Science of the Total Environment 771 (2021) 145256

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Towards advancing the earthquake forecasting by machine learning of satellite data

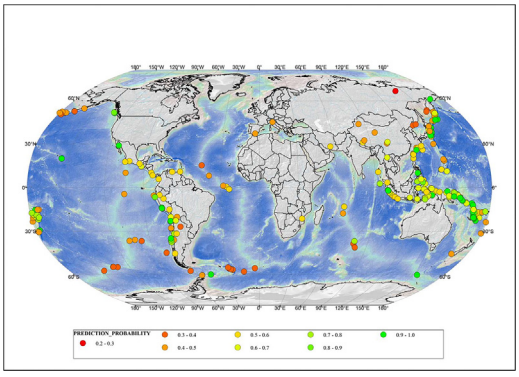
Pan Xiong [a,h], Lei Tong [c], Kun Zhang [i], Xuhui Shen [b,*], Roberto Battiston [e,f], Dimitar Ouzounov [g], Roberto Iuppa [e,f], Danny Crookes [h], Cheng Long [d], Huiyu Zhou [c]

[a] Institute of Earthquake Forecasting, China Earthquake Administration, Beijing, China
[b] National Institute of Natural Hazards, Ministry of Emergency Management of China, Beijing, China
[c] School of Informatics, University of Leicester, Leicester, United Kingdom
[d] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[e] Department of Physics, University of Trento, Trento, Italy
[f] National Institute for Nuclear Physics, the Trento Institute for Fundamental Physics and Applications, Trento, Italy
[g] Center of Excellence in Earth Systems Modeling & Observations, Chapman University, Orange, CA, USA
[h] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom
[i] School of Electrical Engineering, Nantong University, Nantong, China

HIGHLIGHTS

- An AdaBoost-based ensemble framework is proposed to forecast earthquake.
- Infrared and hyperspectral global data between 2006 and 2013 are investigated.
- The framework shows a strong capability in improving earthquake forecasting.
- Our framework outperforms all the six selected baselines on the benchmarking datasets.
- Our results support a Lithosphere-Atmosphere-Ionosphere Coupling during earthquakes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 November 2020
Received in revised form 5 January 2021
Accepted 14 January 2021
Available online 28 January 2021

Editor: Fernando A.L. Pacheco

Keywords:
Earthquake forecasting
Earthquake precursors
Machine learning
Infrared and hyperspectral parameters

ABSTRACT

Earthquakes have become one of the leading causes of death from natural hazards in the last fifty years. Continuous efforts have been made to understand the physical characteristics of earthquakes and the interaction between the physical hazards and the environments so that appropriate warnings may be generated before earthquakes strike. However, earthquake forecasting is not trivial at all. Reliable forecastings should include the analysis and the signals indicating the coming of a significant quake. Unfortunately, these signals are rarely evident before earthquakes occur, and therefore it is challenging to detect such precursors in seismic analysis. Among the available technologies for earthquake research, remote sensing has been commonly used due to its unique features such as fast imaging and wide image-acquisition range. Nevertheless, early studies on pre-earthquake and remote-sensing anomalies are mostly oriented towards anomaly identification and analysis of a single physical parameter. Many analyses are based on singular events, which provide a lack of understanding of this complex natural phenomenon because usually, the earthquake signals are hidden in the environmental noise. The universality of such analysis still is not being demonstrated on a worldwide scale. In this paper, we investigate physical and dynamic changes of seismic data and thereby develop a novel machine learning method, namely Inverse Boosting Pruning Trees (IBPT), to issue short-term forecast based on the satellite data of 1371

* Corresponding author.
E-mail address: shenxh@seis.ac.cn (X. Shen).

https://doi.org/10.1016/j.scitotenv.2021.145256

earthquakes of magnitude six or above due to their impact on the environment. We have analyzed and compared our proposed framework against several states of the art machine learning methods using ten different infrared and hyperspectral measurements collected between 2006 and 2013. Our proposed method outperforms all the six selected baselines and shows a strong capability in improving the likelihood of earthquake forecasting across different earthquake databases.

1. Introduction

Over 500,000 earthquakes are recorded per year. Many of these are undetected or unnoticed because of their small magnitude, while others cause devastation to buildings, bridges and mountains. Earthquake damage mitigation is an active topic in geological research. The majority of earthquakes occur due to the sudden release of stress in the earth's crust that gradually builds up from tectonic movement. However, the response of the crust to the changing stress is non-linear and is dependent on the compression capability of the crust, which is highly variable and complex (Council, 2003). Satellite remote sensing enables us to detect large-range and continuous changes of the near-surface thermal field (Frick and Tervooren, 2019; Niu et al., 2012; Ouzounov et al., 2006; Pulinets et al., 2006; Tramutoli et al., 2005; Tronin, 2007). It can be utilized to monitor thermal anomalies caused in the process of earthquake preparation, which provides useful hints learnt from the measurements for short term and imminent forecasting of earthquakes. Moreover, some researchers believed that thermal anomalies might be related to the variation in the composition of atmospheric gas mixtures above the near-surface fault (Pulinets and Ouzounov, 2018). Microfractures of rock and surface within earthquake regions may expand due to the intensification of pre-earthquake tectonic activities, which helps release underground high-concentration gas, such as H_2, CO, O_3, CO_2 and water vapour to space, and produce additional atmospheric electricity, which stimulates the infrared electromagnetic radiation (Liperovsky et al., 2011).

Existing studies mainly aim at abnormality identification and analysis of a single physical parameter or a specific earthquake; the results of analyzing abnormalities are short of universal, and cannot confidently explain the pre-earthquake multi-parameter anomalies. In recent years, in order to study the pre-earthquake multi-parameter anomalies from the perspective of the energy balance of the Earth, Pulinets and Ouzounov have revised the physical model of lithosphere-atmosphere-Ionosphere-coupling (LAIC) (Ouzounov et al., 2018a; Pulinets and Ouzounov, 2011; Wu et al., 2012) which explains the synergy of different physical processes and variations, known as short-term pre-earthquake anomalies. Other scholars modify LAIC into lithosphere-coversphere-atmosphere (LCA) concept supporting a wide range of remote sensing applications (Wu et al., 2012). It has been pointed out that the effect of seismic crustal stress enables mechanical energy to be transformed to heat energy, and then transmitted to the earth surface through pores in rocks and tiny pre-earthquake ruptures, resulting in enhancement of stress in the rock mass and increase of surface temperature in strike-slip parts. At the same time, the local stress enhancement and the occurrence of micro-earthquakes can cause numerous microcracks and micro-fractures inside the lithosphere, so that the geogases and fluids within the lithosphere, such as H_2, Rn, CO, O_3, CO_2, and water vapour, would spill out along these channels. The increase of geo-gases in the atmosphere stimulates physical processes and chemical reactions from the ground surface up to the troposphere. Along with the effect of electric fields, it stimulates the infrared electromagnetic radiation, and the atmospheric temperature may increase, which can be measured by satellites (Pulinets and Ouzounov, 2018).

With in-depth studies on the LCA coupling model as well as accumulation and utilization of pre-earthquake satellite multi-parameters (Ouzounov et al., 2018a; Wu et al., 2012), some correlation analysis of multi-parameters has been carried out (Table S1). Singh et al. analyzed abnormal variations of sea surface temperature (SST), surface latent heat flux (SLHF), atmospheric temperature and humidity before the earthquake occurred in Sumatra, Indonesia on December 26, 2004 (Singh et al., 2007) and the Wenchuan earthquake on May 12, 2008 (Singh et al., 2010); it was considered that there was a strong land-ocean-atmosphere coupling before earthquakes. Rawat et al. compared anomalies of temperature and long-wave radiation before the earthquakes happened in India and Romania (Rawat et al., 2011). Wu et al. summarized the studies of pre-earthquake anomalies before the earthquake hit L'Aquila, Italy on 22 October 2012, where they selected the parameters of the lithosphere, coversphere, atmospheric and ionized layers to study abnormal synchronization and coupling mechanisms (Wu et al., 2016). Jingfeng et al. studied the abnormal synchronization of multi-parameters such as the latent heat flux, long-wave radiation, atmospheric temperature, humidity and pressure before the Wenchuan Earthquake in 2008 (Jing et al., 2013). Qin et al. utilized multiple satellite parameters to analyze thermal anomalies before the New Zealand Earthquake in 2010–2011, and they studied the physical mechanism behind the earthquake thermal anomaly by investigating the regional tectonics, hydrogeology and meteorological environment (Qin et al., 2012). They also proposed the deviation-timespace-thermal method to analyze the spatiotemporal synchronicity and inherent mechanism among multi-parameters of various thermal anomalies and ionospheric anomalies before the 6.7 magnitude's earthquake occurred in Pu'er (Qin et al., 2013b) and Yushu (Qin et al., 2013a). Ouzounov et al. tested (retrospectively and prospectively) a new approach of integrated satellite and terrestrial framework (ISTF) for detecting atmospheric pre-earthquake signals. The approach is based on a sensor web of coordinated analysis between three physical parameters validated by the LAIC model: OLR (satellite), dTEC (electron concentration in the ionosphere), and atmospheric chemical potential (atmospheric assimilation models). ISTF has been applied for three major earthquakes: M 6.0 Napa of 2014 (USA), M 6.0 Taiwan of 2016, and M 7.0 Kumamoto, Japan of 2016. Molchan's error diagram (MED) for all parameters shows results that are better than random guesses. Prospective tests based on 22 earthquakes over Japan (2014–15) revealed the existence of general temporal-spatial evolution pattern in the atmosphere ahead of the main earthquakes only in cases when a multi-parameter analysis been used (Ouzounov et al., 2018b).

However, the above studies have two several constraints such as (1) earthquake studies are still limited in space and time, and (2) the methodology for multi parameters analysis is non uniform and therefore cannot meet the requirements of the practice of earthquake monitoring. Moreover, the applications of anomaly-evaluation applied to specific earthquakes often are with lack of understanding of the underlying physics, which may cause various or even contradictory conclusions for the same earthquake (Blackett et al., 2011a, 2011b).

With the rapid development of artificial intelligence and machine learning, researchers have made progress in the domain of earth sciences (Sarkar and Mishra, 2018), especially in earthquake forecasting (Asencio-Cortés et al., 2018; Asim et al., 2018a, 2018b; Bergen et al., 2019; Hulbert et al., 2018; Lubbers et al., 2018; Rafiei and Adeli, 2017; Reyes et al., 2013; Rouet-Leduc et al., 2018). At the same time, machine learning are also very effective for spatial remote sensing data handling (Du et al., 2020). For example, artificial intelligence technology may

provide a useful measure for resolving those problems mentioned above. In this paper, a novel Inverse Boosting Pruning Trees (IBPT) based framework is presented for earthquake forecasting, which utilizes satellite data with ten parameters such as infrared sensing, hyperspectral imaging and gas sensing signals collected from worldwide earthquakes during 2006 and 2013. Our proposed method aims to be a general-purpose technology, using the labels generated by time series clustering techniques. This technology simplifies the process of forecasting because the proposed method deals with the sequences of labels instead of real data itself. Four datasets of earthquakes with different magnitudes, collected between 2006 and 2013, are used to forecast earthquakes and compared against other state of the art techniques.

2. Data and processing

2.1. Datasets

Considering the research results reported in the literature (Jing et al., 2013; Qin et al., 2013a; Qin et al., 2012; Qin et al., 2013b; Rawat et al., 2011; Singh et al., 2007; Singh et al., 2010; Wu et al., 2016) and published data, without loss of generality, according to the lithosphere-coversphere-atmosphere (LCA) coupling model (Ouzounov et al., 2018b; Pulinets and Ouzounov, 2011; Wu et al., 2012), we selected ten parameters for the earthquake anomaly analysis. Fig. S1 shows the inherent relations between the selected ten parameters. These parameters were generated from two different satellite data sources (Table S2). The first nine parameters as shown in Table S2 were created from the Atmospheric Infrared Sounder (AIRS) on NASA's spacecraft Aqua and the engaged parameters are recorded with $1.0 \times 1.0°$ resolution, with the frequency of twice per day.

