International Journal of Computer Applications (0975 - 8887)
Volume 176 - No.34, June 2020

Prediction of Lung Cancer Patients Survival Time using Regression Analysis and Image Processing Techniques

Al Hasib Mahamud
Ahsanullah University of Science and Technology
141 & 142 Love Road, Tejgaon Industrial Area
Dhaka-1208, Bangladesh

Mahmudul Islam
Ahsanullah University of Science and Technology
141 & 142 Love Road, Tejgaon Industrial Area
Dhaka-1208, Bangladesh

Raqeebir Rab
Ahsanullah University of Science and Technology
141 & 142 Love Road, Tejgaon Industrial Area
Dhaka-1208, Bangladesh

ABSTRACT

In recent time Lung cancer becomes one of the fatal and common disease in the world. The prediction of survival time will improve the care of patients. In this era of Artificial Intelligence, computer aided detection system can be helpful to estimate more precisely the patient's survival time. Inspired by the ongoing advances in image processing and machine learning in the bio-medical area, we have developed a model for predicting the survival period in tentative patients utilizing regression analysis and image processing techniques to assist doctors with historical data. Our proposed approach involves image acquisition,pre-processing,feature extraction and finally regression analysis to anticipate the survival time. Comparison analysis of three feature extraction techniques namely–Gray level co-occurrence matrix (GLCM) approach, Statistical Parametric approach and Hybrid approach which is the ensemble of both GLCM and Statistical Parametric approach have been performed. For predicting patient's survival time three different regression analysis algorithms have been used and have got best result using Support Vector Regression(SVR) with the lowest Mean Absolute Error (MAE) of 12.44 and Root Mean Square Error of 16.35 for Statistical Parametric approach.

General Terms

Computer Vision, Digital Image Processing

Keywords

Survival time, Computer tomography, Segmentation, Morphological opening, Machine learning

1. INTRODUCTION

Lung cancer is considered as one of the world's most dangerous cancers. Most of the cases the patients who are diagnostic for lung cancer have a low survival rate. Survival time can be defined as the period elapsing between the completion or institution of any procedure and death. It is specifically related to its detection time. The biological behavior of the tumor and its analysis help to choose correct treatment and can improve the consequences of cancer[2]. Accurate estimation of prognosis and survival duration is the most critical aspect of a clinical decision-making phase in patients with malignancies[6]. This research work aims to build a model that can predict the survival time of lung cancer patients. Different regression analysis techniques–linear regression, random forest, and support vector regression are employed for predicting survival rate. Three feature extraction techniques namely Gray level co-occurrence matrix (GLCM) approach, Statistical Parametric approach and Hybrid approach which is the ensemble of both GLCM and Statistical Parametric approaches have used to get the key attributes for comparing the predictive power between various methods.We have also tried to find out the best extraction method for features which can describe a specific pattern in lung CT scanning images. The result showed that Support Vector Regression (SVR) for Statistical Parametric approach got the lowest Mean Absolute Error (MAE) of 12.44 and Root Mean Square Error (RMSE) of 16.35. At the same time, an observation of different techniques of digital image processing like smoothing, enhancement, segmentation, morphological opening etc is taken place for processing CT scan images. The rest of the paper is organized as follows: section 2 summarization of the related researches, section 3 proposed methodology, section 4 result and comparison analysis and section 5 conclusion.

2. LITERATURE REVIEW

To understand the problem and to decide working approach, a number of conference and journal articles relevant to this research are reviewed. Extraordinary studies have been conducted to diagnose lung cancer, but only a handful of them have been focusing on the estimation of lung cancer survival.
Vijay A.Gajdhane et al. [7] proposed a model to detect lung cancer stages. They used the CT scanning images as data in their method. Median filtering, gabor filtering, and watershed algorithm are included in their suggested approach. Features are extracted as scalar value and SVM is applied as classifier. They extracted three features (area, eccentricity, perimeter). Based on their different values

