ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

A Two-Stage Hybrid Fraud Detection Framework with Learnable Fusion and Probability Calibration for Credit Card Transactions

[1] Chitra L, [2] Gopika K L, [3]Jasnavi N, [4] Kavya S, [5] Vaijayanthi M
[1] [2] [3] [4] Dept. of Computer Science and Engineering, Er. Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India.
[5] Assistant Professor, Department of Computer Science and Engineering Er. Perumal Manimekalai College of Engineering, Hosur

Abstract: ***Credit card fraud detection is still a major problem in financial systems because of the extreme class imbalance, changing nature of fraud and the high cost of misclassification. Here we propose a novel two- stage hybrid detection framework that combines a supervised XGBoost classifier and an unsupervised Autoencoder-based anomaly detector. The outputs of both stages are fused by a learnable Logistic Regression meta-model that is trained on out-of-fold (OOF) predictions to prevent data leakage. We use Platt scaling to calibrate the fusion output so that the risk quantification is reliable. The calibrated fraud probability is then translated into a three tier risk bucketing scheme — LOW, MEDIUM, and HIGH — providing for auditable, actionable decisions. Our method outperforms the single-model baselines with a test ROC-AUC of 0.9776, PR-AUC of 0.8186, fraud precision of 0.99, recall of 0.72 and an overall accuracy of 99.97% on the publicly available European Credit Card dataset. We further build an operational dashboard with Streamlit for real- time inference and batch analysis, visualization of behavioural patterns and model health monitoring.***

Keywords: Credit card fraud detection, XGBoost, autoencoder, anomaly detection, ensemble learning, SMOTE, probability calibration, risk stratification, out-of-fold stacking.

I. INTRODUCTION

The widespread adoption of digital payment systems around the world has led to a dramatic increase in both the frequency and complexity of credit card fraud. According to the Nilson Report, total global card fraud losses rose above $34 billion in 2023, the highest level in seven years, demonstrating the urgent need for robust, scalable and real-time fraud detection mechanisms [1]. Financial institutions are under pressure to reduce fraud losses, while keeping false positives in check, which erode customer trust and create unnecessary operational overhead.

Automated fraud detection is now dominated by machine learning (ML). Gradient boosting methods, and especially XGBoost, have been successful on structured tabular transaction data, due to their ability to model complex non-linear relationships, deal with missing values, and incorporate class weighting. Supervised methods, however, are inherently limited by the quality and coverage of the labeled training data. Fraud patterns are constantly evolving and purely supervised models may not generalize to novel attack strategies that are not represented in the historical training sets.

Unsupervised anomaly detection provides a complementary view. Autoencoders are neural networks trained to reproduce their own inputs. They have been successfully applied to fraud detection by learning a compact representation of legitimate transaction behaviour. Anomaly score is computed as the high reconstruction error of the transactions deviating from this learned normality. This approach does not require any fraud labels and in principle is able to detect previously unseen types of fraud. Despite the promise of both paradigms, combining supervised and unsupervised signals in a principled, leak-free manner remains a methodological challenge.

Naive score-level fusion, which is just averaging or putting together outputs, does not learn the best combination and could leak data when the meta- model is trained on in-sample predictions. Furthermore, the raw outputs of ensemble models seldom represent well-calibrated probabilities, which constrains their effectiveness in risk-based decision frameworks that depend on significant probability thresholds.

This paper fills in these gaps with a single four- stage framework: (1) XGBoost with SMOTE-based oversampling for supervised classification; (2) Autoencoder anomaly detection trained only on legitimate transactions; (3) out-of-fold learnable fusion via Logistic Regression to stop leakage; and (4) Platt-scaling calibration to give reliable posterior fraud probabilities. The calibrated output is put into a three-tier risk bucketing scheme (LOW/MEDIUM/HIGH) to help with operational decision-making. A Streamlit dashboard that users can fully interact with is being made for deployment. The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the dataset and preprocessing. Section IV presents the proposed methodology. Section V reports experimental results. Section VI discusses findings and limitations. Section VII concludes the paper.

