DOI:
10.1109/ACCESS.2019.2929677

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 16, 2019, accepted July 11, 2019, date of publication July 18, 2019, date of current version August 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929677

Predictive Analysis of Landmine Risk

WAQAS RAFIQUE, DAWEI ZHENG, JAMIE BARRAS, SAGAR JOGLEKAR, AND PANAGIOTIS KOSMAS
Department of Informatics, King's College London, London WC2R 2LS, U.K.

Corresponding authors: Waqas Rafique (waqas.rafique@kcl.ac.uk) and Panagiotis Kosmas (panagiotis.kosmas@kcl.ac.uk)

This work was supported in part by the UK's Engineering and Physical Sciences Research Council (EPSRC), and in part by the Global Challenges Research Fund (GCRF) under Grant EP/P02906X/1.

ABSTRACT Demining is a highly impactful but complex problem which requires considerable resources and time. Land mine detection is the most hazardous and time consuming of the tasks in the demining pipeline. Currently, the risk of landmines being present in an area is estimated on the basis of non-technical surveys which are expensive and slow. This paper presents a novel spatial landmine risk prediction model to help and improve the allocation of resources in demining operations and even predict future areas of interest. Our approach is based on training predictive models on geographical and social development data for areas recorded to have been demined in the past. We then use this model to predict areas with high chance of mine presence in the vicinity of the demined area so as to progressively expand the area of operations. We explore weighted classification and biased scoring methods to improve the performance of our base logistic regression and support vector machine models. Refinement of conventional models allows us to tackle the problem of unbalanced datasets in our application. The resulting pipeline is then characterized in terms of various performance metrics. The results show that the pipeline has a potential to provide reliable predictive information based on historic demining data, which can help organizations plan their resource allocations in future demining operations.

INDEX TERMS Social implications of technology, logistic regression, support vector machine, predictive analysis, landmine risk.

I. INTRODUCTION

Landmines are a blight on regions recovering from conflict. Un-cleared landmines claimed more than 8000 casualties[1] in 2016 alone, and the numbers unfortunately have been more or less steady year on year for the past 20 years [1]. Since landmines are cheap to produce, easy to deploy, maintenance free and extremely durable, huge amounts were excessively deployed in countries such as Cambodia, Mozambique, Afghanistan and other counties during recent civil conflicts [1]. In the post-conflict period, the existence of unexploded landmines results in migration, displacement and casualties, which significantly influences the local communities [2].

Many global non-governmental organizations (NGOs) including those funded by the United Nations, have been participating in demining operations with a positive impact on local economies and communities.[2] The demining process comes with inherent substantial complexity. Typically, a demining cycle includes processes such as: vegetation clearance, detecting a landmine using devices, prodding to find the position of the landmine, and finally, removing the mine [3]. In recent years, the following approaches have been mainly applied to detect landmines: trace/vapour explosive detection, bulk explosive detection, mine casing detection, and infrared/hyper spectral detection. These technologies have shown promise in detecting landmines in various scenarios, with one of the main challenges being to detect mines buried in great depth [4]–[7].

One of the major issues in demining operations is the mismatch between the size of the area to cover and the resources available. Furthermore, due to the lack of clear evidence on the location of the contaminated areas, effective planning of the deployment of limited demining resources is a challenge. Currently, the allocation of demining resources

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

[1]http://www.the-monitor.org/en-gb/reports/2017/landmine-monitor-2017/casualties.aspx#ftn1

[2]https://peacekeeping.un.org/en/mine-action

mainly depends on non-technical surveys, demining dogs, and local knowledge. The cost of conducting a non-technical survey varies anywhere between $1.8 per square feet to $5 per square feet [8]. Moreover in a post-war environment, a non-technical survey can be biased for various reasons, such as locals providing inaccurate information for their own benefit such as to create jobs or increase the value of land. Thus, the intelligence from these techniques is often inaccurate, biased, and misleading [9]. Furthermore, the use of sniffing animals such as dogs or rats can add significant overhead costs, can be extremely time consuming, and have limited accuracy. Therefore, a method which can use historic as well as up-to-date information to triage areas based on probabilities of mine contamination could add great value to these surveys. To the best of authors' knowledge, currently there is no automated method that has been implemented for the prediction of landmine risk, and this work is the first to propose the application of well-known machine learning (ML) algorithms to this problem.

In particular, we show that we can achieve high accuracy in predicting mine contamination in unexplored areas using publicly available historic demining data and conventional ML algorithms. This could widely impact the operational efficiency of demining organizations in terms of cost and speed of demining operations, thus bringing clear societal and economic benefits to mine-affected communities.

II. RELATED WORK

Recently, many industries have had their operation methods greatly advanced through the application of information technology. For example, patterns of interest can be extracted from spatial predictive analysis using platforms such as the geographic information system (GIS). By applying advanced data-mining techniques on criminal events records in police departments' GIS database, the analysis can predict trend of criminal events in the future [10], [11]. Similarly, demining operations can also benefit from application of GIS, with suitable ML techniques applied.

Logistic regression (LR) [12], [13] is widely used in classification problems. It can automatically transfer a binary label result to its probability, which is very suitable to the nature of landmine risk problem. Authors of a recent paper [14], [15] used a GIS-based study to make several modifications to adapt the original LR model to a rare-events dataset – mineral distribution in southeast China. This GIS-based study provided some useful corrections, which, if more-generally applied, could possibly improve the regression model performance in terms of recall and overall accuracy. For example, a variance inflation factor (VIF) was calculated to reduce the dimension of dataset. In our problem, a trade-off between elimination of multi-collinear variables and completeness of information exists. This topic will be explained in details in later section.

Support vector machine (SVM) [18] is another popular algorithm to solve similar problems. Especially when dealing with a non-linear separable problem, the SVM model has unmatchable advantages compared with linear models. Similar to LR, the SVM method has been widely applied to rare event prediction such as credit risk classification [19]–[23]. Moreover, the SVM algorithm is also implemented for spatial predictive analysis with altered non-random sampling in order to overcome influence from imbalanced data [24]–[27].

On the basis of this discussion, our proposed method includes the process of translating traditional binary output of the SVM into a probability result. Probability results are designed to give a broader definition of 'presence', and thus provide demining organizations with more flexibility in decision-making.

