Test-Training Leakage in Evaluation of Machine Learning Algorithms for Condition-Based Maintenance

Omri Matania[1, *], Roee Cohen[1, *], Eric Bechhoefer[2], and Jacob Bortman[1]

[1]*Ben-Gurion University of the Negev, Beer Sheva, 8410501, Israel*
omrimatania@gmail.com / omrimat@post.bgu.ac.il
coroe@post.bgu.ac.il
jacbort@bgu.ac.il

[2]*GPMS International Inc., 93 Pilgrim Place, Waterbury, Vermont, 05676, USA*
eric@gpms-vt.com

ABSTRACT

Many articles have been published utilizing machine learning algorithms for condition-based maintenance through the analysis of vibration signals. One extensively researched topic is the classification of fault types in rolling bearings. There is a fairly widespread problem in the evaluation of these learning algorithms, where the separation of examples between the test and training sets is incorrect, leading to an optimistic conclusion about the algorithm's performance even when it is not the case. In this article, we will review this issue and explain how the data should be properly divided between the test and training sets to avoid this occurrence.

1. INTRODUCTION

Condition-based maintenance of rotating machinery, through the analysis of vibration signals, can significantly reduce maintenance costs and also help prevent catastrophic accidents (Matania et al., 2024; Randall, 2021). Over the years, a wide variety of machine learning algorithms have been developed to enhance traditional signal processing methods for vibration analysis (Lei, 2017).

One of the topics extensively explored in the field is the classification of fault types in bearings using machine learning algorithms (Lei et al., 2020). In this task, the algorithm is required to predict the fault type from four possibilities for a given input record: healthy condition (i.e., no fault), fault in the inner race, fault in the outer race, or fault in the rolling element. To achieve this, the algorithm is provided with examples of input records with various fault types during the training phase, and it predicts the fault type for new input records during the testing phase.

A wide variety of machine learning algorithms have been applied to this task. The first type comprises classical machine learning algorithms, where a domain expert extracts correlated features related to the fault, and the learning algorithm learns the relationship between these features and the fault type (Shalev-Shwartz & Ben-David, 2014c). The second type, developed later during the third wave of deep learning, utilizes deep neural networks to address this problem. Unlike classical algorithms, the neural network autonomously learns features that connect the vibration signals to the fault type, essentially eliminating the need for a domain expert (Goodfellow et al., 2016). In both types of learning algorithms, many studies incorrectly split the training set and the test set, leading to significant test-training leakage (Kapoor & Narayanan, 2023) that results in inaccurate, overly optimistic performance evaluations of the examined algorithms (Hendriks et al., 2022).

The first type of test-training leakage, which is also the more problematic of the two, involves splitting the same input record into different segments and randomly distributing them between the test set and the training set. Figure 1 illustrates this type of splitting. This splitting is fundamentally flawed, as many features in the same input may be unrelated to the fault type, causing the learning algorithm to inadvertently learn them. Often, when disassembling the test rig to change the tested bearing, there is a change in the vibration signature unrelated to the fault type at all. For example, researchers from the SKF group found evidence of this phenomenon in a study on fault severity assessment (Liefstingh et al., 2021). They demonstrated that the learning algorithm learned features from the vibration signature related to the transfer function of the test rig instead of information related to the fault. In such

[*]Equally contributed.

a case, when segments from the same record are divided between the training and test sets, the learning algorithm may seem to predict the fault type well, although it actually relates the segments from the same records based on the characteristics of the transfer function.

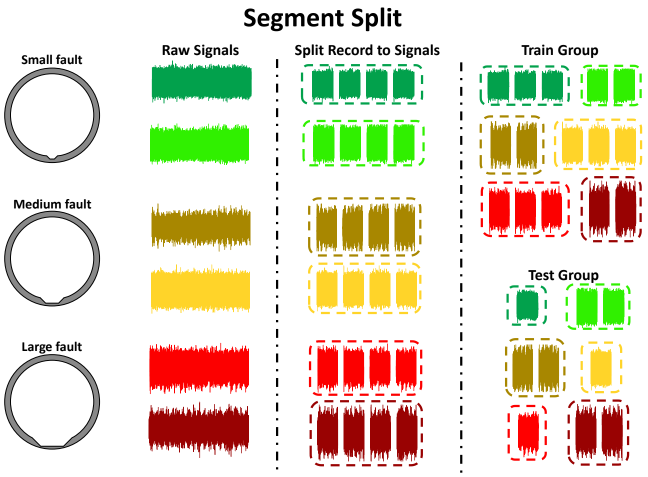

Figure 1. Illustration of random split of segments between the training and test sets.