Specifically, surface skin temperature, temperature of the atmosphere at the earth's surface, water vapour mass mixing ratio at the surface, total integrated column ozone burden, retrieved total column co, retrieved total column CH_4, ARIS outgoing longwave radiation flux, and clear-sky outgoing longwave radiation flux can be obtained from the AIRS3STD v6 product (L3 Standard Daily Product processed using only AIRS radiances in Version 6), and land surface temperatures is from the AIRX3SPD v6 product (L3 Support Daily Product processed using AIRS and AMSU radiances in Version 6), which retrieved from MODIS averaged over MYD11C3 (MODIS/Aqua Land-Surface Temperature/Emissivity Monthly Global 0.05Deg CMG) 0.05 degree (~5 km) pixels. The last parameter was obtained from the National Oceanic and Atmospheric Administration (NOAA) Climate Forecasting Center web site (ftp ftp.cpc.ncep.noaa.gov; cd precip/noaa* for OLR directories), which provided original gridded daily Outgoing Longwave Radiation (OLR) data from NCAR with temporal interpolation. The OLR algorithm for analyzing the Advanced Very High-Resolution Radiometer (AVHRR) data is proposed by Gruber and Krueger (Gruber and Krueger, 1984), which integrates the IR data with the wavelengths between 10 and 13 μm. The data is mainly sensitive to the near surface and/or cloud temperatures. The two data sources provide abundant observation data, and all the ten parameters are of the same spatial resolution and time scales. This study covers worldwide events, including land and submarine earthquakes. In total, 1234 earthquakes with magnitudes between 6 and 7, and 137 earthquakes with magnitude 7 and over, are recorded in the study area, which is spread over the period between 2006 and 2013. In order to verify the reliability and improve the robustness of the proposed model, we generate 1371 artificial non-earthquake events, the same amount as the real earthquakes, and stagger the time and place to match when and where the real earthquakes occur. As can be seen, there are millions of measurement values in this study.

We look at the temporal features that are extracted within N days before an earthquake occurs, and attempt to detect earthquake anomalies during these days. As reported by Tronin et al., anomalies are observed around 6–24 days before an earthquake strikes (Gorny et al., 2020). Given that there is no universal standard for the temporal window, we set the temporal window to be 30 by default in our study and determine the best temporal window in the experiments. The spatial feature is M degrees away from the epicenter where the earthquake occurs. By analyzing the NOAA and Moderate Resolution Imaging Spectroradiometer (MODIS) images before earthquakes, Ouzounov et al. found that thermal anomalies occur approximately 2.5° away from the epicenter (Ouzounov et al., 2007). Again, since there is not any agreed standard, the square region with its center at the epicenter and a deviation of 3° was selected as the spatial feature in our study.

2.2. Features generation

The original satellite data is in a format that is not appropriate for the proposed algorithm to proceed and requires data preprocessing. The first task to be completed is to split the dataset: each dataset is carefully split into two contiguous pieces: 80% for training models, which is 01 January 2006 to 12 May 2012, and 20% for testing purposes and final evaluation, which is from May 12, 2012, until December 25, 2013. The next step is the normalization of data, which is used for the clustering process, the normalization of the datasets needs to be conducted to reduce or eliminate data redundancy, Z-Score normalization was performed on the training data, and normalizing the test data with the normalizing parameters used for training data. Moreover, satellite data is affected by factors such as satellite payload interference and space environment, which can cause occasional errors in continuous data. In our study, we use a “sliding time window” implementation combined with time series clustering techniques (Petitjean et al., 2011) to take overlapping 5-days windows in the time series (separated by a 1-day time lag) features, which ensures that there is a coincidence between the formed series, but also avoids the error of continuous single-point data and has stable robustness.

2.2.1. Standard features' generation

For the purpose of comparison, we generate standard features. Firstly, we choose the well-known scalable K-means algorithm (Lloyd, 1982) to classify the data set, the K-means algorithm requires that the user provides the number of clusters to be created. However, this number is a priori unknown, and its selection and evaluations on the results obtained by clustering are crucial. Thus, the most challenging problem of the clustering realm is to select the right number of clusters of the dataset. For these reasons, the Elbow method (Ketchen and Shook, 1996) has been applied to the data in order to determine how many groups the original continuous dataset has to be split into.

Secondly, we perform k-means clustering for different values of k, for instance, by varying k from 5 to 20 clusters. For each k, we calculate the total within-cluster sum of square (wss). Further, an algorithm has been implemented and applied to estimate the location of a bend (knee) within all the calculated wss with the number of clusters k.

Finally, clustering labels with the standard features (Fig. S2) are generated after the above data processing, which will be used as input in the next step.

2.2.2. Time series based features generation

2.2.2.1. Sliding time window implementation. Fig. S3 shows the principle of the sliding time window. The key variables involve the size as well as the sliding step-length of the time windows for data partitioning. The size configuration of the time windows has an essential impact on computational efficiency. In addition, the sliding step also has a critical influence on the clustering results. Too small steps may lead to redundant repetitive computation because of the overlapping of cross-window time series data sets. In our study, the size of the windows is set to 5-days and the sliding step is set to 4-days, which will form overlapping 5-days windows in the time series (separated by a 1-day time lag).

2.2.2.2. Clustering technique. In order to reduce the data complexity of the series data, we cluster each parameter into several disjoint intervals. Clustering is a difficult task due to the great number of possible geometric shapes for the clusters and distances that can be divided.

For the series-based data, we use classical Dynamic Time Warping (DTW) time series clustering algorithm (Petitjean et al., 2011) to carry out clustering analysis to obtain the demarcation point of each segmentation interval. As the parameters of the sliding time window are configured, the window data of the time series will be extracted and stored for each remote sensing parameter. Then, for each collection of remote sensing time series data, all-time series within the same time window will go through clustering, and each process of clustering within the time window will produce several clusters. As a result, the method will generate a number of cluster labels that cover all the time series data.

However, this number of clusters is a priori unknown, and its selection and later evaluations of the results obtained by the clustering are crucial. Thus, one of the most challenging problems in the clustering realm is to select the right number of clusters for the data sets. For these reasons, the Davies-Bouldin Index (Davies and Bouldin, 1979) has been applied to the data in order to determine the optimal number of clusters by varying the number of clusters k from 10 to 20 clusters. For each k, we calculate and compare the corresponding Davies-Bouldin Indexes. While the Davies-Bouldin index reaches its minimum value, the corresponding clusters number k is generally considered as the number of the clusters (Table S3).

Clustering labels with 5 days as the overlapping time series are generated after the above data processing, which will be feed to machine learning algorithms as input features in the next step.

After the data preprocessing is completed, we get four base data sets (Table S4), which are DataSet I: Satellite data of earthquakes of magnitude 7 or greater, DataSet II: Satellite data of earthquakes of magnitude between 6 and 7, DataSet III: the satellite data of the earthquakes of magnitude 7 or greater with the standard features and DataSet IV: the satellite data of earthquakes of magnitudes between 6 and 7 with the standard features. Moreover, we generate two datasets DataSet I-nonoverlap (Dataset II-nonoverlap) exactly as we generate DataSet I (DataSet II) except that we use non-overlapping sliding windows instead of overlapping ones.

3. Methodology

The methodology carried out in this work is shown in a schematic way in Fig. S4. First, a total of 1234 earthquakes with magnitude between 6 and 7, and 137 earthquakes with magnitude 7 and over, covering a global area, are selected for the study. With a combination of different magnitudes of earthquakes and features, two datasets with ten remote sensing multi-parameters are generated.

Each dataset, is carefully split into the training and test data, and the Z-Score normalization was performed as data preprocessing. Then the "sliding window" technique was implemented for the clustering process. The last step of data preprocessing is clustering, which is first performed on training data, and then cluster the testing sets according to the rules of training data, and finally, we generate time series based features.

We benchmarked eight state of the art methods: Frequent Pattern Learning (FPL) (Cheng et al., 2007), Generalized Linear Models (GLM) (Zeger and Karim, 1991), Gradient Boosting Machines (GBM) (Friedman, 2001), Deep Neural Network (DNN) (LeCun et al., 2015), Random Forests (RF) (Geurts et al., 2006), Convolutional Neural Network (CNN) (Krizhevsky et al., 2012), Logistic Regression (LR) (Walker and Duncan, 1967) and Naive Bayes (NB) (Maron, 1961). In our system, which applies Convolutional Neural Networks, we used a network architecture similar to Thibaut Perol's work (Perol et al., 2018) where 4 layers are used in the study as the input samples are too few; this is implemented in Python (v 3.5) with PyTorch (v 0.4.0), and the other eight methods are implemented in R (v 3.4.1) packages: stats, H2O (v 3.18.0.1), arules (v1.6-1) and RevoScaleR (v9.2.1). Since the methods are sensitive to parameter selection, we choose to use the parameters that enable us to obtain the best performance in the experiments. After we have determined the parameters for each method, the performance of each method based on these parameters was compared with the others. We use eight performance measures to evaluate the performance of each method. Benchmarking was performed on a desktop PC equipped with an Intel® Core™ i5-3470 CPU and 16GB of memory.

Each algorithm is trained through the training dataset to produce a parameterized model, which is then applied to the testing dataset for forecasting labels. The models were then applied to every data in each testing dataset, resulting in forecasting label (votes). For the earthquake forecasting, we use majority voting, which is a reasonable decision rule that treats each alternative equally according to May's theorem (May, 1952), and every element makes a forecasting vote for the input data and the final earthquake forecasting is the one that receives more than half of the total votes. All the earthquake forecastings are compared with their corresponding actual values. This may result in certain deviations. Such deviations are evaluated resulting in ROC curves and Area Under the Curve (AUC).

3.1. The proposed machine learning algorithm

The proposed ensemble model is called Inverse Boosting Pruning Trees (IBPT) scheme, which combines an Adaboost variant with pruning decision trees for classification. In this paper, given the flexibility and ease of use of decision tree, we decide to use decision tree as the boosting base estimator. When an entire tree has a high variance, a decision dump often presents a mismatch problem. Therefore, in order to improve the generalization ability of the model, we consider pruning the tree. Our approach consists of two components: (1) Searching for the best-pruned tree. We applied all the training samples, allowed the decision tree to grow fully, and some branches of the tree are then pruned according to the cost-complexity pruning method mentioned in Breiman (2017). (2) Building an inverse boosting structure. We use an inverse boosting structure with the pruned trees and updated weights. Then, repeat the steps until the maximum number of trees is reached. The proposed framework is summarized in Fig. 1.

3.1.1. Discrete Adaboost

In this paper, the classification algorithm is based on the discrete Adaboost algorithm proposed by Freund and Schapire (1996). Algorithm 1.1 proposes a baseline scheme of discrete Adaboost, which combines many simple assumptions (called weak learners) to form a strong classifier for classification tasks (Leshem, 2005). We summarize the algorithm as follows: (1) Train multiple base classifiers in turn, and distribute the weight $ln(\beta_m)$ according to their training error ε_m. (2) A higher weight $w_{m+1,\,i}$ is assigned to the samples misclassified by the previous classifier, which will make the classifier pay more attention to these samples. (3) At last, all the weak classifiers and their weights are integrated to constitute an ensemble Learner $G(X)$. Normally, Adaboost uses a decision dump (a one-level decision tree) as its weak learner. But, due to its simple structure, decision dumps sometimes do not fit training data well, result in that the integrated boosting learner does not perform well in complex data sets (Leshem, 2005). In this paper, IBPT algorithm is recommended, which outperforms the standard Adaboost algorithm in two aspects: (1) it improves the ability of base classifier fitting and generalization. (2) We propose a new boosting structure to reduce the impact of less contributed data.

