimize classification and calibration performances on class imbalanced datasets. Empirical results have shown that Cal-Net achieves the best classification and confidence calibration performances across four simulated and four real world datasets across a diverse range of class imbalance.

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

II. Related Works

There are several approaches to address class imbalance during training and several procedures to calibrate model outputs after training. However, to the best of our knowledge, there is no published approach that simultaneously addresses class imbalance and calibration during training.

Methods for handling class imbalance are well established in the literature and include sampling strategies and cost-sensitive learning. Examples of sampling strategies include oversampling [7], which re-samples the minority class at random to match the distribution of the majority class and undersampling [8], which eliminates samples from the majority class at random to match the distribution of the minority class. Even though sampling strategies are easily implemented, oversampling can cause overfitting [9] and random undersampling can cause information loss [10]. Furthermore, undersampling one class modifies the priors of the training set and consequently biases the calibration of the final model [11]. Synthetic minority oversampling technique (SMOTE) [12] shows improvements over random oversampling by creating synthetic minority class samples but it still suffers from high variance [13]. Cost-sensitive learning addresses class imbalance via constructing an objective function that assigns different costs to the error for each class [14], [15]. Sample weighting [16] assigns high weights to samples from the minority class that can be incorporated in the entropy calculation for classifiers. In neural networks, class imbalance is usually handled using sampling strategies or cost-sensitive learning or a combination of both techniques.

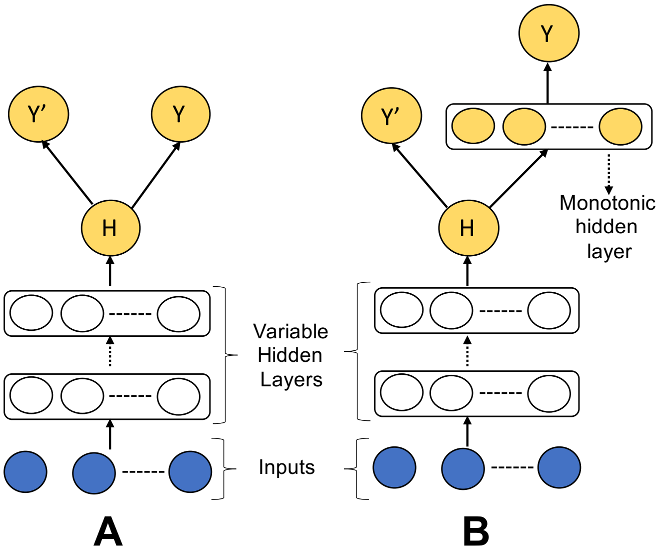

Fig. 2. Variants of the Cal-Net architecture with primary output Y and secondary output Y', both drawing from hidden node H: (A) - Simple Cal-Net with logistic function activated outputs and (B) - Cal-Net architecture enforcing monotonicity between the two outputs.

There are several methods to improve the calibration performance of models after training. For example, there are parametric approaches like Platt scaling [17] and non-parametric approaches based on isotonic regression [18] and binning [19]. Platt scaling [17] applies a logistic transformation to calibrate the outputs of a predictive model using a maximum likelihood estimation framework. Confidence calibration, based on isotonic regression [18], maps the outputs of predictive models using isotonic (monotonically increasing) transformations. Histogram binning divides outputs of a classifier into several bins and uses the fraction of positive samples in each bin as the calibrated probability. Bayesian binning into quantiles (BBQ) [20] is a post-processing confidence-calibration method that considers different binning strategies and their combinations to yield calibrated predictions. Similarly, for multiclass classifiers, extensions to Platt scaling have been proposed [21], though multiclass-classification is out of the scope of this study. These post-processing calibration methods rely on a validation dataset that is held out during training, which means that less data is used during training. Recent studies aim to incorporate confidence calibration during classifier training using a variance-weighted confidence-integrated loss function [22]. However, the challenges of jointly improving classification and calibration performances on class imbalanced data have not been addressed.

Although many studies have focused on developing improved post-processing algorithms, class probability estimates on imbalanced datasets systematically underestimate the probabilities for minority class instances [5], [23]. Moreover, none of these proposed calibration algorithms addresses the particular challenges of class imbalance with improvements to the neural network architecture itself.