II. RELATED WORK

2.1 Supervised Methods for Fraud Detection

Logistic Regression, Decision Trees, and Support Vector Machines were some of the first ML methods used to find fraud. The emergence of ensemble boosting techniques, especially XGBoost, markedly enhanced the field's advancement. Alarfaj et al. [2] did a thorough comparison of supervised ML and deep learning algorithms on the European credit card dataset. They showed that XGBoost always did better than single classifiers when it came to ROC-AUC and precision. Ileberi et al. [3] examined six ML classifiers utilizing SMOTE oversampling, determining that XGBoost attained the optimal equilibrium of precision and recall in scenarios of class imbalance. El Kafhali et al. [4] put forward an improved deep learning method that uses dense layers and XGBoost as a stacking meta-learner.

2.2 Anomaly Detection with Autoencoders

Autoencoders that use reconstruction to find anomalies have been used a lot in fraud and intrusion detection. The fundamental principle — that a model trained exclusively on normal data will demonstrate increased reconstruction error when presented with anomalous inputs — was initially formulated in early research on network intrusion detection and subsequently expanded to financial fraud by Fanai and Abbasimehr [5], who introduced a hybrid deep autoencoder and deep classifier architecture.

Prajesha and Veni [6] specifically combined probabilistic XGBoost thresholding with autoencoder reconstruction error, establishing a direct precedent for the fusion approach adopted in this work. Extensions based on variational autoencoders have shown that they can find new fraud patterns by looking at latent space [7].

2.3 Ensemble and Stacking Approaches

Many writers have looked into using stacking ensemble methods to find fraud. Zhang et al. [8] used LSTM and GRU networks as base learners and an MLP meta-learner in a stacking framework. Li et al.
[9] presented a hybrid deep learning ensemble that combines CNN, LSTM, and Transformers with XGBoost as a meta-learner, achieving a ROC-AUC of 0.972 on the European dataset. Importantly, these studies fail to tackle the issue of data leakage in meta- model training, which occurs when base learner outputs from in-sample data are utilized to train the fusion layer — a problem that this study explicitly resolves through OOF stacking.

2.4 Class Imbalance Mitigation

Fraud datasets usually have a fraud rate of less than 0.2%, which means that they need to be handled in a special way. SMOTE [10] is still the most popular way to oversample. Chawla et al. [10] showed that SMOTE always works better than random oversampling. One important but often ignored problem is where to put SMOTE in cross-validation pipelines. If you use SMOTE before splitting the data, it lets information leak from the validation folds to the training folds, which makes performance metrics look better than they really are. Our implementation solves this by putting SMOTE inside an imbalanced-learn Pipeline, which makes sure it only works on training folds.

2.5 Probability Calibration

classifiers trained on imbalanced datasets that are either undersampled or oversampled give class probability estimates that are always wrong. They demonstrated that Platt scaling effectively mitigates this bias, especially when SMOTE or under sampling techniques have been employed. Even though this is the case, most fraud detection research does not use explicit calibration before making decisions based on thresholds, which could make reported confidence scores less reliable. Our framework directly addresses this deficiency.

III. DATASET AND PREPROCESSING

3.1 Dataset Description

The experiments utilize the publicly accessible European Credit Card Fraud Detection dataset [12], comprising 284,807 transactions executed by European cardholders over two days in September 2013. The dataset has a very uneven class balance, with 492 fraudulent transactions (0.172% of all transactions). Features V1 through V28 are principal components derived from a PCA transformation applied to protect cardholder privacy. The raw features Time and Amount are available in their original form.

Table I. Dataset Statistics

Property	Value	Notes
Total Transactions	284,807	After dedup
Legitimate (Class 0)	284,315	99.83%
Fraudulent (Class 1)	492	0.17%
PCA Features	V1 – V28	Anonymized
Time Feature	Seconds elapsed	Engineered
Amount Feature	Transaction EUR	Log-transformed
Train / Test Split	80% / 20%	Stratified

3.2 Feature Engineering

The raw features Time and Amount are changed to make the model work better. To make the right- skewed distribution of consumer spending smaller, the transaction amount is log-transformed (Amount_log = log(1 + Amount)). The Time feature, which shows how many seconds have passed since the first transaction, is broken down into understandable time signals: hour of day (hour = ⌊ Time / 3600⌋ mod 24), calendar day (day = ⌊ Time / 86400⌋), and day of week (weekday = day mod 7). These time-based features show patterns of fraud that happen every day and every week that are important in the study of

financial fraud. The original Time and Amount columns are discarded after engineering.