III. METHODS

A. LOGISTIC REGRESSION (LR)

LR [12] is designed to construct regression that can fit underlying relations between multiple explanatory variables and dependent variables. The dependent variable is typically binary, with values 0 and 1. In the context of this paper, we have assigned presence and absence of landmine to dependent variable values 1 and 0, respectively. The relation between the dependent and independent variables is generally described by a linear equation of the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d \tag{1}$$

where y denotes the result of regression, and $x_i \forall i \in [0, d]$ represents dimensions of the observation features or, in our case, the features corresponding to the geographical position of a prospective mine. The coefficients $\beta_i \forall i \in [0, d]$ represent the regression coefficients to be learned through the data. Finally, d is the number of explanatory variables for an observation $\boldsymbol{X}$.

By convention, the sigmoid function is used to interpret the output y to a probability value p(X), ensuring that p(X)∈(0,1), and that the value increases sharply from 0 to 1 in monotonous rise. This is an approximation suitable for the binary case. [15]–[17]:

$$p(\boldsymbol{X}) = \frac{1}{1+e^{-y}} \tag{2}$$

where $p(\boldsymbol{X})$ is the predicted probability of the observation belonging to class 1 (presence of landmine). A major task of training the LR model is to find a suitable coefficients vector $\boldsymbol{\beta}$, which can be acquired by maximum likelihood estimation.

If we assume there are n observations:$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, and their observed class labels are $\omega_1, \omega_2, \ldots, \omega_n$, then the likelihood function can be given as,

$$L(\beta) = \prod_{i=1}^{n} (p(\boldsymbol{X}_i))^{\omega_i} (1 - p(\boldsymbol{X}_i))^{1-\omega_i} \tag{3}$$

From equations (1) and (2), $p(\boldsymbol{X}_i)$ is a function of $\boldsymbol{\beta}$ for observed values of vector $\boldsymbol{X}_i$. The goal is therefore to find $\boldsymbol{\beta}$ which maximizes the value of function (3). Applying the natural algorithm in (3) yields the following equation which

can be solved to acquire the optimal $\boldsymbol{\beta}$ by using the gradient descent method [17]:

$$\ln\left(\mathrm{L}\left(\boldsymbol{\beta}\right)\right) = \sum_{i=1}^{n} \left(\omega_i \ln\left(p\left(\boldsymbol{X}_i\right)\right) + (1-\omega_i)\ln(1-p\left(\boldsymbol{X}_i\right))\right) \tag{4}$$

B. NON-LINEAR SUPPORT VECTOR MACHINE

Support vector machine (SVM) techniques [18] are based on linear discriminant functions to solve binary classification problems. In practice, SVM relies on projecting data points to a higher dimension space, as, in most cases, it is easier to use this hyperplane to separate two groups of data points. Non-linear kernel functions can support this process in a non-linearly separable dataset [19]. The training mechanism is shown below.

Similar to LR, this work can be regarded as a two-class classification problem which has a training sample (observation) dataset: $\mathrm{S} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_n\}$. Observation $\boldsymbol{X}_i$ belongs to either one of the classes (absence of landmine, presence of landmine), which is represented by $y_i \in \{-1, +1\}$. As previously, $\boldsymbol{X}$ denotes a feature vector corresponding to a particular geographical data point.

In an ideal case where data points are linearly separable, the separating hyperplane is of similar form to LR and can be written as:

$$f\left(\boldsymbol{X}\right) = \boldsymbol{\beta}^T \boldsymbol{X} + \beta_0 \tag{5}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_d]^T$ denotes the group of coefficients applied to the corresponding observation feature, d represents the number of features (dimensions) of an observation, and β_0 is the intercept. Two hyper-planes $f\left(\boldsymbol{X}\right) = +1$ *or* -1 leave previously mentioned margins in both sides of the separating hyperplane, allowing possible wrongly classified observations in this area. Therefore the final classification criteria is written as [15]:

$$\begin{cases} \boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0 \geq +1 & for\ y_i = +1 \\ \boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0 \leq -1 & for\ y_i = -1 \end{cases} \tag{6}$$

The observations fall onto two margin hyperplanes (when $f(\mathbf{X_i})$ equals to $+1$ or -1). The optimal solution should distinguish the two classes of support vectors as much as possible from the separating hyperplane. For any support vector, the distance to a separating hyperplane is $1/\left\|\boldsymbol{\beta}\right\|$, and is doubled for both sides of margins as $2/\left\|\boldsymbol{\beta}\right\|$. To simplify the derivation, the problem becomes an optimization problem:

$$\mathrm{Minimize} \frac{1}{2} \left\|\boldsymbol{\beta}\right\|^2 \tag{7}$$

Which is subject to:

$$y_i(\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0) \geq 1 \tag{8}$$

Applying a Lagrange function to solve this problem, we can get $\boldsymbol{\beta}$ and β_0 by solving the following group of equations [15]:

$$\begin{aligned} \boldsymbol{\beta} &= \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{X}_i \\ \sum_{i=1}^{n} \lambda_i y_i &= 0 \\ \lambda_i &\geq 0 \\ \lambda_i(y_i\left(\boldsymbol{\beta}^T \boldsymbol{X}_i + \beta_0\right) - 1) &= 0 \end{aligned} \tag{9}$$

where λ_i is the introduced Lagrange multiplier. Here, all $\boldsymbol{X}_i$ are support vectors. For the non-linear separable case, a mapping function of selection $\Phi(\cdot)$ needs to be applied to all observations $\boldsymbol{X}_i$ to work as mapping function, projecting $\boldsymbol{X}$ into higher dimensional space, where the observations represented by $\mathbf{Z}$ become linearly separable [16]:

$$\mathbf{Z}_i = \Phi(\boldsymbol{X}_i) \tag{10}$$

The solution steps previously demonstrated should be applied in the same way, on $\mathbf{Z}_i$ instead of $\boldsymbol{X}_i$.

We note that, in order to unify the setting with the LR model, the SVM predictor's output '-1' for the absence class will be replaced by '0' in our notation below.

C. BAGGING – AN ENSEMBLE METHOD

The ''bagging'' method is designed to enable multiple classifiers into same sets to carry out a ''major vote'' on the final result. It can substantially improve performance compared with a single predictor model as it helps avoiding overfitting and reducing runtime variance [27]–[30].

Assuming the training dataset is S, training m meta-classifiers requires generating m training subsets S1, S2,.., Sm by sampling with replacement. The meta-classifier can be any of the traditional classification models such as SVM or decision tree. The number of observations for each subset will be smaller or equal to the size of S.

