Web Intrusion Classification System using Machine Learning Approaches

Mansi Bhatnagar [1], Gregor Rozinaj [1], Puneet Kumar Yadav [2]
[1] Slovak University of Technology FEI, Ilkovicova 3, 81219 Bratislava, Slovakia
[2] Ambalika Institute of Management and Technology Lucknow, India
mansi.bhatnagar@stuba.sk

2022 International Symposium ELMAR | 978-1-6654-7003-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ELMAR55880.2022.9899790

***Abstract*— Fast growth enhancement of web application has leaded to various security issues regarding intrusion not only in computer network framework but also web application themselves. Most of the techniques now in use in the context of Online Intrusion System (WIS) are unable to keep up with the complex and dynamic nature of cyber- attacks on web applications and issues linked to it. Attackers are using different techniques day by day to exploit the vulnerability of web applications. In order to identify intruder behaviors in order to predict future attacks on any web site, authors created an intrusion detection system (IDS) employing various machine learning and deep learning algorithms. The web application's log files will be used to monitor any forthcoming requests, which will then be forwarded to our eight variously trained detection models. Anomaly detection model will use to identify unseen attacks also known as zero-day attack. The process of feature extraction and data processing is done for each different by itself to detect specific attack. After, experiments the evaluations shows that the average accuracy of all the models is 99.3% on benchmark data set–CSIC 2010 HTTP and ECML/PKDD 2007.**

***Keywords*— Intrusion Detection System (IDS); Web Attack; Machine Learning (ML); Anomalous Detection**

I. INTRODUCTION

In today's world, internet plays an important role in modern life. There is only one click distance between user and rest of world. Like other software, web applications are also given minimum attention on security perspective as a consequence of which there is high number of sites which are influenced by hackers.

From past few years' intrusion attempts has been surprisingly raised to value of about 75% of cyber- attacks [1] and specially targeted as web applications. It was found in a conducted survey that malicious attacks which has expenditure about hundred times more than malware and fifty times more than trojans, viruses and worms per year [2]. ID System has an important task to address security issues of web servers by giving time to time report of malicious activity and prevent the effect of attack.

Many of the intrusion detection system focuses on single event stream like network traffic, which has the direction towards the server, or the access log file produced by server application. Absence of state-full detection model and inability to analyze different event streams in integrated ways restricts the effect of current available intrusion detection system.

Here proposed IDS uses the machine learning classifiers and deep learning technique [3] based on system which can be established by modular approaches and extends an application runtime and dealing with specific domain of application.

In past years, various number of anomaly computer and intrusion detection system have been developed on many machine learning and deep learning techniques for improvement of detection rates. There are various techniques which can be utilize for system malfunctioning because of that, various attacks can be performed in web attacks. as per the ETL report of 2018, majority of web attack was SQL, which was dominating in web applications and was expected around 51% [4] of figure value. Local file inclusion and cross site scripting (XSS) are deliberately holding second and third position respectively. The main quality of an intrusion detection system is to response instantly as soon as the detection of a malicious threat takes place at very less cost and should be accurate enough to raise an alarm when any threat is detected. There can be variations of alarming can be possible from logging the malicious threats and alerting the administrator to raise certain precautional steps [5].

II. RELATED WORK

Sharma et al. [6], have proposed a system consisting of deep neural network for detection of anomalous files as it compares the performance of itself with the classical neural network at the same time. The comparison is done based on method by which the set of features learned at the detection of anomalies takes place.

In this work, they have developed a simple, more reliable, accurate and fast mechanisms for visualization of features at the upper level. Here authors have tried to develop a system which will be able to detect abnormality and leads to reduction in the quantity of data to be processed manually by focusing only on the specified part of the data. For the same purpose good set of feature selection is required which further leads to good abnormality detection. In this method deep learning is used for learning features directly from raw data [7], [8]. Experimentally the authors explored and made comparisons with two types of structures, first one is classical neural network and other is deep belief network. On an average total of 42.4% to 66.5% of improvements achieved on the two data sets in experimentation [9]. One of the algorithms, artificial neural network is a statistical learning algorithm which are generated and developed by Biological neural network. Regarding neural network, first work was published by Professor McCulloch and Pitts in year