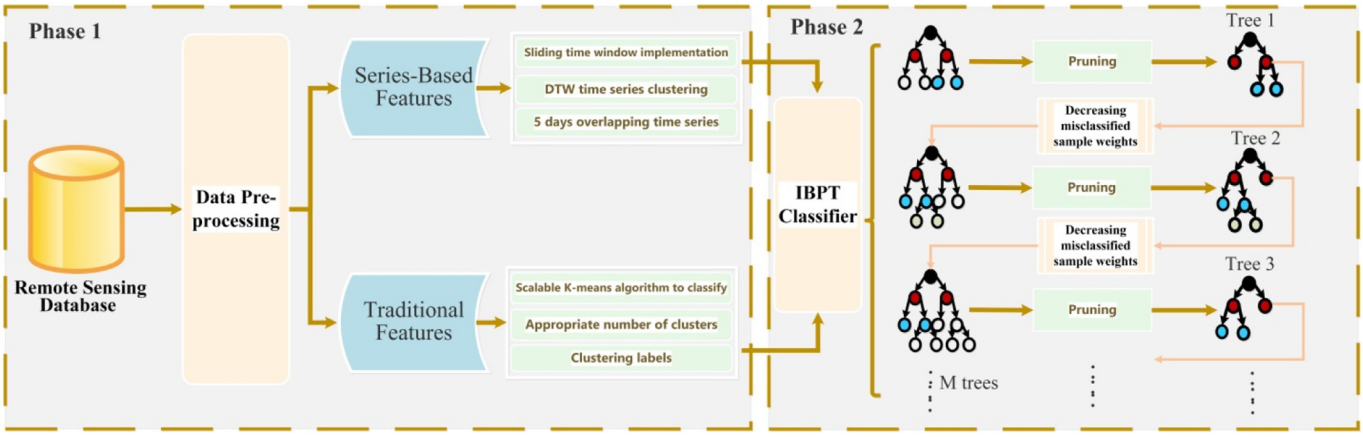


Fig. 1. The flowchart of the proposed IBPT framework.

Algorithm 1 Discrete Adaboost algorithm.
Require: Tree number M, N samples
1: Initialise sample weight distribution $D_m = (w_{mi}\ldots)$, $m = 1, 2, \ldots M$, $i = 1, 2, \ldots N$, and set each sample weight w_{mi} to $\frac{1}{N}$.
2: **for** $m \in (1, M)$ **do**
3: Fit a classifier $G_m(X)$ to the training data with D_m.
4: Let $d_i = 1$ if the i-th case is classified incorrectly, otherwise zero. Then compute training error $\varepsilon_m = \sum_{i=1}^{N} w_{m,i} d_i$.
5: Update sample weight for step $w_{m+1,i} = \frac{w_{m,i}\beta_m^{d_i}}{\sum_{i=1}^{N} w_{m,i}\beta_m^{d_i}}$, where $\beta_m = (1 - \varepsilon_m)/\varepsilon_m$.
6: **end for**
7: **Output** $G(X) = sign\left(\sum_{m=1}^{M} \ln(\beta_m) G_m(X)\right)$.

3.1.2. Inverse boosting pruning trees

In this part, we introduce the IBPT algorithm. When an entire tree has a high variance, decision dump often has a high bias against the training data. Therefore, in order to make the system generalized we decide to trim the tree. In our algorithm, we first use all the training samples and allow the decision tree to grow fully, then pry some branches of the tree using the cost-complexity pruning method mentioned in Breiman (2017), and then use the corrected criteria to evaluate the system performance of the pruned tree and the updated weights. Finally, iterate through the steps until reaching the maximum number of trees. To shape our algorithm, here we declare the symbols used in this formula. Here, we represent the training dataset as $L = X_1, y_1, X_2, y_2. \ldots X_N, y_N$, where, X_n means sample feature vector, y_n means class label and N means sample number. We use $D_m = (w_{m,\,i}, w_{m,\,i+1}\ldots)$, $m = 1, 2, \ldots M$, $i = 1, 2, \ldots, N$ to show the sample weight's distribution in each iteration. M means estimator number (iterations), and in the first iteration of normalization, each sample weight is initialized to $\frac{1}{N}$. In addition, we apply φ_m and $G_{final}(x)$ to represent m-th estimator's weight and the final classifier, respectively.

3.1.2.1. Search for the best pruned tree. In most of the previous boosting algorithms (Chen and Guestrin, 2016; Friedman, 2001; Kokel et al., 2020), except *num_ trees*, *max_depth* and *num_leaves* are two key hyperparameters which affect the classifier's performance significantly. Manually tuning the hyperparameter combinations is a heavy task and it is hard to find the best parameter combinations for different datasets. Therefore, we propose a novel function called resampling weighted pruning to automatically prune redundant leaves and produce robust tree models, where weights are used to establish a relationship between the pruning and boosting practices.

First, we define the original learning sample set as L, and randomly divide it into V subsets, L_v, $v = 1, \ldots, V$, then, the training set of each subset is $L^{(v)} = L - L_v$. T_{max}, represents the tree comes from the original set L, and we build a complete tree in each subset L_v. The decision trees' cost function is defined:

$$\begin{aligned} Gini\,(\{T\},\{w\}) &= \sum_{|\tilde{T}|}\left[Gini\left(\tilde{T}\right)\right] \\ &= \sum_{|\tilde{T}|}\left[1-\sum_{c=1}^{C} P_c^2\right] \\ &= \sum_{|\tilde{T}|}\left[1-\sum_{c=1}^{C}\left(\frac{\sum_{i_o} w_{m,i_c}}{\sum_{i} w_{m,i}}\right)^2\right] \end{aligned} \tag{1}$$

where $|\tilde{T}|$ means leaves' number, C represents the class number, the sample of class c is defined as i_c. The loss of the trees is calculated by summarizing the gini impurity of all the leaves. Because each leaf node contains only the same class samples, the loss of an entire tree is generally zero. However, in the pruning process, $Gini(\{T\},\{w\})$ will increase when the samples of the pruning nodes are combined into their parent node. Since $Gini(\{T\},\{w\})$ always favors large trees, it is not the best method to select a pruned tree. Therefore, we add a penalty term, regularization parameter α and the tree leaves $|\tilde{T}|$ to the cost function. The new equation is shown as follows:

$$R_\alpha(T) = Gini\,(\{T\},\{w\}) + \alpha\cdot|\,\tilde{T}\,| \tag{2}$$

when α is constant and $|\tilde{T}|$ decreases with pruning, the penalty term is the benefit of a smaller tree.

Here, $R_\alpha(T - T_t) - R_\alpha(T)$ defines the variation in the cost function, where T means the complete tree, T_t means the branch with the node at t, so the tree pruned at node t should be $T - T_t$. Next, $R_\alpha(T - T_t)$ is equivalent to the branch at node t, so as to calculate the cost of the pruning on the internal nodes.

$$\begin{aligned} &R_\alpha(T-T_t)-R_\alpha(T)\le 0 \\ &\Rightarrow R_\alpha(t)-R_\alpha(T_t)\le 0 \\ &\Rightarrow Gini\,(\{t\},\{w\})+\alpha-Gini\,(\{T_t\},\{w\})-\alpha\,|\,\tilde{T}_t\,|\le 0 \\ &\Rightarrow \frac{Gini\,(\{t\},\{w\})-Gini\,(\{T_t\},\{w\})}{|\,\tilde{T}_t\,|-1}\le\alpha \end{aligned} \tag{3}$$

where,

$$g(t) = \frac{Gini\,(\{t\},\{w\})-Gini\,(\{T_t\},\{w\})}{|\,\tilde{T}_t\,|-1} \tag{4}$$

when $\alpha \geq g(t)$, the cost value will decrease, and the branch T_t will be pruned. The order in which we pruning the branches begins like this: first, set $\alpha = argmin\, g(t)$ to find the branch, and prune the branch, then repeat the process until the tree is left with the root node. This provides a sequence of pruned trees $\{T_\alpha^{(v)}, \alpha = 0, \ldots\}$ with the associated cost-complexity parameter α.

For α, we use the pruned tree $T_\alpha^{(v)}$ to estimate the $v - th$ subset and obtain the following training error:

$$TE_\alpha^{(v)} = \frac{\sum i_{miss} w_{m,i_{miss}}^{(v)}}{\sum i w_{m,i}^{(v)}} \tag{5}$$

where i_{miss} means the index of the misclassified sample's weight, $w_{m,i}^{(v)}$ means the sample weight of the test set L_v and $TE_\alpha^{(v)}$ means the misclassified rate of set L_v. Therefore, the average misclassified rate of v is obtained as follows:

$$TE_\alpha = \frac{1}{V}\sum_{v=1}^{V} TE_\alpha^{(v)} \tag{6}$$

Meanwhile, we denotes $\alpha^* = argmin_\alpha TE_\alpha$, and through pruning T_{max} till $R_\alpha(T_{max})$ reaches the minimum, we obtained the best pruned tree.

3.1.2.2. Inverse boosting structure. As shown in the 4th and 5th steps of Algorithm 1, Adaboost adopts the training error ε_m as the boosting coefficient of the weak leaner to update the estimator's weights and the sample weights. However ε_m is not a suitable criteria for the pruned trees. Actually, we should give the pruned trees higher estimated weights when they have lower training errors. Therefore, we suggest using a novel boosting structure, which associates the classification outcome with the pruned trees accordingly.

Firstly, we fit the training data L into a complete decision tree and prune it to obtain the best tree structure. Since TE_α truly reflects the result of the pruned tree, we apply TE_α instead of ε_m as the evaluation criteria. The weight of the estimator would be updated as follows,

$$\varphi_m = \ln\frac{1-TE_\alpha}{TE_\alpha} \text{ s.t.} TE_\alpha < \frac{1}{2} \tag{7}$$

The training process stop when $TE_\alpha \geq \frac{1}{2}$, because the current estimator cannot maintain the classification performance at all times.

Next, we propose an inverse structure to update the sample weight. In each iteration, we treat the misclassified samples as 'intractable' items, which may affect the judgment and the robustness of the pruned tree. Therefore, inspired by Tong et al. (2019), the sample weight can be updated:

$$w_{m+1,i} = \frac{w_{mi}}{Z_m}\ exp(\varphi_m I(G_m(X_i), y_i)), i = 1, 2, \ldots N \tag{8}$$

$$D_{m+1} = \left(w_{m+1,1}, w_{m+1,2} \ldots\ldots w_{m+1,N}\right), m = 1, 2, \ldots M \tag{9}$$

In Eq. (8), Z_m is a normalization factor

$$Z_m = \sum_i w_{mi}\ exp(\varphi_m I(G_m(X_i), y_i)) \tag{10}$$

$G_m(X_i)$ is the estimated value of X_i by the pruned tree G_m, y_i means the ground-truth of sample X_i, and $I(G_m(X_i), y_i)$ is defined:

$$I(G_m(X_i), y_i) = \begin{cases} 1, G_m(X_i) = y_i \\ -1, G_m(X_i) \neq y_i \end{cases} \tag{11}$$

The weight of the misclassified samples is decreasing, while the weight of the classified samples is increasing, which reduces the influence of the "intractable" items, which help to optimize the tree building in the subsequent iterations. Then, the next estimator is trained in the dataset L with a new weight distribution D_{m+1}. The training process will execute iteratively until meet the hyperparameter *num_tree* M. The final integrated classifier is:

$$G_{final}(X) = sign\left(\sum_{m=1}^{M} \varphi_m G_m(X)\right) \tag{12}$$

The whole procedure is shown in Algorithm 2.

3.1.3. Hyperparameter optimization

Hyperparameters determine the pre-defined characteristics of a classifier and hyperparameter should be set before the training process begins. Our novel classifier IBPT has few hyperparameters and here we use the Grid search method to tune the hyperparameters of IBPT: number of pruning trees, sample minimum number per leaf and iteration number of the pruning process. More specifically, the minimum number of the samples per leaf controls the pre-defined complexity of the decision trees and the final depth of the trees will be determined by our pruning methods. We create different combinations against these three parameters and use the Grid search method to search for the optimal hyperparameter list of IBPT. The performance of the trained classifiers is compared using five-fold cross-validation: the training data is divided randomly into five subsets and each time we will use four subsets to train a new IBPT classifier with the remaining data as the validation set. Then, comparing the validation results of different IBPT models (using the Grid Search method), the best model is identified and its hyperparameters are selected for the final model in the testing stage.