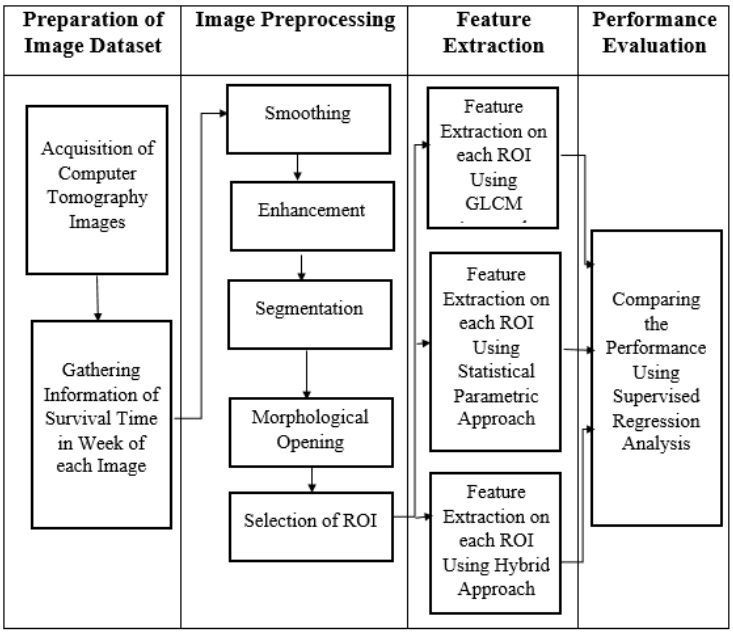

Fig. 1. Workflow of Proposed Methodology.

of features they detected different stages.
Harry B. Burke et al. [3] proposed artificial neural networks based model to improve the accuracy of cancer survival prediction. For this purpose they have used Patient Care Evaluation dataset and SEER breast carcinoma dataset. They have employed artificial neural network with three interconnected layers of nodes and achieved 78.4% accuracy using ANN.
YenChan et al.[6]proposed a model that used Artificial Neural Network to predict the survival rate of cancer patients from microarray and clinical data and showed good accuracy.
Azadeh Bashiri et al. [2] analyzed gene expression using machine learning techniques. They have involved preprocessing, method expression and analyzing gene expression data in proposed method. Using the combination of ANN and Fuzzy logic they have achieved 93% accuracy.
Md Sajib Ahmed et al. [4] proposed a model that can detect tuberculosis type and multidrug resistant tuberculosis. They have utilized CT scan images for their research and have generated region of interest using mask images. They have used GLCM for feature extraction purpose and averaged slice wise attribute values for eight different types of classification techniques. Most of the authors have used categorized attributes for predicting survival time of lung cancer patients.

3. METHODOLOGY

The whole method can be divided into three main phases.

—Image pre-processing

—Feature extraction from processed image

—Applying different regression algorithms for acquiring the final output

Process involved in three phases for prediction of patient's survival time, the whole system of the method can be viewed in Figure 1.

3.1 Image Acquisition

For analysis purpose, the dataset is taken from The Cancer Imaging Archive (TCIA).This site is dedicated for high quality cancer datasets.
The dataset contained images and information about 61 patients. All the images are DICOM format and images dimensions are 512 x 512, which is perfect for implementing as train and test sets in this study.

Fig. 2. Sample Slice of CT Image

3.2 Image Pre-processing

To improve the quality of an image, image pre-processing are required. Mahmudul et al.[9] proposed a model in a target to processing lung ct scan images for predicting stage of cancer patients. From there a method is proposed to process lung ct scan images utilizing different image pre-processing techniques like smoothing, enhancement, morphological opening and segmentation.Fig 3 describes the steps of image processing of our proposed methodology and Fig 4 shows the outcome of each steps of image processing.

3.3 Feature Extraction

Large dimensional input data is difficult to be processed. Feature selection algorithm is used to reduce the dimension of data. The selected features are expected to contain the relevant information from the input data.
Three different feature extraction processes are performed in this model.