978-1-6654-7003-2/22/$31.00 ©2022 IEEE

1943which was entitled as "A Logical Calculus of the Ideas Immanent in Nervous Activity" which describes the concept of ANN from human neural network in history for very first time [9]. As a consequence of which, all the values which are multiplied, are further added together and sum inputs as an activation function and regression analysis [1]. ANN has recently been used in recognition, prediction, and reasoning. According to Althubiti et al., in 2017 has proposed a method, based on machine learning to detect abnormal traffic by conducting an experiment on ECML/PKDD2007 and CSIC 2010 HTTP dataset. There were 36000 normal and 25065 anomalous requests were considered in the experiment. The dataset was containing various attacks like XSS, SQLi, etc. in both the dataset as anomalous request. By using Weka analysis, it was founded that 5 best features out of 9 features were relevant. The best 5 features were request length, number of arguments, argument length, number of special characters and path length. Whole dataset was divided into 40% testing and 60% training part. Various techniques of machine learning [11] like Random Forest, Logistic Regression were used to classify normal and anomalous request, AdaBoost classifier and Naive Bayes. By using above classifier, total accuracy of 99.94% except Stochastic Gradient which is about 99.88%. Although they have achieved good accuracy, they have not classified further to the attacks [5], [10].

III. PROPOSED FRAMEWORK

In this section we have given detailed an Figure1: Shows step by step process of a Simple XSS analysis of about the architecture of our proposed Web Intrusion Detection System, types of attack we have used, Data processing, Feature engineering, and different intelligent modeling techniques we have used to get better monitoring of the malicious HTTP request in real time.

A. SQL Injection (SQLi)

It is a category of an attack which is injected and further controls the web application's database server and executes the malicious SQL statements. An unauthorized user injects vulnerabilities to go through the security measures of application. They can retrieve the data of whole SQL database by going around authentication and can further delete records, add, and modify the database [5]. These attacks are amongst the oldest, very dangerous and for the most part prevalent web application vulnerabilities [8].

TABLE I. MALICIOUS PAYLOADS AND ITS TYPE

First 5 lines of SQL

	payload	is_malicious	injection_type
0	\n	1	SQL
1	a' or 1=1-- \n	1	SQL
2	"a"" or 1=1--"\n	1	SQL
3	or a = a\n	1	SQL
4	a' or 'a' = 'a\n	1	SQL

B. Cross-Site Scripting (XSS)

Cross-site scripting is one of the most popular vulnerabilities out of OWASP few vulnerabilities [8]. It is also known as XSS. Using XSS attacker can bypass the SOP (same origin policy) concept in a vulnerable web application. SOP can be state as the very important principle of security in a web browser. XSS is type of Injection attack. The attacker's target is to execute the malicious code in a web application. An XSS attack takes place by passing two phases.

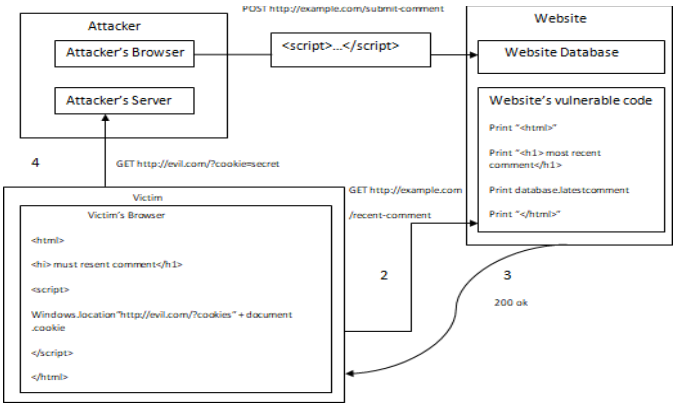


Figure 1. Step by step process of a Simple XSS

C. Shell Scripting

A shell is an interface through which an user works on programs, commands, and scripts and it is accessed by a terminal which runs it. Shell scripting is the process of writing a series of commands for shell so that it can be executed. Shell scripting can combine the repetitive and lengthy sequences of commands into a just a simple and single script. This can be stored and executed anytime whenever required. This concept reduces the effort required by the end user.

IV. DATA SET COLLECTION

The data set is the combination of two different sets. The first one named as the ECML/PKDD 2007 and the second is CSIC 2010.

