Mortality Prediction of Patients with Cardiovascular Disease Using Medical Claims Data under Artificial Intelligence Architectures: Validation Study

Linh Tran, Lianhua Chi, Alessio Bonti, Mohamed Abdelrazek, Yi-Ping Phoebe Chen

Submitted to: JMIR Medical Informatics
on: October 14, 2020

Table of Contents

Mortality Prediction of Patients with Cardiovascular Disease Using Medical Claims Data under Artificial Intelligence Architectures: Validation Study

Linh Tran[1] MSc; Lianhua Chi[2] PhD; Alessio Bonti[1] PhD; Mohamed Abdelrazek[1] PhD; Yi-Ping Phoebe Chen[2] PhD

[1]School of Info Technology Deakin University Burwood AU
[2]Department of Computer Science and Information Technology La Trobe University Bundoora AU

Corresponding Author:
Lianhua Chi PhD
Department of Computer Science and Information Technology
La Trobe University
Thomas Cherry Bldg, 3rd Fl
La Trobe University
Bundoora
AU

Abstract

Background: As stated by WHO, Cardiovascular disease (CVDs) are the number 1 cause of death globally, which means more people die annually from CVDs than from any other cause. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke. In this study, we present a benchmark comparison of various Artificial Intelligence (AI) architectures on predicting mortality of CVD patients using the structured medical claims data.

Objective: This study mainly aims to support health clinicians to accurately predict mortality among patients with CVD using only claims data before a clinic visit.

Methods: The used dataset was joined from Medical Benefits Scheme (MBS) and Pharmaceutical Benefits Scheme (PBS) service information in the period between 2004 and 2014, released by the Department of Health Australia in 2016. It includes 346,201 records corresponding to 346,201 patients. A total of five AI algorithms including four classical Machine Learning (ML) algorithms (Logistic Regression (LR), Random Forest (RF), Extra Trees (ET) and Gradient Boosting Trees (GBT)) and a deep learning algorithm which is a densely connected neural network (DNN) were developed and compared in the study. In addition, due to the minority of 'deceased' patients in the data set, a separate experiment using Synthetic Minority Oversampling Technique (SMOTE) was conducted to enrich the data.

Results: Regarding model performance, in terms of discrimination, GBT and RF are the models with highest AUROC (97.8% and 97.7% respectively), followed by ET (96.8%) and LG (96.4%) while DNN is the least discriminative (95.3%). In terms of reliability, LG predictions are the least calibrated compared to those of four algorithms. In this study, despite increasing training time, SMOTE is proved to further improve model performance of LG while other algorithms, especially GBT and DNN, work well with class imbalanced data.

Conclusions: Compared to other research in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient since we only utilize a smaller number of features but still achieve high performance. And this study could support health professionals to accurately choose AI models to predict mortality among patients with CVD using only claims data before a clinic visit.

(JMIR Preprints 14/10/2020:25000)
DOI: https://doi.org/10.2196/preprints.25000

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Mortality Prediction of Patients with Cardiovascular Disease Using Medical Claims Data under Artificial Intelligence Architectures: Validation Study

Abstract

Background: Cardiovascular disease (CVD) is Australia's greatest health problem, kills more people than any other disease and creates enormous costs for the health care system. In this study, we present a benchmark comparison of various Artificial Intelligence (AI) architectures on predicting mortality of CVD patients using the structured medical claims data. Compared to other research in the clinical literature, our models are more efficient since we utilize a smaller number of features and this study could support health professionals to accurately choose AI models to predict mortality among patients with CVD using only claims data before a clinic visit.
Objective: This study mainly aims to support health clinicians to accurately predict mortality among patients with CVD using only claims data before a clinic visit.
Methods: The used dataset was joined from Medical Benefits Scheme (MBS) and Pharmaceutical Benefits Scheme (PBS) service information in the period between 2004 and 2014, released by the Department of Health Australia in 2016. It includes 346,201 records corresponding to 346,201 patients. A total of five AI algorithms including four classical Machine Learning (ML) algorithms (Logistic Regression (LR), Random Forest (RF), Extra Trees (ET) and Gradient Boosting Trees (GBT)) and a deep learning algorithm which is a densely connected neural network (DNN) were developed and compared in the study. In addition, due to the minority of "deceased'" patients in the data set, a separate experiment using Synthetic Minority Oversampling Technique (SMOTE) was conducted to enrich the data.
Results: Regarding model performance, in terms of discrimination, GBT and RF are the models with highest AUROC (97.8% and 97.7% respectively), followed by ET (96.8%) and LG (96.4%) while DNN is the least discriminative (95.3%). In terms of reliability, LG predictions are the least calibrated compared to those of four algorithms. In this study, despite increasing training time, SMOTE is proved to further improve model performance of LG while other algorithms, especially GBT and DNN, work well with class imbalanced data.
Conclusions: Compared to other research in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient since we utilize a smaller number of features but still achieve high performance. And this study could support health professionals to accurately choose AI models to predict mortality among patients with CVD using only claims data before a clinic visit.

Keywords: mortality prediction; cardiovascular; medical claims data; imbalanced data; machine learning; deep learning.

Introduction

Background

In Australia, Cardiovascular Disease (CVD) is the most concerning health problem, killing more people than any other disease and placing heavy burdens for the health care system due to enormous costs and individuals and the community due to resulting disabilities. CVD was the leading cause of death among Australians in 1997, accounting for 52,641 deaths, 41% of all deaths [1]. An estimated 1.2 million (5.6%) Australian adults aged 18 years and over had one or more conditions related to

heart or vascular disease, including stroke, in 2017–18, based on self-reported data from the Australian Bureau of Statistics (ABS) 2017–18 National Health Survey. The prevalence of CVD by Age group and Sex, 2017-2018 is shown in Figure 1.

The major risk factors for CVD are tobacco smoking, high blood pressure, high blood cholesterol, overweight, insufficient physical activity, high alcohol use and type 2 diabetes [1]. CVD treatments are usually prescribed in combination, with the prevalence of use of drugs such as anti-diabetics, anti-hypertensives, lipid-lowering drugs, anticoagulants and anti-platelet agents [2]. Besides eating a healthy diet and maintaining fitness with regular physical activity, medication use is an important management factor for patients diagnosed with heart disease conditions. Medications are used to minimize symptoms, reduce the risk of exacerbation and improve quality of life.

Many methods have been approached to predict the mortality of patients with CVD, utilizing many algorithms and predictor variables. There are three main methods of mortality forecasting: explanation, expectation and extrapolation [3]. Out of these, the most common basis of mortality forecasting is extrapolation which assumes future state highly correlates to the past. In the clinical literature, historical EHR are widely used to develop AI models that can predict health outcomes of patients contracting with a disease. Information commonly extracted from EHR as input for AI models includes patient demographics, health indexes, medical conditions, biomedical images or clinical notes while structured medical claims data are rarely utilized. Even though medical claims data little inform patient health conditions, this source of information is crucial in reflecting patient healthcare access frequency and level of participation in disease prevention/treatment, holding great impact in determining patient health outcomes.

In this study, we present a benchmark comparison of the performance of different AI architectures such as four classical Machine Learning (ML) algorithms (Logistic Regression (LR), Random Forest (RF), Extra Trees (ET) and Gradient Boosting Trees (GBT)) and a deep learning algorithm which is a densely connected neural network (DNN) in using medical scheduling and pharmaceutical dispensing information from historical claims data to predict mortality of patients with CVD. Compared to other researches in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient since we utilize a smaller number of features but still achieve high performance. Furthermore, we also propose SMOTE, a technique to enrich training data and handle class imbalance, as an approach to improve the performance of developed AI models.

Related Work

Recent day trends involve using artificial intelligence models to learn patterns from large datasets in order to predict mortality with higher accuracy [4]. The American College of Cardiology Foundation’s National Cardiovascular Data Entry (NCDR) conducted a study that used statistical analysis to predict rate of risk in per-cutaneous coronary intervention. The study results show that ML models performed better in terms of accuracy than classical statistical models [5]. One conducted study shows that machine learning models such as random forest, decision tree and logistic regression perform exceptionally well due to today’s computational capacity which allows them to process data from the electrical health records [6] of patients. Deploying machine learning models on routine clinical data performed better than standard cardiovascular risk assessment models and found great merits in terms of preventive treatment and avoidance of mistreatment for cardiovascular disease according to a study conducted on a large sample of patients in the UK [7]. Moreover, using neural networks for predictive analysis of illnesses have been shown to be fruitful as early as 2005 [8]. Z. Wang et al. predicted the mortality rate due to heart failure by deploying a convolutional layered neural network which inculcated feature rearrangement to select the best

features [9]. In another instance, studies have shown that deep neural networks performed better than traditional machine learning models with respect to accuracy and available sample size [10].

Many factors have been considered to predict health outcomes of patients suffering from heart disease. Some techniques used to extract learning features are automated imaging interpretation [11-12], natural language processing or text mining [13-14] and electronic health records extraction [15-18]. Imaging interpretation has been carried out by deep neural networks [12] with promising results. Natural language processing of clinical notes has been able to correctly identify risks of cardiovascular disease patients [13] while systematic application of text mining to the EHR has had variable success for the detection of cardiovascular phenotype [14]. It is proven that applying ML helps finds clinically relevant patterns in the data [19]. Features extraction from EHR allows the utility of many factors such as patient demographics, characteristics and health conditions including cardiovascular health (CVH) indexes [20] or per-cutaneous coronary interventions (PCI) indexes [16-17] in predicting mortality risks.

Based on these studies in the literature, it can be reviewed that mortality rate of patients in cardiology cohort has been accurately predicted using a variety of algorithms, methods and predictor features. However, there has been little focus on utilizing medical claims for predicting health outcomes of patients with CVD. This information reflects patient medication usage as well as health care access frequency and level of participation in disease prevention/treatment which hold great impact in determining patient health outcomes [21]. Hence, in order to contribute to closing this literature gap, in this paper, mortality will be predicted based on patient medical schedule information and pharmaceutical dispensing history acquired from medical claims.

The Pharmaceutical Benefits Scheme (PBS) and Medicare Benefits Schedule (MBS) claims data collected by the Department of Human Services and held by the Department of Health has great potential to provide further insight into medical scheduling and pharmaceutical dispensing history for patients with CVD. This study utilizes the PBS and MBS claims data in the period between 2004 and 2014 in investigating the mortality rate of patients having heart disease conditions in Australia and building and comparing five AI models to predict the mortality risk of a patient being in these conditions. We built the prediction models based on patient's age, gender, relevant medication prescriptions, medical schedule information and pharmaceutical dispensing history obtained from the dataset. We then assessed and compared the performance of each model and suggested recommendations for future work.

Objectives

The primary aim of this research is to support health clinicians to accurately predict mortality among patients with CVD using only claims data before a clinic visit. Compared to other researches in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient since we utilize a smaller number of features but still achieve high performance. This study has applications to supporting health clinicians to accurately predict mortality among patients with CVD using only claims data before a clinic visit.

Methods: Artificial Intelligence Architectures

In this study, four classical machine learning algorithm architectures which are logistic regression (LR), random forest (RF), extra trees (ET) and gradient boosting trees (GBT) along with a deep learning algorithm called densely connected neural network (DNN) are used to develop mortality prediction models. The MBS & PBS dataset is a well-structured and very informative one which allows simple algorithms to better learn. Since our study is a probabilistic prediction problem, we put

more emphasis on the discrimination and calibration of model performance. Through initial experiments, we found that LR, RF, ET and GBT are classical machine learning algorithms that produce best performance in terms of these two criteria. On the other hand, we were also curious on how the state-of-the-art deep learning might perform on the dataset, we developed the simplest neural network which is a densely connected neural network for further comparison and insights. The reason why we do not choose to develop more complex deep learning architectures such as RNN or CNN is because these algorithms are not necessary for such structured dataset to highly perform. In this section, these experimental algorithms are described and their architectures are proposed.

Logistic Regression (LR)

LR is a supervised machine learning algorithm. It is a powerful and well-established method for binary classification problem [24]. LR is extended based on linear regression and can be used to calculate the probability of an event that has two possible outcomes by assigning weights to a number of predictor variables (features). Let's say, given a set of independent variables

$$x_1, x_2, x_3, \dots, x_n \quad (1)$$

and a dependent variable y which takes values between 0 and 1. First, LR is designed to find a set of weights

$$b_1, b_2, b_3, \dots, b_n \quad (2)$$

for each of the independent variables so that the following linear equation is able to output a $logit$ score:

$$logit = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \quad (3)$$

Then from this $logit$ score, probability y is derived by the following formula:

$$y = p = \frac{1}{1+e^{-logit}} = \frac{e^{logit}}{1+e^{logit}} \quad (4)$$

To use the LR as a binary classifier, a threshold needs to be assigned to differentiate two classes. Normally, LR will classify an input instance with a probability value higher than 0.50 as positive class; otherwise, negative class. Depending on the problem, 0 and 1 can be translated into different meanings.

Random Forest (RF)

Before going to the description of RF algorithm, it is important to understand the concept of decision tree algorithm [25]. DT is one of the simplest and earliest machine learning algorithms. It structures the decision logic into a tree-like model. The nodes in a DT tree are partitioned into different levels where the up-most node is called the root node while other nodes which have at least one child represent tests on input variables/features [26]. Depending on some criterion of the test, higher nodes are split into lower nodes repeatedly towards the leaf nodes [27] which have no child at all and correspond to the decision outcomes. An illustration of a simple DT is depicted in Figure 2. According to Figure 2, three circles Sex, Age and $A10$ are tests on corresponding input variables while the rhombuses at the end are the classification outcomes ('deceased' or 'alive').

A random forest (RF) is an ensemble classifier consisting of many DTs similar to the way a forest has many trees [28]. Different DTs in an RF are trained using different parts of the training dataset and tested on different subsets of input variables. To classify a new instance, the input vector of the instance is pushed through each of DTs in the forest. Each DT makes decisions on a different part of that input vector and gives a classification outcome. The forest then makes final prediction by majority vote in classification problems and arithmetic average in regression problems. Since the RF algorithm aggregates outcomes from many different DTs to make decision, the result has smaller variance compared to the consideration of a single DT for the same dataset. In addition, similar to other tree-based ensemble, variables for each tree in RF is randomized while node splitting cut-points are locally optimized according to the criterion [28]. Figure 3 shows an illustration of the RF algorithm. According to Figure 3, the training dataset is randomly split into the desired number of trees in the forest; next, each random sub-sample is used to train a decision tree that is tested on a randomly selected subset of input variables.

Extra Trees (ET)

The Extremely randomized trees or Extra-Trees algorithm is also an ensemble classifier consisting of many single DTs similar to RF. ET method also uses a random subset of features to train each base estimator [29]. However, its two main differences from RF and other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random (or random selection of threshold) and it uses the whole learning sample to grow each tree in the ensemble rather than subset of training data [30]. The final prediction produced is the aggregated predictions of all trained trees, yielded by majority vote and arithmetic average in classification problems and in regression problems respectively. In terms of bias-variance, ET is able to reduce variance more effectively than the weaker randomization schemes used by other ensemble methods. On the other hand, full training sample rather than bootstrap batches is used to train each base estimator in an attempt to minimize bias [30]. A simple illustration of an ET model is depicted in Figure 4.

