DOI: 10.62897/COS2023.1-1.54

DETECTING ANOMALIES IN THE FM FREQUENCY BAND USING STATISTICAL METHODS

Szilárd L. Takács
Széchenyi István University, Egyetem tér 1, 9026 Győr, Hungary
takacs.szilard.laszlo@sze.hu

Based on Act C of 2003 on electronic communications, in Hungary, the National Media and Infocommunications Authority is responsible for ensuring harmful interference-free frequency usage and electromagnetic compatibility. Continuous measurements are conducted nationwide in order to reach this goal, but the evaluation and analysis of anomalies are time-consuming.

This research focuses on the detection of anomalies in the FM radio frequency spectrum. Within that, the study was concerned with the outages of radio transmission and the outages of modulation. The goal of this study is to automate the detection process, providing real-time alerts for potential anomalies and saving valuable time for spectrum monitoring engineers.

In order to solve the problem, statistical learning was used, including classification algorithms. Comparing the following algorithms: k nearest neighbor classification method, logistic regression, linear discriminant analysis, quadratic discriminant analysis, naive Bayes classification, support vector machines, and random forests. The most efficient method for this is Support Vector Machines, which can identify the phenomena with 93.28 % accuracy.

Statistical machine learning is highly efficient at identifying known phenomena in spectrum monitoring and generating real-time alerts. Alerts can be generated within a minute, effectively providing real-time information.

1. Introduction

In the field of infocommunication, establishing uninterrupted communication is of critical importance and falls under the jurisdiction of the National Media and Infocommunications Authority (NMHH). The NMHH is responsible for frequency allocation and regularly announces calls for frequency tenders. The winner of such a tender, who obtains the allocated frequency, is obligated to utilize it. It is essential to continuously monitor and maintain the assigned frequency, as NMHH is responsible for the harmful interference-free frequency.

The NMHH (National Media and Infocommunications Authority, 2015) provides rules for governing broadcasting. This decree specifies the distribution of national frequency bands and governs frequency usage. Radio broadcasting (FM) takes place between 87.5 MHz and 108 MHz in the VHF band, and in this thesis, FM broadcasting was exclusively discussed. The national frequency allocation and the guidelines for the use of frequency bands, as outlined in the 7/2015 (XI. 13.) NMHH decree (NFFF), are visualized and managed using the software named STIR.

Previous research has been conducted on the topic, where the spectrum is treated as time series (Yokkampon et al., 2020). Based on these studies, it can be concluded that treating the spectrum as time series leads to more efficient anomaly detection. There were times when the spectrum was monitored in time using a

sliding window solution, and then machine learning was applied (Peng et al., 2022). The radio spectrum is extensive and diverse, making it unclear what type of anomaly detection we would like to perform, as there are known phenomena that occur regularly, as well as isolated occurrences that may not be repeated. Therefore, it is crucial to precisely define the type of anomaly we want to detect to determine whether supervised or unsupervised learning should be employed.

Articles have been published regarding the application of unsupervised machine learning. Since labeling is exceptionally challenging, in most cases, it is easier to train on good samples and request alerts for deviations. An example is the article "Unsupervised Wireless Spectrum Anomaly Detection with Interpretable Features," which achieves 80 % accuracy in anomaly detection and provides only a 1 % false alarm rate in the system (Rajendran et al., 2019). AnoFSTSCNN approach has already been proposed in a published article. An alternative for radio monitoring and spectrum anomaly detection, the AnoFSTSCNN's performance was confirmed on a simulated data set, and it was demonstrated that the accuracy is significantly higher than that of the ST- and STFT-based techniques (Wang et al., 2022).

Nowadays, in the field of spectrum monitoring, there are measurements running all the time, which engineers evaluate days later. To simplify the workflow and to achieve real anomaly detection, it is advisable to improve the current unsupervised methods. Since some of the detected anomalies are recurrent, supervised machine-learning algorithms can be applied.

During the current research, the study focused on identifying two anomalies: radio transmission interruption and radio broadcast modulation shutdown. As these are well-distinguished phenomena, they cannot be considered classical anomaly detection cases. Therefore, supervised machine learning methods were employed in this study.

The **transmission outage** phenomenon occurs when a radio program abruptly terminates. This behavior is depicted in Figure 1, where the color change signifies a high electric field strength represented in the figure, which then rapidly diminishes beyond a certain point. The color intensity and electric field strength will increase again after a predefined period of time when the transmission resumes.