III. Materials and Methods

A. *The Cal-Net Architecture*

Cal-Net works by transforming the binary classification problem into a multitask problem with a primary and a secondary output(Figure 2). The primary output (Y) is tuned to produce well-calibrated probabilities. The secondary output (Y'), used only during training, is tuned to maximize the classification performance by upweighting the samples from the minority class to be equally prevalent to samples from the majority class. The network architecture enforces a monotonic relationship between these two outputs. This multitask architecture enables Cal-Net to learn a hidden state that is tuned to maximize both classification and calibration performances, allowing both training tasks to cross-talk and refine the model simultaneously.

The primary output, $Y = \{y_i\}$, and the secondary output, $Y' = \{y'_i\}$, indexed by instance i, are computed using a logistic activation function. The Cal-Net network architecture enforces a monotonic relationship between the two outputs Y and Y'. This relationship is enforced by using a multitask architecture that funnels to a hidden layer H, a layer with a single node. The outputs are computed as monotonic functions of this layer, which ensure that they are monotonic transformations of each other.

We have considered two variants. In the first variant, the simplest "Cal-Net" defines both outputs Y and Y' as logistic functions of H without using a monotonic hidden layer between H and Y. This requires four weights in total, with two scaling weights and two biases. Two outputs are, thus, constrained to be linear transforms of each other in logit space.

In the second variant, "Monotonic Cal-Net," the primary output Y is computed from H using a single layer monotonic network. Here, the weight matrices are re-parameterized so that they are always positive [24]. The secondary output is still computed as a logistic function of H. The two outputs are constrained to be monotonically related, but this relationship can be non-linear in logit space. This architecture enables Cal-Net to learn a hidden state H, that is tuned to maximize both classification and calibration performances, allowing both training tasks to cross-talk and refine the model simultaneously.

B. Two Outputs and Four Loss Components

The primary output, $Y = \{y_i\}$, indexed by instance i, is tuned with three loss components to produce well calibrated predictions. The output node uses a logistic activation function. Three components of the loss function are computed based on this output and the target class labels $T = \{t_i\}$. The secondary output Y' is tuned with the fourth loss component L_B.

The first loss component, L_X, is the commonly used cross entropy error between Y and T. On class imbalanced training sets, instances in the majority class are far more common, so they contribute more to the loss than instances in the minority class.

The second loss component, L_H, computes the "histogram loss", based on histogram binning [19]. Conceptually, examples are binned by their respective prediction values. In a well-calibrated model, the proportion of positive examples in each bin should match the midpoint of the bin. So the loss is computed as the RMSE between these values,

$$L_H = \frac{1}{N}\sum_{n=1}^{N}(p_n - m_n)^2, \tag{1}$$

where p_n is the proportion of positive examples in bin n, m_n is the midpoint of bin n, and N is the number of bins. In this study, we used $N = 10$ in all assessments. The proportion is computed by summing up over examples i,

$$p_n = \frac{m_n \cdot \lambda_p + \sum_i t_i \cdot M(i,n)}{\lambda_p + \sum_i M(i,n)}, \tag{2}$$

where λ_p is the strength of a prior on the proportion and $M(i,n)$ is the membership of instance i in bin n. For this study, we used $\lambda_p = 10$. The membership is thus computed using the function

$$M(i,n) = \max\left[0, 1 - 2n \cdot |m_n - y_i|\right]. \tag{3}$$

This loss is minimized when the proportion of positive instances in a bin matches the midpoint of that bin. Unfortunately, this loss alone is not sufficient to train a well calibrated model. A degenerate minimum occurs when all examples are assigned the same score, equal to the proportion of positives in the whole dataset.

The third loss component, L_T, is the "t-test loss", which addresses this degenerate solution by penalizing poor separation between the distribution of positive and negative instances. We define this loss as the negative of the two sample t-test score between the positive and negative examples,

$$L_T = \frac{\sum_i (1 - t_i) \cdot y_i - t_i \cdot y_i}{\max[S, \epsilon]}, \tag{4}$$

where the summation is over examples i, S is the pooled variance and $\epsilon = 0.0001$, which is used to prevent division by zero. The pooled variance is computed as $S = \sqrt{s_1/n_1 + s_0/n_0}$, where n_1 and n_0 are the number of positive and negative instances, respectively, and s_1 and s_0 are the sum of the squared deviations of the positive and negative instances' predictions from their respective means. This loss function is minimized when the positive examples are all assigned an output of 1 and the negative examples are all assigned an output of 0.