—Gray Level Co-occurrence Matrix Approach

—Statistical Parametric Approach

—Hybrid Approach

3.3.1 ***Gray Level Co-occurrence Matrix Approach:*** . The Gray Level Co-occurrence Matrix (GLCM) is a statistical calculation which calculates how often a pixel with gray-level value i occurs with the value j [9] .It can occur either horizontally, vertically, or diagonally to adjacent pixels. GLCM is a matrix that describes how a gray level pixel occurs within another gray level pixel under an area of interest by following a specified linear relationship [4].By calculating GLCM we got the following features.

—**Contrast**
Contrast calculates the intensity contrast between a pixel and its neighbor pixel over the whole image [9].

$$Contrast = \sum_{i,j}^{N-1} |i-j|^2 P_{ij} \quad (1)$$

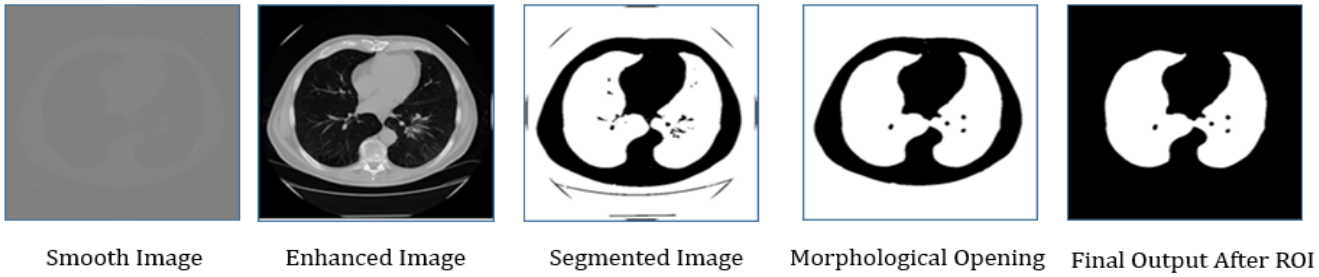

Fig. 4. Outcome of Each Steps Using Image Preprocessing.

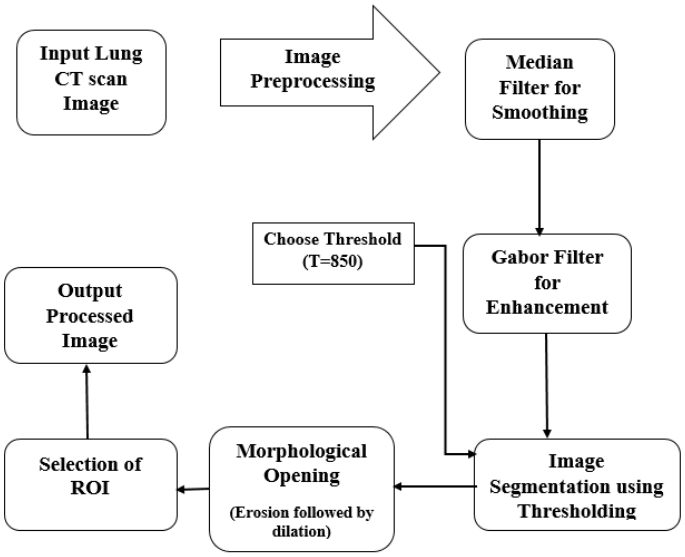

Fig. 3. Flow Diagram of Image Preprocessing.

Here, P_{ij} is the element i,j of the normalized GLCM and N is the number of gray levels in the image .

—**Correlation**
Correlation calculates how a pixel is correlated with its neighbor over the whole image [9].

$$Correlation = \sum_{i,j}^{N-1} \frac{(i-\mu_i)(j-\mu_j)P_{ij}}{\sigma_i \sigma_j} \tag{2}$$

Here μ is the GLCM mean can be calculated as:

$$\mu = \sum_{i,j=0}^{N-1} iP_{ij} \tag{3}$$

And σ is the variance of the intensities of all reference pixels in the relationships that contributed to the GLCM, calculated as:

$$\sigma = \sqrt{\sum_{i,j=0}^{N-1} (i-\mu)^2 P_{ij}} \tag{4}$$

—**Energy**
Energy calculates the summation of the squared elements of the whole image [9].

