OIL SPILLS DETECTION BY MEANS OF INFRARED IMAGES AND WATER QUALITY DATA USING MACHINE LEARNING

***Vladislavs Zavtkevics, Dmitrijs Gorelikovs**
Riga Technical University, Latvian Maritime Academy, Latvia
*Corresponding author's e-mail: vladislavs.zavtkevics@rtu.lv

Abstract
The paper presents the results of the research on oil spill detection using machine learning methods such as Support Vector Machine (SVM) for classification of infrared images and Logistic regression for water quality parameters. This paper focuses on real time detection of oil spills using infrared images and water quality data obtained by RPA equipped with multi-sensor payload. The developed Naïve Bayes (NB), SVM and Logistic regression classification models for prediction of oil spill have been successfully tested in real experiment conditions. All developed classification models were tuned using grid search method and main tuning parameters to determine the optimal parameters. The proposed complex algorithm for identification of oil spills using infrared images and water quality parameters is evaluated by experiments in real environment conditions. The proposed algorithm is based on the binary SVM and NB classification of infrared images and the classification of water quality parameters using the machine learning method logistic regression allows to rapidly and with high accuracy identify any oil pollution of water. Proposed complex algorithm achieves higher accuracy and efficiency; moreover, the developed machine learning models will further reduce the probability of human error and save man-hours of work.
Key words: RPA, machine leaning, SVM, oil spill.

Introduction

Oil pollution as a result of oil transportation activities and accidental oil spills is the main factor that influences water quality in an area contaminated by oil. This paper focuses on real time detection of oil spills using infrared images and water quality data obtained by RPA equipped with a multi-sensor payload. Operational monitoring of large sea aquatorium areas with the aim of detecting oil pollution is now carried out using various technical devices – satellites, seagoing vessels and various aircraft (Zavtkevics & Urbaha, 2022). Taking into account that oil spills cannot be completely eliminated, the main objective is to develop multi-sensor RPA payloads for more efficient and sustainable water monitoring and management in line with the WFD 2000/60/EC requirements (Zavtkevics & Urbaha, 2022). The main objective of the paper is to consider the problem of oil spill detection using machine learning methods such as Support Vector Machine (SVM) for classification of infrared images and logistic regression for water quality parameters. Taking into consideration the complexity of monitoring with RPA equipped with an infrared camera and water quality sensors, it would be feasible to develop a complex algorithm for identification of oil spills using infrared images and water quality parameters. Moreover, performance of machine learning models and complex algorithm for identification of oil spills need to be evaluated by experiments in real environment conditions.

Materials and Methods

Oil, which is optically thick, absorbs solar radiation and reemits a portion of this radiation as thermal energy, primarily in the 8 to 14 mm region (Urbahs & Zavtkevics, 2017). Remote sensing using an RPA equipped with an infrared camera can detect thin oil film temperature that due to the contrast may appear different than the surrounding water both during the day and night. For oil spill detection, an image clustering is required, since the image of the water surface in the case of an oil spill is not homogeneous, and it is necessary to identify the oil slick as a separate object (Xing *et al*., 2015). Infrared images of water surface allow using different processing techniques for machine learning to define oil slicks and films. The main reason of creating the dataset by experiments in laboratory was the necessity to develop an innovative oil spill detection system based on machine learning techniques using infrared images of oil spills. A sufficient number of training samples and their representativeness are critical for image classifications (Hubert-Moy *et al*., 2001). Experiments for creating a data set were conducted out of the laboratory on sunny days under the real environment conditions according to experiment settings. Oil and water temperatures were measured using the infrared thermography system, and infrared images were taken in the real time. Infrared images were processed using segmentation, and pixel values were extracted as features (He *et al*., 2015). Processed data were used for creating a data set in csv format using developed python software. On the first step, method's K means was proposed for segmentation of an RPA infrared image for using to segment object of interest such as oil slick from water surface background. In the dedicated case of oil slick detection on water surface, the main objective is to select abnormally measurable features

to a group and collect other features to another group (Al-Ruzouq *et al*., 2020). Because of the high efficiency of the K means algorithm, it is widely used in the clustering of large-scale data (Capó, Pérez, & Lozano, 2017). This approach allows to segment an image into the following categories: oil slick and clean water to simplify the image and visualize oil slick's sectors using labelling for future processing for determination of oil spill border coordinates. For achieving maximum efficiency of K means method, optimal number of clusters K, into which data set will be clustered, was determined using Elbow method. The Elbow method was implemented using Python integrated development environment and indicated that K=4 is the optimal number of clusters for infrared images data set. K means is quantization simple method that is optimal for initial data analysis of an unlabelled real data set obtained during experiments. On the second step, two machine learning classification methods were used for classification: NB and SVM.