The fourth loss component, L_B, is the "balanced loss", and is defined as the class-weighted cross entropy loss between T and Y'. Instances from the minority class are upweighted to be equally prevalent as the samples from the majority class. By equally weighting each class, this output is tuned to maximize the classification performance.

The total loss function (L) for Cal-Net, then, is computed as,

$$L = L_X + \lambda_H L_H + \lambda_T L_T + L_B, \tag{5}$$

where the lambdas are hyper-parameters that can be used to tune the histogram loss and the t-test loss, respectively. Empirical analyses show that all four loss components are necessary to optimize accuracy and calibration (section IV, subsection C) in high class imbalanced scenarios. It is possible that other formulations of the second and third loss component could be effective and exploring options are left for future work.

C. Training and Assessment Protocol

In this study, we assessed models for both calibration and accuracy on several datasets using a standardized validation protocol. For the synthetic datasets, a stratified sample of 10% of the data was held out as a test set. Another stratified validation set of 10% was also held out and, thus, each model was trained on the remaining 80% of the data. The validation set was rotated through the data, so that nine models were trained. The average performance of the nine trained models on the test set have been reported. For the benchmark datasets, a stratified sample of 20% of the data was held out as a test set, another stratified validation set of 20% was also held out and the remaining 60% of the data was used for training, thereby training four models. The average performance of the four models on the test set have been reported. Since the benchmark datasets have a low number of minority samples, this strategy ensured that some minority samples are present in each of the sets.

For all the experiments, both Cal-Net versions used a hidden layer of 10 units with ELU activation and L2 regularization.

In other datasets and architectures, it may be worthwhile to explore other ways of regularization such as L1 regularization and dropout. Monotonic Cal-Net also included a monotonic hidden layer of 10 units between H and the primary output Y. The validation set was used to optimize the hyper-parameters. We expected that the hyper-parameters, alternatively, can be analytically derived as a function of dataset size and imbalance ratio. In practice, we found that performance was not tightly dependent on the hyper-parameters. In principle, for this reason, the validation set could be merged into the training set, exposing the model to more data during training. We have left the question of how to determine the hyper-parameters without a validation set for future study.

As a baseline for comparison, we trained a neural network (NN) with a single hidden layer, consisting of 10 ReLU activated units and L2-regularization. The NN was trained using the class-weighted cross-entropy loss(L_B), where samples from the minority class were upweighted to be equally prevalent as the samples from the majority class. Making use of the validation set, the output of the NN was calibrated using non-parametric and parametric post-processing calibration techniques: isotonic regression [18], BBQ [20], [25] and Platt scaling [17].

D. Evaluation Metrics

To evaluate the classification performance of Cal-Net and the baseline models on imbalanced datasets, we reported the Brier Score, maximum F-measure and area under the receiver operating characteristic (AUROC) curve. F-measure is often preferred when the class distribution is highly skewed, since it measures the trade-off between precision and recall. AUROC, alone, is not sufficient since it is often insensitive to class imbalance [26]. To evaluate the calibration performance of Cal-Net and the baseline models on imbalanced datasets, we reported the expected calibration error (ECE) [20], [25], average calibration error (ACE) [27] and also examined the reliability diagrams [3]. As we will see, ECE and ACE can often be misleading in quantifying the calibration performance of a classifier in scenarios with high class imbalance, since calibration models are often assigned a low ECE and ACE in the presence of degenerate solutions (assigning low scores to most examples), as is evident from the reliability diagrams. For each dataset, we reported the imbalance ratio (IR), calculated as $\frac{n_0}{n_1}$, where n_1 is the number of minority (positive) samples and n_0 is the number of majority (negative) samples. A classifier with lower Brier score, higher F-measure and higher AUROC along with lower ECE, lower ACE and not providing degenerate solutions was preferred.

E. Synthetic Datasets

We generated four imbalanced binary classification datasets using Sklearn's [28] make_classification function. Each dataset was generated with 50 continuous features (15 informative and 35 redundant features), one binary outcome (2 classes) and 5 clusters per class. The binary outcomes were not linearly separable, with a varying degree of class imbalance, controlled using the weights parameter inside the make_classification function: 5.8% (IR=16.8), 1.1% (IR=87.5), 0.5% (IR=180.6) and 0.11% (IR=906.8) minority class.

F. Benchmark Datasets

We used four imbalanced binary classification datasets (Table I): (1) the UCI abalone-6 dataset [29], (2) the UCI abalone-7 dataset, (3) the UCI abalone-8 dataset and (4) the UCI adult census income dataset [30].