The second type of test-training leakage involves the random separation of different records of the same fault type with the same fault shape precisely between the test and training sets. Figure 2 illustrates this type of improper separation. Each fault type can exhibit a wide variety of shapes. For example, a fault in the outer race can manifest in numerous different shapes and sizes, potentially even an infinite number. In practice, the likelihood that the exact shape of the fault in a real-world scenario matches one of the faults the algorithm learned from in the training set is very low. Many datasets record each fault multiple times. Randomly distributing these records between the test and training sets is incorrect and does not represent reality. In such a scenario, the algorithm may learn features related to the shape of the fault rather than its type, leading to overly optimistic evaluated performances. Furthermore, in some cases, the records of the same fault shape do not include the assembly of the test rig. Consequently, the algorithm may learn features from the vibration signature related to the transfer function of the test rig rather than information related to the fault, similar to the previous case of random segment split.

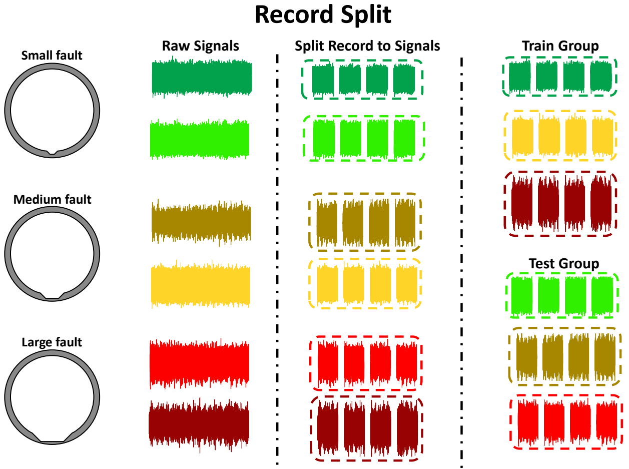

Figure 2. Illustration of random split of records between the training and test sets.

Figure 3 illustrates the correct splitting for evaluation learning algorithms: all records of each fault shape are either sent to the test set or to the training set. Following this separation, each record can be further divided into smaller segments if necessary. In this approach, to achieve an accurate estimation of performance, it is recommended to use K-fold testing. For example, in this study, performance evaluation in the test is implemented using the leave-one-out procedure, which is an extreme form of K-fold testing.

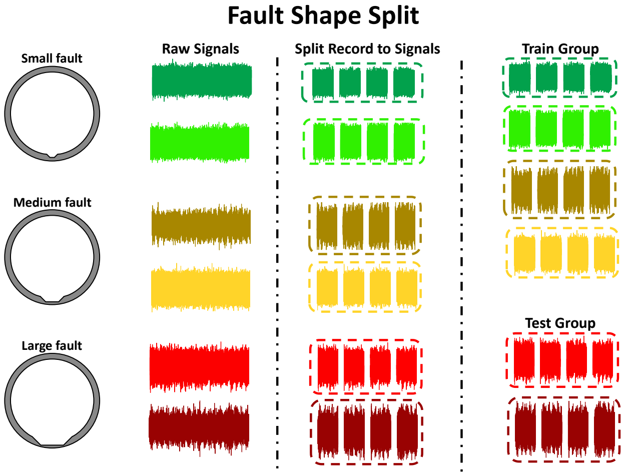

Figure 3. Illustration of split by fault shape between the training and test sets.

Section 2 will discuss the datasets analyzed in the article, and Section 3 will cover the learning algorithms. Section 4 will demonstrate that indeed, both segment split and record split lead to optimistic results compared to the correct way of fault shape split. Section 5 will summarize the article and present the conclusions.

