UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Pedro Henrique Oliveira Silva

Multiclass Classification using Logistic-NARX Modeling: Methods and Engineering Applications

Juiz de Fora

2025

Pedro Henrique Oliveira Silva

Multiclass Classification using Logistic-NARX Modeling: Methods and Engineering Applications

Thesis submitted to the Graduate Program in Electrical Engineering of the Federal University of Juiz de Fora as a partial requirement for obtaining a Doctors's degree in Electrical Engineering. Concentration area: Electronic Systems.

Supervisor: Prof. Dr. Augusto Santiago Cerqueira

Co-supervisor: Prof. Dr. Erivelton Geraldo Nepomuceno

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Silva, Pedro Henrique Oliveira.
Multiclass Classification using Logistic-NARX Modeling:
Methods and Engineering Applications / Pedro Henrique Oliveira Silva.
– 2025.
105 f. : il.

Supervisor: Augusto Santiago Cerqueira
Co-supervisor: Erivelton Geraldo Nepomuceno
Tese (Doutorado) – Universidade Federal de Juiz de Fora, Faculdade de Engenharia Elétrica. Programa de Pós-Graduação em Engenharia Elétrica, 2025.

1. Machine learning. 2. system identification. 3. railroad dynamics. I. Cerqueira, Augusto Santiago, orient. II. Nepomuceno, Erivelton Geraldo, coorient. III. Título.

FEDERAL UNIVERSITY OF JUIZ DE FORA
RESEARCH AND GRADUATE PROGRAMS OFFICE

Pedro Henrique Oliveira Silva

Multiclass Classification using Logistic-NARX Modeling: Methods and Engineering Applications

Thesis submitted to the Graduate Program in Electrical Engineering
of the Federal University of Juiz de Fora as a partial
requirement for obtaining a Doctors's degree in Electrical Engineering.
Concentration area: Electronic Systems

Approved on 19 of September of 2025.

EXAMINING BOARD

Prof. Dr. Augusto Santiago Cerqueira – Academic Advisor
Federal University of Juiz de Fora

Prof. Dr. Erivelton Geraldo Nepomuceno – Academic Co-Advisor
Maynooth University

Prof. Dr. Leandro Rodrigues Manso Silva
Federal University of Juiz de Fora

Prof. Dr. Rafael Antunes Nóbrega
Federal University of Juiz de Fora

Prof. Dr. Danton Diego Ferreira
Federal University of Lavras

Prof. Dr. Eduardo Mazoni Andrade Marçal Mendes
Federal University of Minas Gerais

Juiz de Fora, 09/19/2025.

Documento assinado eletronicamente por **Augusto Santiago Cerqueira**, **Professor(a)**, em 19/09/2025, às 12:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Danton Diego Ferreira**, **Usuário Externo**, em 19/09/2025, às 12:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leandro Rodrigues Manso Silva**, **Professor(a)**, em 19/09/2025, às 12:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Erivelton Geraldo Nepomuceno**, **Usuário Externo**, em 19/09/2025, às 12:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Mazoni Andrade Marçal Mendes**, **Usuário Externo**, em 01/10/2025, às 09:07, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2589136** e o código CRC **99C20710**.

ACKNOWLEDGEMENTS

To my family, especially my parents, Elisabete and Adélio, and my siblings, Marina and Gabriel, for their unconditional love, constant support, and unwavering trust throughout my journey. To my advisors, Professor Augusto Santiago Cerqueira and Professor Erivelton Geraldo Nepomuceno, for their dedicated guidance, continuous encouragement, and valuable contributions during the development of this work. To my colleagues from the Graduate Program in Electrical Engineering, for their collaboration and enriching companionship. To the Federal University of Juiz de Fora, for the institutional support and for providing the resources necessary for the completion of this research. To CAPES, CNPq, and FAPEMIG, for the financial support that made this study possible.

"The creation of a single world comes from a huge number of fragments and chaos." (Cavallaro, 2015, p. 5).