TABLE I
STATISTICS FOR BENCHMARK DATASETS

Dataset	Size	% +ve	Features	IR
Census income	32561	24.08	14	3
Abalone-6	4177	6.2	8	15.13
Abalone-7	4177	9.36	8	9.68
Abalone-8	4177	13.5	8	6.35

IV. RESULTS, ABLATIONS AND DISCUSSION

A. Performance on Synthetic Data

Cal-Net variants outperformed the baseline models in classification performance across all class imbalanced synthetic datasets (Table II) in terms of the Brier Score, F-measure and AUROC. In some cases, the "Monotonic Cal-Net" performed slightly better, at the cost of a more complex architecture.

Cal-Net variants achieved the best all-round calibration performance by obtaining strong performance in terms of reliability plots (Figure 3), ECE and ACE (Table II). Although Cal-Net variants achieved similar ECE and ACE scores to the baselines, they avoided the pitfall of degenerate solutions which assign all examples the same score, as shown in the reliability plots (Figure 3). Consistent with these results, Cal-Net variants exhibited far better reliability diagrams than other methods (Figure 3) across all the four synthetic datasets. The post-processing calibration methods such as BBQ and Platt scaling achieved lower ECE and ACE scores with degenerate solutions, assigning low probabilities to most examples and with several empty bins. In contrast, the Cal-Net variants did not achieve low ECE and ACE with degenerate solutions, making predictions across the whole range from 0 to 1, assigning at least some examples to most bins, with close correspondence between bin centers and the proportion of positives in each bin. Although ECE and ACE were not significantly lower for Cal-Net in many cases, the reliability diagrams indicated that Cal-Net had the best all-around performance. Although Cal-Net variants did not always achieve the best possible scores in a single metric, yet they provided overall better solutions.

B. Performance on Benchmark Datasets

We observed similar results on the benchmark datasets. In terms of classification and calibration performances, the Cal-Net variants often outperformed the baseline models by achieving higher F-measure and AUROC scores along with low Brier Scores, ECE and ACE (Table III). For the abalone-8 dataset, we observed that the ECE scores of the Cal-Net

TABLE II

On the synthetic datasets, Cal-Net variants achieved best all-round classification and calibration performances.

	5.8% minority class (IR: 16.8)					1.1% minority class (IR: 87.5)				
	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓
Cal-Net	**0.0208**	**0.77**	**0.91**	0.0039	**0.061**	0.0051	**0.69**	**0.9381**	0.0012	0.076
Monotonic Cal-Net	0.0209	**0.77**	**0.91**	**0.0035**	0.065	**0.0050**	**0.69**	0.9355	**0.0008**	**0.044**
NN + Iso. Reg.	0.0263	0.72	0.89	0.0053	0.068	0.0058	0.64	0.9241	0.0012	0.122
NN + BBQ	0.0265	0.72	0.89	0.0065	0.121	0.0059	0.64	0.9184	0.0013	0.094
NN + Platt	0.0285	0.72	0.90	0.0173	0.113	0.0071	0.65	0.9230	0.0055	0.129
NN Uncalibrated	0.0912	0.72	0.90	0.2037	0.341	0.0698	0.65	0.9230	0.1764	0.426
	0.5% minority class (IR: 180.6)					0.11% minority class (IR: 906.8)				
	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓
Cal-Net	**0.0017**	**0.78**	0.952	**0.0003**	0.103	**0.00017**	**0.912**	**0.939**	**0.00003**	0.235
Monotonic Cal-Net	0.0019	0.774	0.953	0.0004	0.092	0.00021	0.892	0.935	0.00007	0.211
NN + Iso. Reg.	0.0019	0.770	0.956	0.0005	**0.063**	0.00021	0.891	0.935	0.00005	**0.048**
NN + BBQ	0.0019	0.769	0.952	0.0006	0.127	0.00021	0.892	0.933	0.00005	0.082
NN + Platt	0.0031	0.771	**0.957**	0.0039	0.164	0.0003	0.892	0.935	0.0005	0.173
NN Uncalibrated	0.0467	0.771	**0.957**	0.1448	0.437	0.0410	0.892	0.935	0.1022	0.422

