

Statistical and Machine Learning Methods for Physical Signal Analysis

Sophia Yan

*Portola High School, Irvine, USA
starrynight37964@gmail.com*

Abstract. Physical signal analysis is crucial to scientific work and engineering applications since present-day sensing technologies very often generate huge amounts of measurement data which must be accurately interpreted. The extraction of useful information from complex physical signals has been greatly enhanced in recent decades by advances in statistical signal processing and machine learning. A thorough overview of statistical and machine learning techniques for physical signal analysis is given in this article. It presents basic concepts of physical signals and measurement data, and then explains the traditional statistical signal processing techniques like time-frequency analysis, feature extraction and signal preprocessing. The review then presents an overview of some representative machine learning and deep learning techniques such as Support Vector Machines (SVM), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks and Transformer models and their applications in industrial fault diagnosis, biomedical signal analysis and human activity recognition. With special focus on data quality, model interpretability, computational efficiency, and generalization ability, recent research trends, present constraints, and future research objectives are also covered. In summary, this paper highlights the symbiotic relationship between machine learning and statistical signal processing today in physical signal analysis. It provides readers with a comprehensive understanding of how data-driven learning, signal processing and physical measurement can be integrated to enable intelligent signal analysis and future scientific and engineering applications.

Keywords: physical signal analysis, statistical signal processing, machine learning, deep learning, measurement data

1. Introduction

Physical signals convey information which is critical to comprehending and monitoring natural and engineered systems. They are produced when physical properties vary during different physical processes. Sensors capture changes in the environment or system state through vibrational, acoustic, electrical, thermal, optical or biomedical signals and convert them into measurable information. These signals are important for determining the operating state of the system and are used in scientific research, industry, medical care and intelligent manufacturing. Recent advances in sensing and measurement technologies, along with the Internet of Things, have enabled modern systems to

use machine learning efficiently and accurately in signal-based applications. Thus, extracting useful and precise information from physical signals has become an important research topic [1, 2].

Traditional signal processing techniques have been used in physical signal analysis for decades. The Fourier transform and wavelet transform are classical techniques which are commonly used for denoising, frequency analysis and feature extraction. These methods, with solid mathematical foundations, can be used in several different areas where the properties of physical signals are well understood. However, these techniques rely on handcrafted features and require expert knowledge, and therefore lack flexibility when dealing with large and complex datasets [1].

The field of machine learning has been rapidly evolving, drastically altering the way physical signals are analyzed. Machine learning algorithms can automatically find meaningful patterns in the data and can be used for classification, regression, anomaly detection and prediction. Traditional machine learning techniques such as Support Vector Machines (SVM) and Random Forests can perform well with engineered signal features. In addition, deep learning architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models, have further enhanced the capacity to learn hierarchical representations directly from raw measurements of the signals. These approaches based on data processing have been applied successfully in fault diagnosis, biomedical signal interpretation and intelligent monitoring systems [3-5].

There is a tendency to address the problem using either statistical signal processing or machine learning; however, combining the two can be more effective than using them independently. Both statistical signal processing and machine learning provide the tools for signal representation, denoising, feature extraction and uncertainty analysis, and allow for data-driven models which can recognize complex nonlinear relations. Recent research shows that signal-processing expertise and recent advances in machine learning can help to combine domain knowledge and data-driven modeling. Thus, the hybrid techniques that integrate signal processing and artificial intelligence have been an important research direction in the field of physical signal analysis [1].

This review attempts to provide a summary of the recent developments of statistical and machine learning methods for physical signal analysis. First, some basic principles of physical signal and measurement data are explained. Then, a comparison of the statistical signal processing techniques and machine learning algorithms is made. Then, the research trends and applications in recent years are discussed. Finally, an overview of this field is given by analyzing the current limitations and future opportunities.

2. Basic concepts of physical signals and measurement data

Physical signals are measurable quantities that can describe the behavior or the state of physical systems with respect to time and/or space. Physical signals are used to gain information about natural phenomena and to monitor engineered systems. Physical signals can be generated by physical phenomena like mechanical vibration, acoustics, electromagnetic fields, biomedical activity and environmental variation and are sensed by a specific sensing device. With the advancement of sensing technology, it has been possible to obtain a large amount of measurement data with high precision; this is beneficial information for scientific research. The quality of the signal, however, is of critical importance as it will be the determinant of the usefulness of the data [1, 2].

Physical signals can generally be classified as continuous and discrete signals. Continuous signals are continuous at every point in time and typically come from real world physical processes. However, modern computers deal with digital information; before analysis, these continuous signals are converted to discrete signals via analog-to-digital sampling. Sampling frequency is important

because information can be lost and aliasing can occur if it is too low, whereas an unnecessarily high sampling frequency increases data volume and storage costs.