After m meta-classifiers are generated, the classification result of an observation $\boldsymbol{X}$ would be:

$$f_{final}\left(\boldsymbol{X}\right) = \sum_{i=1}^{m} \frac{1}{m} f_i\left(\boldsymbol{X}\right) \tag{11}$$

where f denotes the classification function of any kind of model, such as SVM for this particular work. The final result is the average value of the meta-classifiers' result. There are different ways to interpret the final result. In this paper we propose to take the average value itself as a final result, thus the final result of SVM's output could be treated the same as a probability output from LR.

IV. PROPOSED MODEL FOR LANDMINES PREDICTION

In this work we are proposing a pipeline shown in Figure 1, which could apply a combination of heuristics and mapping to a dataset from active demining operations, in order to extract usable features. These features could then be used

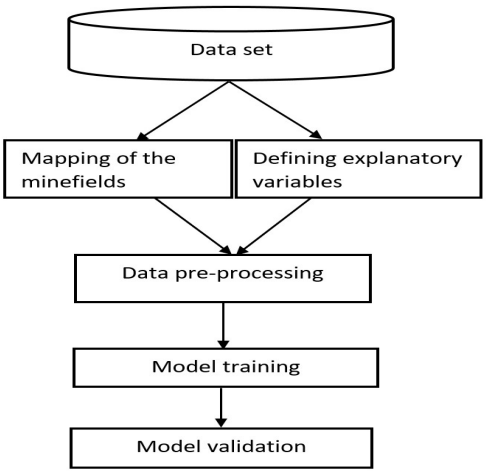


FIGURE 1. Proposed model for prediction of landmine risk.

to train a variety of models for the purpose of predicting landmine distributions in mine fields. Our aim is to demonstrate that even the most accessible ML algorithms, such as SVM or LR, could have significant positive impact on demining operations around the world by predicting landmine distributions.

The main steps of the proposed model comprise defining the explanatory variables of our dataset and the mapping of the minefields, a pre-processing step, training of the prediction algorithms, and performance evaluation. The core focus of this work is to make a prediction on the presence or absence of landmines within a particular study area.

A. STUDY AREA AND DATASET

Cambodia is one of the countries that has suffered most from explosive remnants of war (ERW). A high proportion of the country's area remains contaminated, and hence an accurate landmine risk prediction tool would help expedite operations in this country. To test our approach, landmine operations data in Cambodia was provided by HALO Trust. This data mainly contains 2 parts:

- Demining operations done by HALO Trust in 1992-2012
- Demining operations done by all other organizations in 2010-2017

Among the aforementioned dataset, all ERW including different types of landmine and other unexploded devices are taken into consideration. The operations that have found landmine and other ERW at certain geographical locations are combined with the country map (Positions shown in Figure 2). The study chose an area close to the city Phsar Prum (12.93 N, 102.50 E). This area covers 113 square kilometres. Furthermore a larger area is also included as Extended Area in the study, which covers 248 square kilometres and is shown in Figure 3. The terrain in this region is mainly flat, with a small portion of hills. The study area and extended surrounding area share the same type of terrain; the minefield densities are also similar. Transportation networks are also

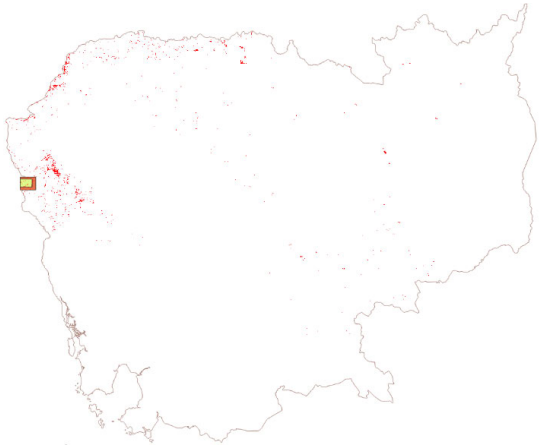


FIGURE 2. Distribution of landmine and ERW in cambodia based on demining operations in 1992-2017 and location of study area (yellow rectangle covers study area, orange rectangle represents surrounding area).

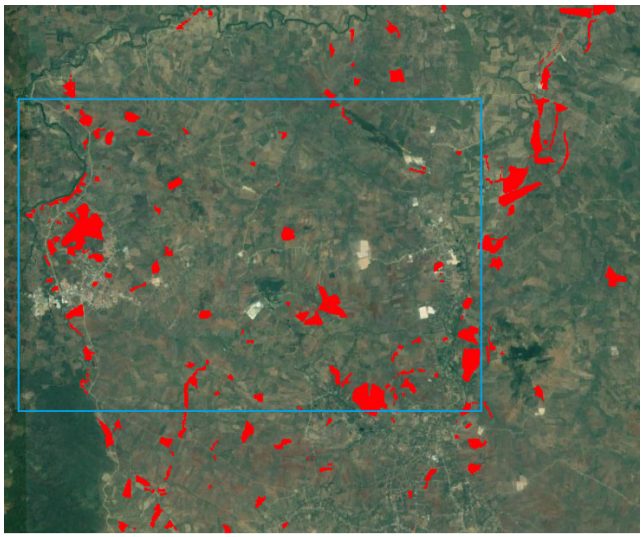


FIGURE 3. 3 Landmine distribution in study area (inside blue frame) and surrounding area (red shadowed areas are minefields).

present within this area. In our implementation, we train the models on data sampled from the study area (train set), and then validate the models on both the test set from the study area and on the extended area.

B. MAPPING OF MINEFIELDS AND EXPLANATORY VARIABLES

1) MINEFIELD DATA CONVERSION

The minefield data were provided in shapefile format, which is a popular serialization format for GIS applications. In order to simplify the process linking dependent variable (minefield presence) with other explanatory variables, all information was converted to raster layer format. The resolution of the finest observation, which is also regarded as a datum point, is 900 square meters. Based on experience, previous studies [18]–[21], and availability, in total 11 features were selected as candidate explanatory variables, which are summarized in Table 1.

TABLE 1. Information of explanatory variables.