2. Tested Datasets

Two datasets that are frequently used for evaluating machine learning algorithms for fault classification in rotating machinery are discussed in the article. The first dataset, Case Western Reserve University dataset (CWRU), is accessible via the link (*Case Western Reserve University Bearing Data Center Website*, n.d.) and is extensively described in the work by Smith and Randall (Smith & Randall, 2015). It is important to note that this dataset has several issues, as explained by Smith and Randall, yet for unknown reasons it is still widely utilized. The CWRU test rig is illustrated in Figure 4. The CWRU dataset comprises a total of 416 distinct records, but in practice, only 12 truly different faults exist.

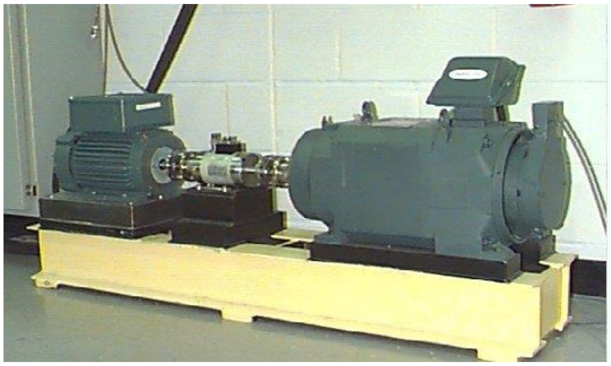

Figure 4. CWRU dataset test rig. Reproduced from (*Case Western Reserve University Bearing Data Center Website*, n.d.).

The Paderborn University (PU) dataset also serves for the evaluation of various learning algorithms and is extensively described in the study of Lessmeier et al. (Lessmeier et al., 2016). Our review of this dataset led to the conclusion that it also has several issues, such as unclear sources of interferences in the spectrum. In total, the PU dataset contains 2493 recordings, with 26 truly distinct faults in practice. Figure 5 depicts the experimental setup.

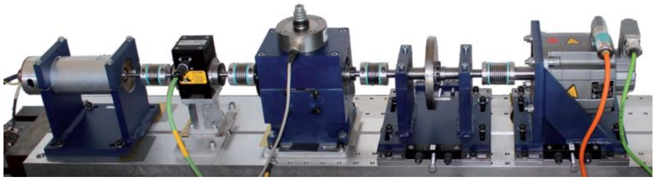

Figure 5. PU dataset test rig. Reproduced from (Lessmeier et al., 2016).

3. Tested algorithms

In the current section, two learning algorithms will be described, which are used to demonstrate the effect of test-training leakage. The first is K-nearest neighbors (KNN) (Shalev-Shwartz & Ben-David, 2014a) and the second is Random Forest (Shalev-Shwartz & Ben-David, 2014b). All tested algorithms used the following features: mean, variance, kurtosis and absolute mean.

KNN operates by determining the class of a data point based on the majority class among its k-nearest neighbors within the feature space. The algorithm computes the distance between the given data point and its neighbors. The parameter K, denoting the number of neighbors taken into account, is a crucial factor that can significantly influence the model's performance. Small K values may result in overfitting, while large K values may lead to inadequate fitting of the training data. In the current study, K was set to 1 to prevent additional issues with training-validation splitting. Figure 6 provides a visualization of the KNN process for classification.

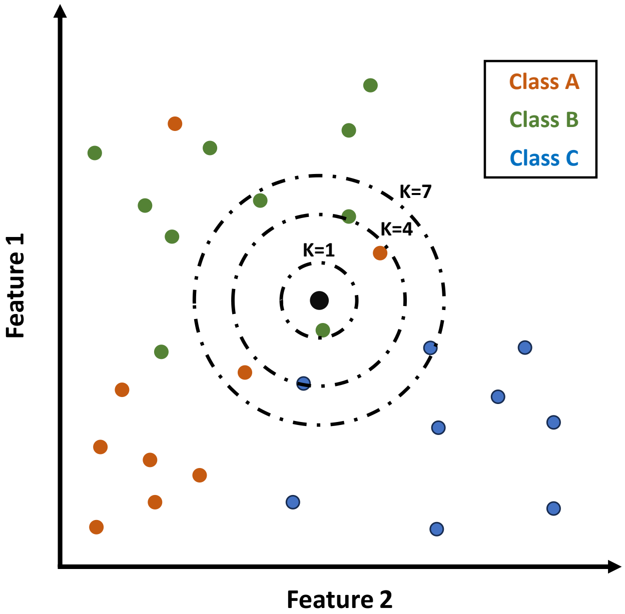

Figure 6. Illustration of KNN.