Two stage approach for infrared images of oil spill analysis with probability machine learning method using data set of an oil spill is more precise because NB classification is a probabilistic algorithm. On the second step, the NB method was used to predict classification from pixel values based on the training data set, created by classification model. It is likely that data obtained using remote sensing for oil spill detection was distributed according to the Gaussian probability density. For a reliable and efficient classification of oil slicks on water aquatorium, surface and probably oil pollution spills are required to implement probabilistic classification methods after clustering an image. NB model was developed for the classification of pixels and implemented for oil spill prediction using an experiment data set with infrared images. NB is based on Bayes theorem, but it defines that all features are independent. It is considered that NB classifiers, which make the simplifying assumption that the features are independent of each other, given the label (Duda & Hart, 1973). Therefore, advantages of NB classification are simple models that can achieve high level of accuracy. The implementation of Naïve Bayes is a simple algorithm used in the machine learning based on assumption that variables in the data set are not correlated. Correlation coefficient of variables is zero. For increased performance, simplification and increased accuracy, a data set is split into two random and independent subsets. First subset will be used for training of a model and second for evaluation and testing of performance of classifier model on real online data that were not applied for training of a model. For optimal oil slick detection on the second stage using NB classification, a training data set was created in real experimental conditions.

SVM is a reliable machine learning classification method for binary classification when classification labels in the data set are known (De Kerf *et al*., 2020). The data set developed in the laboratory by experiments in real conditions consists of different oil types labelled infrared images according to performing procedure. The main advantage of using SVM in an oil pollution detection missions is effective processing of high amount of data from different sensors of multi-sensor payloads. SVM finds a hyperplane that is a hyperplane in a feature space induced by a kernel K (the kernel defines a dot product in that space (Wahba, 1990)). The advantage of using SVM methods for oil spill detection is reliability to perform dividing data into two classes and possibility to create a model on small data sets. Pixel values as features were extracted to create a data set labelled with assumption that there are only two classes and further applied in the developed classification SVM model. For implementation of the developed SVM model, infrared oil pollution images, created according to the experiment procedure, these were reshaped to a two-dimensional array and three colour values (RGB). The data set A used in binary classification is a two-dimensional data set consisting of n points with assumption that there are only two classes yi $\in$ {+1,−1}, where yi is class label of data. The liner model was formulated by equation (1).

$$y_n [w^T x_i + b] \begin{cases} \geq 0 \text{ if oil spill detected} \\ < 0 \text{ if oil spill not detected} \end{cases} \quad (1)$$

where w^T is a weight vector;
T is size of training data set;
w_i is input vector;
b is bias.

In case of water contamination by oil, as a result of chemical and physical processes, the complex water quality parameters change. Measurement of complex water quality parameters in real time will provide the detection of oil spill and water quality parameters in the area of the oil spill as well as the degree of pollution. As a criterion for the quality of water in the oil spill area, physical and chemical parameters that correlate with oil pollution and can be used for classification, are taken for detection of an oil spill. The proposed method for classification of an oil spill is based on the concept of electrical conductivity of oil. Deviation of water conductivity from standard range could indicate water pollution

from oil products. In the oil spill zone, conductivity of water reduces from standard value due to lower conductivity of oil.