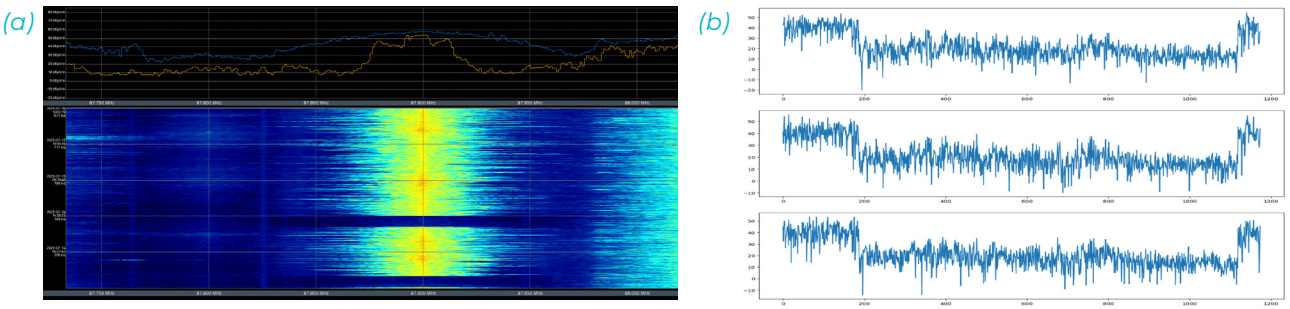

Figure 1: Radio transmission outages. a) Spectrogram about radio transmission outages. The x-axis is the frequency, y-axis is the time, and the color represents the magnitude of the electric field strength. b) A time series of outages. x-axis is the time, y-axis is the electric field strength value of the signal

Modulation outages is the other phenomenon under study, as depicted in Figure 3. A sinusoidal signal can carry information thanks to various modulation techniques. The three essential components of a sinusoidal signal—amplitude, phase, and frequency—can all be modulated. When modulation is stopped, the carrier frequency and the subcarriers are also discernible.

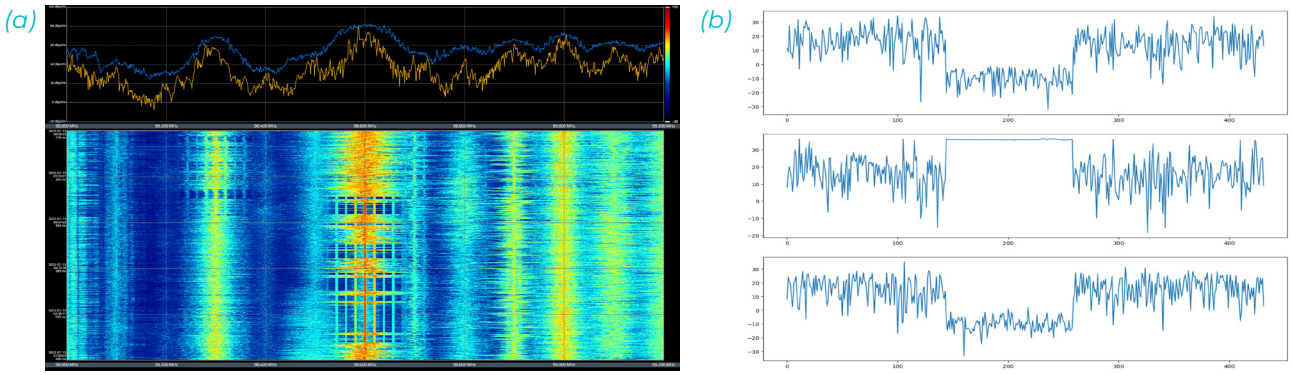

Figure 2: Modulation outages. a) Spectrogram about modulation outages. The x-axis is the frequency, the y-axis is the time, and the color represents the magnitude of the electric field strength. b) A time series of modulation outages; the x-axis is the time, and the y-axis is the electric field strength value of the signal.

The aim of the results is to develop a supervised machine learning-based signaling system compatible with the national spectral monitoring network used in Hungary. To implement the training, statistical machine learning will be used. Different algorithms will be applied and compared to determine which one yields the most accurate results.

2. Technology prospection and technology roadmap

Two hypotheses will be proven, which are:

- Supervised machine learning (only for known phenomena) can be used to develop much more efficient programs than unsupervised anomaly detection programs.
- The program can replace some of the engineering work, saving considerable time for the National Media and Communications Authority's staff.

In the following, several statistical-based classification methods have been used. Initially, a database was created, followed by a correlation-based analysis. After that, several algorithms were used to classify the phenomena. Finally, the accuracy of the algorithms was compared.