Random Forest stands out as a robust ensemble learning algorithm widely applied in machine learning for classification tasks. It generates numerous decision trees during training and outputs the mode of the classes. The key innovation of Random Forest lies in its incorporation of randomness—each tree is trained on a random subset of the data, and during each split, a random subset of features is taken into consideration. This randomness aids in mitigating overfitting and enhancing the model's generalization performance. Furthermore, for classification, the predictions from multiple trees are consolidated through majority voting, resulting in a resilient and accurate final prediction. In the current case, the number of trees was set to 300. This is a standard number of trees intended to prevent overfitting. Once again, this parameter was not set based on the validation set to avoid additional issues with training-validation splitting. Figure 7 provides a visualization of the random forest process for classification.

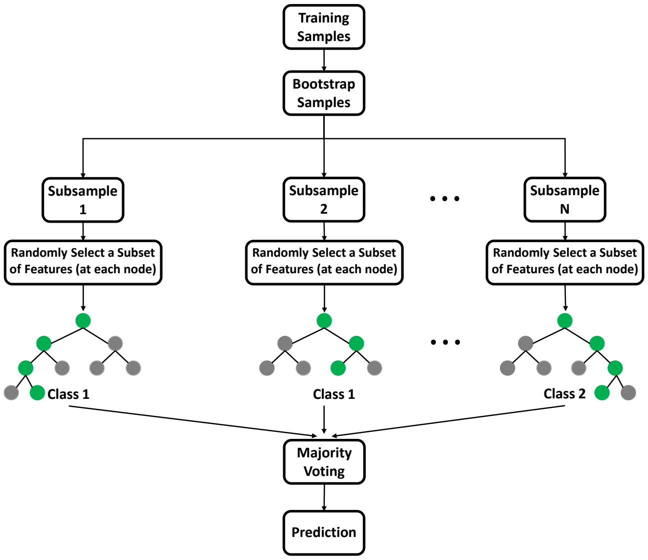
Training Samples
Bootstrap Samples
Subsample 1
Subsample 2
Subsample N
Randomly Select a Subset of Features (at each node)
Randomly Select a Subset of Features (at each node)
Randomly Select a Subset of Features (at each node)
Class 1
Class 1
Class 2
Majority Voting
Prediction

Figure 7. Illustration of random forest.

4. Results

The results of KNN and random forest on the two tested datasets, CWRU and PU, are depicted in Figure 8 for the three types of splitting: segment split, record split, and fault shape split. The results of the segment and record splits were determined using a 10-fold cross-validation technique to calculate the average accuracy. The fault shape split results were obtained through a leave-one-out procedure. Furthermore, to compare the performance with a degenerated algorithm, two lines were added to the figure representing predictions of the test examples in the fault shape splitting for CWRU and PU, based on the most prevalent label in the training set. This degenerated algorithm disregards the features and, for any new unseen examples, returns the mode of the classes from the training set.

As can be seen from the figure, for the CWRU dataset, when changing from segment split to record split, the accuracy significantly decreases. For both datasets, when the correct splitting method is utilized, namely the fault shape split, the results are significantly worse. In the case of CWRU, they are even lower than the accuracy of the degenerated algorithm, which predicts the training mode constantly.

These results demonstrate that incorrect random splitting leads to overly optimistic conclusions. For the CWRU dataset, based on segment split, it seems that the very straightforward approach of using simple signal features and classic machine learning algorithms like KNN and random forest enable achieving good accuracy, close to 90%. However, when the record split is applied, the results are much less optimistic, and when the correct method is applied, the results are worser than constantly predicting the training mode, indicating that both algorithms probably learn nothing related to the fault type. For PU datasets, even when record split is utilized, the results are still optimistic, and only when the correct splitting of fault shape is utilized can we again conclude that the algorithm did not learn too much information related to the fault type.