A. ECML/PKDD 2007

The dataset contains different subsets generated from the ECML/PKDD 2007 Challenge. They provided the real time traffic of http request in xml format stored in .txt file. The total number of samples contain by the dataset are 50,116, from which 35,006 request are classified as valid request and 15110 are different types of attack request. All the samples in the dataset includes a full HTTP request which are divided into six categories [13]. Method, Protocol, URL, Query, Header, Body.

B. CSIC 2010

The HTTP dataset contains two types of HTTP request labelled as normal and anomalous, where 36,000 normal requests and more than 25,000 requests are anomalous. Each HTTP request is defined by the following features: method, URL, protocol, user Agent, pragma, cache Control, accept, accept Encoding, accept Charset, accept Language, host,

connection, content Length, content Type, cookie, and payload[14]. This dataset includes attacks such as SQL Injection, Buffer overflow, Information gathering, File disclosure, CRLF Injection, XSS, Server side include, Parameter tampering HTTP request is defined by the following features: URL, user Agent, cache Control, accept, accept Encoding, pragma, accept Charset, accept Language, protocol, methods, host, connection, content Length, content Type, cookie and payload. This dataset includes various attacks like: SQL Injection, Buffer Overflow, Information gathering and File Disclosure.

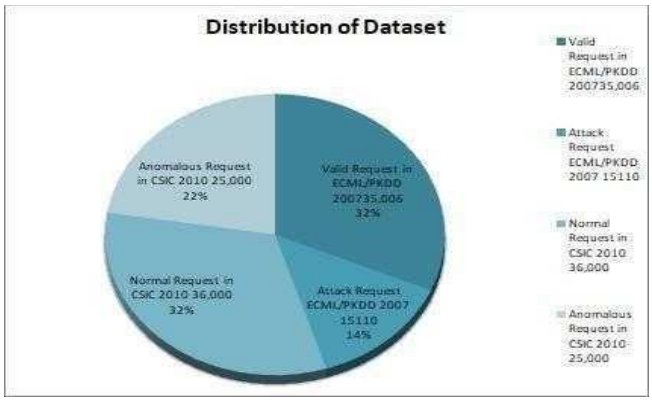


Figure 2. Data Distribution

V. Experimental Setup and Outcomes

This section represents the experimental setup in this work and the results of used different machine learning techniques.

For training purpose, here we are using four classifiers where the model is predicting average score value from each classifier, which is about greater than 90%. First, automate hyper parameter tuning and out of sample testing using train model, then uses the train model function to train, optimize and retrieve out of sample testing results from a range of classifiers. Classifiers tested using our custom feature space and bag of words feature space. In this work we take Classifiers are as follows:

A. MLP Classifier-

MLP stands for Multi-layer Perceptron classifier, which is basically used for connecting Neural Network. MLP Classifier is having characteristic to rely on an underlying Neural Network which perform the purpose of classification which is not possible by any other classifier.

B. Random Forest-

It is a Meta estimator which fits the number of decision tree classifier on variety of a lot of sub-samples of dataset and uses averaging which leads to the improvement of the predicting accuracy and leads to control over fitting. A powerful tool is cross-validation. We are able to use our data more effectively and learn considerably more about the effectiveness of our algorithm thanks to it[12]. Using sophisticated machine learning models. The performance of machine learning models when generating predictions on data not utilized during training is estimated using the k- fold cross- validation approach. This process can be used to compare and choose a model for the dataset as well as to optimize a model's hyper parameters on a dataset.

C. Multi nominal Classifier-

It is a basic term referring to a conditional independence of each and every feature and model with a specific instance which uses multinomial distribution for each individual of the features. Divide sample S into n- terms. Do the following for every k-th class Ck k=1..m Calculate the n- feature vector of the form Wki, where wki is the frequency with which the corresponding i-th term appeared in the Ck. As the whole probability that a document happened in the documents from class Ck, evaluate the prior p(Ck). Calculate the posterior Pr(Ck | W) by adding the prior p(Ck) to the sum of the wi, given Ck, probabilities p(wi | Ck) for each term. It is possible to improve the performance of multinomial models by semi- supervised learning.

D. Decision Tree-

Decision Tree is a supervised non-parametric machine learning technique which is basically used for both of regression and classification purposes.