ABSTRACT

Classification techniques are essential for interpreting complex data across various engineering domains, including signal processing, manufacturing, and environmental monitoring, providing actionable insights and informing critical decision-making processes. System identification and data-driven modeling methodologies offer robust frameworks for analyzing these complex engineering processes. Parametric modeling approaches, notably Nonlinear AutoRegressive models with eXogenous inputs (NARX), are particularly advantageous due to their linear-in-the-parameter form, enabling interpretability and model transparency. Although recent developments have extended NARX methodologies to classification tasks, their application remains predominantly limited to regression and binary classification, leaving multiclass scenarios relatively unexplored. To bridge this gap, this thesis introduces a novel classification algorithm, Logistic-NARX Multinomial, combining logistic regression with NARX modeling principles. This integration facilitates direct interaction among input terms, resulting in sparse, interpretable models where the significance of individual and combined input variables is clearly discernible. Extensive evaluations using benchmark and real-world datasets confirm that Logistic-NARX Multinomial achieves competitive predictive performance compared to traditional classifiers. Additionally, this thesis proposes a practical engineering methodology for railway infrastructure maintenance. Utilizing acceleration data from railway vehicles and multibody simulations, the developed Logistic-NARX Multinomial framework provides interpretable and transparent predictive models capable of effectively assessing track conditions, significantly enhancing geo-referenced decision-making and contributing to improved safety and maintenance practices.

Keywords: machine learning; system identification; NARX models; power quality; multiclass classfication; wheel–rail contact dynamic forces; railroad dynamics.

RESUMO

Técnicas de classificação são fundamentais para interpretar dados complexos em diversas áreas da engenharia, incluindo processamento de sinais, manufatura e monitoramento ambiental, fornecendo insights úteis e auxiliando processos críticos de tomada de decisão. A identificação de sistemas e a modelagem orientada a dados oferecem estruturas robustas para analisar esses processos complexos da engenharia. Entre essas abordagens, destacam-se os modelos paramétricos, como os modelos AutoRegressivos Não Lineares com entradas eXógenas (NARX), particularmente vantajosos devido à sua estrutura linear nos parâmetros, que facilita a interpretabilidade e a transparência dos modelos resultantes. Embora desenvolvimentos recentes tenham estendido os modelos NARX para tarefas de classificação, sua aplicação permanece predominantemente restrita a cenários de regressão e classificação binária, deixando relativamente inexploradas as situações de classificação multiclasse. Para preencher essa lacuna, esta tese propõe um novo algoritmo denominado Logistic-NARX Multinomial, integrando a regressão logística aos princípios da modelagem NARX. Essa integração permite interações diretas entre os termos de entrada, resultando em modelos esparsos e interpretáveis, nos quais a relevância das variáveis individuais e suas interações fica claramente identificada. Avaliações extensivas utilizando conjuntos de dados clássicos e reais confirmam que o Logistic-NARX Multinomial apresenta desempenho preditivo competitivo em relação a classificadores tradicionais. Adicionalmente, a tese propõe uma metodologia prática aplicada à manutenção da infraestrutura ferroviária. Utilizando dados de aceleração provenientes de veículos ferroviários e simulações multicorpo, o modelo Logistic-NARX Multinomial gera modelos preditivos interpretáveis e transparentes, capazes de avaliar efetivamente as condições da via férrea, aprimorando significativamente a tomada de decisões georreferenciadas e contribuindo para práticas mais seguras e eficientes de manutenção.

Palavras-chave: aprendizado de máquina; identificação de sistemas; modelos NARX, qualidade de energia; classificação multiclasse; contato roda-trilho; dinâmica ferroviária.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NARX	Nonlinear AutoRegressive with eXogenous inputs
PQ	Power Quality
OFR	Orthogonal Forward Regression
ERR	Error Reduction Ratio
OVA	One-Versus-All
HOS	Higher Order Statistics
RF	Random Forest
LASSO	Least Absolute Shrinkage and Selection Operator
LS	Least Squares
GD	Gradient Descent
OLS	Orthogonal Least Squares
SVM	Support Vector Machines
UCI	University of California Irvine
ML	Machine Learning
KNN	K-Nearest Neighbors
MATLAB	MATrix LABoratory
CV	Cross-validation
L-NARX M	Logistic-NARX Multinomial
RMS	Root Mean Square
IEEE	Institute of Electrical and Electronics Engineers
SNR	Signal-to-Noise Ratio
LDA	Linear Discriminant Analysis
IIR	Infinite Impulse Response
FDA	Fisher Discriminant Analysis
FRA	Federal Railroad Administration
VTI	Vehicle/Track Interaction
L/V	Lateral-to-Vertical force ratio
ES	European Standards
ECS	European Committee for Standardization
ABNT	Brazilian Association of Technical Standard
VAMPIRE	Vehicle Analysis Modeling Package in the Railway Environment
GPS	Global Positioning System
FN	False Negatives
FP	False Positives
FP	False Positives