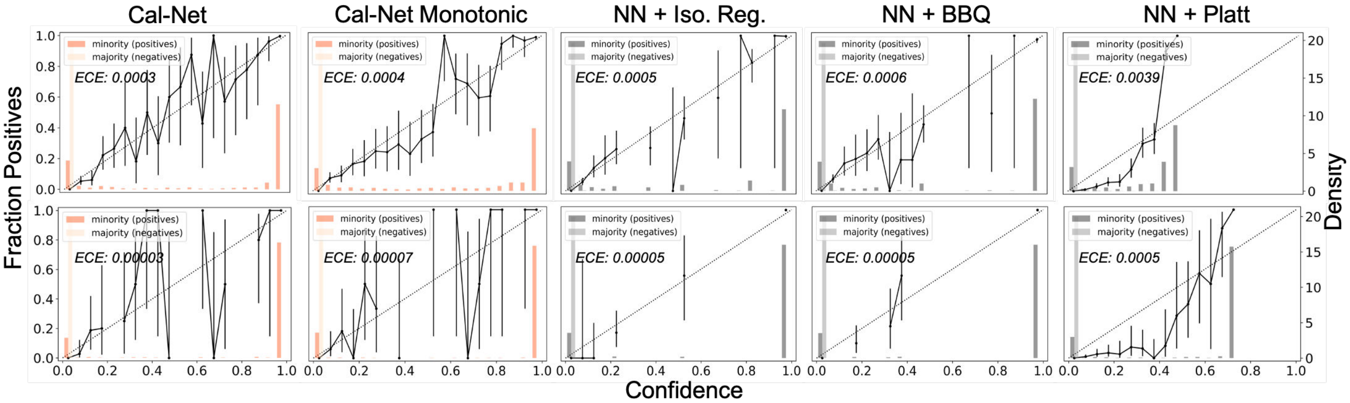

Fig. 3. Reliability diagrams (bins=20) for the synthetic datasets showing Cal-Net variants are far better calibrated, even though standard calibration techniques have nearly equivalent ECE. Top row: 0.5% minority class; Bottom row: 0.11% minority class. The x-axis or "confidence" is the average prediction of each bin, the primary y-axis or "fraction positives" is the fraction of minority (positive) samples in each bin and the secondary y-axis or "density" is the histogram density that shows the distribution of predictions for the classes.

TABLE III

On the benchmark datasets, Cal-Net variants achieve best all-round predictive performances.

	Abalone-6 : 6.2% (IR: 15.13)					Abalone-8: 13.5% (IR: 6.35)				
	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓
Cal-Net	**0.046**	**0.464**	**0.914**	0.025	0.179	0.102	**0.423**	0.781	**0.021**	0.129
Monotonic Cal-Net	**0.046**	0.456	0.908	0.023	0.169	**0.101**	**0.423**	**0.788**	0.025	0.139
NN + Iso. Reg.	0.051	0.413	0.883	0.024	0.264	0.110	0.4	0.76	0.045	0.099
NN + BBQ	0.048	0.421	0.882	0.021	0.063	0.108	0.407	0.76	0.041	**0.064**
NN + Platt	0.048	0.419	0.886	**0.013**	**0.062**	0.106	0.407	0.765	0.039	0.068
NN Uncalibrated	0.131	0.419	0.886	0.162	0.384	0.286	0.407	0.765	0.341	0.378
	Abalone-7 : 9.36% (IR: 9.68)					Census income: 24.08% (IR: 3)				
	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓	Brier ↓	F-measure ↑	AUROC ↑	ECE ↓	ACE ↓
Cal-Net	**0.07**	0.45	0.877	0.024	0.219	**0.105**	**0.7**	**0.9**	0.015	0.023
Monotonic Cal-Net	**0.07**	**0.46**	**0.88**	0.029	0.248	**0.105**	**0.7**	**0.9**	**0.014**	**0.021**
NN + Iso. Reg.	**0.07**	0.45	0.872	0.025	0.158	0.113	**0.7**	0.89	**0.014**	0.027
NN + BBQ	**0.07**	0.45	0.867	0.020	**0.039**	0.113	**0.7**	0.89	0.015	0.042
NN + Platt	**0.07**	0.45	0.875	**0.017**	0.091	0.114	**0.7**	0.89	0.023	0.037
NN Uncalibrated	0.1436	0.45	0.875	0.181	0.326	0.142	**0.7**	0.89	0.158	0.146