Physical signals have several basic properties that are used to describe them as well. Amplitude is the "strength" or size of a signal. Frequency is the number of cycles of an oscillation in a given unit of time. Phase describes the position of a waveform within its cycle, while phase difference describes the relative position of two waveforms. The difference between signal power and signal energy is that signal energy is the total energy accumulated over a period of time while signal power is the average energy per unit time. It is possible to analyze these characteristics in the time or frequency domain, depending on the intended research. Time-domain analysis emphasizes how the amplitude of a signal changes with time, and frequency-domain analysis emphasizes how the energy of a signal is distributed across frequencies. Time-frequency representations can provide a more detailed description of the behavior of many non-stationary signals, which have time-varying frequency content.

Typically, the data obtained from the measurements is not optimal. Signal quality can be influenced by several factors such as sensor limitations, environmental interference, electronic noise, errors during transmission and missing observations. This means that pre-processing is important prior to statistical analysis or machine learning. These common preprocessing tasks involve signal filtering, normalization, denoising, baseline correction, and feature extraction. While keeping the essential information needed for further analysis, these steps enhance the quality of the data. Measurement quality is frequently evaluated by the signal-to-noise ratio (SNR), and a higher SNR generally indicates a stronger useful signal relative to background noise [3].

With the use of machine learning becoming more popular, physical signals can now be analyzed by employing a larger set of representations. The data collected from measurements can be expressed using statistical descriptors, spectral features, wavelet coefficients, graph-based representations, or automatically learned feature embeddings via deep neural networks. Appropriate data representation is an important consideration that affects the performance of a model, since it determines how useful physical information is extracted from raw measurements. An understanding of the basic properties of physical signals and measurement data thus provides the theoretical foundation for both traditional statistical signal processing and recent data-driven learning approaches that are discussed in the sections below [4-6].

3. Statistical signal processing and machine learning methods

Physical signal analysis relies on mathematical methods in statistical signal processing to extract useful information from noisy measurements. Traditional statistical approaches have been adopted extensively, due to their high theoretical interpretability and computational efficiency. By contrast, machine learning methods focus on learning patterns straight from the data and allow models to solve complex classification and prediction problems without having to explicitly specify mathematical rules. Recent studies have shown that combining statistical methods and machine learning can lead to more reliable and accurate analytical frameworks without replacing traditional signal processing methods [1, 2].

One of the most popular statistical signal processing methods is the Fourier Transform (FT), which is used to convert data from the time domain to the frequency domain. With this transformation, researchers may be able to identify periodic behavior that may not be immediately seen in the original time-domain signal and may be able to discover dominant frequency components. Fourier-based techniques have been widely used in many fields including vibration analysis, speech processing and biomedical signal interpretation, since many physical processes

have characteristic frequency patterns. The Fourier Transform, however, does not have the ability to localize a signal in time, and thus is less informative for signals whose frequency characteristics change over time [5].

To overcome this limitation, the Wavelet Transform (WT) is widely used for time-frequency analysis in research. Unlike the Fourier transform, wavelet analysis can provide both the frequency and the time, and thus gives information about the localization of the signal properties at different resolutions. Therefore, wavelet-based techniques have proved to be very useful in the analysis of non-stationary signals such as mechanical fault vibration signals, electrocardiograms (ECG), and electroencephalograms (EEG). Besides frequency analysis, statistical signal processing also includes statistical feature extraction, filtering and autocorrelation analysis. Mean, variance, root mean square (RMS), skewness, kurtosis and entropy are common features extracted and used to summarize the properties of a signal prior to further analysis [3, 5].

These statistical characteristics are manually derived features and are usually inputs to models from conventional machine learning approaches. Of these techniques, Support Vector Machines (SVM) are well suited to high-dimensional classification problems and generally have good generalization performance. Random Forests (RF) are a collection of decision trees and are able to decrease the variance by ensemble averaging. They are usually noise tolerant, but are susceptible to overfitting, depending on the model parameterization and data quality. Other popular algorithms include Artificial Neural Networks (ANNs), Gradient Boosting techniques and k-Nearest Neighbors (k-NN), which all have different trade-offs between computational complexity, interpretability, and prediction effectiveness. The quality of the extracted signal features has a major impact on how effective these algorithms are [4, 7].