The electrical conductivity of water is a physical parameter by which oil pollution of water can be estimated. The level of dissolved oxygen that is the chemical parameter, is effected by oil sheen, that reduces oxygen diffusion process through air water surfaces rate. The oil layer directly can also cause the dissolved oxygen level in the water to become low due to the obstruction of the diffusion process from the air (Ifelebuegu *et al*., 2017). In the oil spill area, pH critical chemical parameter of eco system is considered acidic due to presence of heavy metals from oil. The lower pH in zones 1 and 2 could be attributed to either the presence of heavy metals from the oil spills or mainly from humic acid which resulted from the decomposition of forest materials in the river (Beadle, 1974). The experiments conducted in the laboratory confirmed the existence of the above mentioned dependences. Physical and chemical parameter changes have been observed conducting experiments in the laboratory in real conditions according to the experiment protocol. Logistic regression (LR) analysis is widely regarded as the statistic of choice for situations in which the occurrence of a binary outcome is to be predicted from one or more independent variables (Hosmer & Lemeshow, 2000). The mathematical formulation of logistic regression model for the oil pollution detection is proposed as a binary classification problem with the following predictor variables: pH, electrical conductivity, dissolved oxygen and dependent variable Y that can take the values 1 or 0. Taking into account the small dimensions off the data set, non-linearity, minimal number of sampling points and high volume of water quality data in the oil spill area a logistic regression model was developed to improve efficiency of monitoring. The deviation of quality parameters from standard seasonal values is a sign of oil pollution and a criterion for assessing the threat to the ecosystem in the area of oil pollution. The objective is analysing three parameters, such as conductivity pH, DO correlated with oil spill; therefore, a number of features taken is three.

In the developed complex machine learning algorithm (Figure 1), first, IR images segmentation using K means method is performed to identify segments with possible features of an oil spill from water surface background.

Using water quality parameters algorithm to detect oil pollution uses machine learning logistic regression using the developed model. At the final stage of the proposed algorithm, accuracy of classification results of the IR image and water quality parameters are evaluated.

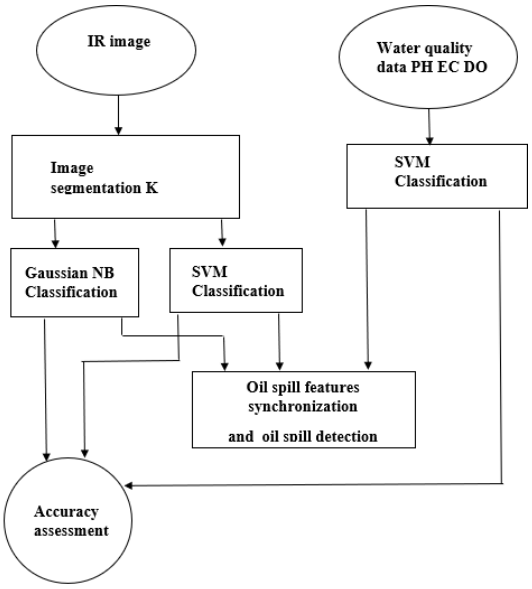

Figure 1. The developed complex machine learning algorithm.

Results and Discussion

The developed NB, SVM and logistic regression algorithms were run using Python environment. The NB and SVM model validation has been performed using the datasets that have been divided into training subset 70% and testing subset 30%. NB and SVM algorithms were used for IR image classification and detection of an oil spill. The logistic regression algorithm was used for the water quality parameters classification and prediction of an oil spill. The logistic regression algorithm validation has been performed using water quality dataset that has been divided into training subset 70% and testing subset 30%. In the oil spill experiment in the real environmental conditions, findings of NB, SVM and logistic regression were recorded and analyzed to estimate the performance of each developed classification model. The developed classifier tools were tuned according to appropriate procedures and parameters to achieve high accuracy results. A series of developed model runs was carried out using grid search approach to determine optimal parameters of SVM, NB and logistic regression for optimization performance. The developed NB and SVM oil spill classification models have been successfully tested in real experiment conditions in the laboratory using thermal camera images for classification of oil spill in a glass tank. In both NB and SVM models, image processing was started with K-means clustering. Afterwards, clusters of images were processed separately by NB and SVM models. The IR image of a controlled oil spill in the laboratory after segmentation was processed by NB model using

the data set, which was created to test model. Figure 2 demonstrates the result of processing using the developed NB model.

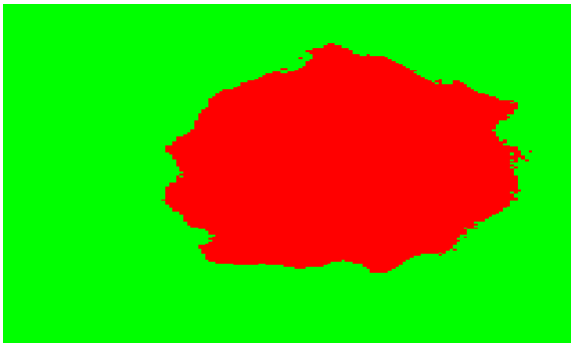

Figure 2. Result of processing using the developed NB model.