The StandardScaler is only used on the engineered features {Amount_log, hour, day, weekday} in a ColumnTransformer. The V1–V28 features stay the same because the upstream PCA transformation has already standardized them.

IV. PROPOSED METHODOLOGY

The suggested framework has four steps that happen one after the other, as shown in Fig. 1. Each stage of high-stakes fraud detection deals with a different problem: supervised pattern learning, unsupervised anomaly detection, leakage-free fusion, and probability calibration for accurate risk quantification.

Fig. 1. System Architecture of the Proposed Two- Stage Fraud Detection Framework

STAGE 1: XGBOOST + SMOTE → P(FRAUD|X) → STAGE 2: AUTOENCODER → E(RECON ERROR) → STAGE 3: OOF FUSION (LOGISTIC REG.) → STAGE 4: PLATT SCALING CALIBRATION → RISK BUCKET: LOW / MED / HIGH

Stage 1: XGBoost Supervised Classifier

XGBoost [13] is the main supervised classifier used in the first stage. XGBoost is trained using an imbalanced-learn Pipeline that includes a Column Transformer preprocessor and SMOTE oversampling. This makes sure that oversampling is only done in training folds during cross-validation to avoid data leakage. The structure of the pipeline is:

Pipeline: [ColumnTransformer → SMOTE → XGBClassifier]

To make up for the big difference in the number of positive and negative training samples, the scale_pos_weight parameter is set to the ratio of negative to positive samples (scale_pos_weight = N_neg / N_pos ≈ 578).

This makes the cost sensitive to each class. To find the right balance between model complexity and generalization, we set the XGBoost hyperparameters as follows: n_estimators = 700, max_depth = 5, learning_rate = 0.03, subsample =

0.8, colsample_bytree = 0.7, min_child_weight = 5, $\gamma = 0.5$, $\alpha = 2$ (L1 regularization), and $\lambda = 6$ (L2 regularization). The evaluation metric is Area Under the Precision-Recall Curve (AUCPR), which is better than ROC-AUC when there is a lot of class imbalance. 5-fold Stratified K-Fold cross-validation is used to choose the best model. This keeps the class distribution the same in each fold.

Stage 2: Autoencoder Anomaly Detector

The second stage uses a feedforward Autoencoder neural network that was only trained on real (Class 0) training transactions. The model architecture follows a design with a symmetric bottleneck:

Encoder: Input(31) → Dense(64, ReLU) → Dense(32, ReLU)

Decoder: Dense(32) → Dense(64, ReLU) → Dense(31)

Using the Adam optimizer (lr = 0.001), the network is trained for 100 epochs with a batch size of 256 and Mean Squared Error (MSE) loss. A different StandardScaler is only fitted on real training transactions and used on all inputs before the autoencoder processes them. The autoencoder learns a compact manifold of legitimate transaction behavior by only training on the normal class. The reconstruction error $\varepsilon(x)$ for any test transaction x is calculated as:

$$\varepsilon(x) = (1/d) \times \Sigma(x_i - \hat{x}_i)^2, \text{ where } \hat{x} = Decoder(Encoder(x))$$

Fraudulent transactions that show feature patterns that do not fit with the learned normal manifold usually cause reconstruction errors to be much higher than normal. The 95th percentile of reconstruction errors on legitimate training data ($T_{95} \approx 0.0052$) is used as a reference point for scoring anomalies.