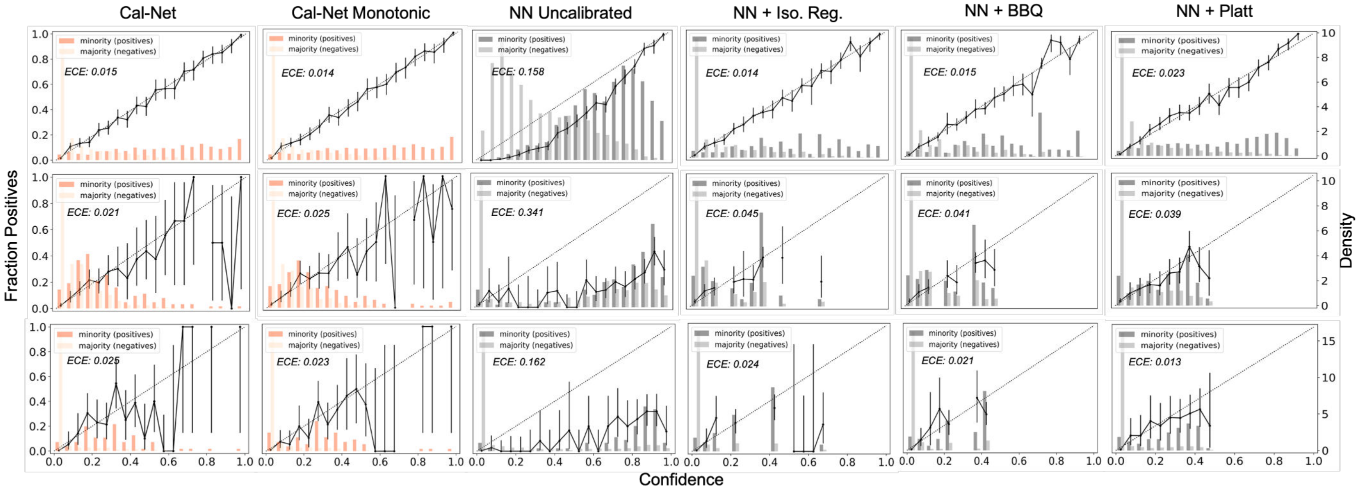

Fig. 4. Reliability diagrams on benchmark datasets (bins=20) showing that Cal-Net variants have the best calibration performance.Top row: UCI Adult Census dataset; Middle row: Abalone-8 dataset; Bottom row: Abalone-6 dataset. The x-axis or "confidence" is the average prediction of each bin, the primary y-axis or "fraction positives" is the fraction of minority (positive) samples in each bin and the secondary y-axis or "density" is the histogram density that shows the distribution of predictions for the classes.

variants were the lowest among all the methods. Although the baselines achieved lower ACE scores, the reliability plots (Figure 4) indicated that this was an artifact. Apart from the Cal-Net variants, almost all baselines excluding the uncalibrated neural network failed to assign predictions over the entire range from 0 to 1. The post-processing calibration methods achieved low ECE and ACE scores with degenerate solutions, assigning low probabilities to most examples and with several empty bins. A recent study [31] has highlighted several drawbacks in using popularly used calibration metrics such as ECE and ACE.

Notably, the Cal-Net architecture outperformed all other baselines on the Abalone-6 dataset. Although the baselines achieved lower ECE and ACE scores, the reliability diagrams indicated that this was an artifact. Cal-Net variants achieved the best overall classification performance by achieving higher F-measure, AUROC and lower Brier scores. These results suggest that Cal-Net's multitask architecture enables the calibration task to provide information relevant to improving classification accuracy and vice-versa. Results from these empirical assessments showed that Cal-Net's performance was stable and robust across multiple class imbalance levels with diverse amounts of minority samples.

C. Ablation Analyses

We performed ablation analyses to highlight the importance of Cal-Net's multitask architecture to incorporate balanced loss (L_B), histogram loss (L_H), t-test loss (L_T) and cross-entropy loss (L_X), which are necessary to improve classification and calibration performances in class imbalanced datasets. We trained multiple variants of Cal-Net by removing components such as the multitask architecture along with L_B, L_H, L_T and L_X, and compared the classification and calibration performances with the standard Cal-Net architecture while keeping all other hyper-parameters fixed. We used the synthetic dataset with 0.11% minority samples (IR=906.8) for these analyses.