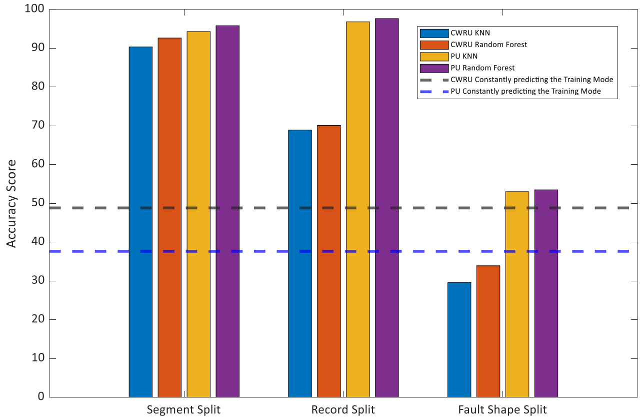

Figure 8. Accuracy score for bearing fault type classification on CWRU and PU datasets by KNN and random forest for different split approaches.

5. Conclusion

Many machine learning algorithms have been suggested for vibration analysis of rotating machinery for condition-based maintenance. As demonstrated in this paper, improper splitting of data between the training and test sets may lead to test-training leakage and, consequently, to an overly optimistic evaluation of the machine learning algorithm performances.

In the current study, this problem was tested on the prevalent task of fault type classification in rolling bearings. It was shown that when improper segment splitting is utilized, overly optimistic conclusions can be drawn regarding a simple approach that combines straightforward signal features with basic machine learning algorithms, as they achieve accuracy close to 90%. However, when the right splitting is utilized, reflecting the real scenario in which records of the exact same fault shape should not be present in both the training and test sets, the results are very poor and, in some cases, worser than constantly predicting the training mode, indicating that the algorithms have not learned anything.

Three further comments regarding machine learning studies in the vibration analysis field are worth discussing. First, most of the currently available datasets, such as CWRU and PU, contain many contaminated records. The research community would benefit greatly from newer datasets without contaminated records, which would also encompass a broader range of fault shapes. Second, it is not clear why so many papers attempt to solve the problem of fault type classification in bearings, as classic approaches in signal processing are adept at solving it (Randall & Antoni, 2011). We recommend that future papers focus on addressing fault severity and estimating remaining useful life tasks (Matania et al., 2023), or alternatively, focus on fault classification of components that currently lack well-established classic approaches. Another option is to examine cases where the signal-to-noise ratio is so low that signal processing algorithms are unable to classify the fault type. The last comment worth noting is that a maintainer or operations manager doesn't really care if a bearing has a ball, inner, or outer race fault – as they will probably replace the entire bearing regardless. The more important issue is fault detection, determining whether the bearing is healthy or not. Fault classification is more interesting if it helps to better estimate severity or remaining useful life.

Acknowledgement

Omri Matania is supported by the Adams Fellowships Program of the Israel Academy of Sciences and Humanities.

References

Case Western Reserve University Bearing Data Center Website. (n.d.). Retrieved November 23, 2022, from https://engineering.case.edu/bearingdatacenter/welcome

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org/

Hendriks, J., Dumond, P., & Knox, D. A. (2022). Towards better benchmarking using the CWRU bearing fault dataset. *Mechanical Systems and Signal Processing*, *169*, 108732. https://doi.org/10.1016/J.YMSSP.2021.108732

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/J.PATTER.2023.100804

Lei, Y. (2017). Intelligent fault diagnosis and remaining useful life prediction of rotating machinery. In *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery* (1st ed.). Butterworth-Heinemann. https://doi.org/10.1016/C2016-0-00367-4

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, *138*. https://doi.org/10.1016/j.ymssp.2019.106587

Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. *PHM Society European Conference*, *3*(1). https://doi.org/10.36001/PHME.2016.V3I1.1577

Liefstingh, M., Taal, C., Restrepo, S. E., & Azarfar, A. (2021). Interpretation of Deep Learning Models in Bearing Fault Diagnosis. *Annual Conference of the PHM Society*, *13*(1). https://doi.org/10.36001/PHMCONF.2021.V13I1.3047