Gradient Boosting Trees (GBT)

GBT is another popular ML algorithm using tree-based ensemble method, first proposed by Friedman [31]. This approach trains learners (decision trees) based upon minimizing the loss function which is computed by gradient descent method [32]. To train a GBT, the algorithm first builds a very simple decision tree from the learning sample with equal weights. Based on results of this weak learner, it tries to create a new learner that gives higher weights to nodes that are more difficult to split and lower weights to those that are easier to split [32]. By doing this, the new learner is able to minimize the errors of the previous leaner. While this process continues, the loss function is optimized [31] making each new model has better goodness of fit to the observation data. Figure 5 illustrates the mechanism of GBT algorithm.

Densely connected neural network (DNN)

Artificial neural network (ANN) [3] is a deep learning architecture that replicates the neuron system inside human brain. McCulloch and Pitts first proposed ANN in 1943 [33] and the concept was later popularized by the research work of Rumelhart et al. in the 1980s [34]. In human brain, neurons are linked together by numerous axon connections [35] and are responsible for adapting, processing and storing information towards (inputs) and away (outputs) from the brain. Likewise, an ANN has hundreds or even thousands of artificial neurons called processing units, which are interconnected by nodes. In ANN architecture, nodes are grouped into layers depending on the activation they implement on the data. In ANN, the output of one node goes as input to another node. Subsequently, the input node after receiving information from previous output node, based on an internal weighting

system, attempts to produce the next output node. Through repeated training, the weight system can amplify or weaken the level of communication between nodes. After mature training, which has optimized the weight system, a trained ANN can make predictions on the test data. Due to the fact that ANN can be constructed by many layers and neurons, this method is considered deep learning algorithm. There are many kinds of ANN currently used in the literature including feedforward neural network, recurrent neural network, convolutional neural network, modular neural network and so on. In this study, since our input data is well structured allowing neural network to effectively learn, we present the simplest form of ANN which is a densely connected feedforward neural network (DNN). Figure 6 shows the illustration of our proposed DNN with three hidden layers.

Results

Benchmark Data

On 1 August 2016, the Department of Health released approximately 1 billion lines of anonymous historical health data relating to approximately 3 million Australians on data.gov.au. The information released includes details on medical services provided to Australians by health professionals along with details of subsidized information. Claims data for a random 10% sample of Australians are made available for research institutions, health professionals and universities. The data release includes historical Medicare data (from 1984) and PBS data (from 2003) up to 2014. The release comprises two files corresponding to the two types of service information (MBS and PBS), and a separate patient demographic file. The dataset used in this study was joined from MBS, PBS service information and patient demographic data by patient ids. It originally includes 346,201 records corresponding to 346,201 patients; however, there are 19 patients that have inadequate information being removed. Following this exclusion, the final dataset comprises of a total number of 346,182 patients.

The data set included 4 classes of variables (ie, features):

- **Demographic variables**: year of birth, sex and age (calculated until 1/1/2015)
- **Numerical variables**: A total of 13 continuous measurements are presented in the dataset, including number of MBS records, number of states, total amount of medical fee charged, total amount of medicare schedule fee, total amount of medical rebates paid, total number of MBS services, total length of patient accessing medicare services, number of PBS records, number of patient's PBS codes, total amount of medication cost paid by government, total amount of medication cost self-paid, total number of prescriptions and total length of patient accessing PBS services.
- **Categorical variables**: They are three relevant medications classified by the Anatomical Therapeutic Chemical (ATC) code and patient state. The medications presented are drugs used in diabetes (code: A10), drugs used for cardiovascular system and hypertension (code: C0) and lipid modifying agents or drugs used for patients with high cholesterol (code: C10).
- **Date variables**: Four date variables include date of first medical schedule, date of last medical schedule, date of first PBS claim and date of last PBS claim.

Among these variables, except for year of birth, age and numerical variables which were kept as they are, other variables were transformed as follows: Sex and medication variables were mapped into binary values while patient state was converted into 6 binary variables corresponding to 6 states. Year of birth, date of first medical schedule and date of first PBS claim were used to calculate age at which patient had first medical schedule and first PBS claim respectively and then removed. Regarding the prediction target variable, since PBS and MBS claims data on their own do not include information about patient's health outcome, the labels must be inferred. Between date of last medical schedule and date of last PBS claim, the later was used to calculate the length of patient discontinuing PBS and MBS services until 1/1/2015. Following this calculation, any patient with

more than 180 days (6 months) discontinuing PBS and MBS was labeled ‘deceased’, otherwise ‘alive’. After pre-processing, the dataset has 26 features and one label that is used for model development.

In terms of features scaling, each feature values were standardized to center around its mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation [22]. This step allows algorithm to effectively learn as it eliminates sensitivity towards multiple features spanning varying degrees of magnitude, range, and units.

In terms of class distributions, there are only 93,164 patients out of the total number of 346,182 classified into ‘deceased’ group while the rest are ‘alive’ patients. This reflects a highly imbalanced class distributions which might affect learning performance of the infrequent class [23] due to lacking samples. To handle this issue, a separate experiment using synthetic minority oversampling technique (SMOTE) is conducted as a proposal to enrich the training set.

Evaluation Metrics

Descriptive statistics is used to learn characteristics of the study population, stratified by health outcome status (ie, alive or deceased). Models are derived from training set and then assessed on testing set by calculating the traditional accuracy, precision, recall scores with an addition to brier loss. Other than that, reporting discrimination and calibration is important for assessing a prediction model [36]. Area Under the Receiver Operating Characteristic curve (AUROC) score and plotting reliability diagram (calibration curves) will also be calculated to assess the performance of AI models.

- **Brier loss** (from https://scikit-learn.org/) measures the accuracy of probabilistic predictions by calculating the mean squared difference between the predicted probability assigned to the possible classes and the actual classes. It is composed of refinement loss and calibration loss so that the lower the Brier score is for a set of predictions, the better the predictions are calibrated or the better the model is.
- **AUROC** score is used to measure the probability that the model ranks a random deceased patient more highly than a random alive patient in terms of mortality rate. Higher AUROC score means that the model has better ability to discriminate between the deceased and alive populations.
- **Calibration curve**, a reliability diagram, is a line plot of the relative frequency of what was observed versus the predicted probability frequency. The closer the points appear along the main diagonal from bottom left to top right, the better calibrated a forecast or more reliable a model [37].

Hyper-parameters

To develop the models, the study population was stratified into a training set, in which the mortality risk algorithms were derived, and a testing set, in which the algorithms were applied and tested. The training set consisted of 90% of the study dataset, and the testing set consisted of the remaining 10%. Training and testing sets were split at the patient level and in a stratifying manner according to class ratio so that patients do not appear in both the training and testing sets and ratio of patient labels (‘deceased’ or ‘alive’) in both sets are equivalent to that of the study population. After stratified assignment, hyper-parameters were determined by using a grid search of 5-fold cross validation to determine the values that leads to the best accuracy. After grid search, each algorithm was re-fitted to the training set with its best hyper-parameters to derive final models. Table 1 presents the parameter search space of four algorithms and the grid results.

Algorithms	Parameter Name	Search Space	Optimal
Logistic Regression	*penalty*	[‘l1’, ‘l2’, ‘none’]	l2
	C	[0.01, 0.1, 1.0]	1.0
	tol	[0.0001, 0.001, 0.01]	0.0001
	solver	[‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’]	lbfgs
	multi_class	[‘auto’, ‘ovr’, ‘multinomial’]	auto
Random Forest	*n_estimators*	[5, 10, 50, 100, 150]	100
	max_depth	[1, 2, 3, 5, None]	None
	max_features	[’auto’, ’sqrt’]	auto
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[1, 2, 4]	1
Extra Trees	*n_estimators*	[5, 10, 50, 100, 150]	100
	max_depth	[1, 2, 3, 5, None]	None
	max_features	[’auto’, ’sqrt’]	auto
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[1, 2, 4]	1
Gradient Boosting	*loss*	[‘deviance’, ‘exponential’]	deviance
	n_estimators	[5, 10, 50, 100, 150]	100
	max_depth	[1, 2, 3, 5]	3
	learning_rate	[0.001, 0.01, 0.1]	0.1
	criterion	[’friedman_mse’, ’mse’, ’mae’]	friedman_mse

Table 1: Hyper-parameters for Grid Search

After the grid search, it is found that LR with L2 regularization, which is also known as Ridge Regression [38], produces most accurate predictions in cross validation, its C value and tolerance rate are 1.0 and 0.0001 respectively. This can be explained by the fact that our dataset has a small number of features making L1 regularization, which is Lasso Regression and works well for feature selection in data set with high dimensionality [39], less favorable. Next, RF and ET both achieve optimal accuracy after grid search with max_depth 'None' scheme (from https://scikit-learn.org/). According to scikit-learn team, this scheme means that nodes are expanded until all leaves are pure or until all leaves contain less than $min\_samples\_split$ samples, which is optimized at 2 in both cases. Besides, the number of trees grown in both algorithms is the same, 100 ($n\_estimators$). Lastly, errors in GBT is minimized using $deviance$ loss function, there are also 100 trees built with the maximum number of nodes equal to 3.

To develop DNN model, the study population was also stratified into training and testing sets with ratio 90% and 10% respectively. Then, the training set was one more time broken down into training and validation sets with the same ratio. The purpose of validation set is to provide an unbiased evaluation of the model while tuning model’s weights [40]. The input layer has 26 units corresponding to the number of features while the output layer has one unit. Activation function used in output layer is $sigmoid$. The architecture of DNN used is composed of three fully connected hidden layers. The numbers of neurons in each hidden layer are 128, 64 and 32 respectively, and Rectified Linear Unit (ReLU) is used as the activation function. During the training process, the parameters of DNN are initialized using the uniform initialization [41]. For each batch of training data, parameters of DNN were modified gradually to decrease the cross entropy of loss function. A

callback was set to stop the training process after 10 epochs since when the model reaches the highest value of AUROC.

All models after the training process were evaluated using the holdout (10%) testing set. Final results were then compared and used to make recommendations.

Model Performance

In our experiments, we trained the models using the original learning sample first then applied SMOTE in order to further improve their performance.

Performance without SMOTE

Details of model performance without SMOTE is presented in Table 2. After adjusting for multiple comparisons, there was no significant difference in accuracy among Random Forest (RF 98.5%), Gradient Boosting Trees (GBT 98.4%), Logistic Regression (LR 97.8%), Extra Trees (ET 97.9%) and Densely Connected Neural Network (DNN 97.1%). In terms of discrimination, GBT and RF achieved highest AUROC (97.8% and 97.7% respectively), followed by LR and ET (96.4% and 96.8% respectively) while DNN was the least discriminative (95.3%). In terms of brier loss, GBT and RF produced the smallest difference between the probability assigned to the predicted classes and the probability of the actual class (both 0.012) while DNN predictions seen the biggest difference (0.024) yet still a good result.

Algorithms	Accuracy	AUROC	Precision	Recall	Brier Loss
Logistic Regression	97.8	96.4	**98.5**	93.4	0.016
Random Forest	**98.5**	97.7	98.1	**96.1**	**0.012**
Extra Trees	97.9	96.8	98.1	94.2	0.016
Gradient Boosting Trees	98.4	**97.8**	97.5	96.5	**0.012**
Artificial Neural Network	97.1	95.3	96.6	91.8	0.024

Table 2: Performance metrics of Machine Learning models without SMOTE

According to the table with training time (Table 3), LR turns out to be superior compared to other models with only less than 1-minute training time. However, DNN takes up to half an hour to train. This could be explained by the complexity level of two algorithms, while LR is a very simple and straight forward model which is based on a linear regression equation, DNN is an architecture that is composed of many neurons, layer and more complex activation functions.

Algorithms	Training Time (seconds)
Logistic Regression	**6.6**
Random Forest	106.8
Extra Trees	46.8
Gradient Boosting Trees	186
Artificial Neural Network	1277.4

Table 3: Training time of Machine Learning models without SMOTE

Clearly, all of our models show a very similar behavior for two classes (see Figure 7. According to the confusion matrices, RF and GBT managed to identify 'deceased' patients with higher performance than other algorithms. Meanwhile, compared to other models, there are a larger number of cases where DNN classifies 'deceased' patients to be 'alive'.

In terms of prediction reliability, calibration curves for the five models appear in Figure 8 showing LG was the least calibrated compared to other four algorithms, highly overestimating patient death risks in all level of probabilities. RF was well-calibrated for patients with lower mortality rate while overestimated the risk of death when the probability of risk is over 50%. ET's good of fit was only seen in the probability of death at 30% while underestimation and overestimation appeared for patients with lower and higher probabilities of death than 30% respectively. Predictions by GBT and DNN were the most well-calibrated while DNN slightly overestimated patients with probabilities of death greater than 10% and below 90%.

Performance with SMOTE

Details of model performance with SMOTE is presented in Table 4 and their calibration plots are displayed in Figure 9. As can be seen in Table 4, SMOTE just slightly improves performance (in italic) of five models. However, using SMOTE helps significantly calibrate predictions of LR. After up-sampling, LR model no longer overestimates death risks of patient, its predictions are more closely aligned with the perfectly calibrated line. Meanwhile, ET is now seen goodness of fit in predictions of patients with death risk between 50-60% but still underestimates and overestimates those with low and high death risks respectively. On the other hand, RF predictions change from being well-calibrated in under 50% probabilities of death risk and overestimating higher ones into being well-calibrated in over 80% probabilities of death risk and underestimating the rest. More interestingly, DNN and GBT receive adversarial affects from up-sampling technique, becoming generally risk underestimating.

Algorithms	Accuracy	AUROC	Precision	Recall	Brier Loss
Logistic Regression	98.2	97.4	**97.3**	95.9	0.015
Random Forest	**98.4**	**98.0**	96.8	97.3	**0.012**
Extra Trees	98.1	97.4	97.1	95.8	0.016
Gradient Boosting Trees	98.1	97.9	95.2	**97.7**	0.014
Artificial Neural Network	96.7	96.2	93.0	95.1	0.026

Table 4: Performance metrics of Machine Learning models with SMOTE

In short, SMOTE is only helpful to further improve model performance and prediction calibration of LG. Meanwhile, using or not using SMOTE makes no difference to the performance of RF and ET in predicting mortality of patients with CVD. Lastly, SMOTE introduces an adversarial effect into GBT and DNN models, making their predictions less reliable, and these two models already work well with class imbalanced data.