1) Performance with and without L_B: To analyze the significance of the balanced cross entropy loss (L_B), a variant of Cal-Net was trained by eliminating L_B thereby resulting in elimination of the secondary output Y'. In class imbalanced datasets, L_B is usually preferred since it prevents the classifier from being biased towards the majority class and results in better classification performance by enforcing a high penalty for misclassifying minority samples. Under these settings, analyzing the classification performance revealed that there was a drop in AUROC and AUPRC by 0.006 and 0.019 respectively (Figure 5A) when L_B was eliminated from the standard Cal-Net architecture. Precision recall curves are particularly successful at quantifying the classification performance in the case of imbalanced datasets [32]. This finding suggested that using a multi-task architecture to incorporate L_B in the loss function was necessary to improve classification performance.

2) Performance with and without L_H: We analyzed the importance of the histogram loss (L_H) which was computed using Cal-Net's primary output (Y) to minimize the mean difference between the proportion of positives and predictions across all the bins. A variant of the standard Cal-Net architecture without L_H was trained and its calibration performance was compared to that of the standard Cal-Net architecture. Analysis of the calibration performance revealed that the Cal-Net architecture without L_H was unable to reduce the deviations between fraction positives and the average prediction in several bins, as shown in the reliability plot (Figure 5B). The Cal-Net architecture improved the calibration performance by assigning examples in bins with close correspondence between fraction positives and bin centers for most bins. In terms of calibration metrics, the standard Cal-Net architecture achieved lower ECE and ACE scores than the modified architecture

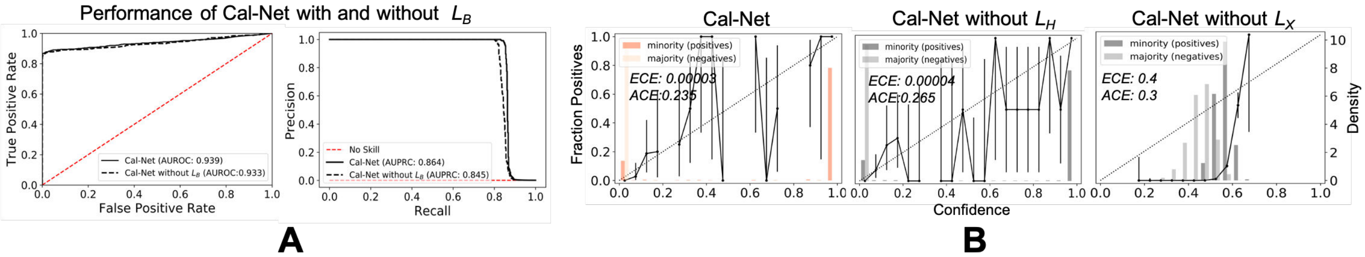

Fig. 5. Ablation analyses using the synthetic dataset with 0.11% (IR: 906.8) minority class. (A): AUROC and AUPRC plots showing classification performance in presence or absence of loss component L_B. (B): Reliability diagrams (bins=20) showing how the presence or absence of loss components L_H and L_X affect the calibration performance.

while making predictions over the entire range from 0 to 1. This finding suggested the potential utility of L_H in improving the calibration performance in the Cal-Net architecture.

3) Performance with and without L_X: We analyzed the reliability plots for the standard Cal-Net architecture and a modified variant without the cross-entropy loss (L_X). Figure 5B shows that the modified variant failed to achieve good calibration performance without L_X. The modified variant did not predict over the entire probability range from zero to one and had several empty bins. The calibration curve also did not closely correspond to the ideal calibration curve and achieved high ECE and ACE scores. This finding highlighted that using the unweighted cross-entropy loss L_X on the primary output Y is necessary to improve classification and calibration performances.

4) Performance with and without L_T: We trained a variant of Cal-Net without the t-test loss (L_T) to analyze its utility. However, empirical analyses indicated that there were no significant drops in classification and calibration performances when compared to the standard Cal-Net architecture. L_T penalizes poor separation of classes and thereby helps in preventing degenerate solutions. In datasets, where the separation between classes is poor, tuning L_T may result in better classification performance. It is plausible that other formulations of this loss function could be effective and is left for future studies.

5) Effects of size of training and validation sets: One plausible reason why the post-processing calibration techniques performed poorly may be attributed to the number of samples in the validation set that was used for calibrating the outputs of the underlying neural network classifier. However, it was observed that adding more samples in the validation set did not improve the calibration performance of the post-processing calibration techniques. Cal-Net variants continued to achieve highest AUROC and F-measure as well as the low ECE and ACE scores while assigning samples in most bins thereby avoiding the pitfalls of degenerate solutions. In cases where there is a shortage of available data-points for training due to data unavailability, it may often be challenging for neural network architectures to generalize well in classification tasks [33]. Furthermore, previous research have shown that neural network classifiers tend to perform poorly depending on the degree of class overlap between the minority and the majority classes and availability of training data [34], [35]. Hence, Cal-Net may be often susceptible to problems commonly faced by neural network architectures. Furthermore, since confidence calibration is a function of confidence and accuracy, poor class separation in class imbalanced datasets may often adversely affect confidence calibration.