LIST OF SYMBOLS

$y(k)$	Output value in terms
f^l	Nonlinear function of the model with a degree of nonlinearity
$u(k)$	System input at discrete time
$e(k)$	Represents uncertainties and possible noises at discrete time
n_u	Maximum delays of the system output
$\phi_i(\varphi(k))$	Linear combination of functions that depend on the regressor vector
θ_i	Coefficients to be estimated
l	Degree of nonlinearity
$\varphi(k)$	Regressor vector
M	Total number of potential terms
q_i	Orthogonal columns of Q
w_i	Intermediate orthonormalized vector (Gram–Schmidt)
g_i	Element of the auxiliary vector g
m	Total number of candidate terms
m_0	Number of terms selected in the final model
α_s	Significant term selected at step s
j_s	Index of the term selected at step s
(n_y, n_u)	Maximum delays
Y	Matrix form vector of estimates
Φ	Regressor matrix
θ	Estimated parameters
ξ	Vector of residuals
Q	Orthogonally matrix
A	Upper triangular matrix
D	Positive definite diagonal matrix
g	Auxiliary parameter vector
$\mathcal{M}$	Search space with all possible regressors
ε	Tolerance value
$f(x)$	Logistic function
$p(x)$	Probability model
j	Selector term with the highest accuracy
f_v	Outcome given by the model corresponding to class v
n_{max}	Maximum number of selected terms
$u_n(k)$	Model terms
$\overline{x}$	Average accuracy
$v(n)$	Power system voltage signal
N	Samples
f_s	Sampling frequency
$f(n)$	Fundamental component, harmonics, interharmonics
$A_0(n)$	Amplitude of the voltage signal

$f_0(n)$	Fundamental frequency
$\theta_0(n)$	Phase of the voltage signal
$r(n)$	Noise with normal distribution
$(c_{2,x}, c_{4,x})$	Second and fourth-order cumulants
J_c	Fisher Discriminant criterion
T	Tangential force
L/V	Lateral-to-vertical force ratio
RMS	Root Mean Square value
SNR	Signal-to-Noise Ratio
$\phi(k)$	Generic regression vector at discrete time k
$\hat{y}(k)$	Estimated output at discrete time k

TABLE OF CONTENTS

1 INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Understanding physical phenomena through observed signals is essential in engineering and scientific disciplines, as these signals convey the behavior and interactions of dynamic systems (Oppenheim et al., 1997). In many applications, input-output relationships govern how systems respond to external stimuli, and modeling these interactions allows for both interpretation and prediction of complex processes (Rogers and Girolami, 2011; Billings, 2013).

In recent decades, the exponential growth in data availability has transformed the modeling landscape. With an estimated 90% of the world's data generated in just the last two years (John Walker, 2014), a major challenge has emerged: transforming vast, unstructured, and often noisy datasets into actionable insights (Manyika et al., 2011). This shift has accelerated the development of empirical and data-driven models in domains ranging from finance and healthcare to energy systems, transportation, and environmental monitoring (Abu-Mostafa, 2012; Liu et al., 2023; Jayaprakash and Balamurugan, 2021). These models aim to uncover latent patterns and relationships directly from measurements, enabling what is commonly referred to as knowledge discovery or learning from data (Shu and Ye, 2023; Pazzani, 2000).

Data-driven approaches, particularly those grounded in machine learning and system identification, have gained prominence due to their adaptability and potential for automation (Wu et al., 2014; Janiesch et al., 2021). However, despite substantial advances, significant challenges remain, particularly in ensuring the transparency, interpretability, and reliability of such models in critical engineering contexts. Many existing techniques act as black boxes, offering little understanding of how predictions are generated (Witten et al., 2016). This lack of interpretability can hinder their adoption in domains where accountability, regulatory compliance, and human-in-the-loop decisions are vital (Kuhn and Johnson, 2013; Zhang and Lang, 2022).