Stage 3: Out-of-Fold Learnable Fusion

One major challenge with stacked ensembles is that there can be data leakage at the meta-level. When the meta-model is trained on the base learner predictions from training data that was also used to train the base learners, those predictions are too good to be true because the base learners have already seen those instances during their training phase. We can resolve this issue by using Out-of-Fold (OOF) predictions as a method of stacking. By using the same 5-fold Stratified K-Fold splits, we train both the pipeline of XGBoost and the Autoencoder on 4- folds and then generate predictions from the model for the unseen 5th-fold. We repeat this for each of the

folds so we can obtain true out-of-sample predictions for all instances in the training set:

$$X\_fusion\_train = [\hat{p}\_XGB\_OOF(i), \hat{\varepsilon}\_AE\_OOF(i)]$$

$$\forall\ i \in \{1,...,N\_train\}$$

Next, we will build a Logistic Regression meta- model using the balanced class weights and training set predictions, X_fusion_train, with the target labels y_train. This will provide us the ability to learn through linear combinations of supervised probability and anomaly scores that best discriminates actual fraud versus actual non-fraud. The coefficients produced through the learning will be interpretable — whereby if the $\hat{p}$_XGB coefficient is large and positive, then we know that the supervised signal is dominant; conversely, if the coefficient $\hat{\varepsilon}$_AE is large, then we can interpret that the anomaly signal is significantly contributing to the result.

Stage 4: Probability Calibration

Probabilistic outputs provided by most classifiers (especially those trained using either oversampling or undersampling) will be miscalibrated. A well calibrated classifier is one where the predicted probability $P(y=1|\hat{y}=p)$ comes close to being equal to p. So for example, with an expected probability of 0.7 fraud on all transactions, approximately 70% of the transactions with expected fraud probability of 0.7 will turn out to be frauds. The importance of having this type of calibrated classification model is that it allows for meaningful classification under and independent of the way that operational risks are categorized, e.g., as categories of HIGH and MEDIUM.

To perform this type of well calibrated classification, we apply Platt scaling [14], specifically we employ CalibratedClassifierCV with the method='sigmoid' and cv=3. Platt scaling is accomplished by estimating the parameters A and B of the logistic function $\sigma(f(x)) = 1/(1 + e^{-Af(x) + B})$, where A and B are learned parameters and f(x) is the raw score output by the fusion classifier.

Risk Bucketing and Decision Logic

The calibrated probability $P\_calibrated \in [0, 1]$ is mapped to a three-tier operational risk level as shown in Table II.

Table II. Risk Bucketing Scheme

Risk Level	Probabilit y Range	Decisio n	Operationa l Action
LOW	0.00 – 0.20	Approve	Auto-pass
MEDIU M	0.20 – 0.60	Review	Manual analyst
HIGH	0.60 – 1.00	Block	Immediate investigatio n

The decision threshold that delineates MEDIUM and HIGH risks (0.60) was further optimized using precision-recall optimization. The value of the maximum precision obtained with a minimum recall constraint of 0.70 was used as the decision threshold. This threshold was determined to be a value of 0.298, yielding a precision result of 0.986 and a recall result of 0.716 when measured on the test dataset.

V. EXPERIMENTAL RESULTS

Experimental Setup
Using Python v3.10 along with libraries (scikit-learn v1.3, XGBoost v2.0, PyTorch v2.1, and imbalanced- learn v0.11) to run all experiments; training takes place using a CPU-based system. Datasets contain stratified 80/20 splits and use 5-fold stratified K-fold cross validation for model selection and out of fold (OOF) generation. All hyperparameters were fixed as specified in Section IV; they were not tuned any further on the test dataset to eliminate optimistic biases.

Cross-Validation Performance
Table III reports the 5-fold cross-validation results for the XGBoost stage. The mean CV ROC-AUC of 0.9816 with low variance ($\sigma = 0.003$) indicates strong generalization and absence of overfitting.Table III. 5-Fold Cross-Validation Results (XGBoost Stage)

Fold	ROC- AUC	PR- AUC	F1	Recall
Fold 1	0.9831	0.8019	0.82	0.74
Fold 2	0.9808	0.7944	0.81	0.73
Fold 3	0.9820	0.8103	0.84	0.77
Fold 4	0.9793	0.8067	0.81	0.72
Fold 5	0.9830	0.8068	0.83	0.75
Mean ± σ	0.9816 ± 0.003	0.8040 ± 0.006	0.822 ± 0.011	0.742 ± 0.018

Test Set Performance
Table IV presents the test set classification report and key metrics for the final calibrated fusion model.