The main objective to be achieved by using this classifier is to generate a model which makes prophecy about the values of variables which are targeted by learning basic decision rule take from the data features The algorithm seeks to partition the dataset into the smallest subset at each split. The objective of this approach is to reduce the loss function as much as feasible, much like any other Machine Learning algorithm.

E. Performance Measurement

The performance of the system is measured using Receiver Operating Characteristic (ROC) Curve. This curve represents the attack detection rate (true positive) against the false alarm rate (False Positive rate). The number of requests used in the training phase will be the parameter of the ROC curve. In this case ROC curve plot for the top 3 classifiers and bottom 3 classifiers, sorted by F1

Precision = true positive /true positive + false positive= true positive/ total predicted positive

Recall = true positive /true positive + false negative = true positive/ total actual positive.

TABLE II. Measurement of Classifiers

CLASSIFIERS	F1 Score	ACCU-RACY	SENSTI-VITY	SPECIFI-CITY	(AUC)
Random Forest classifier	99.92	99.85	98.83	99.96	99.93
Multinomial NB classifier	99.78	99.60	97.66	99.79	99.91
MLP classifier	99.60	99.27	95.47	99.65	99.76
Decision Tree classifier	99.39	98.89	89.57	99.80	99.09

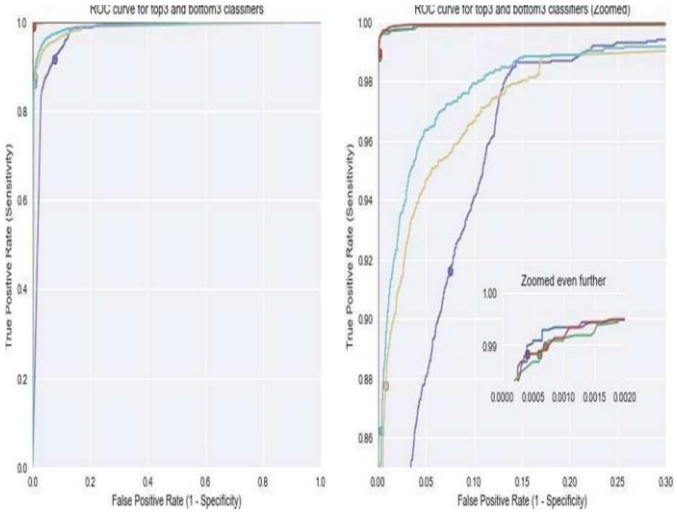


Figure 3. ROC curve of classifiers

VI. CONCLUSIONS

There are significant threats of web application attacks to many of the organization and attacks like various injections and XSS are on the rise.

Therefore, for detecting such malicious attacks, a live intrusion detection system using various machine learning models have been generated. In this work, we have founded the best machine learning classifiers for detecting the malicious attacks. these models are further used in an intrusion detection system.

ACKNOWLEDGMENT

This publication was created thanks to support of the projects: International Center of Excellence for Research on Intelligent and Secure Information and Communication Technologies and Systems - II. stage, ITMS code: 313021W404, co-financed by the European Regional Development Fund and the 2020-1-CZ01- KA226-VET- 094346 DiT4LL ERASMUS+ Innovation Project.

REFERENCES

[1] S. Jo, and B. H. Ahn “A comparative study on the performance of svm and an artificial neural network in intrusion detection,” Journal of the Korea Academia-Industrial Cooperation Society, 17(2), pp. 703-711, (2016).

[2] G. Yan, N. Brown, and D. Kong, “Axploring discriminatory features for automated malware classification,” pp. 41-61, (2013)

[3] A. Agresti, and B. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions", The American Statistician, pp. 119-126, (1998).

[4] Enisa threat landscape report 2018, https://www.enisa.europa.eu/publications/enisathreat-landscape-report, (2018).

[5] H. Bhagwani, "Log based dynamic intrusion detection of web applications," June 2019.

[6] M. Kumar Sharma, D. Sheet, and P. K. Biswas. "Abnormality detecting deep belief network," Proceedings of the International Conference on Advances in Information Communication Technology & Computing. Acm, (2016).

[7] L.P. Jyothsana, E. Anushya, and S. Shantha Kumara, "Ananomaly- Based Approach for intrusion detection in web traffic," International Journal of Advanced Research in Basic Engineering Sciences and Technology vol.3, special issue 25.