To bridge this gap, it is essential to develop models that strike a balance between predictive accuracy and structural clarity. These models must not only perform well but also offer insight into the underlying system dynamics, feature relevance, and interaction mechanisms (Gu and Wei, 2018). The integration of sparse structures with interpretable parameters is especially critical in real-world settings, where decision-making often relies on understanding which variables are most influential.

A promising candidate to address these demands is the Nonlinear AutoRegressive model with eXogenous inputs (NARX), known for its compact, flexible, and interpretable architecture (Vidyalashmi et al., 2024; Aguirre, 2007). Built using measured input-output data, the NARX model facilitates the identification of significant nonlinear relationships

while simultaneously managing redundancy and multicollinearity (Billings and Wei, 2019). Recent advances in NARX modeling have expanded its applications, yet its use has remained largely confined to regression and binary classification tasks (Ayala Solares et al., 2019). This presents an opportunity to extend its capabilities to more intricate classification scenarios, such as those involving multiple discrete classes.

This thesis responds to that opportunity by proposing a hybrid modeling framework that adapts the NARX methodology for multiclass classification problems. By incorporating logistic regression into the NARX structure, the resulting Logistic-NARX Multinomial model achieves transparent and interpretable classification outcomes. This approach is particularly suitable for scenarios where a parsimonious yet expressive model is needed to support both analysis and decision-making.

One practical domain where these methodological advances are highly relevant is the railway sector. With increasing demands on rail networks due to heavier traffic and higher speeds, infrastructure degradation poses significant safety and operational risks (Lasisi and Attoh-Okine, 2018). Traditional diagnostic approaches, such as inspections via specialized vehicles, are costly, infrequent, and can disrupt operations (Malekjafarian et al., 2019). In this context, data-driven models based on acceleration signals and dynamic simulation have emerged as viable alternatives for inferring track conditions (Sun et al., 2024; Marasco et al., 2024). By combining physical insight from multibody dynamics with classification-based modeling, this research enables indirect estimation of critical variables, such as wheel-rail contact forces, and provides actionable information for maintenance planning through georeferenced tools.

Thus, this thesis is motivated by the need to construct models that are not only accurate but also interpretable and adaptable to practical engineering challenges. The proposed methodology seeks to advance the field by offering a robust framework for multiclass classification, particularly suited for applications where explainability and data complexity intersect.

1.2 OBJECTIVES AND SPECIFIC GOALS

This research seeks to investigate and develop hybrid modeling strategies that integrate system identification techniques with machine learning methods to address complex classification challenges in engineering domains. The primary focus is on creating interpretable and efficient models capable of handling multiclass classification tasks, with a strong emphasis on transparency and alignment with the physical behavior of real-world systems.

The **general objective** of this study is:

- To propose and develop a classification methodology that combines the NARX

(Nonlinear AutoRegressive with eXogenous inputs) modeling framework with logistic regression, enabling the analysis of multiclass classification problems through interpretable and structurally simple models.

The research is further guided by the following **specific objectives**:

- To extend the conventional NARX modeling approach, traditionally used for regression and binary classification, to effectively address multiclass scenarios;
- To formulate and validate the Logistic-NARX Multinomial algorithm, employing an accuracy-based criterion for term selection and using cross-validation techniques to ensure model generalization;
- To assess the performance of the proposed methodology using benchmark datasets, focusing on classification accuracy, interpretability of model terms, and parsimony;
- To apply the developed approach to the classification of Power Quality disturbances, using synthetically generated waveform data to demonstrate its practical applicability and discriminative capability in a well-established context;
- To apply and validate the proposed methodology in the context of railway track condition assessment, using data from multibody dynamic simulations and inertial sensor measurements. This application has a dual contribution: it serves as a high-impact validation case and introduces an innovative approach for railway infrastructure monitoring, capable of inferring safety indicators and supporting georeferenced maintenance strategies through interpretable, data-driven models.

By achieving these objectives, the thesis aims to contribute both methodologically, by advancing transparent and hybrid classification models, and practically, by demonstrating their applicability and effectiveness in diverse engineering problems, particularly in power system diagnostics and railway safety monitoring.

1.3 CONTRIBUTIONS

This thesis proposes and evaluates a hybrid modeling approach that integrates machine learning and system identification, specifically tailored to address multiclass classification problems. The primary motivation is to bridge the gap between the high predictive power of data-driven techniques and the need for structural clarity in engineering models. The contributions of this work are twofold: methodological and applicational, both strongly aligned with real-world engineering demands.