Algorithms	Training Time (seconds)
Logistic Regression	**292.9**
Random Forest	497.9
Extra Trees	347.5
Gradient Boosting Trees	648.1
Artificial Neural Network	5480.3

Table 5: Training time of Machine Learning models with SMOTE

In terms of training duration, using SMOTE requires more computing time for all algorithms. However, LR is still the most time-efficient model even when applying SMOTE and produces higher accuracy and better prediction performance in terms of AUROC, recall as well as brier loss compared to LR with original data. Furthermore, SMOTE helps LR outperform ET and become the

second-best algorithm, after RF. Clearly, when bringing SMOTE into the table, ET and LR are those worth considering for this dataset.

Discussion

Principal Results

This study shows that structured medical and pharmaceutical claims data can be used as input for AI models to accurately predict the mortality risk of individuals with CVD. The logistic regression, random forest, extra trees, gradient boosting trees and artificial neural network models trained in this study had high accuracy (i.e., 97-98%) and discrimination (i.e., AUROC, 95-98%) in predicting mortality rate, much higher than traditional statistical models such as the Cox Proportional-Hazards model [42] or the models trained with traditional electrical health record [43-45].

Although there was no statistically significant difference in accuracy among the all experimental algorithms, the random forest model had an advantage compared with other models. Additionally, the random forest model also outperformed other models in recall and brier loss. In terms of discrimination and calibration, the gradient boosting trees is proved to be the most superior. Without SMOTE, logistic regression is unable to produce highly calibrated prediction while using SMOTE significantly improve this model predictions' reliability. All models with SMOTE had very high precision (i.e., 93-97%) and recall (i.e., 95-97%), particularly compared with other logistic regression and random forest prognostic models which did not deal with class imbalance published in the literature [44, 45]. On the other hand, although the artificial neural network had the most moderate performance among experimental algorithms, it was proved to be efficient even with class imbalanced data. It is also suggested that artificial neural network is capable of predicting CVD mortality rate more accurately than other ML algorithms if applied more feature engineering techniques [46, 47], indicating a very promising area of further research.

To our knowledge, this is the first paper comparing AI algorithms using medical and pharmaceutical claims data to predict mortality in a large general cardiology population. Unlike previously developed ML-based prognostic tools in cardiology which utilized clinical information of patients including clinical features [43-45], our models were trained on only claims data of patients with CVD. This claims data primarily provides information about patient’s medical scheduling and pharmaceutical dispensing history which reflects patient’s disease treatment cost, access patterns and medications but not patient’s state of health or other clinical indexes. Furthermore, compared with previously published classifiers in cardiology, our models used fewer features, which is comparatively more efficient than previously trained models in the general cardiology setting.

Limitations

Despite high accuracy and strong discrimination, some models still have not yielded optimal calibration including random forest, extra trees and artificial neural network. This means that the distribution and behavior of the probability predicted is not similar to the distribution and behavior of probability observed in training data. In order to increase reliability of AI algorithms, other techniques should be investigated to better calibrate and improve performance of these models, especially artificial neural network.

Conclusions

In conclusion, we developed, validated and compared five AI architectures to predict the mortality rate of patients with CVD. Based on the evaluation results, we can draw the following conclusions or insights that could help with the choice of AI models: first, without health indexes or health

condition information, AI architectures are able to accurately predict mortality of patients with CVD before a clinic visit using only medical scheduling and pharmaceutical dispensing claims data; second, although there was no statistically significant difference in accuracy among all experimental AI algorithms, the tree-based i.e. random forest and gradient boosting tree models have an advantage compared with other models; third, while the regression-based i.e. logistic regression method produces predictions having the least calibration level due to lack of minority class samples, up-sampling technique SMOTE helps significantly improve the reliability of this algorithm's predictions; fourth, tree-based algorithms and densely connected neural network perform well with class imbalanced data. Finally, this study showed the feasibility and effectiveness of different AI architectures based on structured medical scheduling and pharmaceutical dispensing claims data in identifying patients with CVD who had risk of mortality, which can be a useful tool for precise decision making. In future work, considering the promising potential of the artificial neural network, research should focus on improving prediction performance of this algorithm. It is suggested that artificial neural network is capable of predicting CVD mortality rate more accurately than other ML algorithms if applied more feature engineering techniques, indicating a very promising area of further research.

Acknowledgements

We would like to thank Dr. Dennis Wollersheim and Dr. Shaun Purkiss from La Trobe University for helping us prepare the data set used in this research.

Conflicts of Interest

None declared.

Abbreviations

CVD: Cardiovascular disease
MBS: Medical Benefits Scheme
PBS: Pharmaceutical Benefits Scheme
LR: Logistic Regression
RF: Random Forest
ET: Extra Trees
GBT: Gradient Boosting Trees
SMOTE: Synthetic Minority Oversampling Technique
AUROC: Area Under the Receiver Operating Characteristic

References

1. AIHW, Cardiovascular disease: Australian facts 2011, Cardiovascular dis- ease series. Cat. no. CVD 53.
2. I. Aguilar-Palacio, S. Malo, M. Lallana, C. Feja, J. Gonz´alez, B. Moreno- Franco, M. Rabanaque, Co-prescription patterns of cardiovascular preventive treatments: a cross-sectional study in the aragon worker'health study (spain), BMJ open 9 (4) (2019) e023571.
3. L. Stoeldraijer, C. van Duin, L. van Wissen, F. Janssen, Impact of different mortality forecasting methods and explicit assumptions on projected future life expectancy: The case of the netherlands, Demographic Research 29 (2013) 323–354.
4. H. Booth, L. Tickle, Mortality modelling and forecasting: A review of methods, Annals of actuarial science 3 (1-2) (2008) 3–43.
5. B. J. Mortazavi, E. M. Bucholz, N. R. Desai, C. Huang, J. P. Curtis, F. A. Masoudi, R. E.

Shaw, S. N. Negahban, H. M. Krumholz, Comparison of ma- chine learning methods with national cardiovascular data registry models for prediction of risk of bleeding after percutaneous coronary intervention, JAMA network open 2 (7) (2019) e196835–e196835.
6. T. J. Cleophas, A. H. Zwinderman, Machine learning and unsolved questions, in: Machine Learning in Medicine, Springer, 2013, pp. 205–214.
7. S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, N. Qureshi, Can machine learning improve cardiovascular risk prediction using routine clinical data?, PloS one 12 (4) (2017) e0174944.
8. M. Su, M. Miften, C. Whiddon, X. Sun, K. Light, L. Marks, An artificial neural network for predicting the incidence of radiation pneumonitis, Medical physics 32 (2) (2005) 318–325.
9. Z. Wang, Y. Zhu, D. Li, Y. Yin, J. Zhang, Feature rearrangement based deep learning system for predicting heart failure mortality, Computer Methods and Programs in Biomedicine 191 (2020) 105383.
10. C.-Y. Hung, W.-C. Chen, P.-T. Lai, C.-H. Lin, C.-C. Lee, Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 3110–3113.
11. J. J. Nirschl, A. Janowczyk, E. G. Peyster, R. Frank, K. B. Margulies, M. D. Feldman, A. Madabhushi, A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of h&e tissue, PloS one 13 (4) (2018) e0192726.
12. C. Martin-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baeßler, S. E. Petersen, K. Lekadir, Image-based cardiac diagnosis with machine learning: A review, Frontiers in Cardiovascular Medicine 7 (2020) 1.
13. A. Kilic, Artificial intelligence and machine learning in cardiovascular health care, The Annals of thoracic surgery 109 (5) (2020) 1323–1329.
14. A. M. Small, D. H. Kiss, Y. Zlatsin, D. L. Birtwell, H. Williams, M. A. Guerraty, Y. Han, S. Anwaruddin, J. H. Holmes, J. A. Chirinos, et al., Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease, Journal of biomedical informatics 72 (2017) 77–84.
15. J. Chu, W. Dong, K. He, H. Duan, Z. Huang, Using neural attention net- works to detect adverse medical events from electronic health records, Journal of biomedical informatics 87 (2018) 118–130.
16. M. E. Matheny, L. Ohno-Machado, F. S. Resnic, Discrimination and calibration of mortality risk prediction models in interventional cardiology, Journal of biomedical informatics 38 (5) (2005) 367–375.
17. M. E. Matheny, F. S. Resnic, N. Arora, L. Ohno-Machado, Effects of svm parameter optimization on discrimination and calibration for post- procedural pci mortality, Journal of Biomedical Informatics 40 (6) (2007) 688–697.
18. V. Taslimitehrani, G. Dong, N. L. Pereira, M. Panahiazar, J. Pathak, Developing ehr-driven heart failure risk prediction models using cpxr (log) with the probabilistic loss function, Journal of biomedical informatics 60 (2016) 260–269.
19. T. J. Cleophas, A. H. Zwinderman, Machine learning and unsolved questions, in: Machine Learning in Medicine, Springer, 2013, pp. 205–214.
20. C. Roth, P. R. Payne, R. C. Weier, A. B. Shoben, E. N. Fletcher, A. M. Lai, M. M. Kelley, J. J. Plascak, R. E. Foraker, The geographic distribution of cardiovascular health in the stroke prevention in healthcare delivery environments (sphere) study, Journal of biomedical informatics 60 (2016) 95–103.
21. T. E. Paterick, N. Patel, A. J. Tajik, K. Chandrasekaran, Improving health outcomes through patient education and partnerships with patients, Pro- ceedings (Baylor University. Medical Center) 30 (1) (2017) 112.
22. S. Learn, sklearn preprocessing standardscaler, Available: https://scikitlearn.

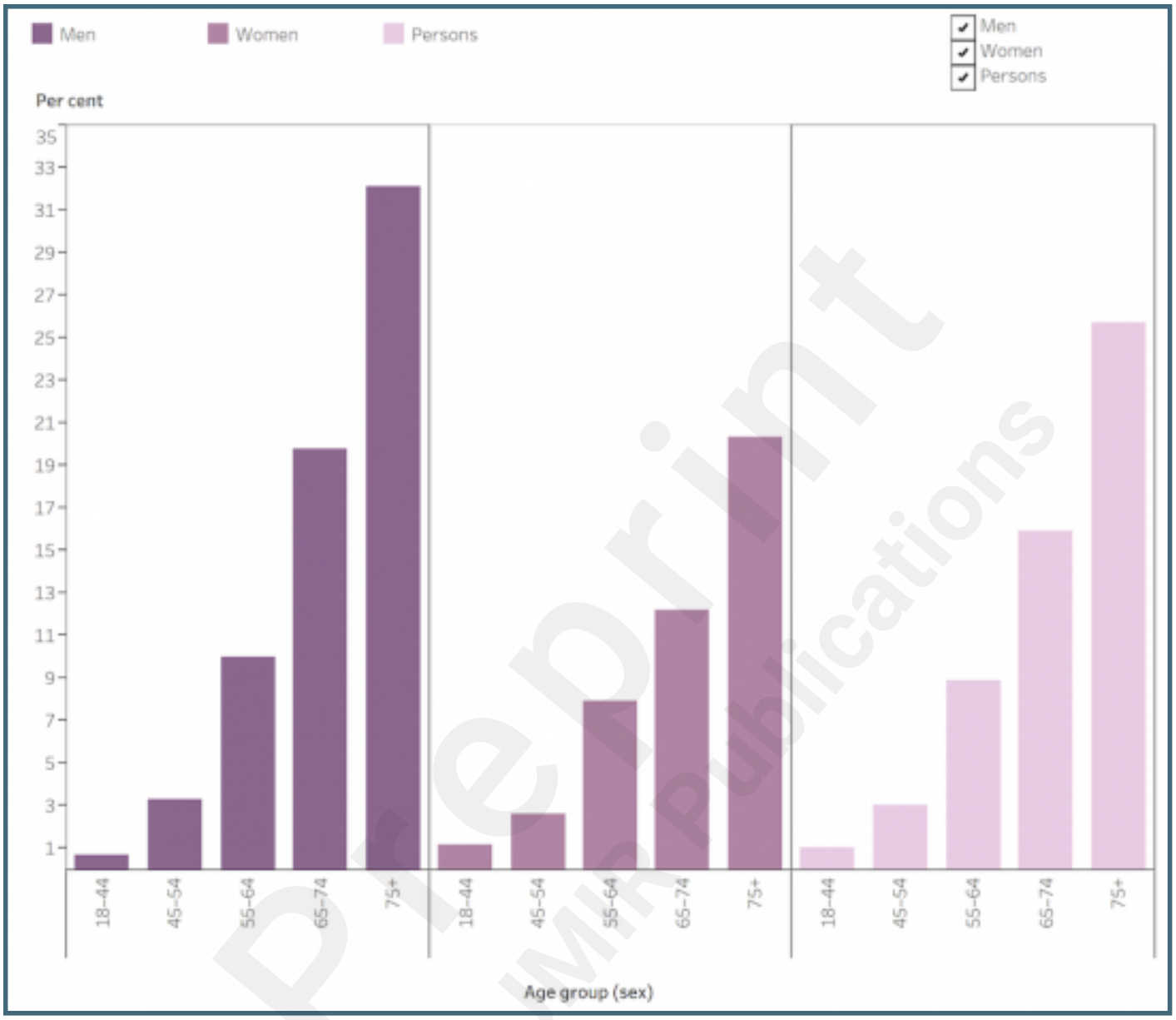
org/stable/modules/generated/sklearn. preprocessing. StandardScaler. html.[Ultimo acceso: Desembre 2019].
23. N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (5) (2002) 429–449.
24. D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, Vol. 398, John Wiley & Sons, 2013.
25. M. A. Friedl, C. E. Brodley, Decision tree classification of land cover from remotely sensed data, Remote sensing of environment 61 (3) (1997) 399– 409.
26. S. Uddin, A. Khan, M. E. Hossain, M. A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Medical Informatics and Decision Making 19 (1) (2019) 1–16.
27. J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.
28. L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
29. V. John, Z. Liu, C. Guo, S. Mita, K. Kidono, Real-time lane estimation using deep features and extra trees regression, in: Image and Video Technology, Springer, 2015, pp. 721–733.
30. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine learning 63 (1) (2006) 3–42.
31. J. H. Friedman, Stochastic gradient boosting, Computational statistics & data analysis 38 (4) (2002) 367–378.
32. S. Rahman, M. Irfan, M. Raza, K. Moyeezullah Ghori, S. Yaqoob, M. Awais, Performance analysis of boosting classifiers in recognizing activities of daily living, International Journal of Environmental Research and Public Health 17 (3) (2020) 1082.
33. W. S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The bulletin of mathematical biophysics 5 (4) (1943) 115- 133.
34. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (6088) (1986) 533-536.
35. R. C. Reid, From functional architecture to functional connectomics, Neuron 75 (2) (2012) 209-217.
36. E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obu- chowski, M. J. Pencina, M. W. Kattan, Assessing the performance of pre- diction models: a framework for some traditional and novel measures, Epi- demiology (Cambridge, Mass.) 21 (1) (2010) 128.
37. J. Brownlee, How and when to use a calibrated classification model with scikit-learn (2020).
38. A. E. Hoerl, R. W. Kannard, K. F. Baldwin, Ridge regression: some sim- ulations, Communications in Statistics-Theory and Methods 4 (2) (1975) 105-123.
39. R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267-288.
40. J. Brownlee, What is the difference between test and validation datasets, Machine Learning Mastery 14.
41. K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026-1034.
42. J. Hata, A. Nagai, M. Hirata, Y. Kamatani, A. Tamakoshi, Z. Yamagata, K. Muto, K. Matsuda, M. Kubo, Y. Nakamura, et al., Risk prediction models for mortality in patients with cardiovascular disease: The biobank japan project, Journal of epidemiology 27 (Supplement III) (2017) S71– S76.
43. J.-m. Kwon, K.-H. Kim, K.-H. Jeon, S. E. Lee, H.-Y. Lee, H.-J. Cho, J. O. Choi, E.-S. Jeon, M.-S. Kim, J.-J. Kim, et al., Artificial intelligence algorithm for predicting mortality of patients with acute heart failure, PloS one 14 (7) (2019) e0219302.
44. D. Chicco, G. Jurman, Machine learning can predict survival of patients with heart failure

from serum creatinine and ejection fraction alone, BMC medical informatics and decision making 20 (1) (2020) 16.
45. D. S. Lee, P. C. Austin, J. L. Rouleau, P. P. Liu, D. Naimark, J. V. Tu, Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model, Jama 290 (19) (2003) 2581–2587.
46. Z. Wang, Y. Zhu, D. Li, Y. Yin, J. Zhang, Feature rearrangement based deep learning system for predicting heart failure mortality, Computer Methods and Programs in Biomedicine 191 (2020) 105383.
47. N. Sadati, M. Z. Nezhad, R. B. Chinnam, D. Zhu, Representation learning with autoencoders for electronic health records: A comparative study, arXiv preprint arXiv:1908.09174.
48. Z.-H. Zhou, A brief introduction to weakly supervised learning, National Science Review 5 (1) (2018) 44–53.