Algorithm 2 Inverse boosting pruning trees based algorithm.
Require: M-Trees' number, N-Samples number, L-Learning samples, and V-Folds.
1: **function** BEST PRUNED SUBTREE (L, V, D_m).
2: Split the learning samples L into V folds, L_v, $v = 1, 2, \ldots, V$, and grow a max tree T_{max} on L.
3: Test sample set $L^v = L - L_v$.
4: **for** $v \in [1, V]$ do
5: Fit a decision tree to L_v training samples.
6: Subtree sequence $\{T_\alpha^{(v)}, \alpha = 0, \ldots\} \leftarrow R_\alpha(T - T_t) - R_\alpha(T) \leq 0$
Recursively repeat till the tree only has root nodes: 1. Calculate $g(t)$ using Eq. (4). 2. Set $\alpha = argmin\, g(t)$ and prune the branchT_t.
7: Calculate $TE_\alpha^{(v)}$ by Eq. (5).
8: **end for**
9: Compute $TE_\alpha \leftarrow \frac{1}{v}\sum_{v=1}^{V} TE_\alpha^{(v)}$
10: Define$\alpha^* \leftarrow argmin_\alpha TE_\alpha$.
11: The best pruned tree $G_m(X)$ is obtained by pruning T_{max} till $R_\alpha(T_{max})$ becomes minimal.
12: **return**$G_m(X)$, TE_α.
13: **end function**
14:
15: **function** INVERSE BOOSTING(L, V, M, N)
16: Initialise sample weight distribution $D_m = (w_{mi} \ldots), m = 1, 2, \ldots M, i = 1, 2, \ldots N$ and set each sample weight w_{mi} to$\frac{1}{N}$.
17: **for** $m \in (1, M)$do
18: $G_m(X)$, $TE_\alpha \leftarrow$ Best Pruned Subtree (L, V, D_m).
19: Update the estimator weight using Eq. (7).
20: Update each sample's weight $w_{m,i}$ using Eqs. (8) and (9).
21: Preserve Di for the next iteration
22: **end for**
23: **return** Final ensemble classifier $G_{final}(X) \leftarrow sign\left(\sum_{m=1}^{M} W_t^m G_m(X)\right)$.
24: **end function**

3.1.4. Feature importance

The feature importance: The importance of a feature is computed as the (normalized) reduction of the errors brought by that feature. It is also known as the Gini importance. The single node importance *NI* is defined as:

$$NI = Gini(\{Node_{split}\},\{w\}) - Gini(\{Node_{left}\},\{w_{left}\}) - Gini(\{Node_{right}\},\{w_{right}\}) \quad (13)$$

where $Node_{split}$ is the splited non-leaf node in the decision tree, $Node_{right}$ and $Node_{left}$ are the right and left children nodes of $Node_{split}$.

The importance for each feature on a decision tree is then calculated as:

$$feature\ importance = \frac{1}{M}\sum_{m}^{M}\frac{\sum_{f}^{F} NI_f^{(m)}}{\sum_{k}^{K} NI_k^{(m)}} \quad (14)$$

So M is the number of the trees in IBPT, F is the number of non-leaf nodes which employ the target feature to split data and *K* is the total number non-leaf in the *m-th* tree.

3.2. Performance evaluation

In this study, we formulate the task of earthquake forecasting as a binary class classification problem and use eight performance measures, namely, Matthews correlation coefficient (MCC), Hanssen–Kuipers discriminant (R score), the Area Under the Curve (AUC), Specificity, Sensitivity, Accuracy and Precision and the Area Under the Recall-Precision Curve (AURPC), and to test the effectiveness of these measures.

Area Under the Curve is the area under the receiver-operating characteristic (ROC), it is a plot of true positive rate (TPR) against false positive rate (FPR). In practice, AURPC is also often used to test the effectiveness, so AURPC can be a good option for the area under the curve (Davis and Goadrich, 2006).

The Accuracy (ACC) is defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

The Sensitivity (TPR) is defined as:

$$TPR = \frac{TP}{TP + FN} \quad (16)$$

The Specificity (TNR) is defined as:

$$TNR = \frac{TN}{TN + FP} \quad (17)$$

The Precision is defined as:

$$PR = \frac{TP}{TP + FP} \quad (18)$$

The Hanssen–Kuipers discriminant (R score) (Hanssen and Kuipers, 1965) is defined as:

$$\text{R } score = \frac{TP \times TN - FP \times FN}{(TP + FN)(FP + TN)} \quad (19)$$

Except for the metrics mentioned above, which stress on positives, it also used Matthews Correlation Coefficient (MCC) (Matthews, 1975). This coefficient is a balanced measure, and it can measure the correlation between the expected class and the obtained class. MCC is calculated as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (20)$$

where *TP* means the true positives, *TN* means the true negatives, *FP* means the false positives and *FN* means the false negatives, respectively.

4. Results and discussion

4.1. Comparison of results with different machine learning methods

As shown in Tables S5 and S6, all the benchmarking methods were used to construct models for forecasting, based on the two datasets with the generated features. No significant unbalance was found in the training and testing datasets, suggesting the credibility and stability of the forecasting models. The performance metrics of MCC, R score and AUC were used to evaluate the performance of all the methods on the testing datasets. For Dataset I, the MCC of each method ranges from 0.4903 to 0.6581, the R score of each method ranges from 0.4643 to 0.6429, and the AUC of each method ranges from 0.5829 to 0.8718. We found that IBFT was the top performer for Dataset I (MCC = 0.6581, R score = 0.6429 and AUC = 0.8718). The ROC curves of the methods on Dataset I are shown in Fig. 2.

For dataset II, IBPT was still the top performer (MCC =0.5958, R score = 0.5942 and AUC = 0.8683). The ROC curves of the methods on dataset II are shown in Fig. 3. IBPT appears to be robust in accuracy on the two datasets. The MCC and accuracy of IBPT were the best on both the datasets. RF performed quite differently over these two datasets, though.

Furthermore, we observed that IBPT performed better for the earthquakes with larger magnitudes (i.e., Dataset I). This can be explained by the fact that the features selected for larger earthquakes are more supportive in discriminating pre-earthquake perturbations. RF achieves the second-best accuracy in Dataset I (accuracy = 0.7679) and Dataset II (accuracy = 0.785). This suggests that a tree-based classifier is capable of producing better performance. Although MLP and CNN are constructed with four layers, e.g. fully connected and convolutional layers, their performance is worse than those of RF and GBM. This might be due to the fact that deep learning architectures require significantly large training sets (a large number of earthquakes) for system optimization and this is not available in the current research domain with very limited resources.

4.2. Comparison between different features

As illustrated in Tables S7 and S8, DataSet III, and DataSet IV were used to construct the models for forecasting, based on the two datasets (compared with the proposed "sliding window" features based on DataSet I and II). In general, we discover that the datasets with the proposed features (DataSet I and II) lead to better classification performance than the datasets with the standard features (DataSet III and IV) for all the classifiers used. The ROC curves of the methods for DataSet III and IV are shown in Fig. 4 and Figure 5for performance comparison, respectively.

In further analysis, we use IBPT as an example to analyze the experimental results, and generate four datasets based on the spatial and temporal features.

As seen in Table S9, all the benchmarking datasets were used to compare the results of using different features by IBPT. We observe that the datasets with the proposed "sliding window" features lead to better classification outcomes than the datasets with the standard features: The MCC of IBPT on the two datasets with the proposed features are 0.6581 and 0.5958, respectively, and the MCC on the two datasets with the standard features are 0.6429 and 0.5258, respectively. From this observation, we interpret that the datasets with the proposed features enable the earthquake forecast to achieve better accuracy than those with the standard features. This is because the way of generating the proposed features by a "sliding window" style that covers 5 days (or so) observation data extracts sufficient information while reducing data redundancy. As shown in Fig. 6, total integrated column ozone burden, outgoing longwave radiation flux (NOAA) and retrieved total column CO are the most important features rendered by the trained IBPT

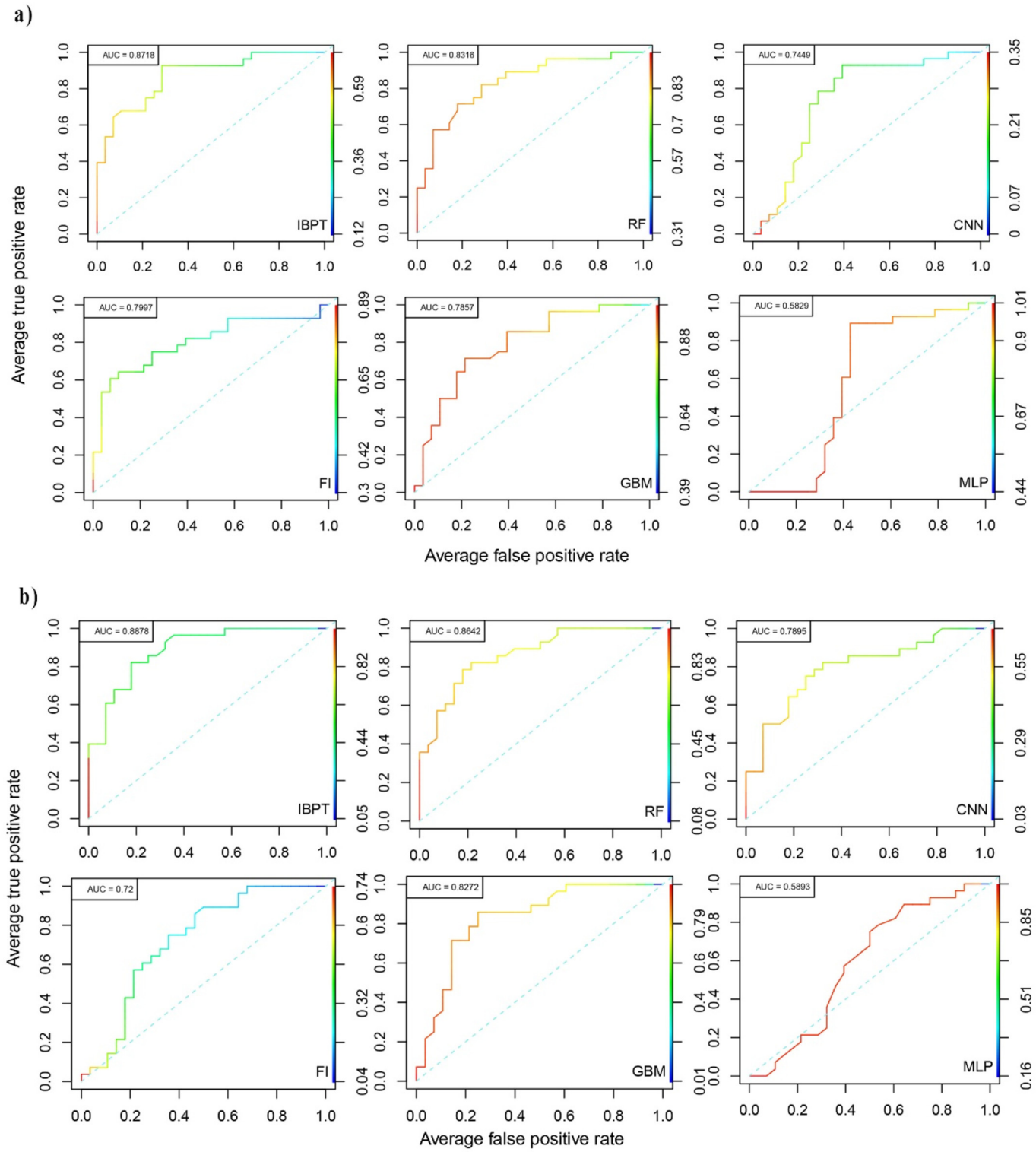


Fig. 2. The ROC curves of the six best-benchmarking methods on the satellite dataset of earthquakes of magnitude 7 or greater with the proposed features a) with aftershocks (Dataset I) b) with aftershocks dropped (Dataset V).

model when it is used to discriminate the earthquake and non-earthquake data.