[8] Top 10, 2013, Owasp foundation, https://www.owasp.org/index.php/top_10_2013-top_10,

[9] P.V. Bro, "A system for detecting network intruders in real- time," Proceedings of the 7th Usenix security symposium, San Antonio, 1998.

[10] W. Mcculloch and W. Pitts, "A logical calculous of the ideas immanent in nervous activitiy, the bulltin of mathematical biophysics," 1943.

[11] S. Chung, and K. Kim, "A heuristic approach to enhance the performance of intrusion detection system using machine learning algorithm," Proceedings of Korea Institutes of Information Security and Cryptology Conference, 2015.

[12] L. Breiman, "Random forests. Machine learning," Statistics Department, University of California, Berkeley,(2001.

[13] CSIC (2010) HTTP and ECML/PKDD (2007).

[14] CSIC2010, https://www.isi.csic.es/dataset. https://github.com/msudol/Web-Application-Attack- Datasets (2010) (2007)

Published by: Croatian Society Electronics in Marine - ELMAR

Edited by: Mario Muštra [1], Branka Zovko-Cihlar [2], Josip Vuković [2]
[1] University of Zagreb, Faculty of Transport and Traffic Sciences Vukelićeva 4, 10000 Zagreb, Croatia
[2] University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3 / XII, 10000 Zagreb, Croatia

Front Cover: Painting by artist Mrs Ljerka Njerš

Printed by: Ispis Ltd., Zagreb

Print ISBN: 978-1-6654-7002-5, CFP22825-PRT

XPLORE ISBN: 978-1-6654-7003-2, CFP22825-ART

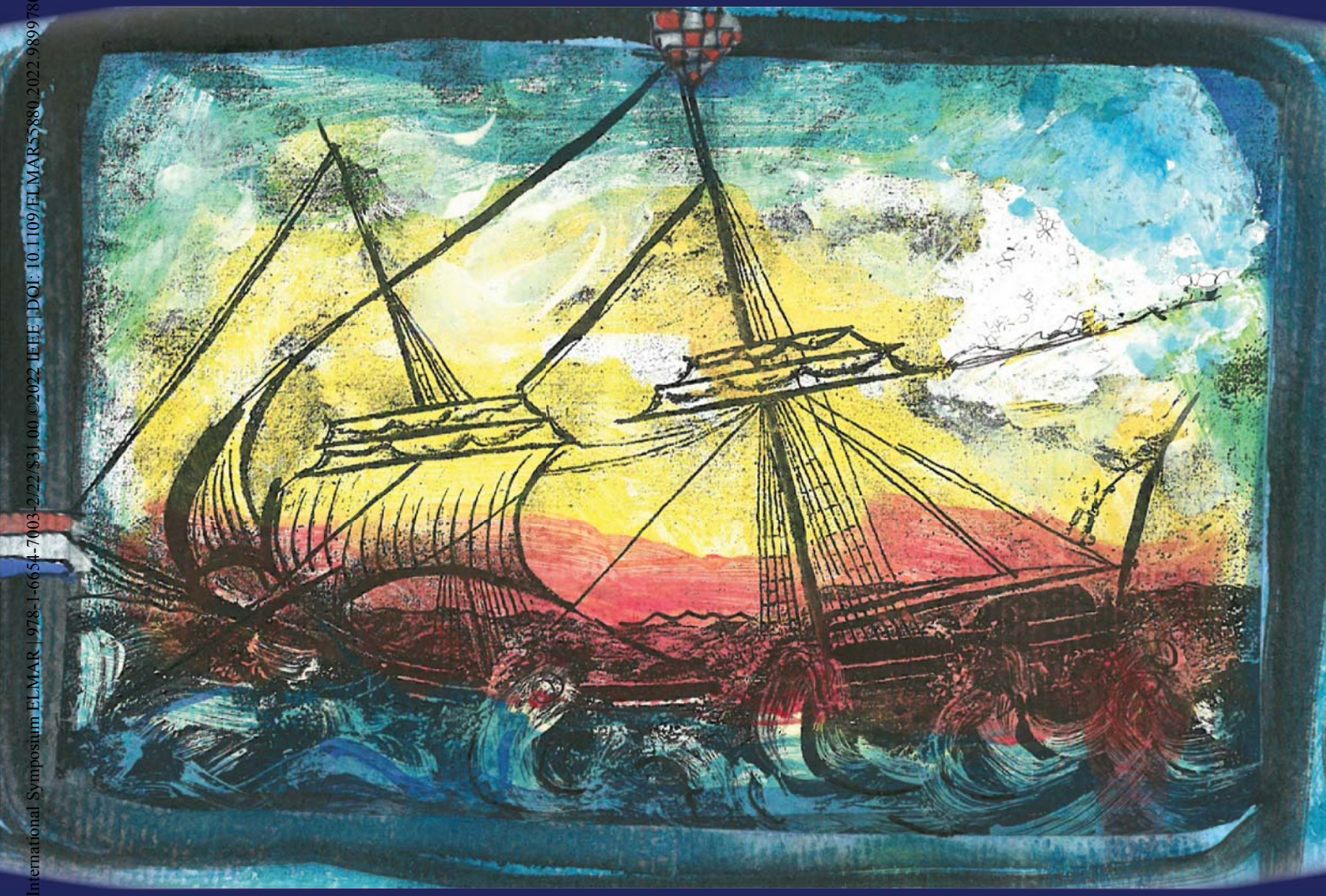
Printed in Croatia.