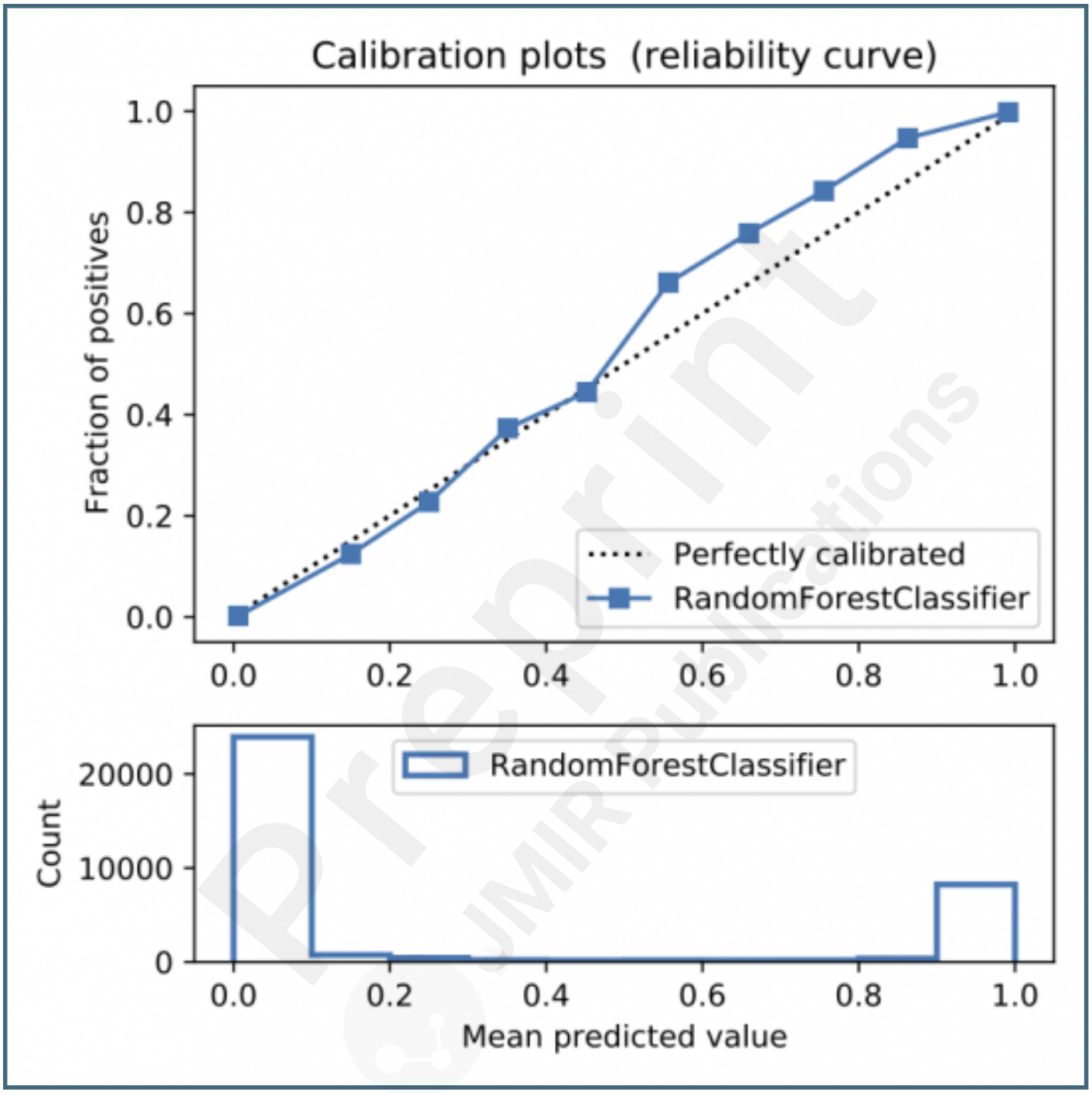
Supplementary Files

Figures

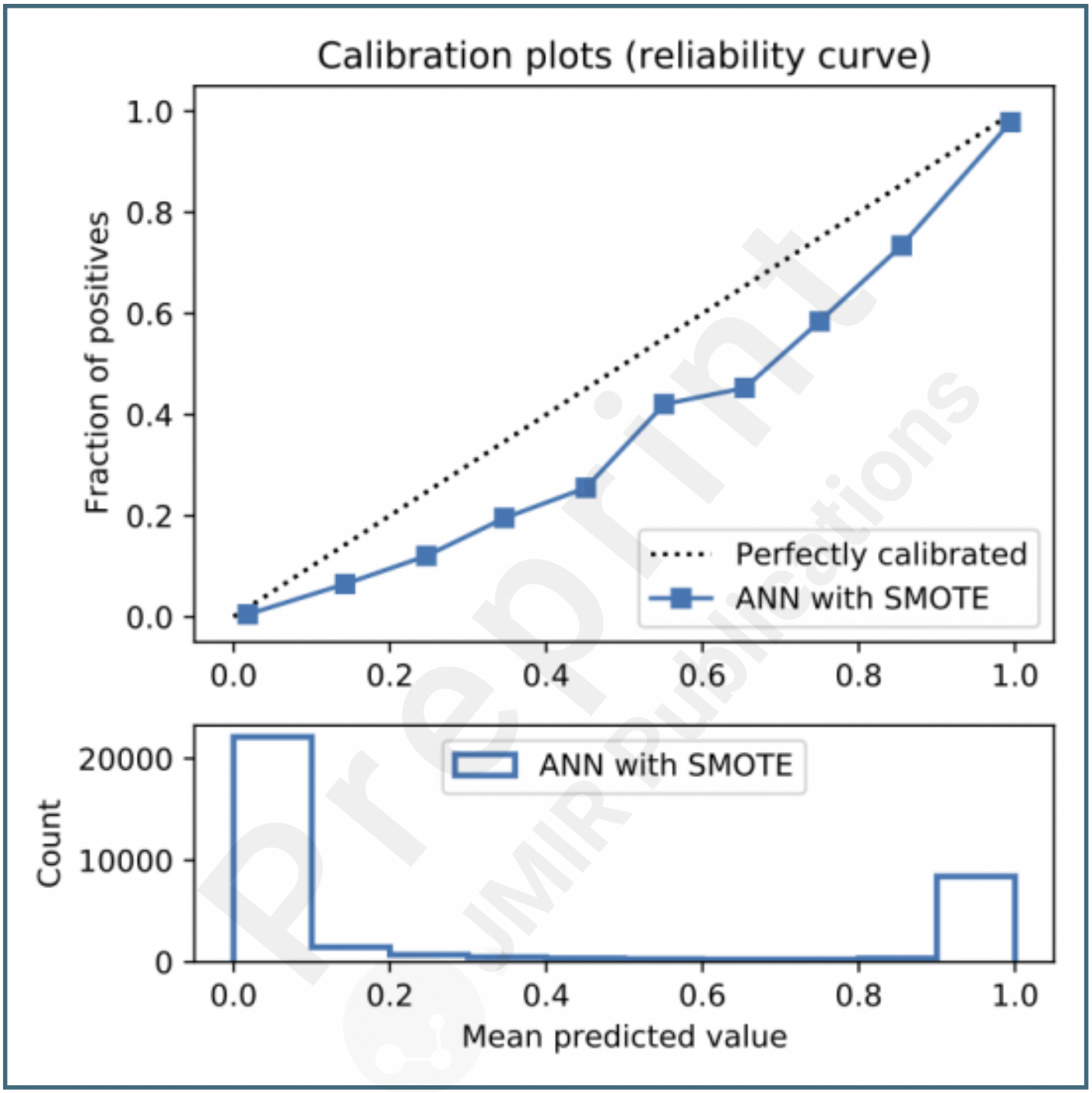
Prevalence of CVD by Age group and Sex, 2017-2018, from AIHW analysis of ABS 2019.



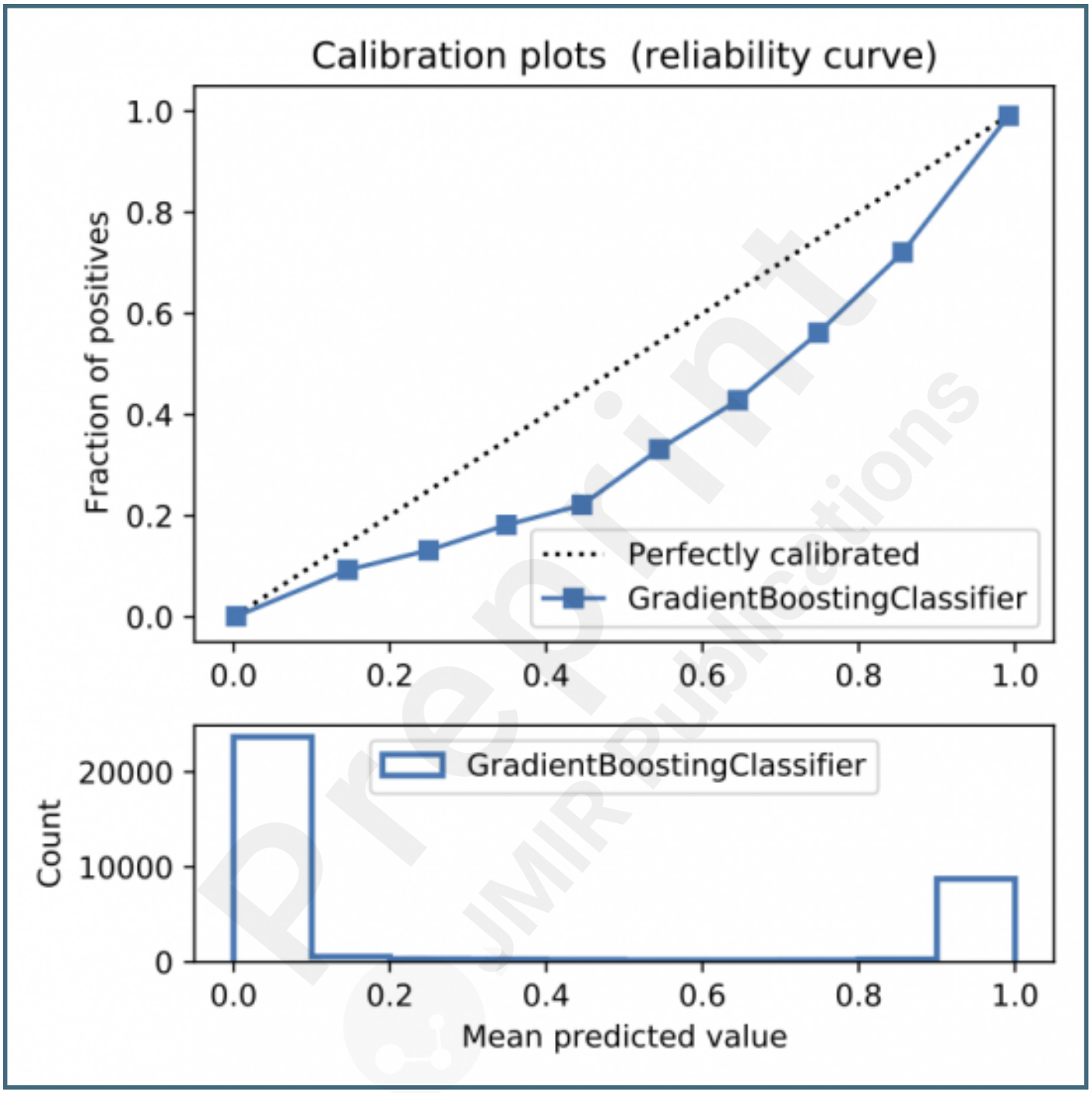
(b) Calibration curves of Random Forest without SMOTE.