As presented in Table 1, the dataset was manually compiled, encompassing seven electric field intensity values and one label for each time event. The rows in the dataset represent various time periods with a precision of 25 seconds. The third column corresponds to the field strength measured at the intermediate frequency (IF), while the subsequent columns represent electric field intensity values measured at frequencies spaced 1 kHz apart from the IF. Each of these field strength values serves as a valid measure to characterize the particular transmission since it operates at a frequency of 150 kHz, which is typical for FM transmissions. The provided labels describe the specific activities that occurred during the respective radio transmission at a given time (Takács, 2023).

Table 1: Input data set, mixed (Takács, 2023)

0	1	2	3 (IF)	4	5	6	Class
-0.04	11.66	13.66	11.36	9.06	-1.03	9.77	switch-off
40.86	53.26	55.66	49.06	28.66	9.07	9.77	modulation switch-off
13.63	15.73	12.83	11.23	14.23	16.73	13.23	nothing
34.96	27.06	27.86	32.06	27.96	33.27	38.77	nothing
-13.67	-6.37	-10.67	-7.37	-12.97	-16.47	-14.77	switch-off

3. Application of supervised machine learning for anomaly detection

Data analysis was performed on the prepared dataset presented above. First, the correlation was examined. The purpose of correlation analysis is to find out which statistical learning algorithms can be applied. Then, statistical machine learning algorithms were trained with the dataset.

3.1. Correlation-based analysis

In the initial step, correlation analysis was conducted by examining the correlations between vectors constructed from the electric field intensities in each row. This analysis was performed within identical classes and across different classes. The average correlations are presented in Table 2. Based on these findings, it is evident that only the modulation switch-off exhibited strong correlations, rendering correlation-based discrimination inapplicable for the other two cases. The visual representation of this observation is shown in Figure 3, where the x-axis represents the correlation values and the y-axis denotes their corresponding frequencies. The figure exclusively focuses on correlations within the same group.

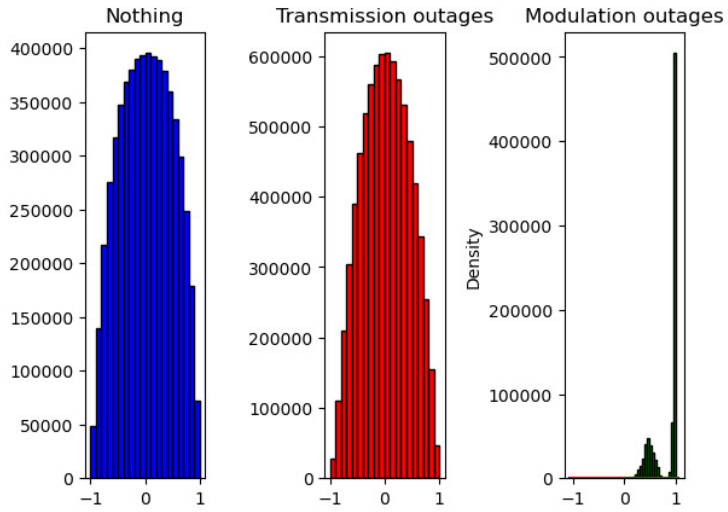

Figure 3: A visual representation of the variance of correlation between the same classes

Table 2: Size of correlation on average between classes

	Radio transmission outages	Modulation outages	Nothing
Radio transmission outages	0.01728	0.11547	0.01496
Modulation outages	0.11547	0.89382	0.12291
Nothing	0.01496	0.12291	0.03970

3.2. Statistical learning and application

k Nearest Neighbors classification method (kNN): The fundamental principle of k-nearest neighbors (kNN) is to compare vectors containing electric field intensities and determine, using the Euclidean distance, which class the newly arrived data belongs to (Hastie et al., 2009). For this purpose, the following equation is used:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{1}$$

where $N_k(x)$ is the neighborhood of x defined by the k closest points x_i in the training sample. We locate the k observations in the input space with x_i closest to x and we average their results.

When using 9 nearest neighbors for cross-validation, the obtained values are as follows:

[0.641079 0.744813 0.565398 0.766089 0.844982 0.781314]

Based on the k-nearest neighbors (kNN) method and using 9 nearest neighbors for cross-validation, the best-achieved result is 84.5 %.

Logistic regression: The essence of logistic regression is to model the relationship between the dependent variable (output variable), and the independent variables using a linear association and then transform this linear relationship with a non-linear logistic function (Alpaydin, 2020).

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \quad (2)$$

where:

- $P(y = 1|x)$ is the probability of the dependent variable (output variable), y being 1 given the input variables x.
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ are the regression coefficients (model parameters).
- $x_1, x_2, \cdots, x_n$ are the values of the independent variables (input variables).