After processing using binary classification method, we have received resulting image with two classes: oil and water. It is likely that binary classification gives reliable results that can be used for determination of the oil slick border. When the RPA finds some areas of the border of an oil slick, these coordinates can be used to calculate the probability of spreading (Urbahs & Zavtkevics, 2020).

The IR image which was used for the NB model was processed by the SVM model using the data set which was created to test the model. Figure 3 shows the result of processing using the developed SVM model. After processing using the SVM method, we have received the resulting image with two classes: oil and water. It can be seen that the SVM model gives reliable results that can be used for detection of the oil slick border.

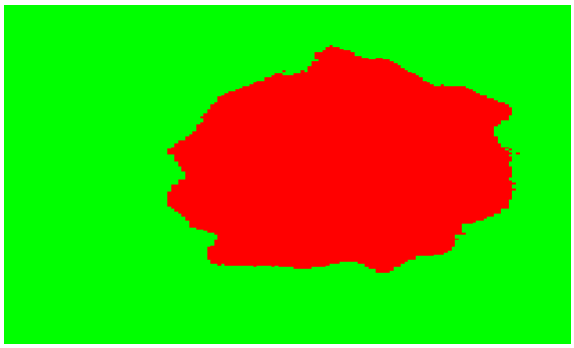

Figure 3. Result of processing using the developed SVM model.

It can be seen that oil spill has been correctly classified using the developed SVM and NB model; however, there is a small difference between the SVM and NB results. It is expected that there are differences between resulting images obtained after classification by two independent methods, as the NB is a generative model, while the SVM is a discrimination model.

To perform evaluation of accuracy and testing of the proposed model, IR images of engine and hydraulic oil slicks were created according to the experiment protocol in the real environmental conditions. The initial preprocessing of IR images for the NB and SVM methods was performed using the K means method for segmentation of oil. Based on the analysis of the training and test data sets, as the main parameter RBF kernel of the SVM method was selected, and the classification of infrared oil spill images was performed using it. For the machine learning SVM method, the accuracy of the classification results for IR images of engine and hydraulic oil slicks are presented in Table 1. The grid search using various pairs of penalty parameter C and hyperparameter gamma was conducted with the objective to define values with the best accuracy of the cross validation. The results of the experiments allow to derive the following dependencies: the best results of accuracy are obtained by C 150 with a gamma from 0.1 to 100. Taking into account the absence of recommendations on the choice gamma, the optimal value of gamma for model was set 1/n where n is the sample size.

Based on the analysis of the training and test data sets, as the main parameter variances smoothing of the NB method was selected, and the classification of infrared oil spill images was performed using it. For the machine learning NB method, the accuracy of the classification results for IR images of engine and hydraulic oil slicks are presented in Table 2. Variances smoothing is using float value provided to calculate the largest variances of each feature and adding it to the stability calculation variance. The results of the experiments allow to derive the following dependencies: it is likely that the best results of accuracy are obtained by variances smoothing 1e-9. Variance smoothing is used to improve model stability by adding a portion of largest variances for each feature to variances. Variance smoothing 1e-9 will be used to increase model stability. Based on accuracy assessment, it is determined that SVM method has better results than NB Gaussian classification method. It should be pointed out that the performance of the SVM is better in comparison with NB, as NB is a probabilistic machine learning method.

The logistic regression model of water quality data was run using hydraulic oil and engine oil spills modulated in plastic containers in the lab according to the experiment in the real conditions. For the machine learning logistic regression method, the accuracy of the classification of oil spill using parameters such as conductivity, dissolved oxygen, pH of engine and hydraulic oil slicks results are presented in Table 3.