$$Energy = \sum_{i,j}^{N-1} P_{ij}^2 \tag{5}$$

—**Entropy**
Entropy is a scalar value to characterize the texture of an image [9].

$$Entropy = \sum_{i,j} -\ln(P_{ij})P_{ij} \tag{6}$$

—**Homogeneity**
Homogeneity calculates how close the distribution of the elements in the GLCM diagonal [9] .

$$Homogeneity = \sum_{i,j}^{N-1} \frac{P_{ij}}{1+|i-j|} \tag{7}$$

In this approach some others features are also considered.

—**Convexity**
Convexity evaluates tumor shape and it is computed as a proportion of tumor size to raised body. It records the morphology change in tumor and anticipate by and large survival pace of patient when dichotomized is the middle worth [8].

—**Mean Intensity Value**
The mean of a particular pixel is just simply the pixel value, since there is only one sample [9].

$$Mean = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} P(i,j) \tag{8}$$

Here $M \times N$ is the size of the image.

—**Area**
Area gives the total number of nodule pixels in a given ROI. Transformation function is used to calculate the area and creates a matrix contains pixel of 255 values from extracted ROI [12].

3.3.2 ***Statistical Parametric Approach:*** . In a classical statistics, to make any sort of measurable surmising is done by assuming that the data originates from a specific distribution. Among all distribution in such a family, a specific one can be recognized by a set of parameters [9].
In the statistical parametric approach, the following major parameters considered:

—**Mean**
Mean estimates the average of the intensity value of pixel over the entire image [10].

$$Mean = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}P(i,j) \tag{9}$$

Where $P(i,j)$ is the intensity value of the pixel at the point $(i,j).M \times N$ is the size of the image.

—**Standard Deviation**
Standard deviation describes the measure of variety or dispersion from the given arrangement of intensity values [9].

$$StandardDeviation = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(P(i,j)-\mu)^2} \tag{10}$$

Where μ is the mean value.

—**Higher Order Moments**
Higher order moments is utilized as the third or higher power of sample and it uses constant, linear and quadratic terms [10]. The fifth and sixth central moments are given respectively.

$$FifthOrderMoment = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[\frac{P(i,j)-\mu}{\sigma}]^5} \tag{11}$$

$$SixthOrderMoment = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[\frac{P(i,j)-\mu}{\sigma}]^6} \tag{12}$$

For this approach, we have also considered Convexity Score and Entropy Ratio to improve the performance. Having six features an observation have described that if new features are added then the performance of the model decreases. Convexity score is already described in GLCM approach.Entropy Ratio is described below.

—**Entropy Ratio**
Entropy ratio is the measure of variation computed on the pixel histogram distribution within a given ROI [8]. It is defined as:

$$EntropyRatio = \sum_{i=1}^{n=255} -p_i log_2 p_i \tag{13}$$

3.3.3 ***Hybrid Approach:*** . Some features from both of the approaches of GLCM and statistical parametric approach considered for Hybrid Approach. For feature selection Standard Deviation, Fifth Central Moment, Sixth Central Moment, Convexity Score, Entropy Ratio , Contrast ,Correlation,Energy, Homogeneity and Area are considered for possible features.

3.4 Regression Analysis to Predict Survival Time

To predict survival time of patients, regression analysis have been performed. It is the process for estimating the relationship among variables.As the dependent variable for survival time is real valued and continuous in nature so regression analysis is the best way for prediction. Linear regression, Support Vector regression,Random Forest regression and Lasso regression are used to predict survival time.

3.4.1 ***Linear Regression:*** . Linear regression is a linear approach to deal with the relationship between independent variables and dependent variable. If X is independent variable and Y is dependent variable then linear equation is represented as $Y = mX + c$, where m is the slope and c is co-efficient.

3.4.2 ***Random Forest Regression:*** . In random forest regression is an ensemble technique where multiple decision trees are combined to predict value. The basic random forests algorithm for regression and for classification are identical.The differences are– MSE criterion that grow with the individual decision trees and the predicted target variable is calculated as the average prediction of over all decision trees.