Variable	Value Range(in Figure 2 Area)	Unit	Source
Elevation	50 - 442	meter	ASTER GDEM Version 002 [32]
Hill Slope (percentage)	0 - 157	%	ASTER GDEM Version 002 [32](calculated)
Distance to Border	0 - 14712	meter	GADM by University of California, Davis [33]
Distance to Roads	0 - 4038	meter	Open Street Map [34], Save Cambodia's Wildlife [35]
Distance to Railways	53377 - 74433	meter	Save Cambodia's Wildlife [34]
Distance to Water Channels	0 - 5816	meter	Onemap Cambodia [36]
Distance to Persistent Forest	0 - 1897	meter	Earth Science Data Records of Global Forest Cover and Change V001 [37]
Distance to Lost Forest	0 - 6138	meter	Earth Science Data Records of Global Forest Cover and Change V001 [37]
Distance to Gained Forest	0 - 1601	meter	Earth Science Data Records of Global Forest Cover and Change V001 [37]
Probability of Built-up	0 - 100	%	Global human built-up and settlement extent (HBASE) dataset [31]
Population Density	0 - 38	persons per km^2	Gridded population of the world, version 3 [38]

2) EXPLANATORY VARIABLES

Topographical elevation data are collected from ''ASTER GDEM Version 002'' [31]. This was a project developed collectively by NASA and Japan's Ministry of Economy, Trade, and Industry, posted on a 30 meter grid with vertical root-mean-squared-error between 10 and 25 meters. The hill slope in percentage data are calculated from elevation in quantum geographic information system (QGIS) platform using geospatial data abstraction library (GDAL)'s slope package, the output of which is also posted to a 30 meter grid raster layer. There are some variable raster maps which don't have that resolution, such as population density which was based on a 4500 meter grid, from which we extracted data to match our 30 meter grid. Having the finest possible grid gave us more data samples, which can increase the accuracy of our prediction.

The location of the national border, transportation networks such as roads, railways, and water channels data are collected from various sources, with details shown in Table 1. As logistic regression works well for continuous variables instead of categorical variables [16], we converted all categorical explanatory variables to continuous numerical by calculating distance-to value in meter unit with GDAL's Proximity package. The module measures distance from the centre of a random pixel to the centre of its nearest pixel which belongs to the target dataset.

Trend of forest coverage data are extracted from Earth Science Data Records of Global Forest Cover and Change V001, a work completed by University of Maryland and NASA. It is mainly derived from enhanced Global Land Survey (GLS+) data sets and surface reflectance ESDRs [29]. In the study we split the data into three parts – Persistent Forest, Forest Loss, and Forest Gain. Generally, areas where landmines are found would experience a change of vegetation coverage. Changes between 1990 and 2000 are collected based on this dataset, as mine operation data is not available before 1990.

We have extracted probability data from a project by NASA and Columbia University targeting year 2010, the Human Built-up and Settlement Extent. This dataset maps levels of urban development using a scale of 1 to 100 with 30m spatial resolution [32]. This work is based on classification of global surface reflectance data from previous projects. We also collected population density data using ''Gridded Population of the World v3'', to get a population density grid of Cambodia in 1990 according to total population statistics of the UN [39]. The mapping has a resolution of 4500 meter which is converted to 30 meter grid in order to have a uniform grid size.

C. DATA SAMPLING

Within the study area, 10,000 data points are randomly sampled from the whole dataset as the training set. In addition 5,000 data points are sampled for the validation and test sets respectively. Statistics of the training set suggest that there are only around 450 points with values indicating presence of a landmine, which counts for 4.5% of the total data points. This indicates that the model is trained on an imbalanced dataset. This could further exacerbate the risk of overfitting during training, especially for the SVM algorithm. Therefore, in order to tackle the imbalance, over-sampling was performed on the dataset to remedy the imbalanced training dataset in accordance with a previous study [33]. Moreover, a large number of existing studies claimed that under-sampling, which means selecting equal numbers of samples from both classes of an imbalanced dataset, will result in losing part of the information [34], [42]. However over-sampling, i.e. filling the gap between the minority class and the majority class by adding minority samples repetitively into the training dataset,

TABLE 2. VIF values of explanatory variables in each iteration of dimension reduction (bold values are maximums of the iteration which are greater than 10, and variable will be excluded from next iteration).

Variable	VIF of 1st iteration	VIF of 2nd iteration
Elevation	2.2531	2.0315
Hill Slope	1.4957	1.4744
Distance to Border	**15.8236**	----
Distance to Roads	1.5022	1.4003
Distance to Persistent Forest	1.428	1.3939
Distance to Gained Forest	1.5943	1.4435
Distance to Lost Forest	2.495	2.3403
Distance to Water Channel	1.8295	1.6223
Distance to Railways	13.7935	2.4853
Population Density	2.788	2.043
Built-up	1.1658	1.1658

could increase the risk of overfitting. Taking these concerns into consideration, we further explored remedies in the model by tweaking performance criteria and decision thresholds.

D. DATA PRE-PROCESSING

This study used GDAL's Translate package in QGIS to transform and import data into ML programs written in Python. Although LR is able to overcome large scale variance by adjusting the β coefficients, it is still preferable to transform variables into the same scale to speed up convergence, especially since SVM is not scale-invariant. Therefore, the pre-processing step also considers the scaling process. Scikit-learn [40] is the main package that is used in model training including scaling of data and prediction.

E. FEATURE SELECTION

A previous study showed that if the explanatory variables used for training an LR algorithm are correlated, the model cannot compute regression coefficients with confidence [41]. The relationship can be detected by calculating VIF values for the explanatory variables. If the VIF value for an explanatory variable is greater than 10, it will likely have a strong dependence on other variables, and would lead to errors in coefficient estimations by the algorithm. We have used the 'statsmodels' package [42] to calculate the factor values of each variable and remove high VIF features iteratively.

Results in Table 2 suggest that the ''Distance to Border'' variable should be removed, as it has multi-collinearity with ''Distance to Railways''. Moreover, the variable value range in Table 2 indicates that the study area is located close to the border and far from major railways of Cambodia. The VIF values of these variables, however, are not extremely large and are based on acquired knowledge. We do know for a fact that landmines were planted along the border during the conflict period to prevent people from escaping the country. Thus, it is difficult to decide if excluding ''Distance to Border'' would benefit the performance, as it might cause loss of important information as well. Therefore, this suggestion will be tested in later sections.