V. Conclusion

As neural networks are increasingly being used in critical decision-making scenarios, improving classification and calibration performances in class imbalanced datasets is a challenging problem of high practical interest. In this work, we developed Cal-Net, a neural network architecture to simultaneously learn classification and calibration in class imbalanced datasets. Empirically, we showed that Cal-Net outperforms commonly used post-processing calibration methods and cost-sensitive neural network architectures both in classification and calibration tasks across four synthetic and four real world datasets by achieving higher F-measure, higher AUROC and lower Brier Score among all the methods. While this study only examined datasets using feed-forward neural networks, Cal-Net may be incorporated in complex classification architectures as the final state to handle class imbalance for binary classification tasks. We are optimistic that Cal-Net may address challenges in classification tasks involving complex neural network architectures and imbalanced datasets.

VI. Acknowledgement

This research was supported by the National Library Of Medicine of the National Institutes of Health under award numbers R01LM012222 and R01LM012482 and the National Institute of General Medical Sciences under award number R01GM140635.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

[2] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Y. Ng, and Nigam H. Shah. Improving palliative care with deep learning. *CoRR*, abs/1711.06402, 2017.

[3] Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983.

[4] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

[5] L. Huang, J. Zhao, B. Zhu, H. Chen, and S. V. Broucke. An experimental investigation of calibration techniques for imbalanced data. *IEEE Access*, 8:127343–127352, 2020.

[6] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, Mar 2019.

[7] Charles X. Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 73–79. AAAI Press, 1998.

[8] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, 1997.

[9] Robert C. Holte, Liane E. Acker, and Bruce W. Porter. Concept learning and the problem of small disjuncts. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'89, page 813–818, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[10] Y. Tang, Y. Zhang, N. V. Chawla, and S. Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009.

[11] Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi. When is undersampling effective in unbalanced classification tasks? In Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, Carlos Soares, João Gama, and Alípio Jorge, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 200–215, Cham, 2015. Springer International Publishing.

[12] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.

[13] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284, September 2009.

[14] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, page 155–164, New York, NY, USA, 1999. Association for Computing Machinery.

[15] Charles Elkan. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[16] Kai Ming Ting. Inducing cost-sensitive trees via instance weighting. In Jan M. Żytkow and Mohamed Quafafou, editors, *Principles of Data Mining and Knowledge Discovery*, pages 139–147, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[17] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[18] 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, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery.

[19] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[20] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 04 2015.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *34th International Conference on Machine Learning, ICML 2017*, 3:2130–2143, 2017.

[22] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 9022–9030, 2019.

[23] Johanna Schwarz and Dominik Heider. GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making. *Bioinformatics*, 35(14):2458–2465, 11 2018.

[24] S. Joshua Swamidass, Bradley T. Calhoun, Joshua A. Bittker, Nicole E. Bodycombe, and Paul A. Clemons. Enhancing the rate of scaffold discovery with diversity-oriented prioritization. *Bioinformatics (Oxford, England)*, 27(16):2271–2278, Aug 2011. 21685049[pmid].

[25] Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[26] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA, 2006. Association for Computing Machinery.

[27] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *Machine Learning for Intelligent Transportation Systems Workshop, NIPS*, 2018.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] Warwick J Nash and Tasmania. Marine Research Laboratories. The population biology of abalone (haliotis species) in tasmania. 1, blacklip abalone (h. rubra) from the north coast and the islands of bass strait. 1994. CIP confirmed.

[30] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[31] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning.

[32] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21, 03 2015.

[33] X. Cui, V. Goel, and B. Kingsbury. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1469–1477, 2015.

[34] Sakhawat Hosain Sumit and Shamim Akhter. C-means clustering and deep-neuro-fuzzy classification for road weight measurement in traffic management system. *Soft Comput.*, 23(12):4329–4340, June 2019.

[35] Vicente García, Jose Sánchez, and Ramon Mollineda. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In Luis Rueda, Domingo Mery, and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, pages 397–406, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.