By reducing the need for manually designed features, recent developments in deep learning have substantially changed physical signal analysis. Convolutional Neural Networks (CNNs) have proven to be excellent for automatically learning the hierarchical spatial or temporal features of a raw signal in biomedical signal classification and fault diagnosis. Long Short-Term Memory (LSTM) networks are specifically designed for sequential data and can capture long-term temporal dependencies. Recently, transformer architectures have received increasing attention because self-attention can model long-range dependencies in signal sequences. Under some challenging signal-analysis tasks, deep learning models can outperform conventional methods when sufficient data are available, but they typically need larger training datasets and more computational resources [5, 8, 9].

The current research has been focused on integrating the two approaches, statistical signal processing and machine learning. In many applications where machine learning algorithms are used to perform classification or prediction, statistical preconditioning methods are often used for noise removal, normalization of measurements and for feature extraction. At the same time, signal processing techniques can contribute to deep learning models by improving data quality and model robustness while supporting automatic learning of feature representations. Consequently, hybrid statistical and data-driven methods have become an important paradigm for modern physical signal analysis with improved interpretability and predictive power across a range of scientific and engineering applications [1, 4].

4. Recent studies of statistical and data-driven physical signal analysis

Recent studies are compared in terms of the application domain, signal modality, dataset, model family, evaluation, and reported limitation. Although statistical methods play an important role in signal preprocessing and feature engineering, many recent works have shown that data-driven models can also directly learn complex properties of signals from raw measurements. This is

associated with rapid advances in deep neural network architectures, improvements in processing power, and the abundance of large-scale sensing data. For the sake of enhancing the accuracy, robustness and computational efficiency, hybrid systems combining statistical signal processing and artificial intelligence have been playing an increasingly important role in the current physical signal processing [1, 2].

Biomedical signal analysis is one of the key research fields where machine learning techniques have been actively used for physiological monitoring and diagnosis of diseases. Electromyography (EMG), photoplethysmography (PPG), electrocardiogram (ECG), and electroencephalogram (EEG) are commonly employed for the evaluation of neurological, muscular and cardiovascular disorders. Recent studies have shown that deep learning models, especially CNNs and LSTMs, can also be used to reduce reliance on manually designed features and automatically find useful features in raw biological signals. These models have been found to be highly effective in various applications such as heartbeat classification, arrhythmia detection, sleep stage recognition, and physiological condition monitoring. In addition, complicated relations among physiological signals can now be modeled with graph signal processing to enable new opportunities in improving the diagnostic accuracy and understanding of interactions between different biological systems [3, 4, 6].

Wearable sensing and human activity recognition (HAR) is another fast-developing research field. Accelerometers, gyroscopes, and other embedded sensing devices have made it possible to continuously monitor human mobility thanks to the increasing use of smartphones, smartwatches, and wearable sensors. Based on recent review studies, in the presence of high-quality handcrafted features, traditional machine learning techniques such as Random Forests and Support Vector Machines still achieve good results. However, deep learning methods are becoming popular since they learn hierarchical representations from raw sensor data automatically. LSTM networks are suitable to model long-term temporal dependencies in sequential activities such as walking and running, and daily living behaviors, while CNNs are suitable to model local motion patterns. Though they are more likely to require larger data sets and higher computational expenses, Transformer-based designs have recently demonstrated interesting performance by leveraging self-attention mechanisms to capture long-range correlation in multivariate sensor data [7-10].

Furthermore, statistical and machine learning techniques have recently been extended to industrial monitoring and defect diagnostics. In modern production systems, vibration, temperature, pressure and acoustic emission signals are continuously collected to be used in predictive maintenance and equipment health monitoring. Conventional statistical signal processing techniques are still useful for denoising and for using Fourier and Wavelet Transforms to obtain frequency-domain representations for feature extraction. These signals are then fed into deep learning models or machine learning classifiers for the detection of abnormal operating conditions and prediction of equipment failures. The combination of statistical preprocessing and deep learning methods can enhance the performance and robustness of fault diagnosis in noisy industrial environments. In addition, transfer learning approaches have been developed to deal with the situation where the operating conditions of the machinery change, so as to realize the adaptation of intelligent monitoring systems between operating conditions [5].

Despite these significant advances, there are still many issues that remain unsolved as identified by current research. Recent studies are compared with respect to application domain, signal modality, dataset, model family, evaluation metric and reported limitation. Signal reliability and generalization capacity can both be affected by environmental noise, sensor variability, and missing observations. To overcome these limitations, there is growing interest in current research on self-supervised learning, transfer learning, continual learning, graph neural networks, and multimodal

data fusion. The aim of these techniques is to reduce the need for manually labeled training data and improve the learning efficiency. Overall, the reviewed works show that in order to create more precise, comprehensible, and adaptable intelligent systems, future advancements in physical signal analysis will probably depend on combining statistical signal processing with cutting-edge machine learning approaches [4, 6, 10].