F. ML IMPLEMENTATION APPROACH

1) SCORING METHODS TO FIND OPTIMAL PARAMETERS

As previously mentioned, the classification of landmine presence is a heavily imbalanced problem. To generate an informative model, we aim at the highest possible true positive rate (TPR), or recall, while ensuring that other performance measurements are controlled within an acceptable range. The study firstly utilized a package's mode to set the weight of samples corresponding to the proportion of its own class in the whole dataset. This ensured that the observations of a landmine presence will have much larger weight than of those an absence in the training set, to balance the fact that landmine presence data represents only 4.5% of all observations.

To compare performance between models, the area under the curve (AUC) is calculated for receiver operating characteristic (ROC) curves resulting from the different models. A ROC Curve is one of the most important evaluation metrics for checking any classification model's performance, as it allows calculating the true positive rate (TPR) and false positive rate (FPR) corresponding to a specific decision threshold of a binary classification. Generally, if the ROC curve is almost full in its feature space, the AUC is close to 1, suggesting that the model is stable and optimal. The calculations of TPR and FPR are as follows:

$$\text{TPR} = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Actual\ Positive} \tag{12}$$

$$\text{FPR} = \frac{False\ Positive}{False\ Positive + True\ Negative} = \frac{False\ Positive}{Actual\ Negative} \tag{13}$$

To select the best probability threshold for binary prediction accuracy, a selection process is programmed to find the threshold corresponding to the point on the ROC curve which has the shortest Euclidean distance to (FPR = 0, TPR = 1).

2) IMPLEMENTATION OF THE LR MODEL

As mentioned above, we use the scikit-learn package, which implements an LR model with regularization controlled by a parameter 'C' [37]. 'LogisticRegressionCV' is used to find the best model parameter. This mainly applies three-fold stratified cross-validation to calculate performance of each selected parameter combination. The scoring method for the LR model is the AUC value of the ROC Curve.

As discussed in Section 2.E, calculating VIF values for the explanatory variable ''Distance to Border'' suggests multi-collinearity that would lead to potential errors. To resolve this issue, the AUC of the ROC curves was calculated for two datasets with and without the variable. By analysing the results, as shown in Table 3, it is safe to state that including ''Distance to Border'' does not de-stabilise the trained mode.

TABLE 3. AUC of ROC curves for logistic regression algorithm on different training sets.

	Training set with Distance to Border	Training set without Distance to Border
AUC of ROC Curve	0.8828	0.8824

TABLE 4. Confusion matrix of predictive models.

			Actual Class	
			Presence	Absence
Predicted Class	Logistic Regression	Presence	**194**	952
		Absence	32	**3812**
	SVM	Presence	**199**	311
		Absence	27	**4463**

We therefore decided to keep this feature in the training set, so as to process as much information as possible in our model.

3) IMPLEMENTATION OF THE SVM MODEL

The SVM algorithm requires defining more parameters compared to the LR algorithm. First, the radial basis function is chosen as the kernel of the algorithm, as it allows a stable and simple optimal selection for model training. Moreover, in addition to the parameter C defining the regularization strength, a parameter γ (gamma) captures the influence of a single observation in its surrounding dimension space, with larger values corresponding to a smaller range of influence. By implementing the built-in package GridSearchCV from scikit-learn [40], we can traverse all pre-defined combinations of parameter C and gamma. In order to emphasize TPR's importance in the training, this study chose 'recall_macro' as a scoring method to pick suitable parameters for the SVM model; 'recall_macro' calculates the arithmetic average recall value of positive and negative classes.

This study also transforms the SVM's binary result to probabilities with bagging method using the BaggingClassifier function. Furthermore, the bagging method is set to generate 100 meta-classifiers where each classifier will make use of 100% data from training set to perform training.

V. RESULTS, ANALYSIS AND EVALUATION

A. PERFORMANCE COMPARISON IN STUDY AREA

Using trained models of the two algorithms to predict the test dataset sampled from the Study Area, the confusion matrix generated for the LR and SVM models is shown in Table 4. Results show that the SVM model is slightly better than the Logistic Regression model in terms of TPR ($199 > 194$). The SVM model has 10.12% FPR, whereas LR's FPR is 23.62%. This indicates that the LR model generates more than double the number of false alarms compared to the SVM algorithm. A lower FPR is meaningful in reducing waste of demining resources in our case.

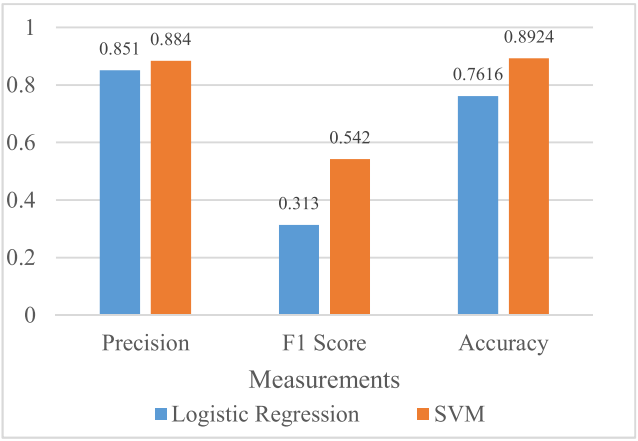


FIGURE 4. Measurements comparison of predictive models.

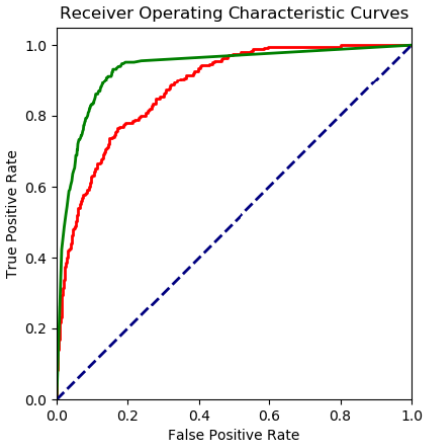


FIGURE 5. ROC curves of predictive models (green line - SVM red line - logistic regression).

Additional performance measures were calculated to gain better understanding of the results from both algorithms. First, the precision score was calculated as follows:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{14}$$

We also calculated the F1 score, which is a balanced evaluation based on the Recall (TPR) and the Precision score, which equals to:

$$\text{F1score} = \frac{2 * precision\ *\ recall}{(precision + recall)} \tag{15}$$

The graph shown in Figure 4 demonstrates that F1 scores are relatively lower than normal due to manual adjustment for TPR. In practice, we believe this issue won't increase the demining workload substantially since the ratio of actual landmine area to entire area is very small. It is obvious that the SVM algorithm outperforms LR in all the measurements shown. Moreover, accuracy and precision for both methods provides excellent insight into the performance of both algorithms in making predictions with high accuracy.