First, the thesis introduces the *Logistic-NARX Multinomial* framework, a novel classification methodology that adapts the NARX (Nonlinear AutoRegressive with eXogenous inputs) paradigm to multiclass settings. Unlike conventional approaches where NARX models are mostly employed in regression or binary classification tasks, the proposed method extends its utility to problems where the output consists of categorical labels. Through the use of logistic regression and orthogonal forward regression for structure selection, this method enables the extraction of sparse, interpretable, and high-performing models. Each selected model term is analytically associated with a specific input variable or interaction, thus revealing the role of features in the classification decision. This makes the methodology not only accurate but also explainable, an essential requirement in many engineering applications. The method was validated using benchmark datasets and a practical case study involving power quality disturbance classification, demonstrating competitive performance and valuable insights into feature relevance and system behavior.

Second, the thesis contributes to the railway domain by proposing and applying the developed methodology to the problem of railway track condition assessment. This is achieved through an innovative modeling strategy that employs acceleration data from multibody dynamic simulations to classify different safety conditions of the railway infrastructure. In this context, the Logistic-NARX Multinomial model is applied to explore potential associations between dynamic behaviors and vehicle–track interaction patterns that may influence proximity to critical safety thresholds. By enabling the indirect inference of wheel–rail contact forces and relating them to stability thresholds, the method supports the early detection of structural irregularities and enhances decision-making through a georeferenced application.

1.4 LIST OF PUBLICATIONS

The following manuscripts were published during the course of this work:

- P. H. O. Silva, A. S. Cerqueira, E. G. Nepomuceno, e A. F. Oliveira, “Classificação de Distúrbios na Qualidade de Energia Usando Modelagem Logística-NARX Multinomial,” in *Anais do Congresso Brasileiro de Automática (CBA)*, 2020. (Silva et al., 2020)
- P. H. O. Silva, A. S. Cerqueira, e E. G. Nepomuceno, “Hybrid Method Based on NARX models and Machine Learning for Pattern Recognition,” in *Anais do XV Simpósio Brasi- leiro de Automação Inteligente (SBAI)*, 2021. (Silva et al., 2021)
- P. H. O. Silva, R. D. Marotta, A. S. Cerqueira, E. G. Nepomuceno, e L. A. S. Lopes, “Avaliação da Condição da Via Permanente usando Dados de Dinâmica de Veículos Ferroviários: Uma Abordagem de Aprendizado de Máquina,” in *Anais do Congresso Brasileiro de Automática (CBA)*, 2022. (Silva et al., 2022)

- P. H. O. Silva, A. S. Cerqueira, E. G. Nepomuceno, "Insightful Railway Track Evaluation: Leveraging NARX Feature Interpretation," in *Anais do Congresso Brasileiro de Automática (CBA)*, 2024. (Silva et al., 2024)

1.5 OVERVIEW

This thesis is organised as follows:

- **Chapter 2** provides the theoretical foundation of nonlinear system identification, with particular emphasis on NARX (Nonlinear AutoRegressive with eXogenous inputs) models. It introduces the key concepts required for understanding the remainder of the thesis, including model structure determination, parameter estimation, and the Orthogonal Forward Regression (OFR) algorithm using the Error Reduction Ratio (ERR) criterion.
- **Chapter 3** introduces the proposed Logistic-NARX Multinomial framework, which extends the conventional NARX model to address multiclass classification problems through the integration of logistic regression and the One-Versus-All (OVA) decomposition strategy. The chapter details the term selection strategy based on k-fold cross-validation accuracy, enabling a transparent and interpretable model structure. The methodology is evaluated using classical benchmark datasets, emphasizing both predictive performance and feature relevance.
- **Chapter 4** applies the proposed methodology to the classification of Power Quality (PQ) disturbances. The chapter encompasses the simulation of electrical events, preprocessing of voltage signals, and extraction of features using Higher Order Statistics (HOS). Feature selection is performed through Fisher's Discriminant Analysis. The results are compared with classical classifiers, demonstrating the method's capability to discriminate between PQ events and highlighting the interpretability of the selected model terms.
- **Chapter 5** focuses on a practical application in the railway domain. A novel approach is developed for assessing railway track conditions using acceleration signals derived from multibody simulations and instrumented railway vehicles. The Logistic-NARX Multinomial model is employed to infer proximity to critical safety thresholds based on dynamic behavior and vehicle–track interaction features. Feature selection and subsampling strategies are used to enhance model generalization. Comparative analyses with traditional machine learning techniques, including Random Forests with hyperparameter tuning, are presented, along with a georeferenced application that supports maintenance decision-making.