Tables S10 and S11 present the forecasting performance of the six best-benchmarking methods on the datasets of the proposed features with non-overlapping windows. From Table S10, the MCC ranges from 0.3299 to 0.6075 and the accuracy ranges from 0.6429 to 0.8036 for different methods. It remains true that IBPT was the top performer for the dataset of earthquakes of magnitude 7 or greater with non-overlapping windows. It has also been found that IBPT outperforms all the selected baselines for the dataset of earthquakes of magnitude between 6 and 7 with non-overlapping windows in Table S11. The ROC curves of the methods are shown in Figs. 7 and 8 for performance comparison, respectively. IBPT was still the top performer for the two datasets. Besides, in most cases, the models work better on datasets when overlapping windows are used for generating time series (i.e., DataSet I and DataSet II) compared to the cases with non-overlapping windows used

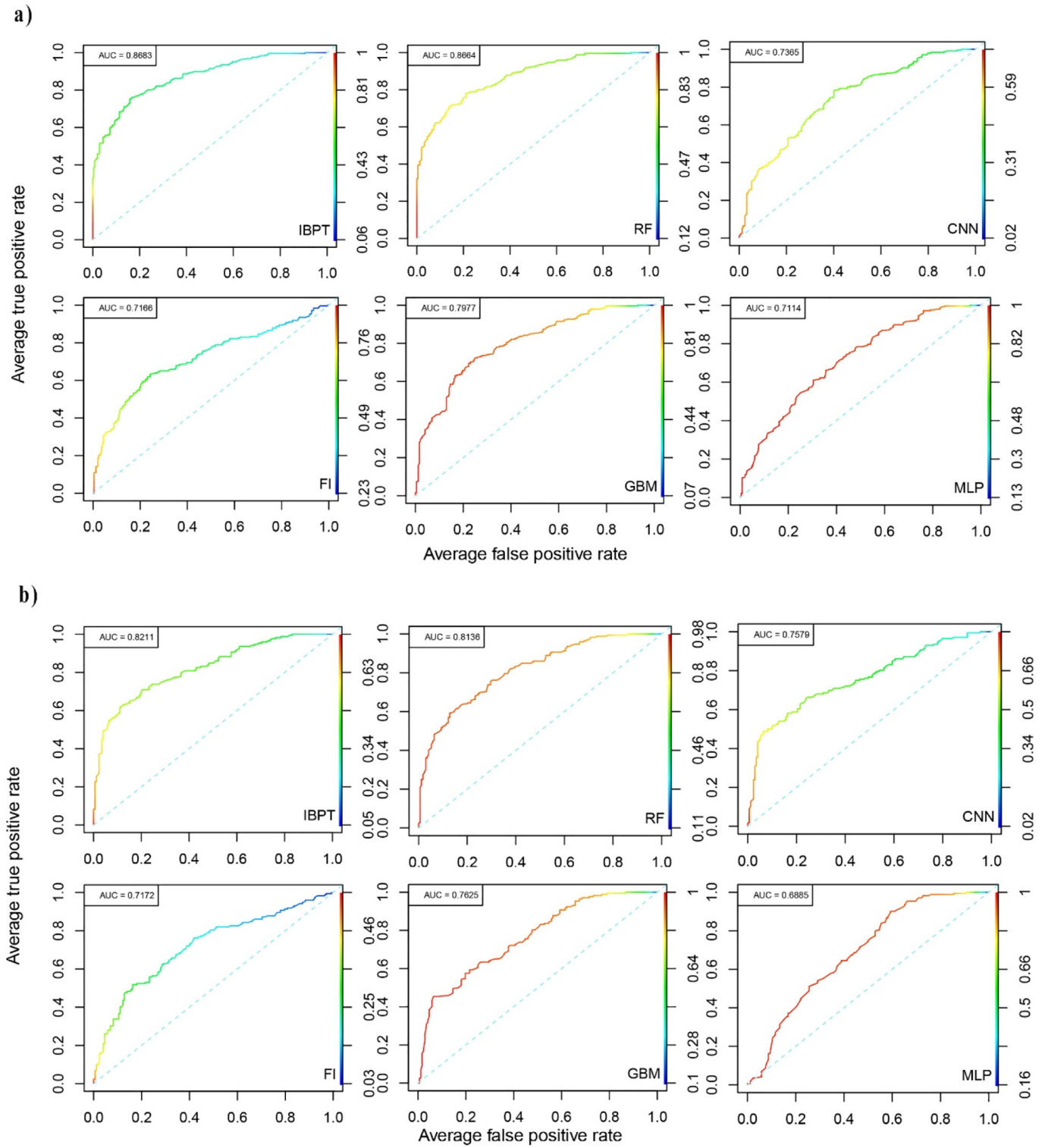


Fig. 3. The ROC curves of the six best-benchmarking methods on the satellite dataset of earthquakes of magnitudes between 6 and 7 with the proposed features a) with aftershocks (Dataset II) b) with aftershocks dropped (Dataset VI).

(i.e., DataSet I-nonoverlap and DataSet II-nonoverlap). A possible reason is that while the features based on non-overlapping windows are less correlated (which have positive effects on the performance), the size of training dataset (generated based on the same raw dataset) are smaller (which have the negative effects on the performance). Besides, IBPT was still the top performer for the two newly considered datasets.

4.3. Considering the aftershock effect

The aftershocks may play an active role on earthquake forecasting. To demonstrate the aftershock effect, we have ruled out the aftershocks and carried out a comparative study. So, it is necessary to delete the data corresponding to aftershocks from the list of

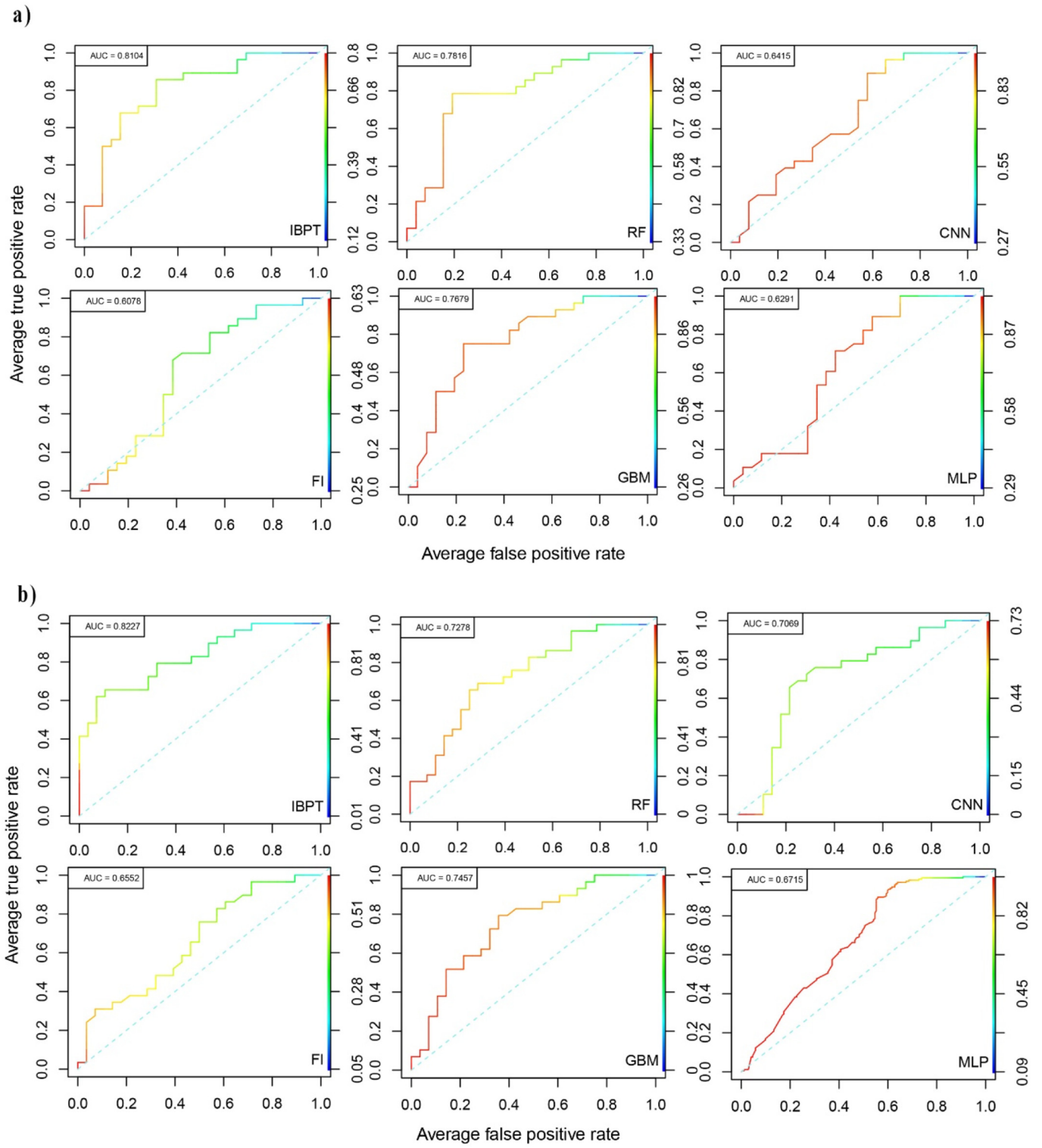


Fig. 4. The ROC curves of the six best-benchmarking methods on the satellite dataset of earthquakes of magnitude 7 or larger with the standard features a) with aftershocks (Dataset III) b) with aftershocks dropped (Dataset VII).

earthquakes (Yan et al., 2017). In our work, we associated an area of 2° × 2° centered on the epicenter for all earthquakes in the list, and to get the result. Practically, we processed the list in the following chronological order: First, select an earthquake (given earthquake) in the list, the setting feature is the time of the earthquake and its related region. Then, from the corresponding data list of the system, to remove any other earthquake occurred in the related area within 30 days after the given earthquake occurred. In our research framework, it is considered that 30 days is the maximum period of anomaly before the earthquake (we set the temporal window to be 30 by default in our study). Finally, we dropped 390 aftershocks. In addition, the data corresponding to the days of the aftershocks are deleted. After these operations, 981 independent earthquakes still remain in the list.

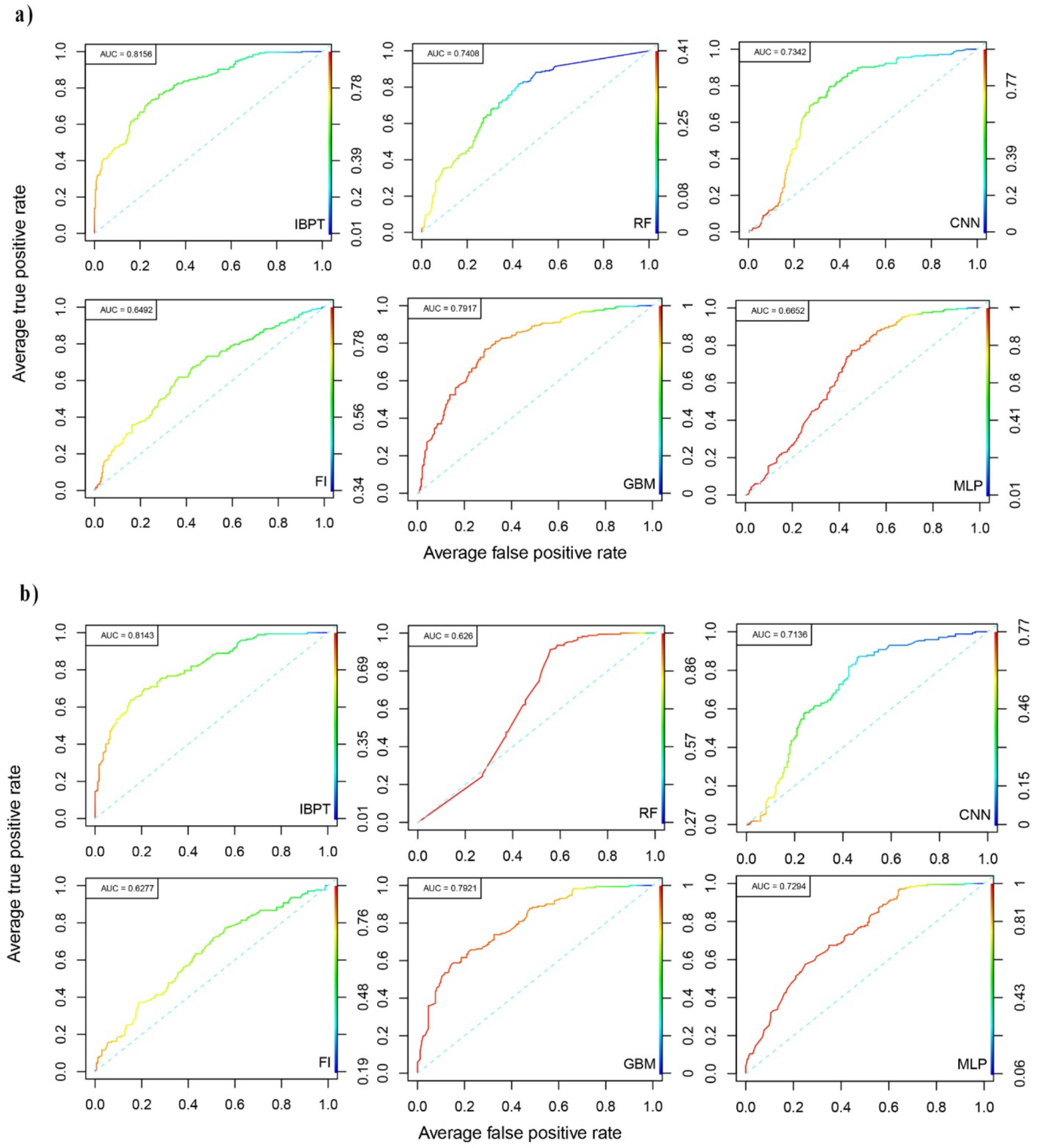


Fig. 5. The ROC curves of the six best-benchmarking methods on the satellite dataset of earthquakes of magnitudes between 6 and 7 with the standard features a) with aftershocks (Dataset IV) b) with aftershocks dropped (Dataset VIII).