ROC curves [43] for both algorithms are shown in Figure 5. The AUC values for the SVM and LR algorithms are 0.93 and 0.88, respectively, and the SVM model also outperforms the

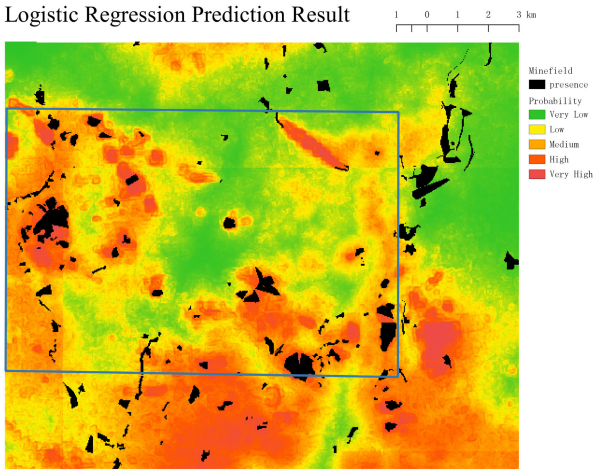


FIGURE 6. Logistic regression prediction result (blue frame shows study area).

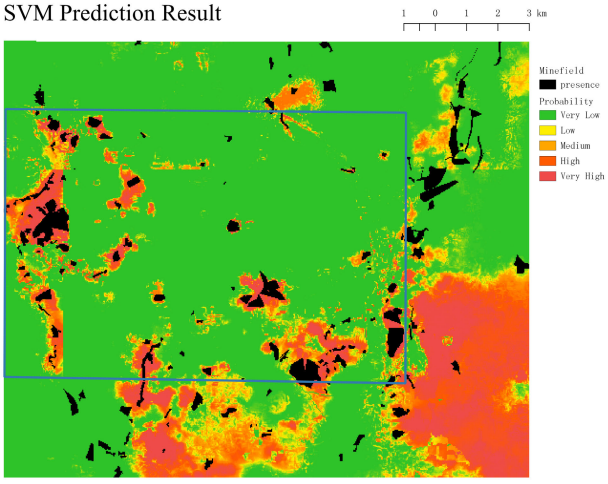


FIGURE 7. SVM prediction result (blue frame shows study area).

LR model in the up-right area where the optimal decision threshold is selected.

The main disadvantage of LR is the sacrificed F1 score to achieve the pre-set goal of TPR higher than 0.85. The usual explanation for this phenomenon is that LR is based on linear separation. Therefore in order to ensure a high TPR, the classification border has to be moved towards negative class data points.

B. MAPPING OF PREDICTION RESULTS IN STUDY AREA AND EXTENDED AREA

Probabilities of landmine presence for all observations within the extended area (including the Study Area) are calculated by both models by repeating the same experiment. This study visualized results within the GIS platform. Using QGIS, results can be shown together with the actual landmine distribution. This generated the scaled heat maps shown in Figure 6 and Figure 7, where the actual landmine distribution is marked as black shadowed areas. With an optimal probability threshold for our binary decision around 0.45 for

TABLE 5. Percentage of observations predicted probability inside study area.

Algorithm	Group of Observations in Study Area	0 - 0.2	0.2 - 0.4	0.4 - 0.75	0.75 – 1
Logistic Regression	All	40.52%	27.30%	24.56%	7.62%
Logistic Regression	Minefield	1.33%	9.29%	36.73%	52.65%
SVM	All	83.30%	2.94%	5.18%	8.58%
SVM	Minefield	8.41%	3.98%	13.72%	73.89%

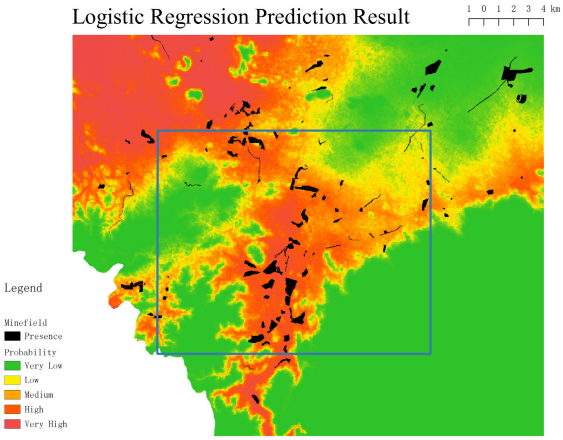


FIGURE 8. Logistic regression prediction result in second chosen area (blue frame shows study area).

SVM and LR, probability 0 is labelled as 'Very Low', 0.2 is labelled as 'Low', 0.4 is labelled as 'Medium', 0.75 is as labelled 'High', and when probability is 1 it is labelled as 'Very High'.

Comparing the two heat maps and statistics in Table 5 provides clear confirmation that, within the study area, the separation performed by the SVM model is more effective. Unlike the SVM, the LR model generates a great number of observation probability values in the uncertainty area of 0.2 – 0.75. In surrounding areas outside the blue frame in the heat maps of Figures 6 and 7, the LR does not perform well in the northeast area, while the SVM algorithm does not correctly predict minefields in the southwest area. These results suggest that the LR model can perform accurate predictions for a known area, but when new areas are tested, its performance degrades. The SVM, on the other hand, performs consistently well for a wide range of different locations.

In order to validate the findings from the above results, another mountainous area of 275 square kilometres, located south of Samlout (12.33 N, 102.51 E), is chosen as a second study area, along with a larger extended area of 885 square kilometres. The same methods are applied for training and prediction processes as in the previous experiments and the results are shown in Figure 8 and Figure 9, which are also in the same scale.

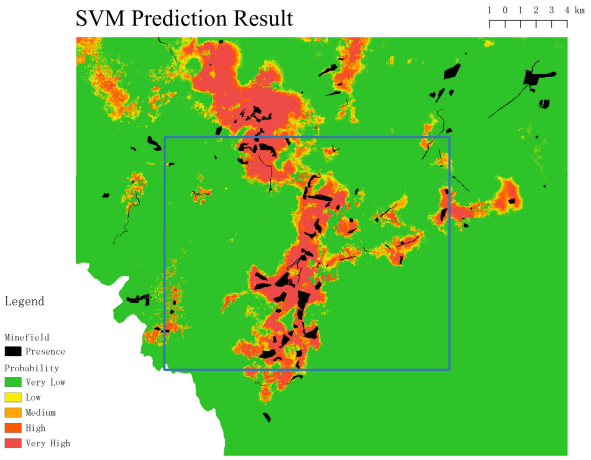


FIGURE 9. SVM prediction result in second chosen area (blue frame shows study area).