Table 1

The accuracy of the classification of results for IR images using SVM method

Penalty parameter (C)	Hyperparameter gamma				
	0.1	1	10	50	100
0.1	95.4545455	95.4545455	95.4545455	95.4545455	95.4545455
1	95.4545455	95.4545455	95.4545455	95.4545455	95.4545455
10	95.4545455	95.4545455	95.4545455	95.4545455	95.4545455
50	95.7118353	95.7118353	95.7118353	95.7118353	95.7118353
100	95.7118353	95.883362	95.883362	96.054885	95.7118353
150	96.1406518	96.1406518	96.1406518	96.1406518	96.1406518

The grid search using various of C known as a 'hyperparameter', and regularizations and l1; l2 was conducted with the objective to define values with the best accuracy of the cross validation. The results of the experiments and grid search allow to derive the following dependencies: the best results of accuracy are obtained by C 100 with a regularization l1; taking into consideration the unavailability of recommendations on the choice of regularizations, the optimal value of regularization for model was set to 11. Due to the fact that the data set was created by real experiments reflecting real water quality data in the oil spill area, the optimal value of the parameter C was set to 100 as to define the data set data is more important.

Table 2

The accuracy of the classification of results for IR images using NB method

Variances smoothing	2e-9	1e-9
Accuracy	93.22	93.4

Table 3

The accuracy of the classification of oil spill using parameters such as conductivity, dissolved oxygen, pH

	Hyperparameter (C)				
	0.01	0.1	1	10	100
Regularization (l1)	97.299509	99.714379	99.924669	99.973017	99.978368
Regularization (l2)	97.188976	99.5494215	99.851346	99.945659	99.975567

Accuracy of the proposed complex algorithm was evaluated by an evaluating metric that takes into account the number of true oil spill detections in relation to the total size of experiment data set formulated in the equation (1).

The accuracy of the proposed complex algorithm for oil spill detection that includes SVM, NB developed models and logistic regression model with data set of defined water quality parameters was evaluated using experimental data set by evaluating metric that takes into account the number of true oil spill detections in relation to the total size of experiment data set formulated in the equation (2).

$$A=TD/N \quad (2)$$

where A is accuracy;
TD is true oil spill detection number;
N is size of experiment data set.

The accuracy of the proposed complex algorithm

is 99.5%; that is a significant value and allows to decrease false detections. It is likely that the extracted information about oil spill by the pixel values as features using SVM model has a clear distinction between oil slick and water. It can be seen that after pre-processing using K means clustering and classification based on the developed SVM model, the resulting image has a clear boundary of the oil spill.

Use of the logistic regression method for water quality parameters is reliable and efficient for detection.

The comparison of accuracy between the SVM model, NB model and logistic regression model allows to conclude that all the developed algorithms have sufficient performance; however, the logistic regression has better accuracy. Each classifier and logistic regression function was tuned using a grid search method and main tuning parameters to determine the optimal parameters of developed classification models.

Logistic regression is one of machine learning classification methods that uses logistic function for prediction of a categorical dependent variable which predicts two maximum values (0 or 1). Therefore, logistic regression function as a binary classification method can be implemented to determine probability of oil spill using the set of independent variables such as water quality parameters.

The state of a water ecosystem during monitoring of oil spills using RPA with multi-sensor payload is determined by many properties such as infrared image of the surface and water quality indicators. The proposed algorithm is based on the binary classification of infrared images and the classification of water quality parameters using the machine learning method logistic regression; the machine learning method allows to identify any oil pollution of water.

During the monitoring, using RPA equipped with thermal camera and water quality sensors, the value of the maximum deviations of water quality parameters and preprocessed infrared images classified by SVM and NB are used to determine the separation of clean water and oil slick. Using the proposed complex algorithm using for monitoring of oil spills RPA, taking into account that currently data collection of oil spill and water quality in the spill area are collected manually and require a lot of human resources, will significantly decrease labor cost of monitoring. The proposed complex algorithm allows to perform data analysis more efficiently; moreover, the developed machine learning models will further replace the manual work of the graphical information systems operators, reduce the probability of human error and save man-hours of work.