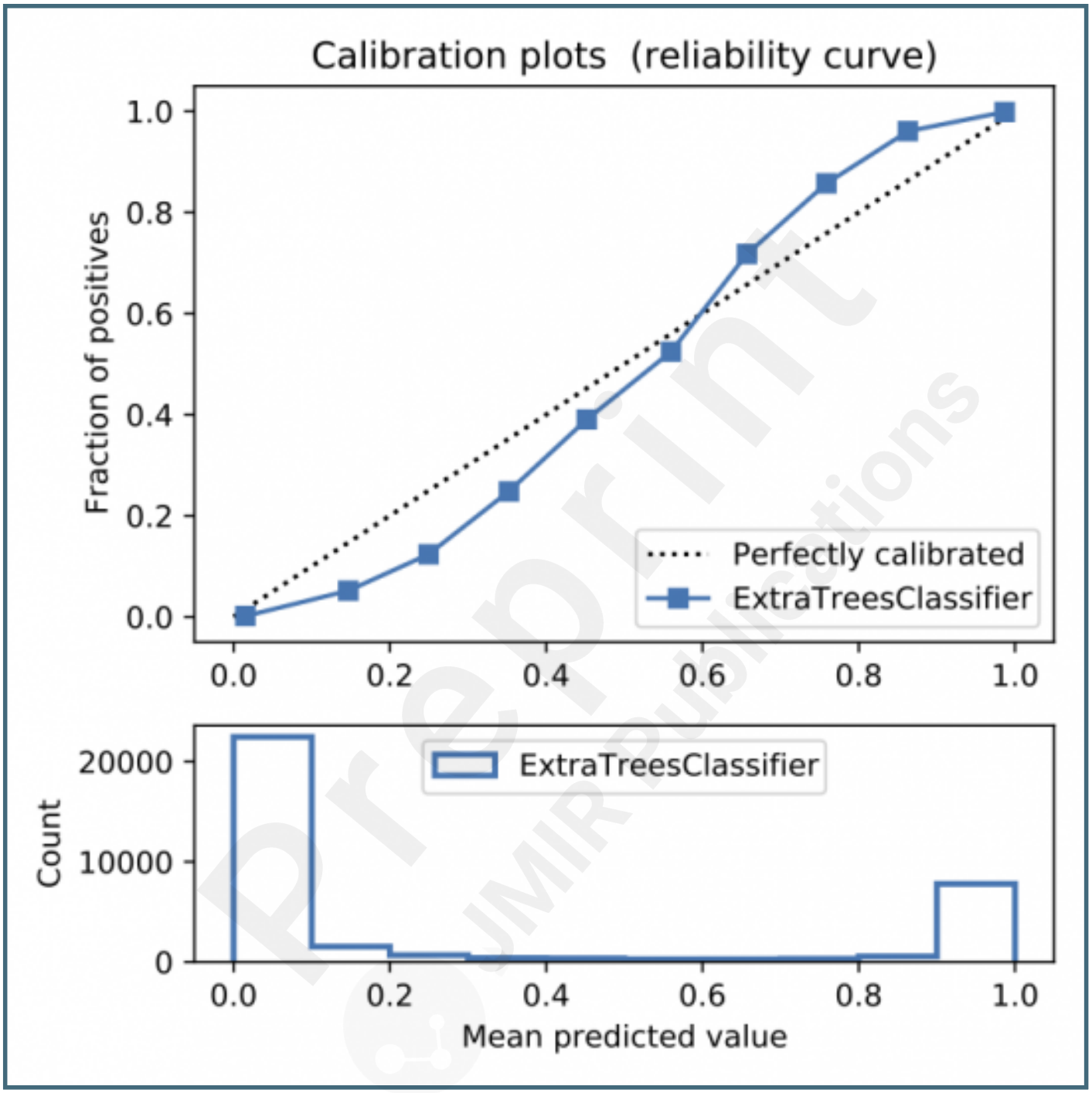
(e) Calibration curves of Artificial Neural Network with SMOTE.



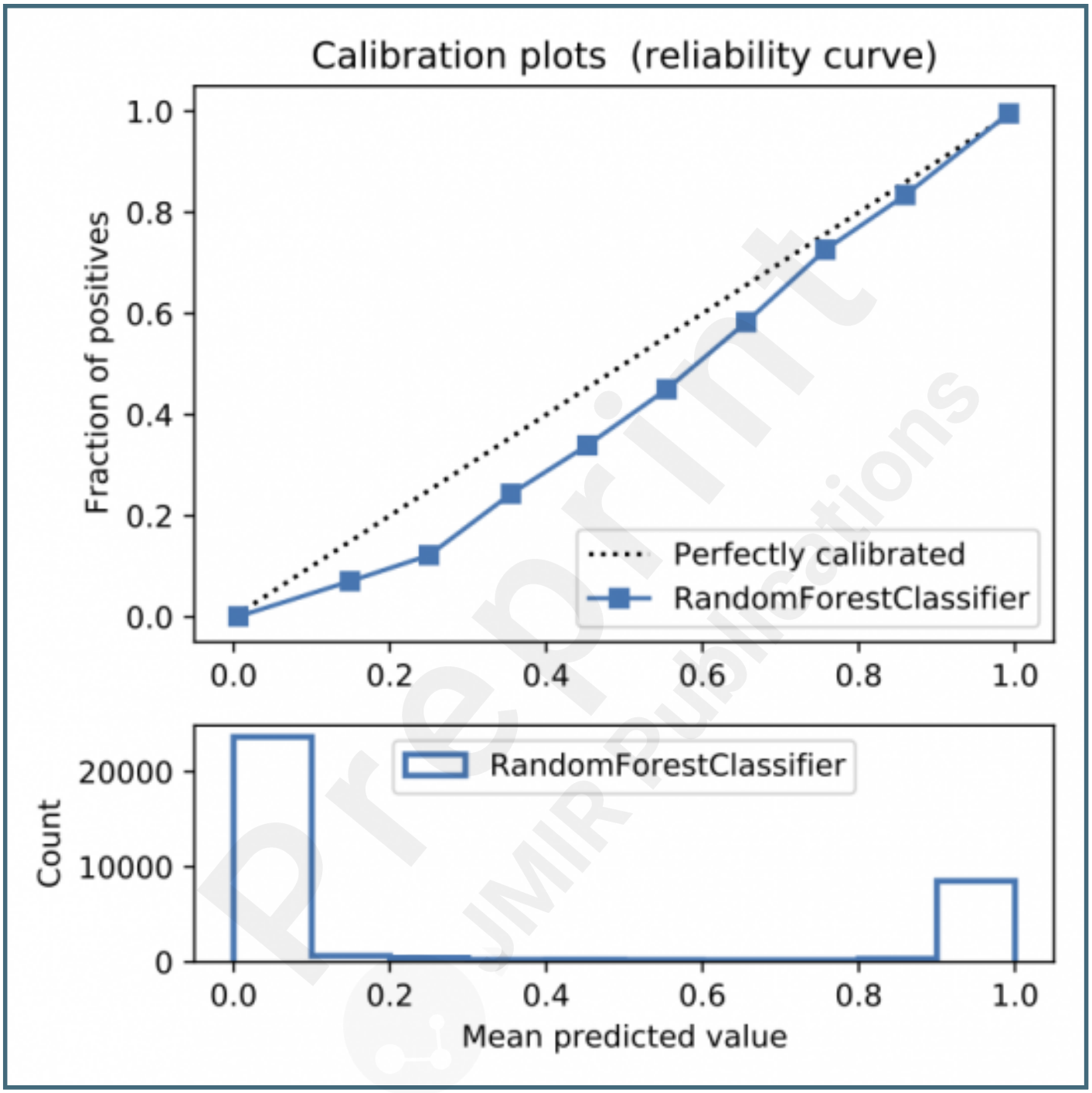
(d) Calibration curves of Gradient Boosting Trees with SMOTE.



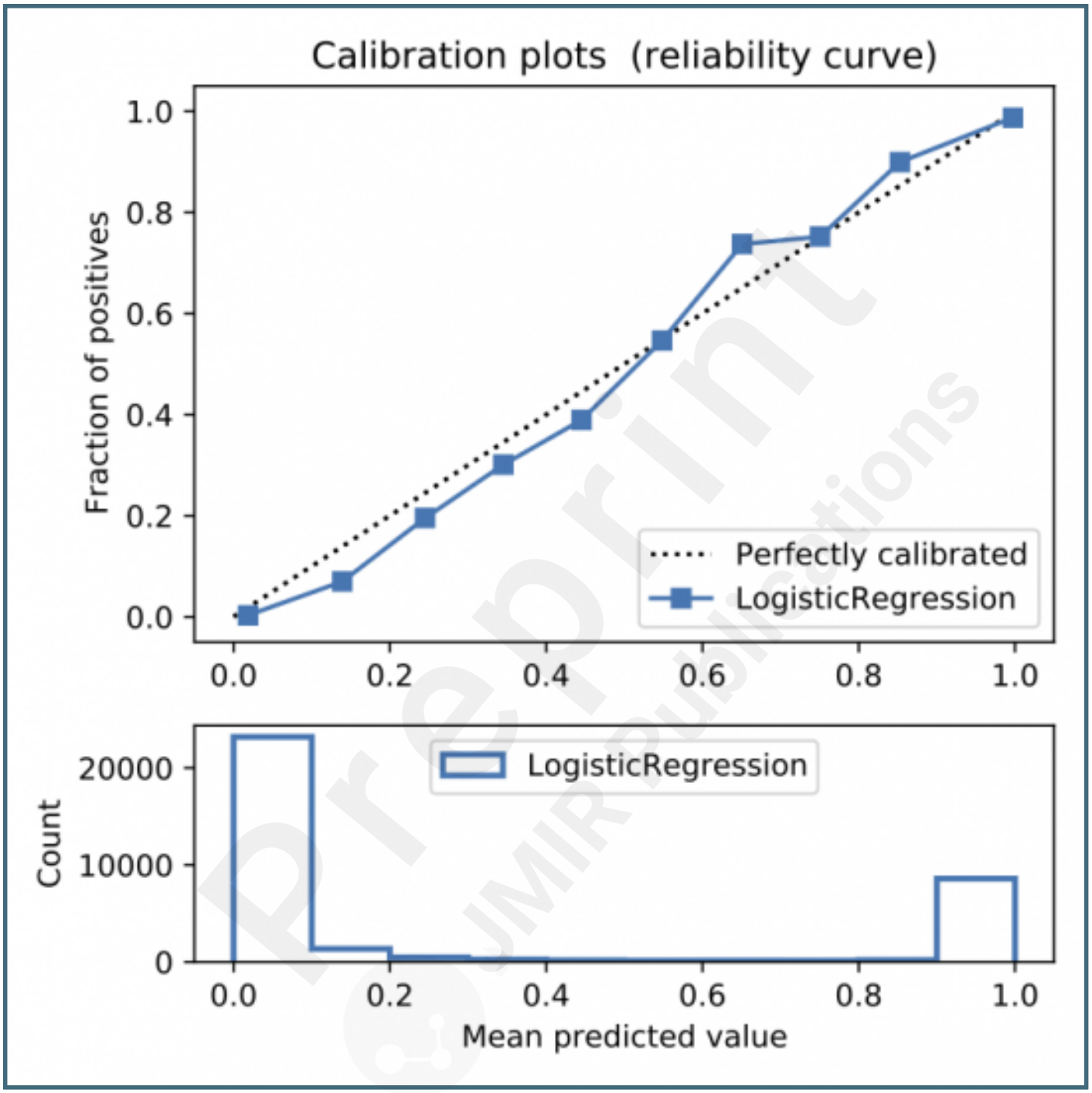
(c) Calibration curves of Extra Trees with SMOTE.



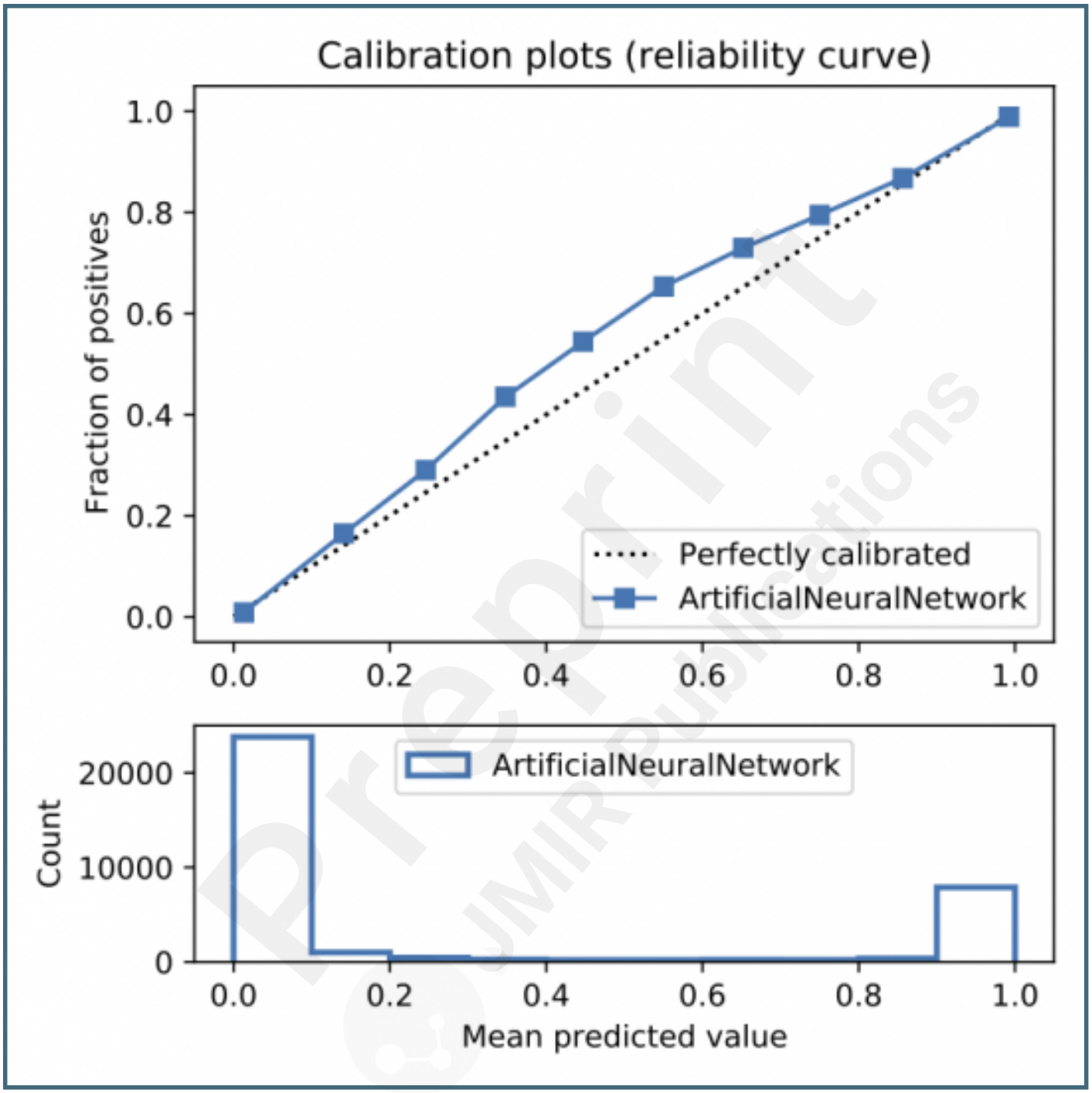
(b) Calibration curves of Random Forest with SMOTE.