According to Table S5, the proposed method IBPT is the top performer for the satellite dataset of earthquakes (with aftershocks dropped) of magnitude 7 or greater with the proposed features (Dataset V) (MCC =0.6429, R score = 0.6429 and AUC = 0.8878). The ROC curves of the methods for Dataset V are shown in Fig. 2b. For Dataset VI (the satellite dataset of earthquakes (with aftershocks dropped) of magnitudes between 6 and 7 with the proposed features), IBPT is still the top performer (MCC = 0.5258, R score = 0.5058 and AUC = 0.8211). The ROC curves of the methods for Dataset VI are shown in Fig. 3b. The MCC and accuracy of IBPT were the best on both the datasets.

As illustrated in Figs. 4b and 5b, DataSet VII (the satellite dataset of earthquakes (with aftershocks dropped) of magnitude 7 or greater with the standard features) and DataSet VIII (the satellite dataset of

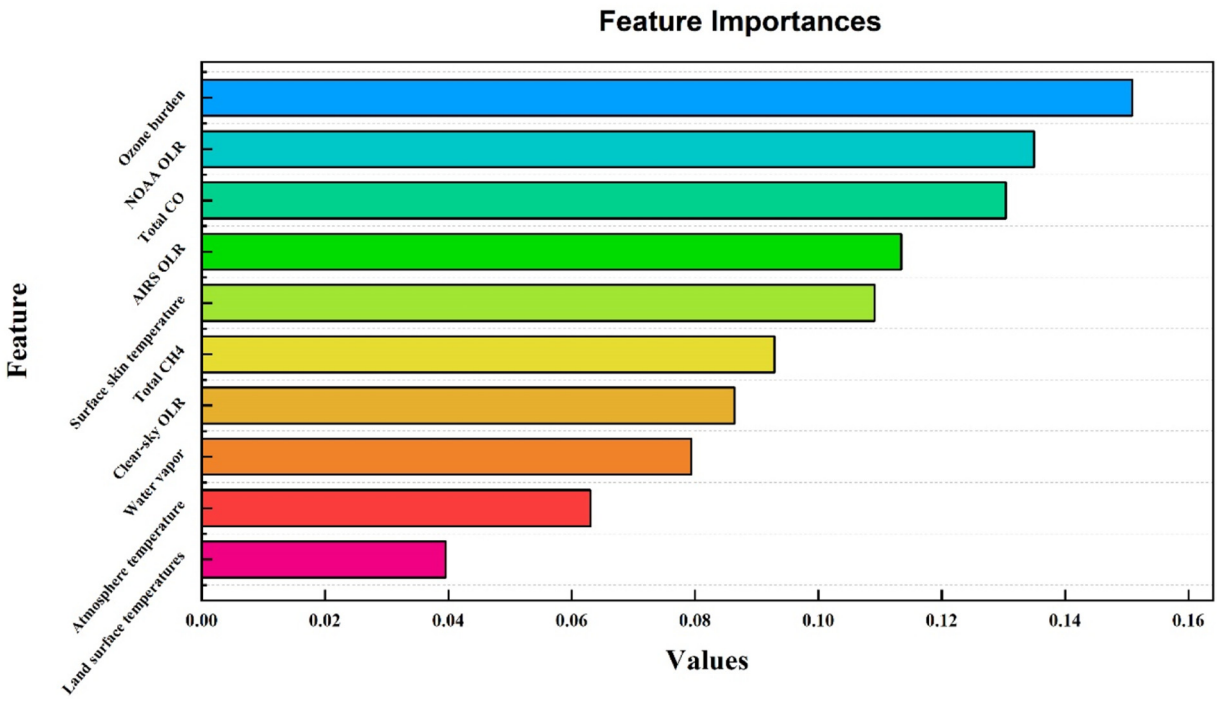


Fig. 6. Features' importance is evaluated by the proposed IBPT model on the satellite dataset of earthquakes of magnitudes with the proposed features. The colours shown here are used merely for better display quality.

earthquakes (with aftershocks dropped) of magnitudes between 6 and 7 with the standard features) were used to construct the models for forecasting, based on the standard features (compared with the proposed "sliding window" features based on DataSet V and DataSet VI). In general, we discover that the datasets with the proposed features (Dataset V and VI) lead to better classification performance than the datasets with the standard features (DataSet VII and VIII) for all the classifiers used. The ROC curves of the methods for DataSet VII and VIII are shown in Figs. 4b and 5b for performance comparison, respectively. IBPT was still the top performer for the two datasets.

In general, although we dropped 390 aftershocks, our work shows that the proposed IBPT framework outperforms the chosen state of the art methods, and becomes the top performer for all the benchmarking datasets. Our work also further proves that the proposed IBPT model in combination with the proposed features performs better than the methods with the standard features, aftershocks had no effect on our result.

4.4. Considering different temporal windows

We have observed that satellite data with a temporal window of 30 days (DataSet V) has good forecasting precision. In order to investigate whether or not our proposed method is capable to predict earthquake with different temporal windows, satellite datasets with temporal windows of 05 days (Dataset IX), 10 days (Dataset X), 15 days (Dataset XI), 20 days (Dataset XII) and 25 days (Dataset XIII) have been generated (shown in Table S12).

Fig. S5a provides the ROC curve of the six datasets with different temporal windows. Table S13 presents the forecasting performance with different temporal windows using IBPT. From Table S13, the MCC ranges from 0.3953 to 0.6429 and the accuracy ranges from 0.6727 to 0.8214 on different datasets. It remains true that by reducing the days of the temporal window, the performances decrease by about 0.24 for MCC and 0.14 for accuracy. That is, the proposed model's performance is worse than that if we reduce the days of the temporal window. Based on these results, we conclude that the choice of the temporal window size is influencing, to a certain extent, the performance of the proposed model, by reducing its capability in predicting earthquakes. Although our proposed method is capable to predict earthquake with different temporal window sizes, it gives the best performance on the dataset with our initial selection of the temporal window of 30 days.

4.5. Considering different spatial windows

Although satellite data with a spatial window with its center at the epicenter and a deviation of 3° (DataSet V) shows a strong capability in earthquake forecasting, satellite datasets of the spatial window with its center at the epicenter and a deviation of 1° (DataSet XVIII), 2° (DataSet XIX), 4° (DataSet XX) and 5° (DataSet XXI) have been generated (Table S12) in order to further find the optimal spatial window.

Table S14 presents the prediction performance with the five datasets of different spatial windows using IBPT. From Table S14, the AUC on each dataset ranges from 0.7389 to 0.8878 and the MCC ranges from 0.4388 to 0.6429. We discover that the best performance is with the dataset with its center at the epicenter and a deviation of 3° (DataSet V, with AUC of 0.8878 and MCC of 0.6429), and by using different distances of the spatial window, the performance of AUC and MCC decreases by about 16.7% and 31.7%, respectively. From this observation, we conclude that although the IBPT model is capable of forecasting earthquake with different spatial window sizes, the dataset with its center at the epicenter and a deviation of 3° enable earthquake forecasting using satellite data to achieve better performance than those with other distance of the spatial window. Fig. S5c provides the ROC curves of the five datasets (including DataSet V) with different spatial windows.

4.6. Considering unbalanced dataset

As the actual earthquake problem is always highly unbalanced, where non-earthquakes instances are always higher as compared to earthquakes. In order to provide the realistic performance overview, we try our proposed method on unbalanced dataset in this section. To investigate whether or not our proposed method is capable to predict

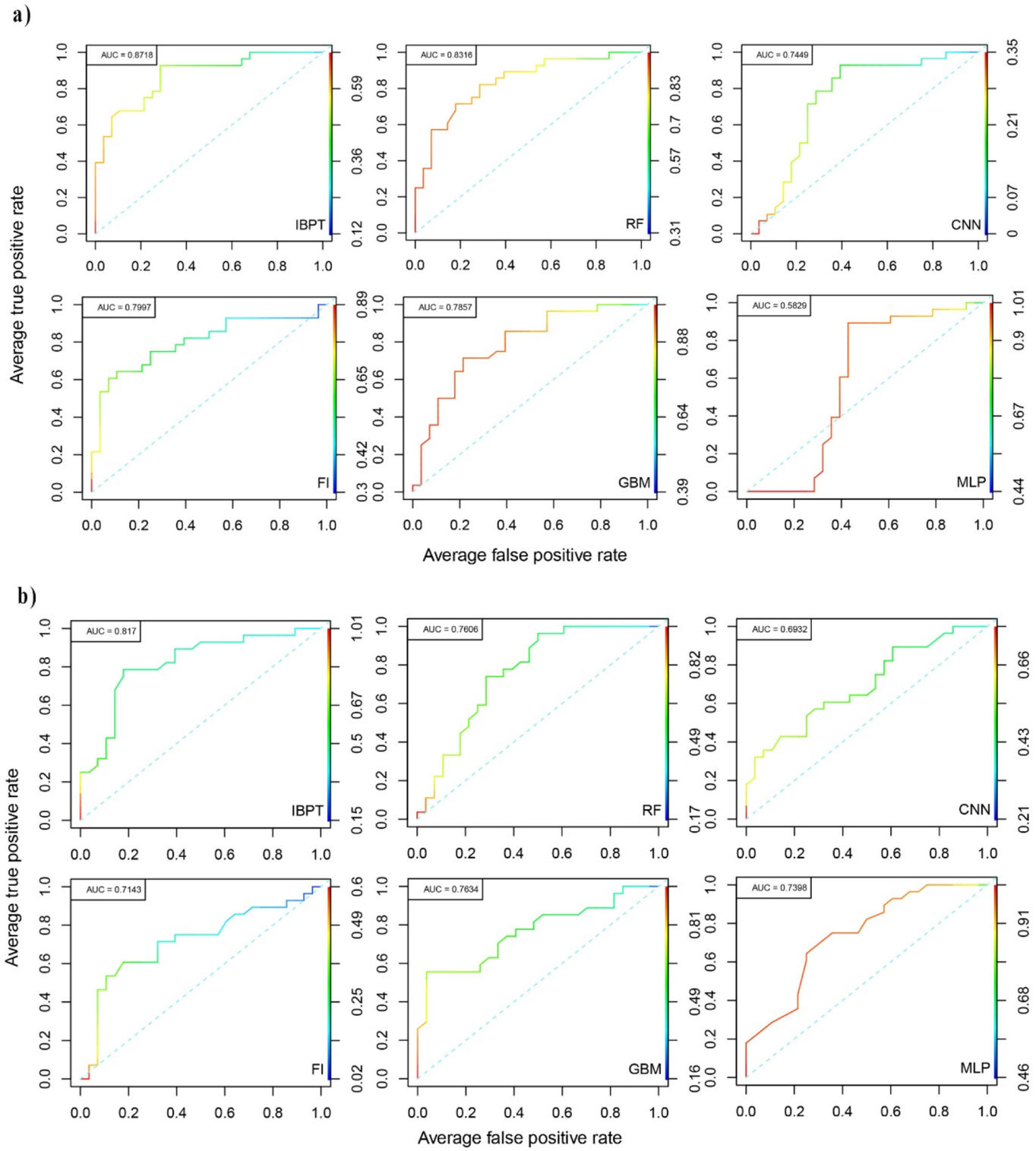


Fig. 7. The ROC curves of the six best-benchmarking methods on the satellite dataset of earthquakes of magnitude 7 or greater with the proposed features a) with overlapped data (Dataset I) and b) with non-overlapped data (Dataset I-nonoverlap).

earthquake with unbalanced datasets, satellite dataset with the positive to negative ratio of 1:2 (Dataset XIV), 1:5 (Dataset XV), 1:10 (Dataset XVI) and 1:15 (Dataset XVII) have been generated (shown in Table S12).