Table IV. Test Set Classification Report – Calibrated Fusion Model

Class	Precis ion	Rec all	F1-Sco re	Supp ort	ROC- AUC
Legiti mate (0)	1.00	1.00	1.00	56,651	—
Fraudulent (1)	0.99	0.72	0.83	95	0.9776
Macro Avg	0.99	0.86	0.91	56,746	—
Overall Accuracy: 99.97%			PR-AUC: 0.8186		Thresh old: 0.698

Comparative Analysis

Table V compares the proposed methodology to baseline and existing methods using the same datasets. Included are both single stage supervised baseline methods (XGBoost only and Multi-Layer Perceptron) and current ensemble methodologies representative from the literature.

The proposed methodology produces the best precision value of any of the methods (0.99), maintains a competitive recall value, and provides the highest PR-AUC value (the most useful metric since there is severe class imbalance).

Table V. Comparative Performance on European Credit Card Dataset

Method	Precis ion	Rec all	F1	RO C- AU C	PR - AU C
Logistic Regression (baseline)	0.73	0.58	0.64	0.921	0.68
XGBoost Only [2]	0.57	0.81	0.67	0.977	0.71
SMOTE + RF [3]	0.87	0.75	0.80	0.967	0.74
LSTM + GRU Stacking [8]	0.91	0.74	0.82	0.975	0.79
CNN+LSTM+ XGBoost [9]	0.93	0.71	0.80	0.972	0.80
Proposed Framework	**0.99** ✓	0.72	0.83	**0.978** ✓	**0.819** ✓

Risk Bucketing Analysis

Table VI presents the distribution of test transactions across risk buckets and the empirical fraud rate within each bucket, validating the calibration quality.

Table VI. Risk Bucket Distribution and Empirical Fraud Rates

Risk Level	Transaction Count	% of Total	Fraud Rate (%)
LOW	56,608	99.76%	0.005%
MEDIUM	70	0.12%	3.57%
HIGH	68	0.12%	97.06%

In the HIGH risk bucket, there is a fraud rate of 97.06%, which means almost all transactions flagged for review require an immediate investigation will be legitimate. This level of precision is of great operational importance for fraud analysts as it provides them with the ability to quickly identify fraudulent transactions in the HIGH risk category; thus, greatly reducing their false positive workload. There are also transactions flagged in the MEDIUM category that are minimum borderline fraud candidates that require human interface to determine if the transaction should actually be reviewed. Transactions identified in LOW are not candidates for fraud.

VI. DISCUSSION

Complementarity of Supervised and Unsupervised Signals

The synergy between XGBoost's supervised probability and Autoencoder's reconstruction error demonstrates how they complement each other in two stages. Transactions flagged as high probability through XGBoost and high reconstruction error on the Autoencoder indicate the clearest examples of fraud. On the other hand, cases with only one signal being elevated can also provide useful information. A transaction that has a high probability score through XGBoost but low reconstruction error may be a fraudulent transaction that has similar patterns to previously successful fraudulent transactions but continues to follow a normal legitimate pattern. Conversely, if a transaction has a low XGBoost probability and high reconstruction error, the transaction may indicate a new form of fraudulent activity that was not previously accounted for during the training phase. The learnable fusion model has been specifically built to learn how to intelligently combine these two types of complementary signals.

Importance of OOF Training

When using the OOF stacking methodology, it adds great credibility to the fusion model. Although we tested the naive version of the fusion model's OOF stacking method, it showed approximately a 4-6% increase in the precision-recall curve (PR-AUC) when producing in-sample predictions with the XGBoost model. The results of the naive version of the fusion meta model do not hold water after using out-of-sample (OOF) predictions, demonstrating how the naive stacking method produces a measurable degree of data leakage which reduces the validity of the evaluation metric results.

Effect of Probability Calibration

The impact of calibration is significant on the operational validity of the risk bucketing scheme. The raw fusion model would provide a higher estimated probability for some test transaction due to the training set being modified via SMOTE. This leads to much larger instances in a MEDIUM bucket and less purity in a HIGH bucket. When the calibration is applied, the fraud rate in the HIGH bucket increases from approximately 84% to 97.06%, indicating that calibration greatly enhances the actionability of risk classifications.