- **Chapter 6** concludes the thesis by summarizing the main findings, discussing methodological contributions, and presenting perspectives for future research. Particular attention is given to possible enhancements in the model selection process and extensions to other domains.

2 NONLINEAR SYSTEMS IDENTIFICATION

This chapter begins with an introduction to system identification, offering a comprehensive overview of the fundamental concepts necessary for understanding the topics discussed throughout this work. The subsequent sections present the main stages of the identification process, with particular emphasis on the NARX model and the Orthogonal Forward Regression (OFR) algorithm, which employs the Error Reduction Ratio (ERR) criterion.

2.1 INTRODUCTION

The development and analysis of models are some of the most important topics in science. Models can be used for system analysis, providing a better understanding of the system. Similarly, models allow one to predict or simulate the behavior of a system. In engineering, models are necessary for designing new processes and analyzing existing ones. Since the quality of the model typically sets an upper limit on the quality of the final solution to the problem, modeling is often a constraint in the development of the entire system. Consequently, there is a strong demand for advanced models and identification schemes (Aguirre, 2007). Different models can be obtained for a specific study, and the choice of model representation depends on the user's knowledge, the system under study, and the objectives of the modeling process. In this context, modeling dynamic and steady-state behavior is fundamental for this type of analysis and is based on system identification procedures.

According to Sjöberg et al. (1995), modeling techniques can be classified into three groups known as white-box modeling, black-box modeling, and grey-box modeling. White-box models are described by the physical laws of the process, and typically the parameters have a physical meaning. Black-box models are obtained entirely from input and output data, without requiring any prior knowledge of the system. Finally, grey-box models are characterized by using auxiliary information that is not present in the dynamic data set used for identification (Corrêa and Aguirre, 2004).

System identification is an experimental approach aimed at identifying and adjusting a mathematical model of a system based on experimental data that records the behavior of the system's inputs and outputs (Billings, 2013; Nelles, 2020). In particular, interest in the identification of nonlinear systems has received considerable attention from researchers since the 1950s, and many relevant results have been developed (Wiener, 1958; Lee and Schetzen, 1965; Schetzen, 1980; Rugh, 1981; Haber and Keviczky, 1999; Billings and Wei, 2005; Pintelon and Schoukens, 2012). A commonly employed model representation is the Nonlinear AutoRegressive with eXogenous input (NARX) model, consisting of a mathematical model based on differential equations. Due to the simplicity and versatility

of NARX models, various systems can be modeled, understood, and controlled using this family of models. One contribution of the present work is to present an efficient alternative for a multiclass classifier by applying NARX models to power quality problems.

The typical stages of the identification process include dynamic testing and data collection, selection of the mathematical representation, determination of the model structure, parameter estimation, and model validation. This chapter provides a set of definitions and preliminary concepts for understanding the field of system identification.

2.2 REPRESENTATION: NARX MODELS

Linear systems are defined as systems that satisfy the principle of superposition, and calculating the relationship between models in a linear scenario is straightforward. For instance, if a system is identified using a state-space model, other types of models can be derived using transformations. However, for nonlinear models, there is no single model to represent all classes of nonlinear systems, and transforming one model into another is generally a challenging task. There are various types of formats and forms to represent a dynamic system in the nonlinear system identification literature, so the choice of representation should ensure the fulfillment of the main objective of the identification process. Thus, a traditional choice is the NARX model, since a substantial portion of nonlinear systems can be represented by this model in the discrete time domain (Billings, 2013).