The following values were obtained for logistic regression:

[0.548409 0.679115 0.759169 0.610381 0.683737 0.912111]

Based on the logistic regression for cross-validation, the best-achieved result is 91.2 %.

Linear discriminant analysis (LDA): LDA is a statistical technique used to discern a linear combination of features that can effectively discriminate between two or more groups of entities or events. The resultant linear combination can be utilized as a linear classifier, or more commonly, to reduce the dimensionality of the data before applying a subsequent classification algorithm. LDA assumes that the distributions within each class possess equal covariance matrices, and adhere to multivariate normal distributions (Ghojogh and Crowley, 2019).

The following values were obtained for LDA:

[0.580221 0.674965 0.697578 0.633218 0.710035 0.901730]

Based on the linear discriminant analysis for cross-validation, the best-achieved result is 90.2 %.

Quadratic discriminant analysis (QDA): QDA is a discriminant method that differs from LDA in its assumption of unequal covariance matrices between classes. Unlike LDA, QDA allows for the consideration of more general covariance matrices for each class. This flexibility allows QDA to handle situations where the data within different classes exhibit varying levels of variability and correlation patterns, making it a powerful tool for classification tasks when the assumption of equal covariance matrices is not valid. By accounting for these distinct covariance structures, QDA can provide more accurate and precise discrimination between groups of entities or events (Ghojogh and Crowley, 2019). The following values were obtained for QDA:

[0.544260 0.706086 0.687197 0.729412 0.880277 0.914187]

Based on the quadratic discriminant analysis for cross-validation, the best-achieved result is 91.4 %.

Naive Bayes classification: In statistics, naive Bayes classifiers belong to a category of simple "probabilis-

tic classifiers" that utilize the Bayes theorem by making strong (naive) independence assumptions between the features. These classifiers merge the Bayes probability model with a decision rule (Kotsiantis, 2013). A common rule is to select the hypothesis that has the highest probability of minimizing the chances of misclassification; this is known as the maximum a posteriori or MAP decision rule (Russell and Norvig, 2005). The corresponding classifier, known as the Bayes classifier, is a function that assigns a class label $\hat{y} = C_k$ for some k, as follows: $\hat{y} = argmax \prod_{i=1} p(x_i \mid C_k)$ (Hart et al., 2000).

The following values were obtained for Naive Bayes classification:

[0.517981 0.673582 0.679585 0.606920 0.477509 0.845675]

Based on the Naive Bayes classification for cross-validation, the best-achieved result is 84.6 %.

Support Vector Machines (SVM): This method works by finding the optimal hyperplane that best separates different classes in the feature space, maximizing the margin between them. SVM's decision boundary is determined by a subset of the training data points, called support vectors, which play a crucial role in defining the optimal separation.

The following values were obtained for SVC:

[0.533195 0.706086 0.555709 0.768858 0.932872 0.773010]

Based on the support vector machines for cross-validation, the best-achieved result is 77.3 %.

Random forest: This method constructs multiple decision trees during training and aggregates their predictions through a voting or averaging mechanism to make final predictions. The key feature of Random Forest is its use of random sampling of both the training data and the features, which introduces diversity among the individual trees and helps to reduce overfitting. Due to its ability to handle high-dimensional data and its robustness against noise and outliers, Random Forest has gained popularity in various domains.

The following values were obtained for Random Forest:

[0.671508 0.769709 0.541176 0.755017 0.831834 0.802076]

Based on Random Forest for cross-validation, the best-achieved result is 83.1 %.

4. Comparison of methods

In statistical learning, cross-validation was employed, which provided six numerical values for each method. Based on these values, the average was examined, and the best achievable value with the method was also considered. Table 3. summarises the results.

Table 3: Spectrum-monitored radio transmitter outages and modulation outages using different statistical learning methods and their accuracy

Method	k Nearest Neighbours classification method	Logistic regression	Linear discriminant analysis	Quadratic discriminant analysis	Naive Bayes classification	Support vector machines	Random Forest
Mean Accuracy	0.72395	0.69882	0.69962	0.74357	0.63354	0.71162	0.72855
Max Accuracy	0.84498	0.91211	0.90173	0.91418	0.84567	0.93287	0.83183