The results in this second area again confirmed the findings that the SVM model achieves higher accuracy than LR within the study area. The accuracy for both algorithms is generally high and can provide substantial information about the contaminated area within the study area. In the surrounding areas, a large proportion is predicted as high risk by the LR model, whereas the ratio of different risk levels in the SVM model's prediction result compared with the study area remains stable. However, some minefields are still not correctly classified by both algorithms. This implies there is a limit of distance from the study area to the surrounding area, within which the accuracy is acceptable.

C. POTENTIAL PRACTICAL USE

In countries like Cambodia which have not been officially declared mine-free, demining operations are still on-going. Hence, the proposed predictive analysis algorithms can help demining operations in the 'clearing' stage. In particular, our study suggests that, in regions where some areas have known landmine distributions, a trained model based on data from the characterised area could help to predict the risk in remaining unknown areas. Our results suggest that the SVM algorithm is generally suitable in any kind of conditions inside the study area by keeping a low FPR, while the LR model achieves a similar TPR but possibly higher FPR.

Another practical use of the algorithms is to predict contamination of adjacent areas based on data from areas that are almost clear. Better performance would be achieved if the demining operations are already completed in a region such as the study areas considered in this paper. Moreover the LR model can be further enhanced by adding some pre-conditions about the training set such as topography information etc. On the other hand, the SVM model can potentially perform better when the training set have complete and accurate information about the study area.

VI. CONCLUSION

This paper argued that it is feasible to use historic demining data along with ML models to evaluate probabilistic risks of land mine contamination. We also assessed the performance of these models and showed that the models are useful at ruling out the presence of landmines. With further work and data, we believe the performance of these models can be improved considerably.

Due to the sparsity of distribution of mines, this particular application poses the challenge of imbalanced datasets. We explored different techniques to mitigate this issue. Moreover, a grid search method with cross validation technique was implemented to optimise parameter selection for both models. The models are adaptive in terms of scoring methods, which enables us to train a model in the favour of gain or cost. The results section of the study compares the SVM and LR methods and results show that the SVM algorithm outperforms LR in most cases. In regions where topographical variance is not very prominent, the LR algorithm can also produce reliable results with much lower cost of time and memory. Combining insights from both models could be the best way forward in maximizing performance. One important limitation of our work can be seen from Table 4. Our pipeline is great at ruling out presence of mines, but performs poorly in ruling in the presence of mines. This limitation could be addressed if more data was available.

As a future work, we aim at acquiring more data from sources who work on site, and improve our models. The explainable nature of these models would allow us to not only evaluate risks of contamination, but also understand the patterns and correlations between geographic and social features and presence/absence of mines. The problem that we are trying to solve affects millions of inhabitants of post-conflict countries around the world, and with further work, the proposed approach can help plan demining operations efficiently and optimize resource allocation.

ACKNOWLEDGMENT

The authors would like to thank The HALO Trust for providing their landmine operations data in Cambodia.

REFERENCES

[1] International Campaign to Ban Landmines, "Landmine monitor report 2003: Toward a mine-free world," Human Rights Watch, New York, NY, USA, Tech. Rep., 2003.

[2] M. Hagenlocher, D. Hölbling, S. Kienberger, S. Vanhuysse, and P. Zeil, "Spatial assessment of social vulnerability in the context of landmines and explosive remnants of war in Battambang province, Cambodia," *Int. J. Disaster Risk Reduction*, vol. 15, pp. 148–161, Mar. 2016.

[3] *Global Mapping and Analysis of Anti-Vehicle Mine Incidents in 2017*. Accessed: Oct. 6, 2018. [Online]. Available: https://www.sipri.org/publications/2018/other-publications/global-mapping-and-analysis-anti-vehicle-mine-incidents-2017

[4] R. Keeley. *Understanding Landmines and Mine Action*. Accessed: Feb. 11, 2019. [Online]. Available: http://mit.edu/demining/assignments/understanding-landmines.pdf

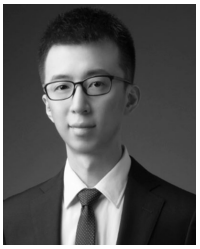
[5] A. Khamis, M. Ashraf, and A. Abdulbaky, "Landmines and UXOs in NWC: A domain review," in *Proc. Int. Workshop Recent Adv. Robot. Sensor Technol. Humanitarian Demining Counter-IEDs (RST)*, Cairo, Egypt, Oct. 2016, pp. 1–6.

[6] F. Lombardi, H. D. Griffiths, and A. Balleri, "Landmine internal structure detection from ground penetrating radar images," in *Proc. IEEE Radar Conf. (RadarConf)*, Oklahoma City, OK, USA, Apr. 2018, pp. 1201–1206.

[7] L. Cardona, J. Jiménez, and N. Vanegas, "Landmine detection technologies to face the demining problem in Antioquia," *Dyna*, vol. 81, pp. 115–125, 2014.