Polynomial NARX representations are discrete-time models that explain the output value $y(k)$ in terms of previous values of the output and input signals:

$$y(k) = f^l(y(k-1), \cdots, y(k-n_y), u(k-1), \cdots, u(k-n_u)) + e(k), \tag{2.1}$$

given that f^l represents a nonlinear function of the model with a degree of nonlinearity $l \in \mathbb{N}$, $y(k) \in \mathbb{R}^{n_y}$ is the system output, and $u(k) \in \mathbb{R}^{n_u}$ is the system input at discrete time $k = 1, 2, \ldots, N$; N is the number of observations. In this case, $e(k) \in \mathbb{R}^{n_e}$ represents uncertainties and possible noise in discrete time k, $n_y \in \mathbb{N}$, and $n_u \in \mathbb{N}$ describe the maximum delays in the output and input of the system, respectively.

Most approaches assume that the function f^l can be approximated by a linear combination of a predefined set of functions $\phi_i(\varphi(k))$, (2.1) can be expressed in the following parametric linear form:

$$y(k) = \sum_{i=1}^{m} \theta_i \phi_i(\varphi(k)) + e(k), \tag{2.2}$$

where θ_i are the coefficients to be estimated and $\phi_i(\varphi(k))$ are the predefined functions that depend on the regressor vector:

$$\varphi(k) = [y(k-1), \cdots, y(k-n_y), u(k-1), \cdots, u(k-n_u)]^{\mathsf{T}}, \tag{2.3}$$

assuming previous outputs and inputs, and m is the number of functions in the set. One of the most commonly used NARX models is the polynomial representation, (2.2) can be denoted in the following form:

$$y(k) = \theta_0 + \sum_{i_1=1}^{n} \theta_{i_1} x_{i_1}(k) + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \theta_{i_1 i_2} x_{i_1}(k) x_{i_2}(k) + \cdots \\ \sum_{i_1=1}^{n} \cdots \sum_{i_l=i_l-1}^{n} \theta_{i_1 i_2 \cdots i_l} x_{i_1}(k) x_{i_2}(k) \cdots x_{i_l}(k) + e(k), \tag{2.4}$$

considering $n = n_y + n_u$,

$$x_i(k) = \begin{cases} y(k-i), & 1 \leq i \leq n_y, \\ u(k-i+n_y), & n_y + 1 \leq i \leq n, \end{cases} \tag{2.5}$$

assuming $n = n_y + n_u$, where n_y is the number of outputs and n_u is the number of inputs, and l is the degree of nonlinearity. The NARX model of order l implies that the order of each term in the model is not greater than l. The total number of potential terms in a polynomial NARX model is given by:

$$M = \frac{(n+l)!}{n! \cdot l!}. \tag{2.6}$$

NARX models can be used to describe a wide variety of nonlinear systems, providing straightforward analytical insights into the model's dynamics. Another advantage is parsimony, meaning that a broad range of behaviors can be represented concisely using only a few terms from the extensive search space formed by candidate regressors. Additionally, a small dataset is required to estimate a model, which can be crucial in applications where acquiring a large amount of data is challenging. There are other advantages of polynomial NARX models (Billings, 2013):

- Polynomial functions are infinitely differentiable (smooth functions) in $\mathbb{R}$;
- According to the Weierstrass theorem Weierstrass (1885), any given continuous real-valued function defined on a closed and bounded interval $[a, b]$ can be uniformly approximated using a power-form polynomial on that interval;
- A wide variety of nonlinear dynamics can be characterized using polynomial NARX models;
- Polynomial NARX models are a well-established topic in the field of system identification, with various algorithms developed for structure determination and parameter estimation;

- NARX models can be employed for prediction and inference.

Example 2.2.1 *Consider a NARX model (Aguirre, 2007) with a nonlinearity degree $l = 2$ and maximum delays $n_y = 2$ and $n_u = 1$, then the following regressor vector is obtained by (2.3):*

$$\varphi(k) = [1 \quad y(k-1) \quad y(k-2) \quad u(k-1) \quad y(k-1)^2 \quad y(k-1)y(k-2) \\ y(k-2)^2 \quad u(k-1)^2 \quad y(k-1)u(k-1) \quad y(k-2)u(k-1)]^{\mathsf{T}}. \quad (2.7)$$

In Example 2.2.1, the total number of potential terms corresponds to the value $M = 10$, as per (2.6). The set of regressors that form the vector $\varphi(k)$ is complete in the sense that it includes all possible regressors given the maximum delays (n_y, n_u) and nonlinearity degree (l). Assuming a specific application with higher values of maximum delays and nonlinearity degree, the number of possible regressors increases significantly. The number of parameters grows exponentially with the polynomial order, and if all corresponding terms of the regression vector are considered unnecessarily, the excessive number of parameters will increase the estimation variance and compromise the model quality. Therefore, a fundamental requirement in the structure determination phase of these models is to identify an optimal subset of regressors.