5. Limitations and future prospects

Even with all the remarkable progress in statistical signal processing and machine learning, intelligent physical signal analysis systems have yet to be fully exploited in practical application because of a number of important challenges. Recent review studies show that the usefulness of the model is not only dependent on the learning algorithm, but also on the quality of the data, the computational capacities, the interpretability of the model and the ability to transfer to other application contexts. As physical sensing systems become more complex and generate more and more types of data, this is becoming one of the most critical research areas [1, 2].

One of the big problems is the quality and availability of the training data. For modern deep learning algorithms to work well, a lot of labeled data is typically needed. However, in many real-world applications, such as industrial fault diagnostics, structural monitoring and healthcare, the acquisition of high-quality labeled datasets is often very expensive and time-consuming, and sometimes impossible. For biomedical applications, access to large data sets may be restricted by privacy considerations and/or annotation requirements, and failure of the system may be relatively uncommon, resulting in class imbalance. In addition, sensor data is often contaminated by environmental noise, missing data, varying sensor characteristics and transmission errors. These factors reduce the robustness of models when applied in real-world situations and can have severe consequences on the predictive performance of the model [3-5, 11].

Another well-known disadvantage of machine learning models is their limited ability to generalize. Many algorithms perform well on benchmark data yet poorly when applied to data collected in different climatic regions, sensor types or operating conditions. This is known as 'data heterogeneity' or 'domain shift'. Data distributions may differ significantly from those used to train a model, due to differences in the placement of the sensor, sampling rate or frequency, individual participants, equipment characteristics, measurement methods, etc. In consequence, models developed for a particular set of data are not necessarily easily transferable to other sets of data without additional modifications or retraining. Recent studies therefore emphasize transfer learning, domain adaptation, and continual learning strategies for improving model adaptability in dynamic scenarios [5, 10].

Although deep learning has demonstrated high predictive accuracy in many physical signal processing applications, its lack of interpretability is still a major concern. Conventional statistical signal processing methods result in features that are mathematically understandable and the physical meaning of many of these features is well known to engineers and researchers. In contrast, deep neural networks can be regarded as "black-box" models, and their decision-making process is difficult to explain. In high-risk applications like medical diagnosis, intelligent transportation, and industrial safety, where trustworthy justifications are frequently needed before automated decisions can be accepted, this restriction is especially crucial. To enhance transparency without hampering predictive performance, research on explainable artificial intelligence (XAI) methods has been active [3].

Another practical constraint is also computational efficiency. Typically, these advanced deep learning architectures, such as Transformer-based models and massive neural networks, require

extensive computational resources for training and inference. However, their high memory footprint, long training times, and greater power consumption might limit their deployment on embedded monitoring devices, wearables with limited hardware resources, or edge devices. Recent research is concerned with lightweight neural network architectures, model compression, model pruning, quantization, and knowledge distillation to reduce computational complexity without sacrificing prediction accuracy. As intelligent sensing devices continue to proliferate into situations with limited resources, these methods are anticipated to become more crucial [9, 12].

Some recent review articles show that hybrid intelligence schemes are a significant trend in physical signal analysis. It is expected that future machine learning systems will integrate domain expertise with data-driven learning algorithms rather than replace statistical signal processing. Methods of statistical preprocessing, feature engineering, and signal manipulation can be used to obtain physically meaningful data, which can improve model interpretability and the learning efficiency. At the same time, deep learning models can automatically identify complex nonlinear interactions that are difficult to identify using conventional analytical methods alone. These complementary advantages are expected to yield more robust, accurate and reliable intelligent systems [1, 4].

The research interest in many new technologies is also on the rise. By enabling models to directly acquire meaningful representations from vast amounts of unlabeled signal data, self-supervised learning seeks to lessen reliance on manually labeled datasets. Multimodal learning involves combining data from multiple sensing modalities – vibration, auditory, visual and physiological – to boost overall system reliability. Moreover, graph neural networks (GNNs) and graph signal processing techniques provide novel approaches to modeling complex dynamics of physical systems and interconnected sensors. To help with data interpretation and decision support, researchers have lately started looking into the use of large language models (LLMs) in wearable sensing and human activity detection. Although there is still research going on in these technologies [4, 8, 10], they present great promise in improving the abilities of intelligent signal processing.

Overall, future studies should focus on developing dependable, comprehensible, numerically efficient, and adaptive learning systems in addition to improving prediction accuracy. In the future, physical signal analysis systems based on statistical signal processing and advanced artificial intelligence techniques are expected to improve in generalization, robustness in real-world situations, and applicability to engineering, healthcare, environmental monitoring and scientific research.