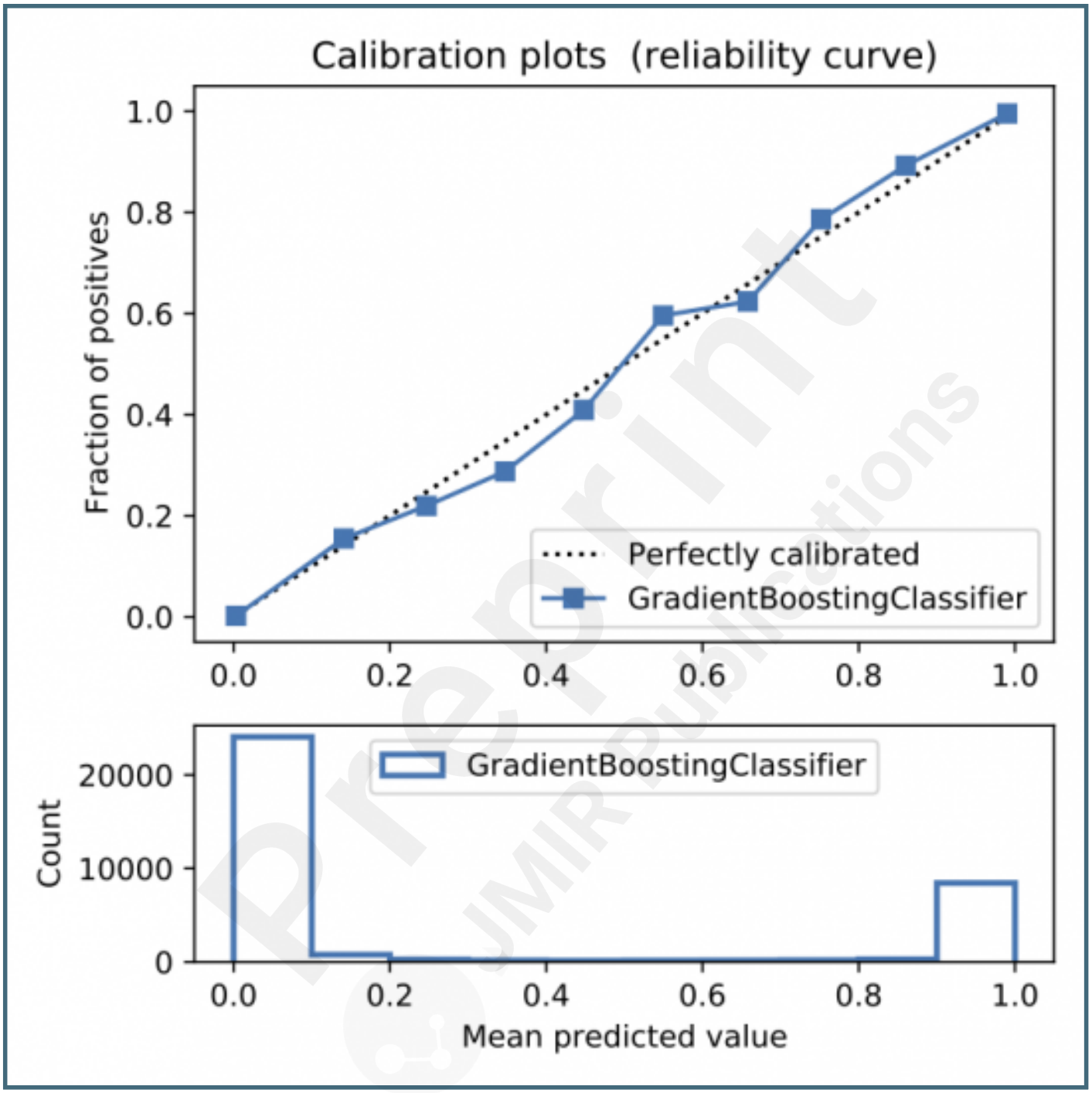
(a) Calibration curves of Logistic Regression with SMOTE.



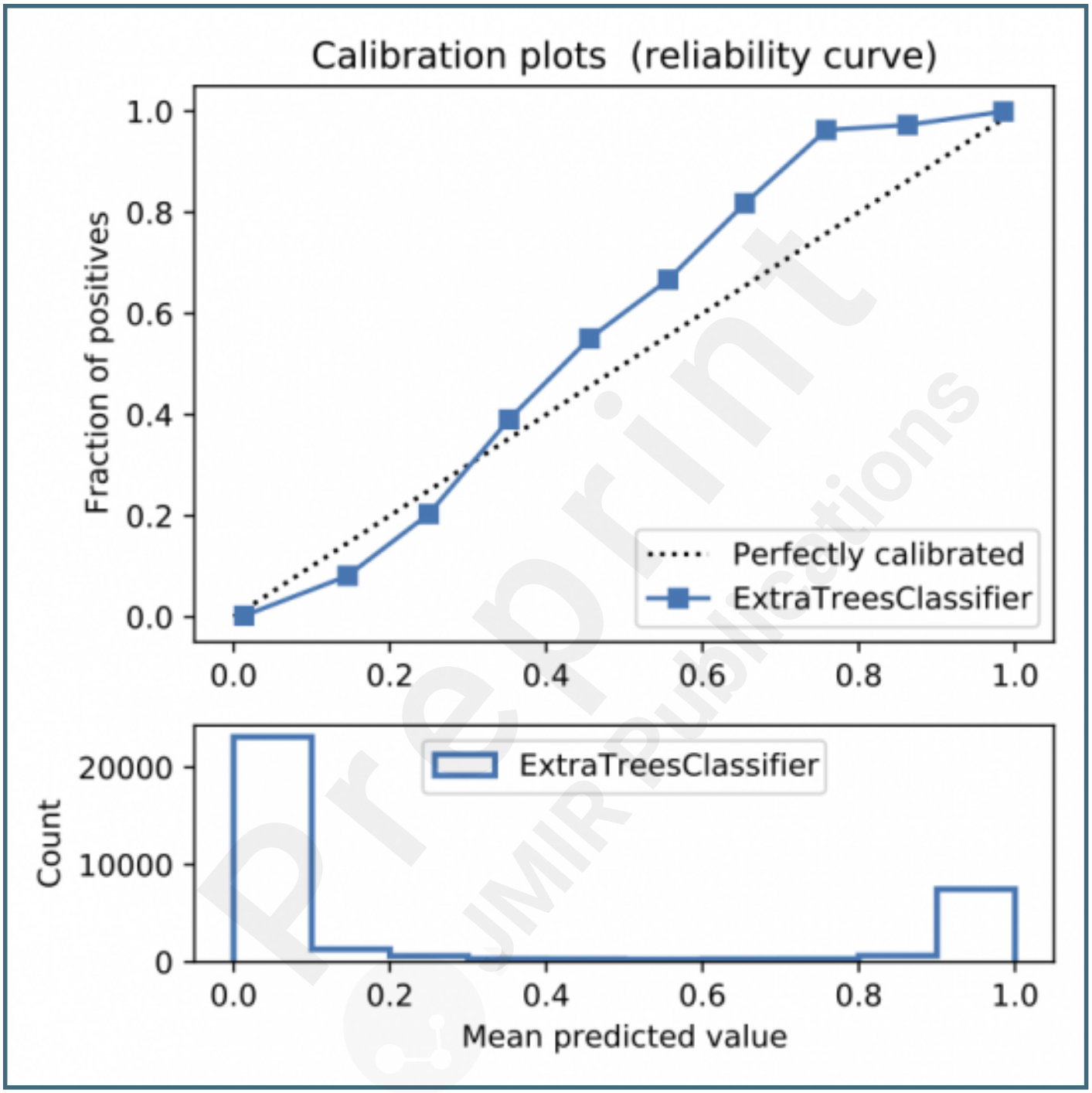
(e) Calibration curves of Artificial Neural Network without SMOTE.



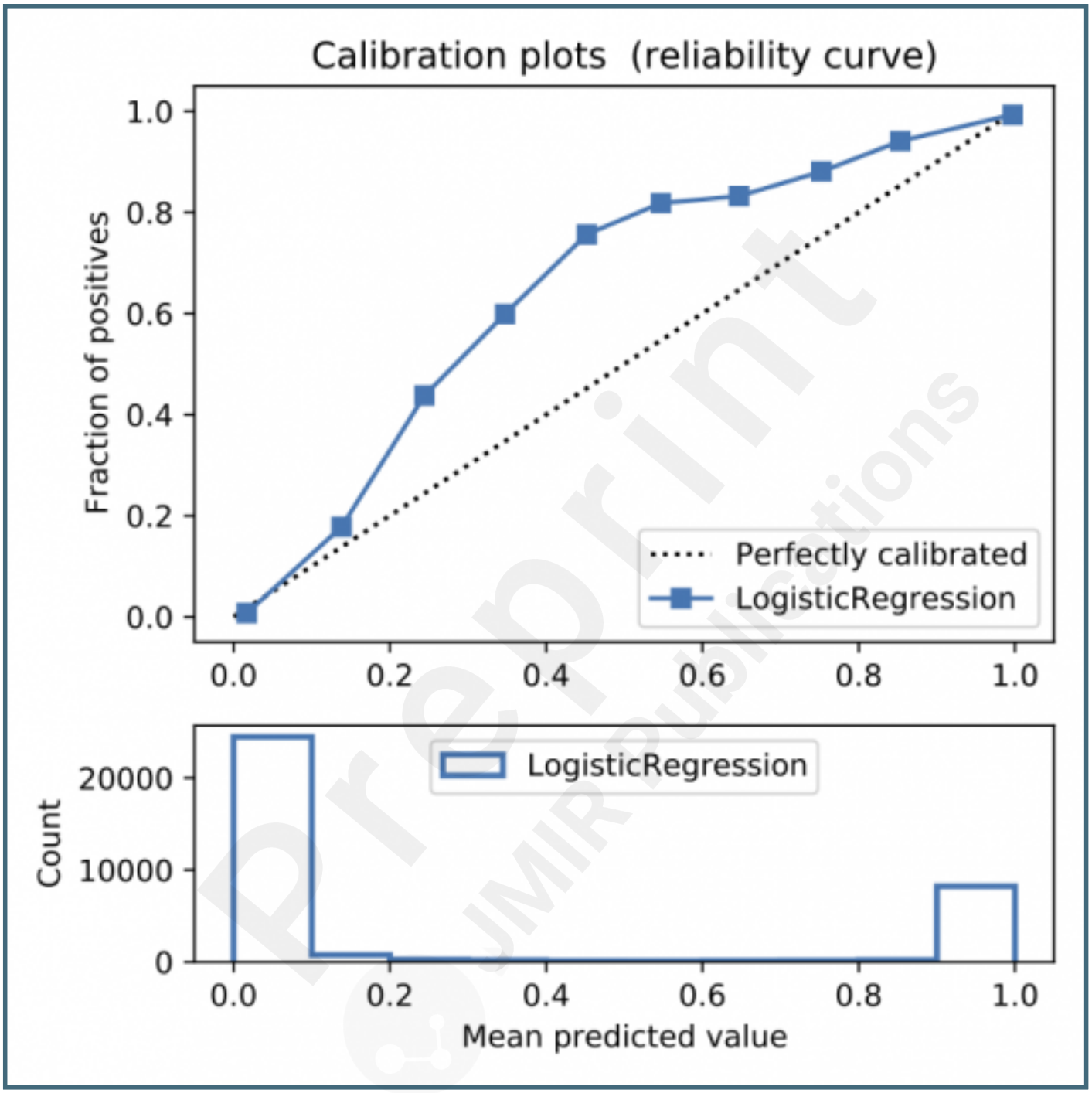
(d) Calibration curves of Gradient Boosting Trees without SMOTE.



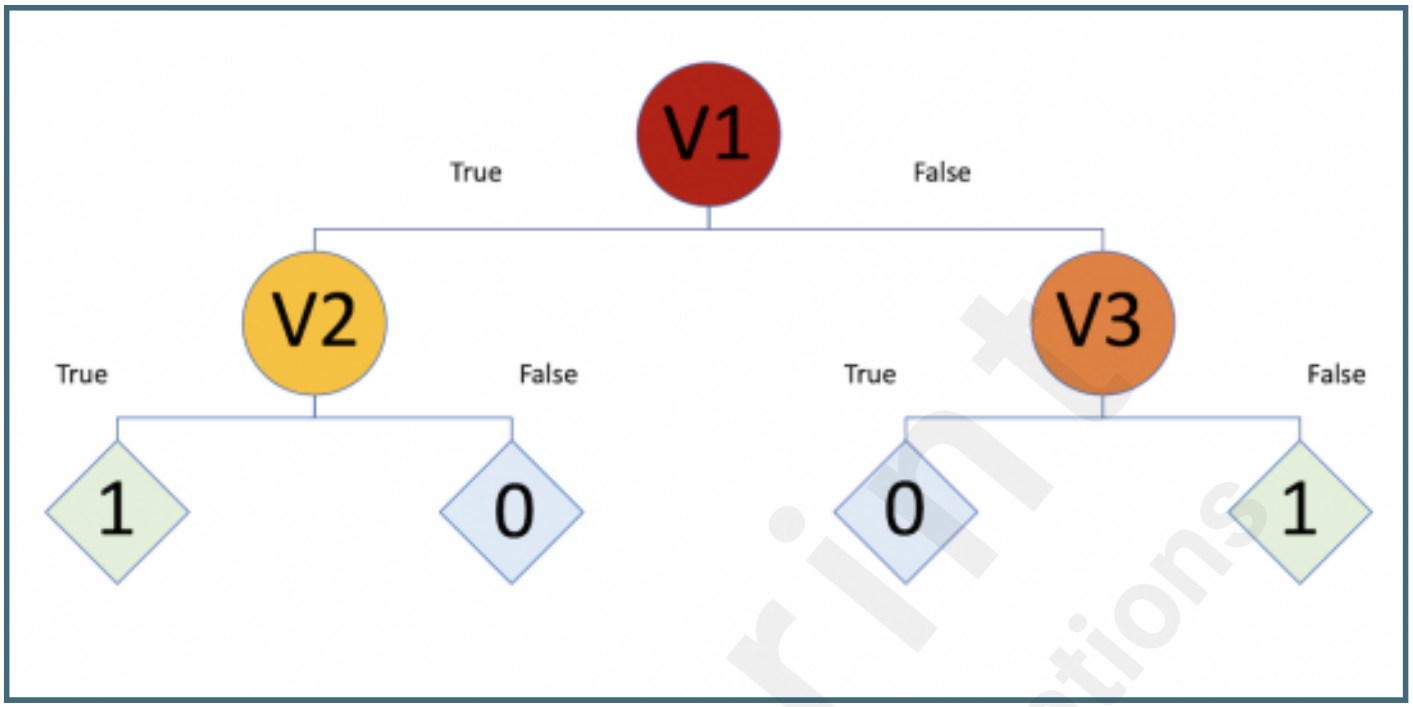
(c) Calibration curves of Extra Trees without SMOTE.