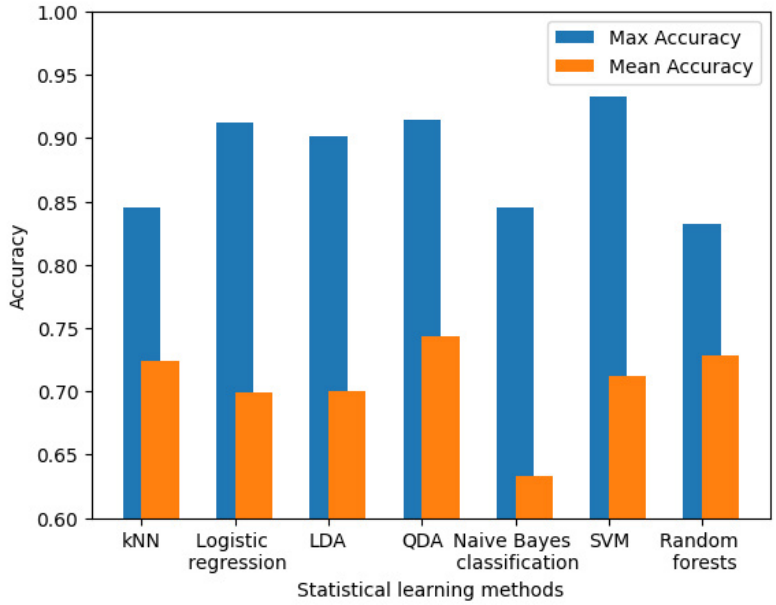

Figure 4: Comparison of statistical learning algorithms based on average and maximum accuracy

Based on these observations, it can be concluded that, on average, the Quadratic discriminant analysis method yields the best results, while the maximum achievement is attainable through the use of Support vector machines.

5. Conclusions

During the research, statistical learning was applied (supervised machine learning) to detect anomalies in the spectrum, specifically radio transmission outages and modulation outages. Supervised machine learning was chosen because these phenomena are distinguishable from each other. For each time point, electric field strength values were associated with the intermediate frequency of radio transmission. The average accuracy of the methods was above 60 %, with some cases achieving an average accuracy of over 70 %. For certain methods, the maximum accuracy after cross-validation has exceeded 90 %. As a result of the study, the best method for solving this problem is SVM, with an accuracy of 93.28 %. The detection can be further improved by treating incoming signals as time series instead of predicting individual time points.

In summary, statistical machine learning can be highly effective in identifying known phenomena in the field of spectrum monitoring, thus creating real-time alerts. The most efficient method for this is the Support Vector Machines, which can identify the phenomena with 93.28 % accuracy. As each phenomenon exhibits for a longer period of time, if the alarm is not triggered on the first anomaly detection, instead it triggers after a certain number of consecutive identical predictions, we achieve a reasonably accurate quasi-real-time signal, in which false alarms can be minimized.

References

- Alpaydin E., 2020, Introduction to machine learning, MIT press, Cambridge, UK, 140-145.
- Ghojogh B., Crowley M., 2019, Linear and Quadratic Discriminant Analysis: Tutorial, arXiv preprint arXiv:1906.02590.
- Hart E. P., Stork, G. D., Duda O. R., 2000, Pattern classification. Wiley, Hoboken, NJ, United States.

- Hastie T., Tibshirani R., Friedman J.H., 2017. The elements of statistical learning: data mining, inference, and prediction, Second edition, corrected at 12th printing 2017. 2nd Ed, Springer series in statistics. Springer, New York, NY, United States, DOI: 10.1007/b94608.
- Kotsiantis S. B., 2013, Decision trees: a recent overview. Artificial Intelligence Review, 39, 261-283.
- National Media and Infocommunications Authority decree - About the national frequency allocation and the rules for the use of frequency bands. In: (7/2015. (XI. 13.)), Hungarian National Legislation-repository.
- Peng C., Hu W., Wang L., 2022, Spectrum anomaly detection based on spatio-temporal network prediction. Electronics, 11(11), 1770.
- Rajendran S., Meert W., Lenders V., Sofie P., 2019, Unsupervised wireless spectrum anomaly detection with interpretable features. IEEE Transactions on Cognitive Communications and Networking, 5(3), 637-647.
- Russell S., Norvig P., 2005, Artificial Intelligence A Modern Approach. Hungarian Translation Panem, HU, 635- 638, 726- 727
- Takács S. L., 2023, Detection of FM transmission outages and modulation outages with classification methods, MSc Dissertation, Széchenyi István University, Győr, Hungary.
- Yokkampon U., Chumkamon S., Mowshowitz A., 2020, Anomaly detection using variational autoencoder with spectrum analysis for time series data. In 2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 1-6). IEEE.
- Wang X. Y., Qian R. R., Huang C. X., Huang M., Yang J. J., 2022, An Anomaly Detector Using Filtering Stockwell Transform and Siamese Convolutional Neural Network in Radio Monitoring (AnoFSTSCNN). URSI Radio Science Letters, 4, 35.