Table S15 illustrates the proposed method's performance on the five datasets. As is shown in, the method has similar performance over the six datasets, e.g., the MCC of the proposed method on all the five datasets is around 0.62 (ranging from 0.6145 to 0.6429), the accuracy of the proposed method on all the five datasets is around 0.83 (ranging from 0.8214 to 0.8588). Fig. S5b shows the ROC curves, we also observe a similar tendency that the performance of our method on the five datasets, suggesting that our method provides satisfactory performance for earthquake forecasting on the unbalanced datasets. Although the five unbalanced datasets are quite different, these results indicate that

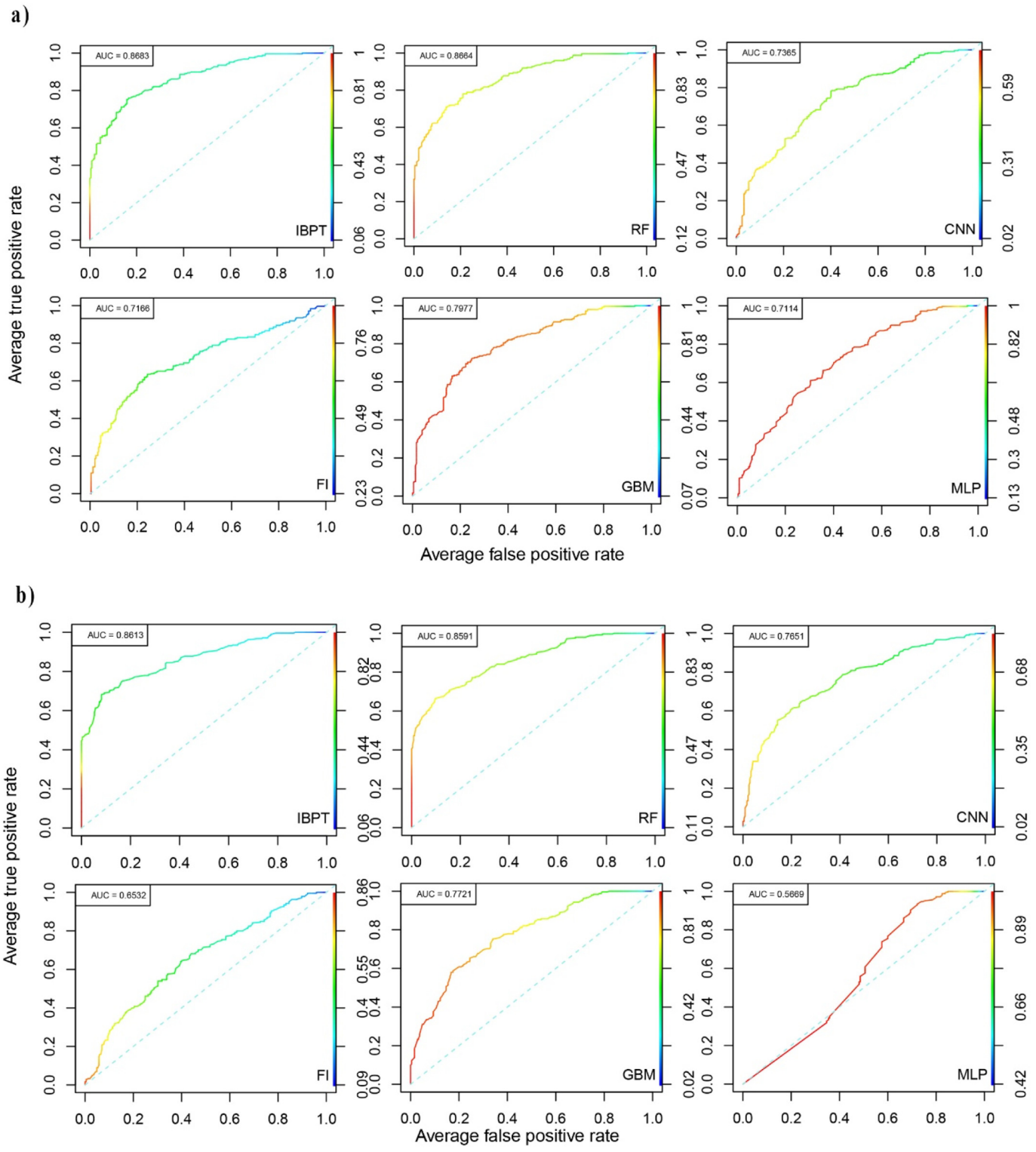


Fig. 8. The ROC curves of the six best-benchmarking methods on the satellite dataset of earthquakes of magnitudes between 6 and 7 with the proposed features a) with overlapped data (Dataset II) b) with non-overlapped data (Dataset II-nonoverlap).

our method is not sensitive to the positive to negative ratio, and that our method can be used to predict earthquakes using unbalanced datasets, and provide good performance.

4.7. Discussion

We summarize the previous studies using machine learning for earthquake prediction and pre-earthquake perturbation analysis from the satellite data, shown in Table S16. Through the performance comparison among these studies, the result shows that among those methods, the IBPT method outperforms the others, i.e., it gives the best performance on all the benchmarking datasets. Moreover, since earthquake is a small probability event, the actual earthquake problem is always highly unbalanced. In this paper, we mainly use Matthew's Correlation Measure (MCC) to evaluate the performance. By comparison, the best MCC of our method IBPT can achieve 0.6581.

There are multiple sources of uncertainty associated with the infrared and hyperspectral satellite data and methods for earthquake forecasting. Uncertainty of data precision is caused by random effects in the data, and uncertainty of data accuracy is caused by systematic effects (Smith et al., 2015). Specifically, uncertainties of infrared OLR data can attribute to several factors, including the magnitudes and the degrees of persistency of the regional OLR diurnal and interannual variations (Gruber et al., 2007), the goodness of fit of the models (Moy et al., 2010), surface emissivity (Clerbaux et al., 2020), the AVHRR OLR's precision is with particularly large uncertainties in the deserts and elevated regions (Gruber et al., 2007). The uncertainty of surface skin temperature obtained from ARIS could be due to the short periods of the satellite based temperature records (Kang et al., 2015). Land surface temperatures uncertainties are affected by the methodologies for the surface retrieval and emissivity first guess (Hulley and Hook, 2012). Pagano et al. (2020) give detailed discussion of measurement uncertainties of AIRS L1B radiances, and note that large uncertainty in the modules at low scene temperatures due to the larger polarization uncertainty, and the larger errors associated with the emissivity degradation in the shorter wavelength modules. Moreover, the wide variety of cloud complexity is an important factor in uncertainty for AIRS error estimation (Kahn et al., 2015; Wong et al., 2015). Furthermore, as parameter tuning of the proposed IBPT model is time-consuming and challenging, we cannot guarantee that optimized parameters were obtained for the models trained in each dataset, though most cases were covered through the grid search method employed in our study. Still, this introduced additional uncertainty to the earthquake forecasting of the proposed model.

One limitation of IBPT is that it has significant computational complexity because of the pruning methods. More specifically, in each iteration, the base tree of the IBPT needs to grow fully with the training data then iteratively prune leaf nodes from bottom to top. This process improves the fitting and generalization ability of the base trees but reduce its training speed. This problem can be handled by deploying more computing resources such as multiple CPUs or GPUs.

5. Conclusions

Hyperparametric optimization and cross-validation are used in our proposed system for earthquake forecasting, which allows us to find the best parameters for our model. In this way, we can perform model selection with high confidence, assuring that a robust model is selected and used. By comparison, our method IBPT improves by 16% in MCC (from 0.5657 to 0.6581) and more than 10% in R score (from 0.5357 to 0.6429) over the next-best CNN. It can be concluded that the proposed IBPT framework outperforms the chosen state of the art methods, and becomes the top performer for all the benchmarking datasets. Moreover, we could observe that infrared and hyperspectral satellite measurements in the circular region with its center at the epicenter and a radius of 3° and 30 days before the times of the shocks are more reasonable in earthquake forecasting. Our work also further proves that aftershocks had no effect on the result performed by the proposed IBPT model.

Our work also indicates that the proposed IBPT model in combination with the proposed features performs better than the methods with the standard features. The proposed time series based are able to help improve the accuracy of the earthquake forecasting task. It can significantly improve performance on the satellite dataset of earthquakes of magnitudes between 6 and 7, the MCC of IBPT with the proposed features improves by 13.3%, which shows the proposed IBPT scheme is effective to some extent on the datasets of a relatively large sample size. It can also be inferred from the feature importance analysis that total integrated column ozone burden, outgoing longwave radiation flux (NOAA) and retrieved total column CO are the most important features rendered by the trained IBPT model when it is used to discriminate seismic and non-seismic data.

Our work shows that the use of big satellite data analytics with machine learning is capable of successfully improving the likelihood of earthquake forecasting. In particular, earthquakes may be forecasted to some extent using the proposed IBPT framework with the proposed spatial and temporal features.

CRediT authorship contribution statement

Pan Xiong: Conceptualization, Methodology, Software. **Lei Tong:** Methodology, Software. **Kun Zhang:** Methodology, Software. **Xuhui Shen:** Methodology, Supervision. **Roberto Battiston:** Conceptualization, Methodology, Supervision. **Dimitar Ouzounov:** Writing – review & editing. **Roberto Iuppa:** Writing – review & editing. **Danny Crookes:** Writing – review & editing. **Cheng Long:** Conceptualization, Methodology, Writing – review & editing. **Huiyu Zhou:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China under Grant No. 2018YFC1503505, and in part by the Special Fund of the Institute of Earthquake Forecasting, China Earthquake Administration under Grant 2020IEF0510 and Grant 2020IEF0705.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.145256.

References

Asencio-Cortés, G., Morales-Esteban, A., Shang, X., Martínez-Álvarez, F., 2018. Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure. Comput. Geosci. 115, 198–210.

Asim, K.M., Idris, A., Iqbal, T., Martínez-Álvarez, F., 2018a. Earthquake prediction model using support vector regressor and hybrid neural networks. PLoS One 13, e0199004.

Asim, K.M., Idris, A., Iqbal, T., & Martínez-Álvarez, F. (2018b). Seismic indicators based earthquake predictor system using genetic programming and AdaBoost classification. Soil Dyn. Earthq. Eng., 111, 1–7.

Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid earth geoscience. Science 363.

Blackett, M., Wooster, M.J., & Malamud, B.D. (2011a). Correction to "Exploring land surface temperature earthquake precursors: A focus on the Gujarat (India) earthquake of 2001". Geophysical Research Letters, 38, n/a-n/a.

Blackett, M., Wooster, M.J., Malamud, B.D., 2011b. Exploring land surface temperature earthquake precursors: a focus on the Gujarat (India) earthquake of 2001. Geophys. Res. Lett. 38.