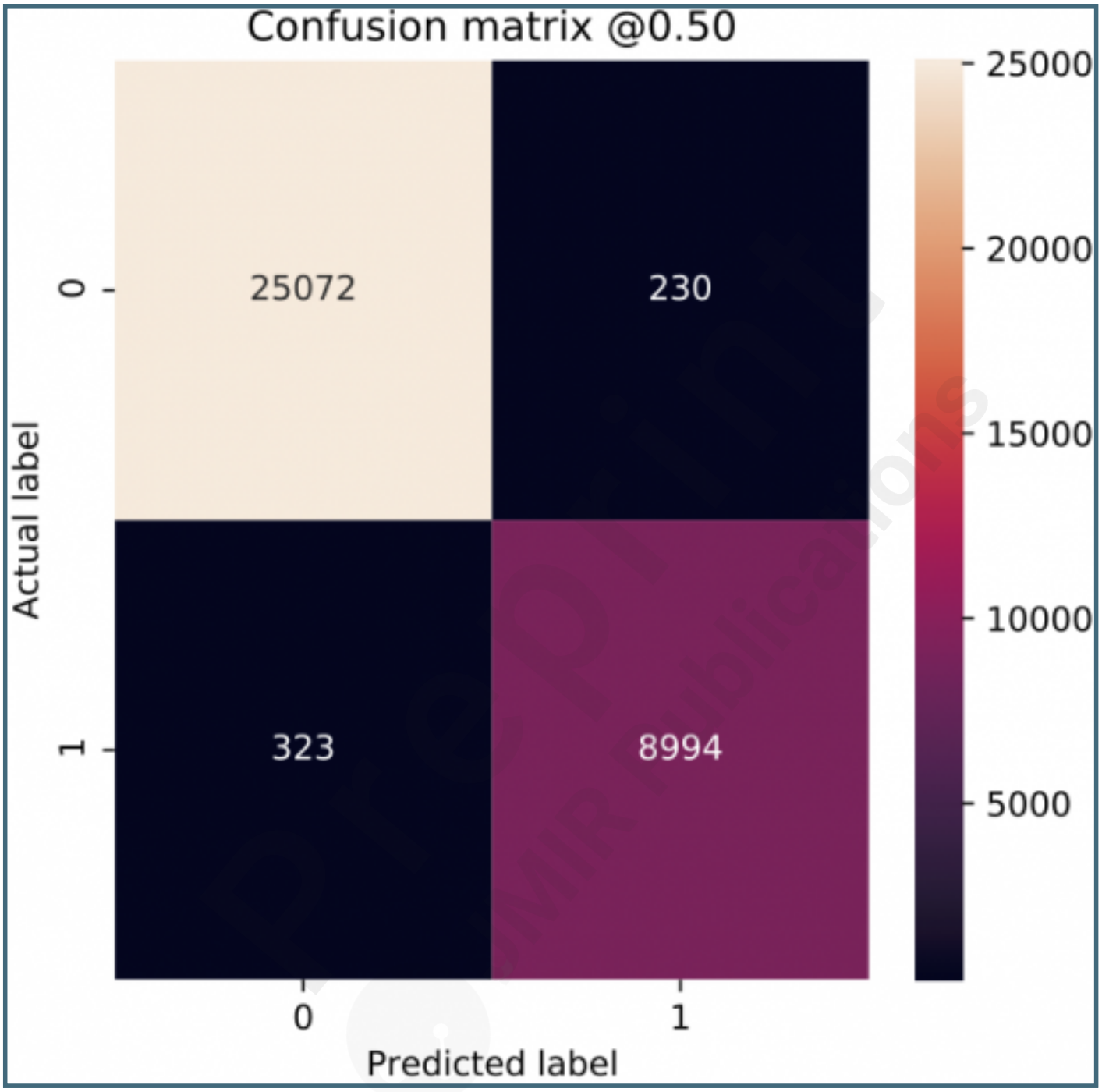
(a) Calibration curves of Logistic Regression without SMOTE.



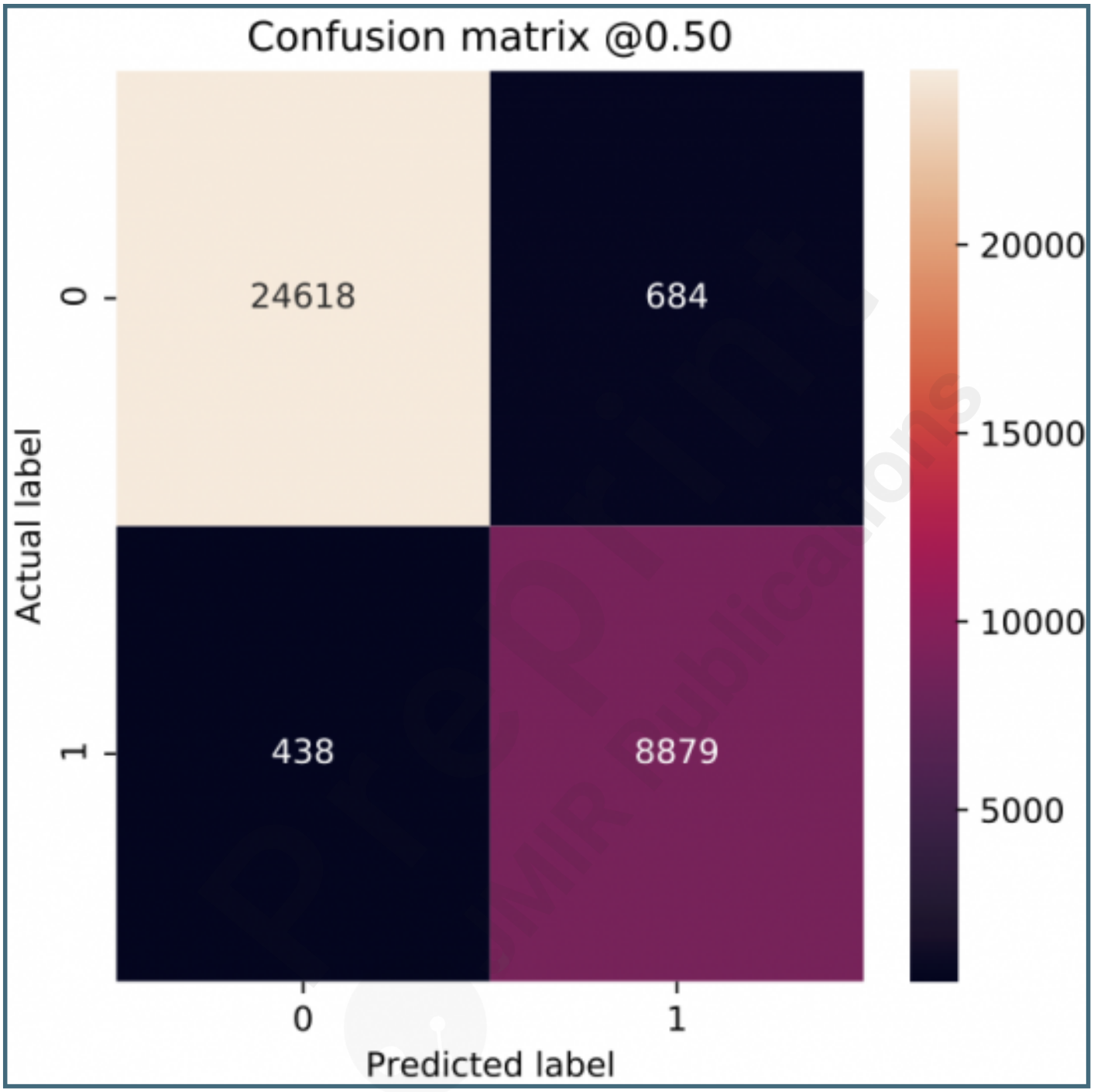
Decision Tree.



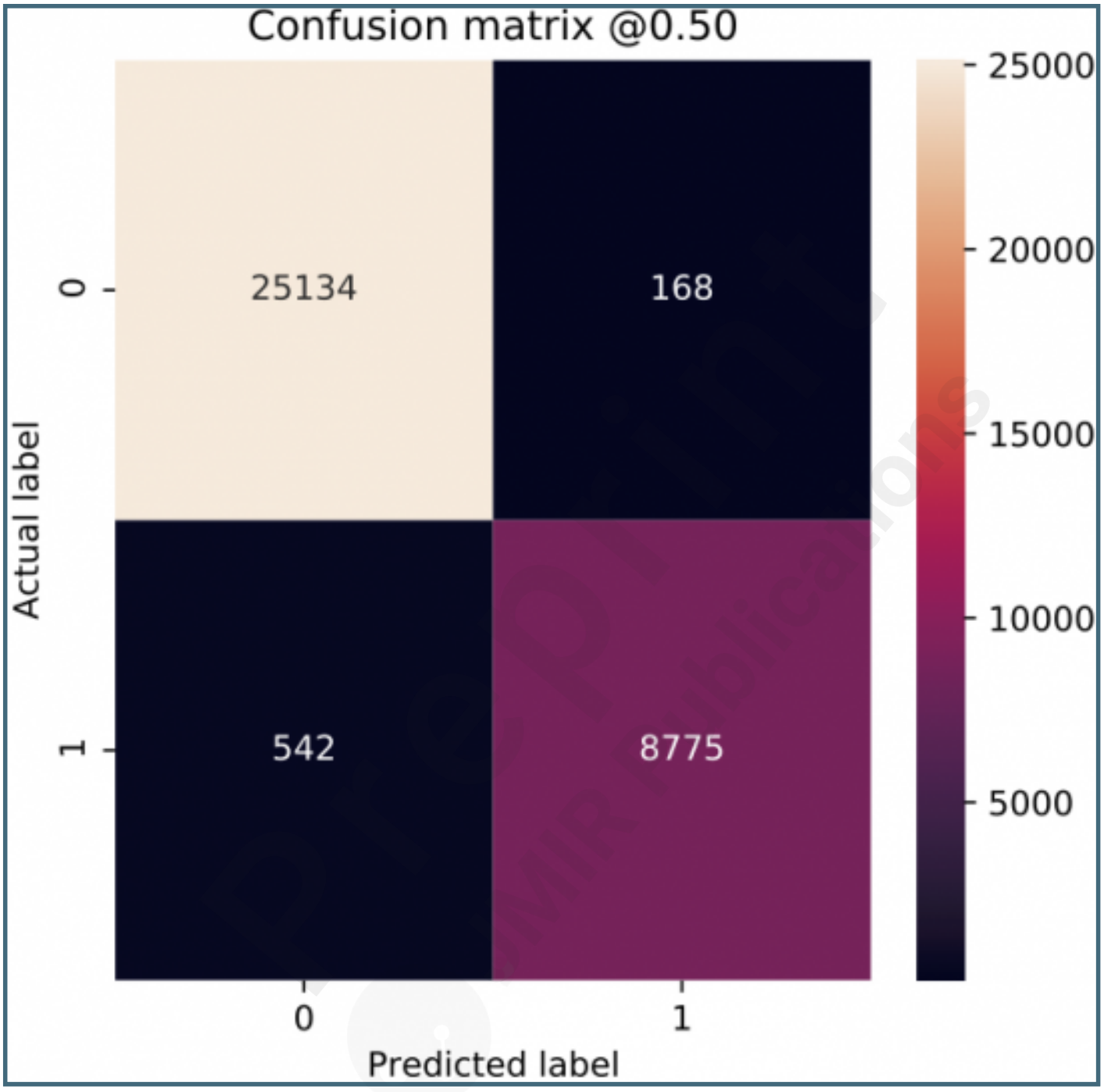
(e) Confusion matrices of Artificial Neural Network.



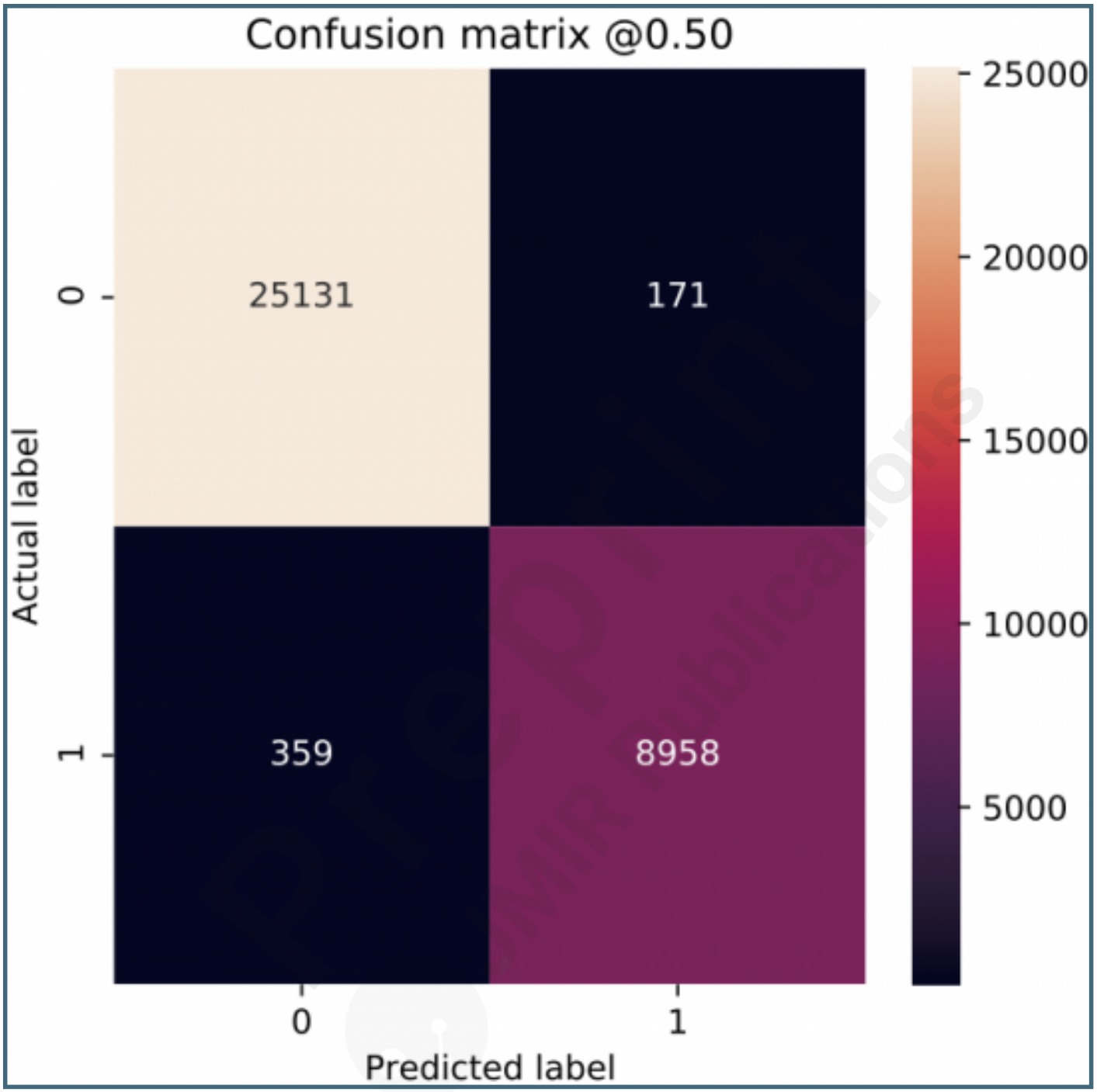
(d) Confusion matrices of Gradient Boosting Trees.