Matania, O., Bachar, L., Bechhoefer, E., & Bortman, J. (2024). Signal Processing for the Condition-Based Maintenance of Rotating Machines via Vibration Analysis: A Tutorial. *Sensors 2024, Vol. 24, Page 454*, *24*(2), 454. https://doi.org/10.3390/S24020454

Matania, O., Bachar, L., Khemani, V., Das, D., Azarian, M. H., & Bortman, J. (2023). One-fault-shot learning for fault severity estimation of gears that addresses differences between simulation and experimental signals and transfer function effects. *Advanced Engineering Informatics*, *56*, 101945. https://doi.org/10.1016/J.AEI.2023.101945

Randall, R. B. (2021). *Vibration-based condition monitoring : industrial, automotive and aerospace applications* (2nd ed.). WILEY. https://www.wiley.com/en-sg/Vibration+based+Condition+Monitoring%3A+Industrial%2C+Automotive+and+Aerospace+Applications%2C+2nd+Edition-p-9781119477556

Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics—A tutorial. *Mechanical Systems and Signal Processing*, *25*(2), 485–520. https://doi.org/10.1016/J.YMSSP.2010.07.017

Shalev-Shwartz, S., & Ben-David, S. (2014a). Chapter 19 - Nearest Neighbor. In *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057135, pp. 258–267). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Shalev-Shwartz, S., & Ben-David, S. (2014b). Section 18.3 - Random Forests. *Understanding Machine Learning: From Theory to Algorithms*, *9781107057135*, 255–256. https://doi.org/10.1017/CBO9781107298019

Shalev-Shwartz, S., & Ben-David, S. (2014c). Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, *64–65*, 100–131. https://doi.org/10.1016/J.YMSSP.2015.04.021

Biographies

Omri Matania is currently a Ph.D. student in BGU-PHM LAB in the department of mechanical engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. Omri is a Talpiot graduate and served nine years in IDF in several roles including algorithm section leader. He completed with honors his bachelor's degree in mathematics and physics in the Hebrew University of Jerusalem and completed his master's degree with honors in mechanical engineering in Ben-Gurion University of the Negev.

Roee Cohen is currently a Ph.D. student in BGU-PHM LAB in the department of mechanical engineering in Ben-Gurion University of the Negev, under the supervision of Prof. Jacob Bortman. Roee completed with honors his bachelor's degree and master's degree in mechanical engineering in Ben-Gurion University of the Negev.

Eric Bechhoefer received his bachelor’s degree in biology from the University of Michigan, his master’s degree in operations research from the naval postgraduate school, and a Ph.D. in general engineering from Kennedy Western University. He is a former naval aviator who has worked extensively on condition-based maintenance, rotor track and balance, vibration analysis of rotating machinery, and fault detection in electronic systems. Dr. Bechhoefer is a fellow of the prognostics health management society, a fellow of the society for machinery fault prevention technology, and a senior member of the IEEE reliability society. Additionally, Dr. Bechhoefer is also a member of the SAE committee covering integrated vehicle health management, and a member of the MSG-3, rotorcraft maintenance programs industry group.

Jacob Bortman is currently a full Professor in the department of mechanical engineering and the head of the PHM Lab in Ben-Gurion University of the Negev. Retired from the Israeli air force as brigadier general after 30 years of service with the last position of the head of material directorate. Chairman and member of several boards: director of business development of Odysight Ltd, Chairman of the board of directors, Selfly Ltd., board member of Augmentum Ltd., board member of Harel finance holdings Ltd., Chairman of the board of directors, Ilumigyn Ltd. Editorial board member of: “Journal of Mechanical Science and Technology Advances (Springer, Quarterly issue)”. Head of the Israeli organization for PHM, IACMM - Israel Association for Comp. Methods in Mechanics, ISIG - Israel Structural Integrity Group, ESIS - European Structural Integrity Society. Received the Israel National Defense prize for leading with IAI strategic development program, Outstanding lecturer in BGU, The Israeli prime minister national prize for excellency and quality in the public service - First place in Israel. Over 80 refereed articles in scientific journals and in international conference.