2022 International Symposium ELMAR | 978-1-6654-7003-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ELMAR55880.2022.9899795

ELMAR

PROCEEDINGS ELMAR-2022

2022 International Symposium ELMAR | 978-1-6654-7003-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ELMAR55880.2022.9899780



12-14 September 2022, Zadar, Croatia

Edited by

Mario Muštra
Branka Zovko-Cihlar
Josip Vuković

region IEEE 8

IEEE CROATIA SECTION

UNIVERSITY OF ZAGREB
Faculty of Electrical Engineering and Computing

PROCEEDINGS OF ELMAR-2022

64th International Symposium ELMAR-2022

12-14 September 2022, Zadar, Croatia

EDITED BY

Mario Muštra
Branka Zovko-Cihlar
Josip Vuković

University of Zagreb
Croatia

Published by:
Croatian Society Electronics in Marine - ELMAR, Zadar, Croatia

ISBN: 978-1-6654-7002-5

IEEE Catalog Number: CFP22825-PRT

ELMAR-2022 SYMPOSIUM
INTERNATIONAL REVIEW COMMITTEE

Goran Bakalar, Croatia
Alen Begović, Bosnia and Herzegovina
Marko Bosiljevac, Croatia
Jelena Božek, Croatia
Aura Conci, Brasil
Emil Dumić, Croatia
Hrvoje Gold, Croatia
Sonja Grgić, Croatia
Edouard Ivanjko, Croatia
Niko Jelušić, Croatia
Juraj Kačur, Slovakia
Hrvoje Leventić, Croatia
Panos Liatsis, United Arab Emirates
Časlav Livada, Croatia
Lidija Mandić, Croatia
Marko Matulin, Croatia
Marta Mrak, United Kingdom
Štefica Mrvelj, Croatia
Mario Muštra, Croatia
Miloš Oravec, Slovakia
Jarmila Pavlovičova, Slovakia
Tomislav Petković, Croatia
Jan Pidanič, Slovakia
Peter Planinšič, Slovenia
Pavol Podhradský, Slovakia
Tatiana Privalova, Russia
Snježana Rimac-Drlje, Croatia
Gregor Rozinaj, Slovakia
Markus Rupp, Austria
Renata Rybárová, Slovakia
Gerald Schaefer, United Kingdom
Chakib Taybi, Morocco
Mario Vranješ, Croatia
Josip Vuković, Croatia
Krzysztof Wajda, Poland
Dominik Žanić, Croatia
Ivana Žeger, Croatia

ELMAR-2022 SYMPOSIUM
ORGANISING COMMITTEE

Mislav Grgić, Croatia
Mario Muštra, Croatia
Jelena Božek, Croatia
Josip Vuković, Croatia
Dominik Žanić, Croatia
Ivana Žeger, Croatia

2022 International Symposium ELMAR | 978-1-6654-7003-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ELMAR55880.2022.9899818