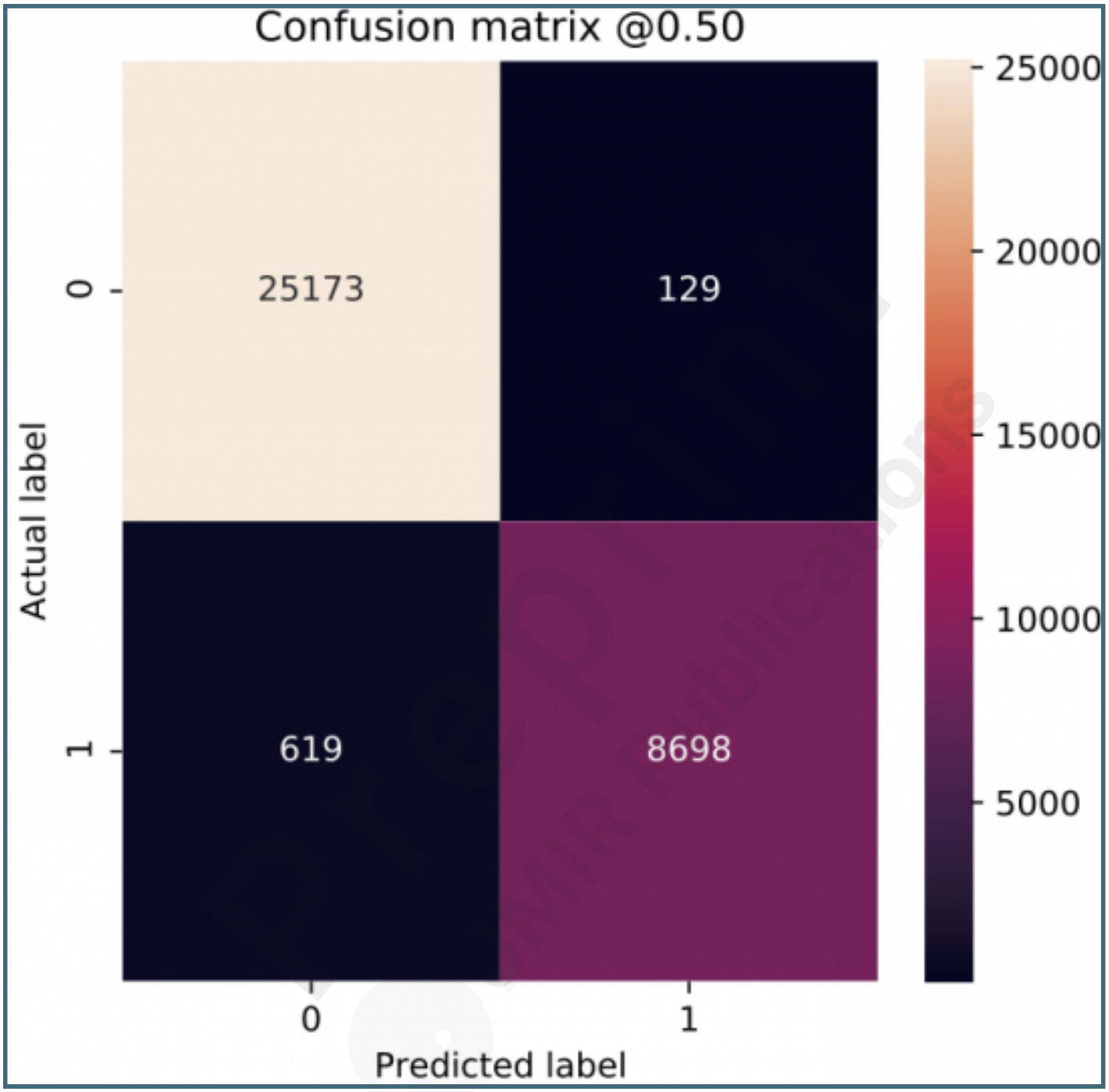
(c) Confusion matrices of Extra Trees.



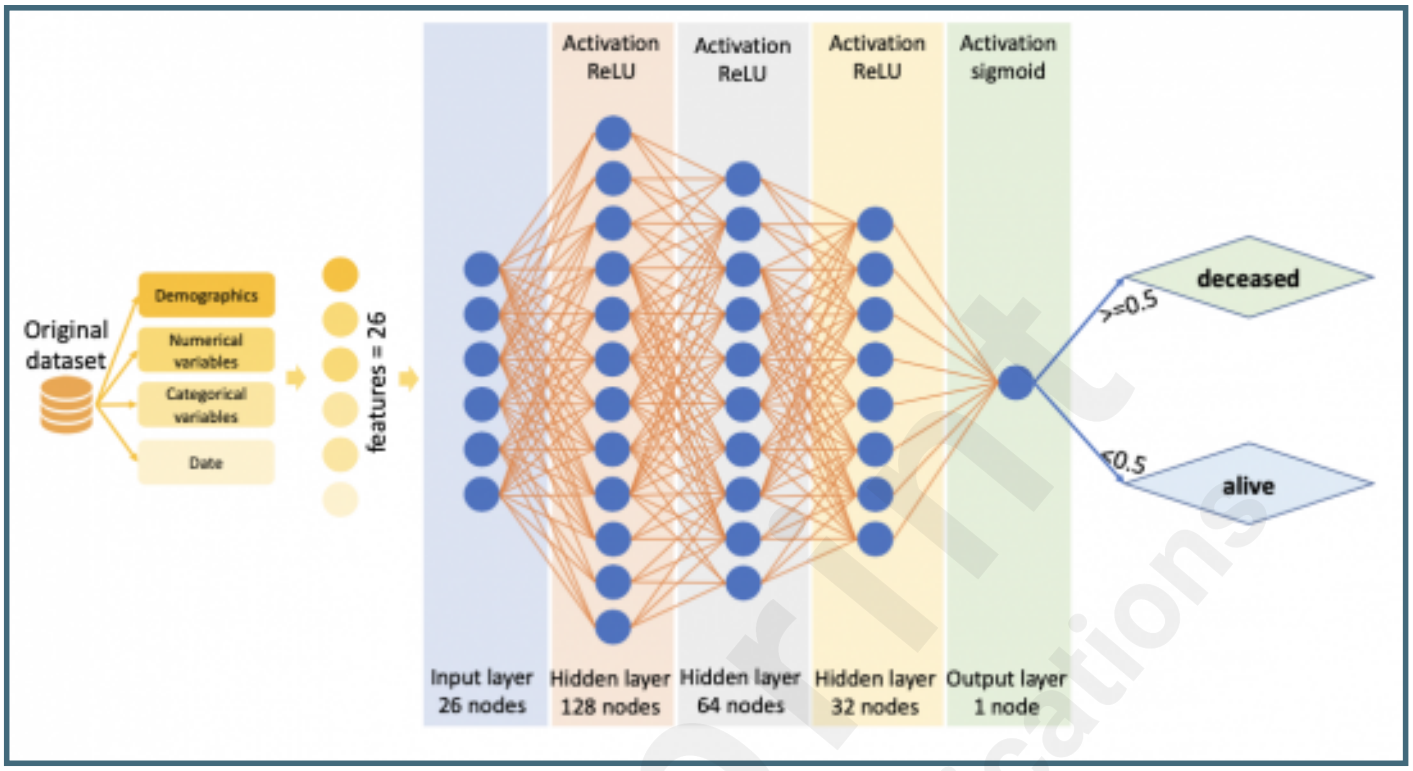
(b) Confusion matrices of Random Forest.



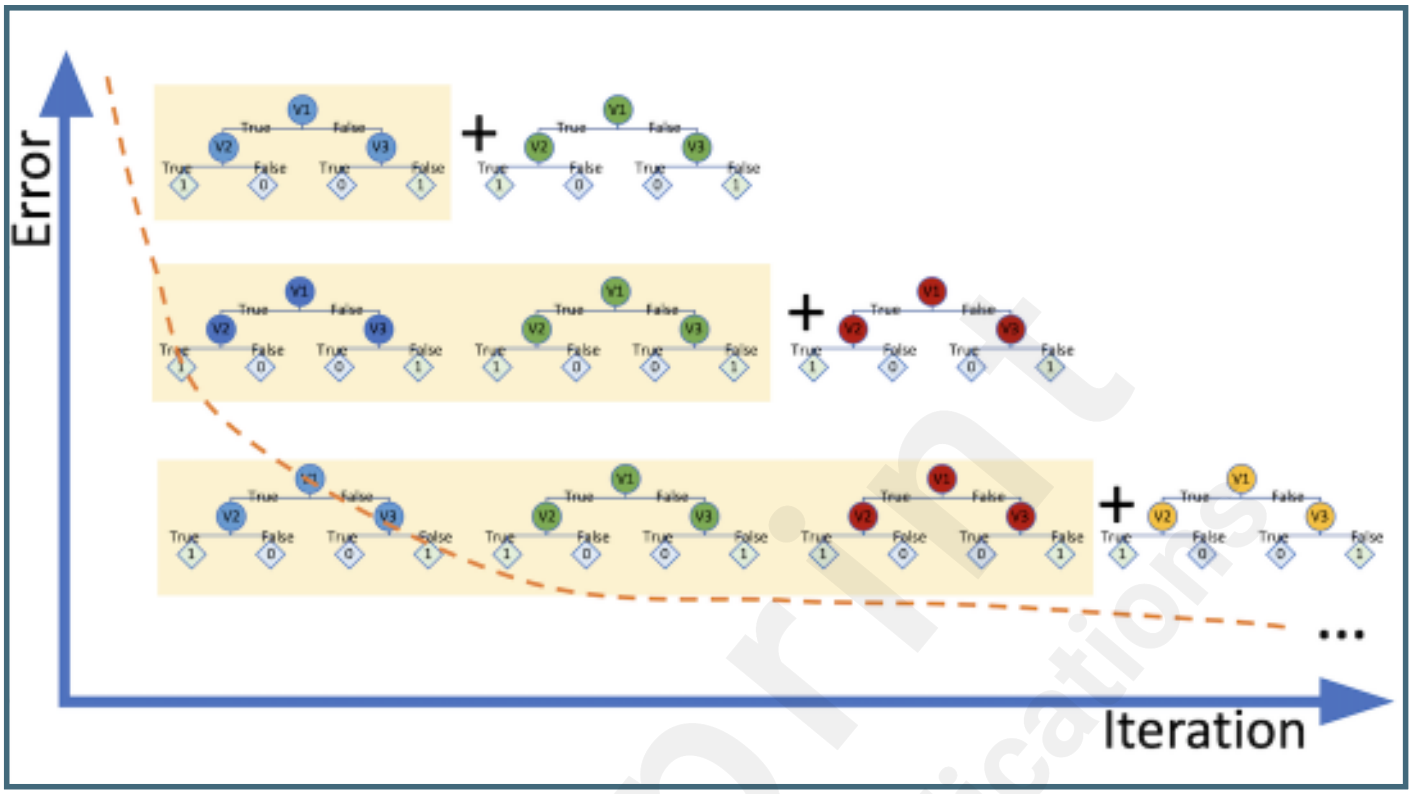
(a) Confusion matrices of Logistic Regression.



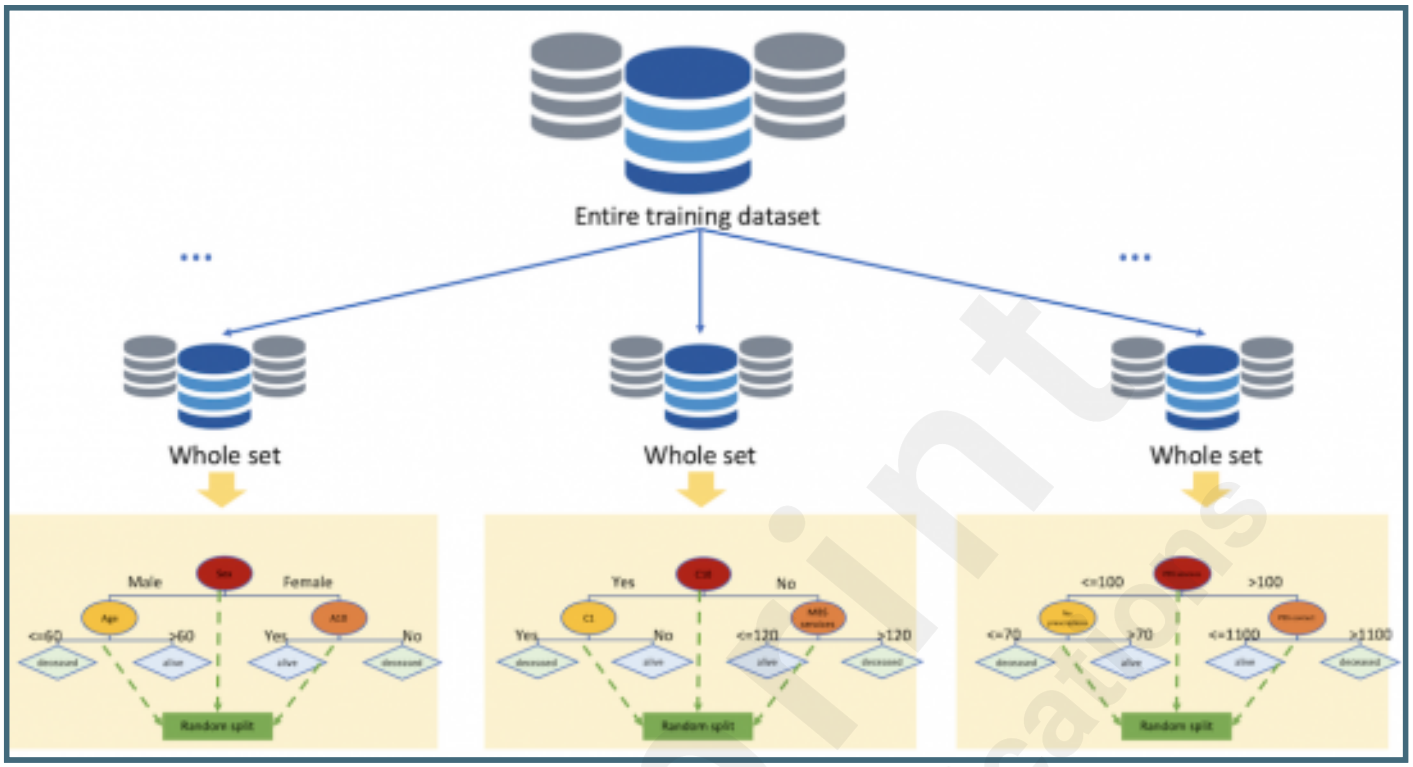
Artificial Neural Network architecture.



Gradient Boosting Trees Workflow.



Extra Trees.



Random Forests.