Limitations and Future Work

While there are many strengths of the current framework, there are also several limitations that should be noted. To begin with, the autoencoder is trained from scratch in each out of fold (OOF) fold and is limited to 20 epochs to reduce compute time and may lead to an anomaly detector that has converged less optimally than previously desired. Future work should investigate the possibility of training the final autoencoder for the full 100 epochs and with the full training set prior to deployment. Second, the current framework does not yet account for temporal sequence modelling, which would provide the ability to recognise behavioural drift over the course of an individual cardholder's transaction history. Third, even though the Streamlit dashboard allows for real-time inference, this implementation does not currently contain a model drift monitoring

feature to trigger retraining based on observations of significant redistribution of scores. Fourth, the framework has evaluated against only one benchmark dataset; validating against real-world transaction datasets that are proprietary and include their true distributions would increase the validity of the generalisation.

VII. CONCLUSION

This research implements a four-stage hybrid credit card fraud detection framework that delivers systematic solutions for the four major issues associated with supervised learning including evolving patterns, anomaly detection with respect to unlabelled data, leakage of meta-model data, and miscalibration of probabilities. By utilizing the discriminative power of XGBoost in conjunction with the anomaly sensitivity of an autoencoder through the use of leakage-free out-of-fold stacking and Platt- scaling calibration, the framework achieved a test ROC-AUC of 0.9776, PR-AUC of 0.8186, fraud precision of 0.99 and a high bucket fraud rate of 97.06% using the European credit card dataset as a test case.

The three-tier risk bucketing scheme (LOW, MEDIUM or HIGH) establishes an effective bridge between the probabilistic model results and actual operational fraud management decision making (allowing for automated approval, analyst review and immediate investigation type workflows). The Streamlit dashboard provided enables operationalization in real time for single transaction analysis, batch processing, behavioural pattern exploration and model health monitoring.

The future direction of the research will encompass time series modelling, online learning for adapting to concept drift, integration with graph- based transaction networks and validation across a broad range of real world data sets. In addition, the modular architecture of the framework allows for simple substitution or augmentation of any stage of the framework as more effective models become available.

REFERENCES

Nilson Report, "Global Card Fraud Losses," HSN Consultants Inc., Carpinteria, CA, Issue 1296, 2024.

F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of-the- art machine learning and deep learning algorithms," IEEE Access, vol. 10, pp. 39700– 39715, 2022.

E. Ileberi, Y. Sun, and Z. Wang, "Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost," IEEE Access, vol. 9, pp. 165286– 165294, 2021.

S. El Kafhali, M. Tayebi, and H. Sulimani, "An optimized deep learning approach for detecting fraudulent transactions," Information, vol. 15, no. 4, p. 227, 2024.

H. Fanai and H. Abbasimehr, "A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection," Expert Systems with Applications, vol. 217, p. 119562, 2023.

T. M. Prajesha and S. Veni, "Probabilistic XGBoost threshold classification with autoencoder for credit card fraud detection," IJRITCC, vol. 11, pp. 528–537, 2023.

M. Adamantis, V. Sokolov, and P. Skladannyi, "Variational autoencoders for detecting anomalous and fraudulent financial transactions," in Proc. xAI 2024 Late- Breaking Work, 2024.

T. Zhang, et al., "A deep learning ensemble with data resampling for credit card fraud detection," IEEE Access, vol. 11, pp. 22561– 22573, 2023.

Y. Li, et al., "A hybrid deep learning ensemble model for credit card fraud detection," IEEE Access, vol. 12, pp. 158901–158915, 2024.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W.

P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

A. Dal Pozzolo, O. Caelen, R. A. Johnson, and

G. Bontempi, "Calibrating probability with under sampling for unbalanced classification," in Proc. IEEE SSCI, Cape Town, 2015, pp. 159–166.A. Dal Pozzolo, O. Caelen, Y. A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," Expert Systems with Applications, vol. 41, no. 10, pp. 4915–4928, 2014.

T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, New York, 2016, pp. 785–794.

J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.

G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," J. Mach. Learn. Res., vol. 18, pp. 1– 5, 2017.

D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, San Diego, CA, 2015.