6. Conclusion

In this review, the integration of statistical signal processing and machine learning for physical signal analysis in industrial, biomedical and wearable sensing systems is studied. Physical measurements are often noisy, have missing values, are non-stationary, and vary between different sensors/operating conditions, making pre-processing and data representation very important prior to model training. Interpretable descriptions of signal behavior are provided through Fourier and wavelet methods, filtering, spectral analysis and statistical feature extraction. These features enable traditional machine learning approaches to be used effectively when labeled data is scarce, whereas convolutional neural networks, recurrent models, and Transformers can learn more sophisticated representations from either raw or lightly processed signals when more data and computing resources are available. The studies examined demonstrate that none of the techniques is consistently best. Model selection depends on the type of signal, the number of samples, the temporal structure of the signal, the required level of computational resources, and physical

interpretability. The main challenges include limited labeled data, class imbalance, domain shift, inconsistent sensor configurations, high computational costs and limited transparency in deep learning models. These issues are particularly relevant in medical diagnosis or industrial monitoring, where unreliable predictions could have serious consequences. Future studies should therefore strengthen research on self-supervised learning, transfer and continual learning, uncertainty quantification, multimodal sensor fusion, and explainable artificial intelligence. There is also a need for lightweight models and model-compression techniques for deployment on wearable and edge devices. Data preprocessing, train-test separation, external validation, computational cost and performance under changing conditions should be reported in more rigorous studies. In general, a promising direction is the combination of physical knowledge, interpretable signal representations and data-driven models in systems that are accurate, robust, efficient and can be readily used in science and engineering in the field.

References

- [1] Shlezinger, N., Whang, J., Eldar, Y. C., & Dimakis, A. G. (2023). Model-based deep learning. *Proceedings of the IEEE*, 111(5), 465–499. <https://doi.org/10.1109/JPROC.2023.3247480>
- [2] Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4), 045002. <https://doi.org/10.1103/RevModPhys.91.045002>
- [3] Lee, Y. J., Park, C., Kim, H., Cho, S. J., & Yeo, W.-H. (2024). Artificial intelligence on biomedical signals: Technologies, applications, and future directions. *Med-X*, 2, Article 25. <https://doi.org/10.1007/s44258-024-00043-1>
- [4] Calazans, M. A. A., Ferreira, F. A. B. S., Santos, F. A. N., Madeiro, F., & Lima, J. B. (2024). Machine learning and graph signal processing applied to healthcare: A review. *Bioengineering*, 11(7), 671. <https://doi.org/10.3390/bioengineering11070671>
- [5] Tama, B. A., Vania, M., Lee, S., & Lim, S. (2023). Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals. *Artificial Intelligence Review*, 56, 4667–4709. <https://doi.org/10.1007/s10462-022-10293-3>
- [6] Li, A., Li, H., & Yuan, G. (2024). Continual learning with deep neural networks in physiological signal data: A survey. *Healthcare*, 12(2), 155. <https://doi.org/10.3390/healthcare12020155>
- [7] Ye, X., Sakurai, K., Nair, N.-K. C., & Wang, K. I.-K. (2024). Machine learning techniques for sensor-based human activity recognition with data heterogeneity—A review. *Sensors*, 24(24), 7975. <https://doi.org/10.3390/s24247975>
- [8] Ferrara, E. (2024). Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. *Sensors*, 24(15), 5045. <https://doi.org/10.3390/s24155045>
- [9] Saad, H. S., Zaki, J. F. W., & Abdelsalam, M. M. (2024). Employing of machine learning and wearable devices in healthcare system: Tasks and challenges. *Neural Computing and Applications*, 36(29), 17829–17849. <https://doi.org/10.1007/s00521-024-10197-z>
- [10] Kaseris, M., Kostavelis, I., & Malassiotis, S. (2024). A comprehensive survey on deep learning methods in human activity recognition. *Machine Learning and Knowledge Extraction*, 6(2), 842–876. <https://doi.org/10.3390/make6020040>
- [11] Farrahi, V., & Rostami, M. (2024). Machine learning in physical activity, sedentary, and sleep behavior research. *Journal of Activity, Sedentary and Sleep Behaviors*, 3(1), Article 5. <https://doi.org/10.1186/s44167-024-00045-9>
- [12] Dantas, P. V., Sabino da Silva, W., Cordeiro, L. C., & Carvalho, C. B. (2024). A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 54(22), 11804–11844. <https://doi.org/10.1007/s10489-024-05747-w>