3.4.3 ***Support Vector Regression:*** . Support vector regression follows the all main features of support vector machine algorithm.For support vector regression equation of the hyper plane is:

$$wx + b = 0 \tag{14}$$

Two equation for the boundary lines are:

$$wx + b = +e \tag{15}$$

$$wx + b = -e \tag{16}$$

In this methodology to perform SVR Gaussian kernel is utilized.

$$K(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right) \tag{17}$$

4. RESULT AND PERFORMANCE ANALYSIS

4.1 Dataset Formation

For predicting survival rate of lung cancer patients 61 patients history are gathered as the highest number of instance container, which contains 4682 number of images and patient's survival time are given in week.So survival time can be used as target column and features which are collected from different feature extraction approaches, are used as input column.

4.2 Evaluation Metrics

The classification performances are expressed by performance metrics by the following evaluation metrics:

—Mean Absolute Error(MAE)
Mean absolute error (MAE) is a measurement of difference between two continuous variables.If x_i is the actual expected output and y_i is the model's prediction then

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - x_i| \tag{18}$$

—Mean Squared Error(MSE)
Mean squared error (MSE) estimates the average of the squares of the errors.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2 \tag{19}$$

—Root Mean Square Error(RMSE)
Root-mean-square error (RMSE) is the measurement of the dif-

Table 1. Regression Analysis on GLCM Approach

Regression Model	MAE	MSE	RMSE
Linear Regression	15.29	316.81	17.79
Support Vector Regression	14.99	316.02	17.77
Random Forest Regression	**14.14**	**276.86**	**16.63**
Lasso Regression	14.54	286.81	16.23

Table 2. Regression Analysis on Statistical Parametric Approach

Regression Model	MAE	MSE	RMSE
Linear Regression	16.51	363.13	19.05
Support Vector Regression	**12.44**	**267.55**	**16.35**
Random Forest Regression	13.33	278.96	16.70
Lasso Regression	14.71	293.43	17.12

Table 3. Regression Analysis on Hybrid Approach

Regression Model	MAE	MSE	RMSE
Linear Regression	15.16	314.37	17.73
Support Vector Regression	13.49	267.16	16.35
Random Forest Regression	**13.47**	**276.38**	**16.62**
Lasso Regression	14.54	286.81	16.93

ferences between the predicted values and the real values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2} \quad (20)$$

4.3 Final Outcome for Survival Time Prediction of Lung Cancer Patients

To predict patients survival time different machine learning algorithms are applied with different feature extraction approaches to identify the best regression algorithm and feature extraction approach.
Table 1 shows the result of regression analysis on dataset which is prepared in GLCM approach, Table 2 shows the result in Statistical Parametric approach and Table 3 shows the result in Hybrid approach.
It is clear from above three tables that using Statistical parametric approach best result can be obtained for our selected dataset to predict patients survival time. Support vector regression (SVR) in Statistical approach performs best with having lowest mean absolute error (MAE) 12.44. For performing SVR rbf kernel is used for creating appropriate feature space and statistical parametric approach doesn't make the model complex compare to other approaches. Figure 5 shows the performance of support vector regression in different approaches.

4.4 Performance Analysis on Proposed Methodology and others Methodology

A number of articles have been reviewed so far. A number of work has been done on forecasting lung cancer and lung cancer phases, but very little work has been done linked to the estimation of survival period and no work has been done to predict survival time dependent on image processing techniques. LWC Chan et.al [5] and Chip M. Lynch et. al [11] used categorical data for predicting patients survival such as age,gender,cancer stage,number of lesions etc. But in this methodology image processing techniques are used on lung CT scan images. P.Basak et. al [1] proposed their model for lung cancer detection using the features of area, perimeter and

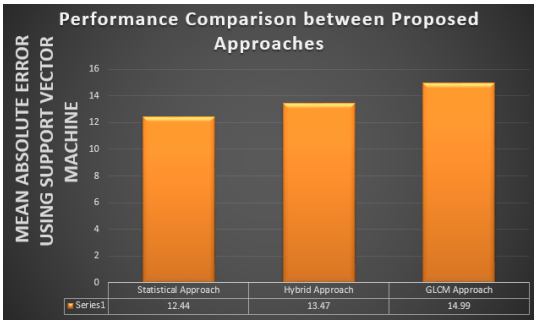

Fig. 5. Performance analysis of support vector regression in different approaches.