Breiman, L., 2017. Classification and Regression Trees. Routledge.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco, California, USA, pp. 785–794.

Cheng, H., Yan, X., Han, J., & Hsu, C. (2007). Discriminative Frequent Pattern Analysis for Effective Classification. In, 2007 IEEE 23rd International Conference on Data Engineering (pp. 716-725)

Clerbaux, N., Akkermans, T., Baudrez, E., Velazquez Blazquez, A., Moutier, W., Moreels, J., Aebi, C., 2020. The climate monitoring SAF outgoing longwave radiation from AVHRR. Remote Sens. 12.

Council, N.R, 2003. Living on an Active Earth: Perspectives on Earthquake Science. National Academies Press.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 224–227.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In, Proceedings of the 23rd international conference on Machine learning (pp. 233-240): ACM.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: a review. Journal of Geovisualization and Spatial Analysis 4.
Freund, Y., & Schapire, R.E. (1996). Experiments with a new boosting algorithm. In, icml (pp. 148-156): Citeseer.
Frick, A., & Tervooren, S. (2019). A Framework for the Long-term Monitoring of Urban Green Volume Based on Multi-temporal and Multi-sensoral Remote Sensing Data. Journal of Geovisualization and Spatial Analysis, 3.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.
Gorny, V., Salman, A., Tronin, A., & Shilin, B. (2020). Terrestrial outgoing infrared radiation as an indicator of seismic activity. arXiv preprint arXiv:2001.11762.
Gruber, A., Krueger, A.F., 1984. The status of the NOAA outgoing longwave radiation data set. Bull. Am. Meteorol. Soc. 65, 958–962.
Gruber, A., Lee, H.-T., Ellingson, R.G., Laszlo, I., 2007. Development of the HIRS outgoing longwave radiation climate dataset. J. Atmos. Ocean. Technol. 24, 2029–2047.
Hanssen, A., & Kuipers, W. (1965). On the relationship between the frequncy of rain and various mateorological parameters: with reference to the problem of objective forecasting. Koninklijk Nederlands Meteorologisch Instituut
Hulbert, C., Rouet-Leduc, B., Johnson, P.A., Ren, C.X., Rivière, J., Bolton, D.C., Marone, C., 2018. Similarity of fast and slow earthquakes illuminated by machine learning. Nat. Geosci. 12, 69–74.
Hulley, G.C., Hook, S.J., 2012. A Radiance-Based Method for Estimating Uncertainties in the Atmospheric Infrared Sounder (AIRS) Land Surface Temperature Product. Atmospheres, Journal of Geophysical Research, p. 117.
Jing, F., Shen, X.H., Kang, C.L., Xiong, P., 2013. Variations of multi-parameter observations in atmosphere related to earthquake. Nat. Hazards Earth Syst. Sci. 13, 27–33.
Kahn, B.H., Schreier, M.M., Yue, Q., Fetzer, E.J., Irion, F.W., Platnick, S., Wang, C., Nasiri, S.L., L'Ecuyer, T.S., 2015. Pixel-scale assessment and uncertainty analysis of AIRS and MODIS ice cloud optical thickness and effective radius. Journal of Geophysical Research: Atmospheres 120, 11,669–611,689.
Kang, H.J., Yoo, J.M., Jeong, M.J., Won, Y.I., 2015. Uncertainties of satellite-derived surface skin temperatures in the polar oceans: MODIS, AIRS/AMSU, and AIRS only. Atmospheric Measurement Techniques 8, 4025–4041.
Ketchen, D.J., Shook, C.L., 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strateg. Manag. J. 17, 441–458.
Kokel, H., Odom, P., Yang, S., & Natarajan, S. (2020). A Unified Framework for Knowledge Intensive Gradient Boosting: Leveraging Human Experts for Noisy Sparse Domains. In, AAAI (pp. 4460–4468).
Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In, Advances in neural information processing systems (pp. 1097-1105).
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Leshem, G. (2005). Improvement of adaboost algorithm by using random forests as weak learner and using this algorithm as statistics machine learning for traffic flow prediction. Research proposal for a Ph. D. Research proposal for a Ph. D. thesis, the Hebrew university of Jerusalem.
Liperovsky, V.A., Meister, C.V., Mikhailin, V.V., Bogdanov, V.V., Umarkhodgaev, P.M., Liperovskaya, E.V., 2011. Electric field and infrared radiation in the troposphere before earthquakes. Natural Hazards and Earth System Science 11, 3125–3133.
Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137.
Lubbers, N., Bolton, D.C., Mohd-Yusof, J., Marone, C., Barros, K., Johnson, P.A., 2018. Earthquake catalog-based machine learning identification of laboratory fault states and the effects of magnitude of completeness. Geophys. Res. Lett. 45, 13,269–213,276.
Maron, M.E., 1961. Automatic indexing: an experimental inquiry. J. ACM 8, 404–417.
Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405, 442–451.
May, K.O., 1952. A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision. Journal of the Econometric Society, Econometrica, pp. 680–684.
Moy, L.A., Knuteson, R.O., Tobin, D.C., Revercomb, H.E., Borg, L.A., & Susskind, J. (2010). Comparison of measured and modeled outgoing longwave radiation for clear-sky ocean and land scenes using coincident CERES and AIRS observations. Journal of Geophysical Research, 115.
Niu, Q., Cheng, W., Liu, Y., Xie, Y., Lan, H., Cao, Y., 2012. Risk assessment of secondary geological disasters induced by the Yushu earthquake. J. Mt. Sci. 9, 232–242.
Ouzounov, D., Bryant, N., Logan, T., Pulinets, S., Taylor, P., 2006. Satellite thermal IR phenomena associated with some of the major earthquakes in 1999–2003. Physics and Chemistry of the Earth, Parts A/B/C 31, 154–163.
Ouzounov, D., Liu, D., Chunli, K., Cervone, G., Kafatos, M., Taylor, P., 2007. Outgoing long wave radiation variability from IR satellite data prior to major earthquakes. Tectonophysics 431, 211–220.
Ouzounov, D., Pulinets, S., Hattori, K., Taylor, P., 2018a. Pre-Earthquake Processes: A Multidisciplinary Approach to Earthquake Prediction Studies. John Wiley & Sons.
Ouzounov, D., Pulinets, S., Liu, J.-Y., Hattori, K., Han, P., 2018b. Multiparameter assessment of pre-earthquake atmospheric signals. Pre-Earthquake Processes 339–359.
Pagano, T.S., Aumann, H.H., Broberg, S.E., Cañas, C., Manning, E.M., Overoye, K.O., & Wilson, R.C. (2020). SI-Traceability and Measurement Uncertainty of the Atmospheric Infrared Sounder Version 5 Level 1B Radiances. Remote Sensing, 12
Perol, T., Gharbi, M., Denolle, M., 2018. Convolutional neural network for earthquake detection and location. Sci. Adv. 4, e1700578.
Petitjean, F., Ketterlin, A., Gançarski, P., 2011. A global averaging method for dynamic time warping, with applications to clustering. Pattern Recogn. 44, 678–693.
Pulinets, S., Ouzounov, D., 2011. Lithosphere–atmosphere–ionosphere coupling (LAIC) model – an unified concept for earthquake precursors validation. J. Asian Earth Sci. 41, 371–382.
Pulinets, S., & Ouzounov, D. (2018). The Possibility of Earthquake Forecasting. In, Learning from nature: IOP Publishing.
Pulinets, S., Ouzounov, D., Karelin, A., Boyarchuk, K., Pokhmelnykh, L., 2006. The physical nature of thermal anomalies observed before strong earthquakes. Physics and Chemistry of the Earth, Parts A/B/C 31, 143–153.
Qin, K., Wu, L.X., De Santis, A., Meng, J., Ma, W.Y., Cianchini, G., 2012. Quasi-synchronous multi-parameter anomalies associated with the 2010–2011 New Zealand earthquake sequence. Nat. Hazards Earth Syst. Sci. 12, 1059–1072.
Qin, K., Wu, L., Zheng, S., Liu, S., 2013a. A deviation-time-space-thermal (DTS-T) method for global earth observation system of systems (GEOSS)-based earthquake anomaly recognition: criterions and quantify indices. Remote Sens. 5, 5143–5151.
Qin, K., Wu, L.X., Ouyang, X.Y., Shen, X.H., Zheng, S., 2013b. Quasi-synchronous ionospheric and surface latent heat flux anomalies before the 2007 Pu'er earthquake in China. Natural Hazards and Earth System Sciences Discussions 1, 2439–2454.
Rafiei, M.H., Adeli, H., 2017. NEEWS: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. Soil Dyn. Earthq. Eng. 100, 417–427.
Rawat, V., Saraf, A.K., Das, J., Sharma, K., Shujat, Y., 2011. Anomalous land surface temperature and outgoing long-wave radiation observations prior to earthquakes in India and Romania. Nat. Hazards 59, 33–46.
Reyes, J., Morales-Esteban, A., Martínez-Álvarez, F., 2013. Neural networks to predict earthquakes in Chile. Appl. Soft Comput. 13, 1314–1328.
Rouet-Leduc, B., Hulbert, C., Johnson, P.A., 2018. Continuous chatter of the Cascadia subduction zone revealed by machine learning. Nat. Geosci. 12, 75–79.
Sarkar, T., Mishra, M., 2018. Soil erosion susceptibility mapping with the application of logistic regression and artificial neural network. Journal of Geovisualization and Spatial Analysis 2.
Singh, R.P., Cervone, G., Singh, V.P., Kafatos, M., 2007. Generic precursors to coastal earthquakes: inferences from Denali fault earthquake. Tectonophysics 431, 231–240.
Singh, R.P., Mehdi, W., Gautam, R., Senthil Kumar, J., Zlotnicki, J., Kafatos, M., 2010. Precursory signals using satellite and ground data associated with the Wenchuan earthquake of 12 May 2008. Int. J. Remote Sens. 31, 3341–3354.
Smith, N., Smith, W.L., Weisz, E., Revercomb, H.E., 2015. AIRS, IASI, and CrIS retrieval records at climate scales: an investigation into the propagation of systematic uncertainty. J. Appl. Meteorol. Climatol. 54, 1465–1481.
Tong, L., Zhang, Q., Sadka, A., Li, L., & Zhou, H. (2019). Inverse boosting pruning trees for depression detection on Twitter. arXiv preprint arXiv:1906.00398
Tramutoli, V., Cuomo, V., Filizzola, C., Pergola, N., Pietrapertosa, C., 2005. Assessing the potential of thermal infrared satellite surveys for monitoring seismically active areas: the case of Kocaeli (İzmit) earthquake, August 17, 1999. Remote Sens. Environ. 96, 409–426.
Tronin, A.A., 2007. Satellite thermal survey—a new tool for the study of seismoactive regions. Int. J. Remote Sens. 17, 1439–1455.
Walker, S.H., Duncan, D.B., 1967. Estimation of the probability of an event as a function of several independent variables. Biometrika 54, 167.
Wong, S., Fetzer, E.J., Schreier, M., Manipon, G., Fishbein, E.F., Kahn, B.H., Yue, Q., Irion, F.W., 2015. Cloud-induced uncertainties in AIRS and ECMWF temperature and specific humidity. Journal of Geophysical Research: Atmospheres 120, 1880–1901.
Wu, L.-X., Qin, K., & Liu, S.-J. (2012). GEOSS-Based Thermal Parameters Analysis for Earthquake Anomaly Recognition. Proceedings of the IEEE, 100, 2891–2907.
Wu, L., Zheng, S., De Santis, A., Qin, K., Di Mauro, R., Liu, S., Rainone, M.L., 2016. Geosphere coupling and hydrothermal anomalies before the 2009 Mw 6.3 L'Aquila earthquake in Italy. Nat. Hazards Earth Syst. Sci. 16, 1859–1880.
Yan, R., Parrot, M., Pinçon, J.-L., 2017. Statistical study on variations of the ionospheric ion density observed by DEMETER and related to seismic activities. J. Geophys. Res. Space Physics 122, 12,421–412,429.
Zeger, S.L., Karim, M.R., 1991. Generalized linear models with random effects; a Gibbs sampling approach. J. Am. Stat. Assoc. 86, 79–86.