Conclusions

The proposed complex algorithm using thermal camera images and water quality parameters obtained by RPA is reliable and allows to detect oil spills in real time. The oil pollution monitoring using RPA requires a qualitative analysis using an IR camera and quotative analysis such as pH, conductivity; therefore, developed complex oil pollution detection algorithm will increase reliability and detection accuracy. The developed NB and SVM oil spill classification models have successfully been tested in real experiment conditions in the laboratory using thermal camera images. The logistic regression model of water quality data was run using hydraulic oil and engine oil spills modulated in plastic containers in the laboratory according to the experiment in the real conditions and demonstrated higher accuracy in comparison with SVM and NB models. Each classifier and logistic regression function were tuned using a grid search method and main tuning parameters to determine the optimal parameters of the developed classification models. The proposed algorithm is based on the binary SVM and NB classification of infrared images, and the classification of water quality parameters using the machine learning method logistic regression allows to rapidly and with high accuracy identify any oil pollution of water. The proposed complex algorithm achieves higher accuracy and efficiency; moreover, the developed machine learning models will further reduce the probability of human error and save man-hours of work. The present research has demonstrated that after experiments and grid search, optimal parameters of the developed models were determined and the detection accuracy of complex algorithm achieved 99.5%.

Acknowledgements

PostDoc Latvia NATIONAL DEVELOPMENT PLAN 2020 EUROPEAN UNION European Regional Development Fund
INVESTING IN YOUR FUTURE

This work has been supported by the European Regional Development Fund within the Activity 1.1.1.2 ‘Post-doctoral Research Aid’ of the Specific Aid Objective 1.1.1 ‘To increase the research and innovative capacity of scientific institutions of Latvia and the ability to attract external financing, investing in human resources and infrastructure’ of the Operational Programme ‘Growth and Employment’ (No.1.1.1.2/VIAA/4/20/650).

References

Al-Ruzouq, R., Barakat A. Gibril, M., Shanableh, A., Kais, A., Hamed, O., Al-Mansoori, S., & Ali Khalil, M.

(2020). Sensors, Features, and Machine Learning for Oil Spill Detection and Monitoring: A Review. *Remote Sensing*. 12. 3338. DOI: 10.3390/rs12203338.

Beadle, L.C. (1974). *The inland waters of tropical Africa. An introduction to tropical limnology*. Longman Group Ltd, Publishers, 74 Grosvenor Street, London: W. 1.

Capó, M., Pérez, A., & Lozano, J.A. (2017). An efficient approximation to the K-means clustering for massive data. *Knowledge-Based Systems*, 117, 56–69.

De Kerf, T., Gladines, J., Sels, S., & Vanlanduit, S. (2020). Oil Spill Detection Using Machine Learning and Infrared Images. *Remote Sensing.* 2020. 4090. DOI: 10.3390/rs12244090.

Duda, R.O., & Hart, P.E. (1974). Pattern classification and scene analysis. *A Wiley-Interscience publication.*

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd Edition, Wiley, New York. DOI: 10.1002/0471722146.

Hubert-Moy, L., Cotonnec, A., Le Du, L., Chardin, A., & Perez, P. (2001). A comparison of parametric classification procedures of remotely sensed data applied on different landscape units. *Remote Sensing of Environment*. 75, pp. 174–187.

Ifelebuegu, A.O., Ukpebor, J.E., Ahukannah, A.U., Nnadi, E.O., & Theophilus, S.C. (2017). Environmental effects of crude oil spill on the physicochemical and hydrobiological characteristics of the Nun River, Niger Delta. *Environmental monitoring and assessment*. 189(4), 1–12.

Urbahs, A., & Zavtkevics, V. (2019). Oil spill remote monitoring by using remotely piloted aircraft. *Aircraft Engineering and Aerospace Technology*. 91. DOI: 10.1108/AEAT-12-2017-0273.

Urbahs, A., & Zavtkevics, V. (2020). Oil Spill Detection Using Multi Remote Piloted Aircraft for the Environmental Monitoring of Sea Aquatorium. *Environmental and Climate Technologies*, 24, 1–22. DOI: 10.2478/rtuect-2020-0001.

Wahba, G. (1990). *Spline Models for Observational Data*. Vol. 59, Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611970128.

Xing, Q., Li, L., Lou, M., Bing, L., Zhao, R., & Li, Z. (2015). Observation of Oil Spills through Landsat Thermal Infrared Imagery: A Case of Deepwater Horizon. *Aquatic Procedia*. 3. DOI: 10.1016/j.aqpro.2015.02.205.

Zavtkevics, V., & Urbaha, M. (2022). Analysis of Remotely Piloted Aircraft Payload for Oil Spill Detection. *Latvian Journal of Physics and Technical Sciences*. 59, 71–82. DOI: 10.2478/lpts-2022-0034.