[8] *Clearing the Mines, Mine Action Review*. Accessed: Oct. 17, 2018. [Online]. Available: http://www.mineactionreview.org
[9] J. Bure and P. Pierre, "Landmine clearance projects: Task manager's guide," World Bank, Washington, DC, USA, Tech. Rep., 2003.
[10] Y. Xue and D. E. Brown, "Spatial analysis with preference specification of latent decision makers for criminal event prediction," *Decision Support Syst.*, vol. 41, no. 3, pp. 560–573, 2006.
[11] X. Wang, D. E. Brown, and M. S. Gerber, "Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, Arlington, VA, USA, Jun. 2012, pp. 36–41.
[12] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. Ser. B, Methodol.*, vol. 20, no. 2, pp. 215–242, 1958.
[13] C. Schultz, A. C. Alegría, J. Cornelis, and H. Sahli, "Comparison of spatial and aspatial logistic regression models for landmine risk mapping," *Appl. Geography*, vol. 66, pp. 52–63, Jan. 2016.
[14] Y. Xiong and R. Zuo, "GIS-based rare events logistic regression for mineral prospectivity mapping," *Comput. Geosci.*, vol. 111, pp. 18–25, Feb. 2018.
[15] S. Sperandei, "Understanding logistic regression analysis," *Biochem. Med.*, vol. 24, no. 1, pp. 8–12, 2014.
[16] M. Tranmer and M. Elliot, "Binary logistic regression," Cathie Marsh Census Survey Res., Manchester, U.K., Tech. Rep., 2008.
[17] C.-Y. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 1–14, 2002.
[18] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[19] L. Yu, R. Zhou, T. Ling, and R. Chen, "A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data," *Appl. Soft Comput.*, vol. 69, pp. 192–202, Aug. 2018.
[20] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[21] R. Zuo and E. J. M. Carranza, "Support vector machine: A tool for mapping mineral prospectivity," *Comput. Geosci.*, vol. 37, no. 12, pp. 1967–1975, 2011.
[22] W. Rafique, S. Erateb, S. M. Naqvi, S. S. Dlay, and J. A. Chambers, "Independent vector analysis for source separation using an energy driven mixed student's T and super Gaussian source prior," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, Aug./Sep. 2016, pp. 858–862.
[23] W. Rafique, J. Chambers, and A. I. Sunny, "An expectation–maximization-based IVA algorithm for speech source separation using student's t mixture model based source priors," *MDPI Acoust.*, vol. 1, no. 1, pp. 117–136, 2019.
[24] L. Wang, D. Wang, and C. Hao, "Intelligent CFAR detector based on support vector machine," *IEEE Access*, vol. 5, pp. 26965–26972, 2017.
[25] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[26] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.
[27] W. Rafique, S. M. Naqvi, and J. A. Chambers, "Mixed source prior for the fast independent vector analysis algorithm," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, Rio de Janerio, Brazil, Jul. 2016, pp. 1–5.
[28] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
[29] W. Rafique, S. M. Naqvi, P. J. B. Jackson, and J. A. Chambers, "IVA algorithms using a multivariate Student's t source prior for speech source separation in real room environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 474–478.
[30] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[31] S. Ding, H. Zhu, W. Jia, and C. Su, "A survey on feature extraction for pattern recognition," *Artif. Intell. Rev.*, vol. 37, no. 3, pp. 169–180, 2012.
[32] NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team, "ASTER global digital elevation model V002," NASA EOSDIS Land Processes DAAC, Washington, DC, USA, Tech. Rep., 2009.
[33] P. Wang, C. Huang, E. C. B. de Colstoun, J. C. Tilton, and B. Tan, "Global human built-up and settlement extent (HBASE) dataset from Landsat," NASA Socioecon. Data Appl. Center, Palisades, NY, USA, Tech. Rep., 2017, vol. 1.
[34] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.
[35] *GADM Data (Version 3.6)*. Accessed: Aug. 5, 2018. [Online]. Available: https://gadm.org/download_country_v3.html
[36] *GADM Data (Version 3.6)*. Accessed: Aug. 25, 2018. [Online]. Available: https://opendevelopmentcambodia.net/dataset/?id=road-and-railway-networks–in-cambodia
[37] *GADM Data (Version 3.6)*. Accessed: Aug. 25, 2018. [Online]. Available: https://opendevelopmentcambodia.net/dataset/?id=map-road-railway-network-market-density
[38] *GADM Data (Version 3.6)*. Accessed: Aug. 25, 2018. [Online]. Available: http://onemapcambodia.blogspot.com/p/cambodia-spatial-data.html
[39] W. Rafique, D. Zheng, J. Barras, and S. Joglekar. (2005). Gridded population of the world, version 3 (GPWv3): Population count grid. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY, USA. Accessed: Aug. 7, 2018. [Online]. Available: https://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count
[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[41] A. Ozdemir, "Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey)," *J. Hydrol.*, vol. 405, nos. 1–2, pp. 123–136, 2011.
[42] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 57–61.
[43] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

WAQAS RAFIQUE received the M.Sc. degree (with distinction) in digital communication systems from Loughborough University, U.K., where he qualified for fully funded scholarship, and also the Ph.D. degree in signal processing and machine learning from Newcastle University, U.K., in 2017, where he was sponsored by ESPRC for his Ph.D. studies. He is currently working with the Department of Informatics, King's College London as a Postdoctoral Researcher with Engineering and Physical Sciences Research Council (ESPRC) funded project focused on machine learning and signal processing. His current research interests include machine learning, signal processing and data analytics, and their application in social and biomedical context.



DAWEI ZHENG received the B.Eng. degree in information engineering from Southeast University, Nanjing, China, in 2014, Engineer's degree in information system from Ecole FRançaise d'Electronique et d'Informatique (EFREI), Paris, France, in 2015, and the M.Sc. degree in advanced computing from King's College London, London, U.K., in 2018. His research interests include data analytics and machine learning.

JAMIE BARRAS received the first degree from the University of Cardiff and the Ph.D. degree from the University of Cambridge. He is currently a Teaching Fellow with the Department of Informatics. Previously, he was a Research Fellow and Technical Manager for EU project CONPHIRMER and three other projects concerned with medicines authentication or the detection of illicit materials, such as contraband drugs or buried landmines. His research interests include quadrupole resonance for detection of landmines, medicines authentication, and security applications.

SAGAR JOGLEKAR received the bachelor's degree in electrical and computer engineering from the University of Pune, India, and the M.Sc. degree from the University of California, Santa Barbara and pursuing the Ph.D. degree with the Department of Informatics, King's College, London and is a King's India Scholar. His research interests include the areas of representation learning and complex networks and their applications to quantify intangible human properties like support, engagement, and aesthetics.

PANAGIOTIS KOSMAS received the Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Northeastern University, Boston, MA, in 2002 and 2005, respectively. He joined King's College London (KCL) as a Lecturer in 2008, and is currently a Reader with the KCL's Department of Informatics. Prior to his appointment at KCL, he held research positions at the Center for Subsurface Sensing and Imaging Systems, Boston, USA, the University of Loughborough, U.K., and the Computational Electromagnetics Group, University of Wisconsin–Madison, USA. He is also a co-founder of the Mediwise Ltd., an award-winning UK-based SME focusing on the use of electromagnetic waves for medical applications. His research interests include radio frequency (RF) engineering with application to sensing and imaging, antenna design, physics-based detection methods, and inverse problems theory and techniques. He has contributed a book chapter and over 90 journals and conference publications, and he has organized and delivered various short courses, special sessions, and workshops in these areas.