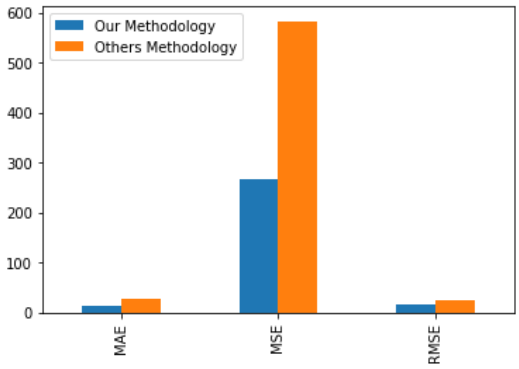

Fig. 6. Performance Analysis between Proposed Methodology and Others Methodology.

eccientricity from lung CT scan images. These features have been taken on the input data of this research and SVR has been implemented to predict survival time. But the MAE, MSE and RMSE scores are much more higher than proposed Statistical parametric approach which is shown in Figure 6. So this can be easily declarable that using this methodology better performance can be achieved for predicting survival time of patients.

5. REFERENCES

[1] Priyanka Basak and Asoke Nath. Detection of different stages of lungs cancer in ct-scan images using image processing techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 5:9708–9719, 06 2017.

[2] Azadeh Bashiri, Marjan Ghazisaeedi, Reza Safdari, Leila Shahmoradi, and Hamide Ehtesham. Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review. *Iranian journal of public health*, 46(2):165, 2017.

[3] Harry B Burke, Philip H Goodman, David B Rosen, Donald E Henson, John N Weinstein, Frank E Harrell Jr, Jeffrey R Marks, David P Winchester, and David G Bostwick. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857–862, 1997.

[4] Carlos Pampulim Caldeira. Atas das oitavas jornadas de informática da universidade de évora. *algorithms*, 17:9, 2015.

[5] Lawrence Chan, T. Chan, L.F. Cheng, and W.S. Mak. Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. pages 467 – 470, 01 2011.

[6] Yen-Chen Chen, Wen-Wen Yang, and Hung-Wen Chiu. Artificial neural network prediction for cancer survival time by gene expression data. In *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4. IEEE, 2009.

[7] Vijay A Gajdhane and LM Deshpande. Detection of lung cancer stages on CT scan images by using various image processing techniques. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 16(5):28–35, 2014.

[8] Olya Grove, Anders E Berglund, Matthew B Schabath, Hugo JWL Aerts, Andre Dekker, Hua Wang, Emmanuel Rios Velazquez, Philippe Lambin, Yuhua Gu, Yoganand Balagurunathan, et al. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PloS one*, 10(3):e0118261, 2015.

[9] M. Islam, A. H. Mahamud, and R. Rab. Analysis of ct scan images to predict lung cancer stages using image processing techniques. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0961–0967, Oct 2019.

[10] Jinsa Kuruvilla and K Gunavathi. Lung cancer classification using neural networks for ct images. *Computer methods and programs in biomedicine*, 113(1):202–209, 2014.

[11] Chip Lynch, Behnaz Abdollahi, Joshua Fuqua, Alexandra deCarlo, James Bartholomai, Rayeanne Balgemann, Victor Berkel, and Hermann Frieboes. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*, 108, 09 2017.

[12] Rajani R Mhetre, Pooja P Kawathekar, Sneha S Kadam, and Megha B Gore. Development of methodology for transforming ct images indicating location and size of lung cancer nodule. In *Techno-Societal 2016, International Conference on Advanced Technologies for Societal Applications*, pages 293–301. Springer, 2016.