2.3 MODEL STRUCTURE AND PARAMETER ESTIMATION

The determination of the model structure is crucial for developing models that can accurately reproduce the system behavior. One of the key factors in the structure determination phase is defining the number of candidate model terms that contribute to the system's output while maintaining an efficient system description (Haber and Unbehauen, 1990). In general, most candidate model terms are redundant or spurious, and their contribution to the system's output is insignificant. Furthermore, a model that includes a large number of terms tends to overgeneralize the problem. Among the advantages of carefully conducting model structure detection are improved prediction or classification accuracy, reduced time and storage costs, and a better understanding of the studied process (Wei et al., 2004).

For nonlinear systems, there are numerous techniques for determining the model structure, such as clustering algorithms (Aguirre and Jácome, 1998), the least absolute shrinkage and selection operator (LASSO) (Kukreja et al., 2006), elastic nets (Zou and Hastie, 2005), genetic programming (Sette and Boullart, 2001), bagging methodology (Ayala Solares and Wei, 2015), and the Orthogonal Forward Regression (OFR) method using the Error Reduction Ratio (ERR) approach (Wei et al., 2004). Once the structure of the model is determined, the parameter estimation is performed, which can be accomplished

using traditional methods such as Least Squares, Gradient Descent, and the Metropolis-Hastings algorithm (Baldacchino et al., 2012).

2.3.1 Orthogonal Forward Regression

In general, determining the model structure and estimating the parameters are carried out together. One of the most popular algorithms for performing both steps for NARX modeling is the Orthogonal Forward Regression (OFR) algorithm (Billings, 2013). The algorithm transforms a set of candidate terms into orthogonal vectors and ranks them based on their contribution to the output data using the Error Reduction Ratio (ERR), identifying and fitting a deterministic and parsimonious NARX model expressible in a generalized linear regression form. The OFR algorithm consists of three main steps:

- Orthogonalize the regressors to remove correlations between variables;
- Select significant terms using ERR as the criterion;
- Estimate the corresponding parameters for the selected terms.

2.3.1.1 Matrix Form of Parameter Representation

In order to present algorithms for parameter estimation in NARX models, it will be convenient to use the more compact representation given by (2.2) in matrix form:

$$Y = \Phi\theta + \xi, \tag{2.8}$$

the vectors and the matrix in (2.8) are represented by:

$$Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix}, \quad \xi = \begin{bmatrix} \xi(1) \\ \xi(2) \\ \vdots \\ \xi(N) \end{bmatrix}, \tag{2.9}$$

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_m \end{bmatrix} = \begin{bmatrix} \phi_1(\varphi(1)) & \phi_2(\varphi(1)) & \cdots & \phi_m(\varphi(1)) \\ \phi_1(\varphi(2)) & \phi_2(\varphi(2)) & \cdots & \phi_m(\varphi(2)) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\varphi(N)) & \phi_2(\varphi(N)) & \cdots & \phi_m(\varphi(N)) \end{bmatrix}, \tag{2.10}$$

given $Y \in \mathbb{R}^N$ as the vector of estimates, $\Phi \in \mathbb{R}^{N\times m}$ is the regressor matrix, $\theta \in \mathbb{R}^N$ is the vector of estimated parameters, and $\xi \in \mathbb{R}^N$ is the vector of residuals.

The parameters presented in (2.8) could be estimated as a result of an algorithm based on Least Squares, but this would require the optimization of all parameters simultaneously, due to the correlation between the regressors and the non-orthogonality feature.

Therefore, the computational cost would be impractical for a large number of regressors. Thus, the problem is addressed by orthogonalizing the matrix Φ, making the regressors uncorrelated and forming an orthogonal basis for the system's solutions. The approach is called Orthogonal Least Squares or OLS, which succinctly transforms a non-orthogonal model into an orthogonal one.

2.3.1.2 Estimador OLS

Assuming that the regressor matrix Φ is of full rank, according to matrix theory, there exists a matrix Q such that Φ can be orthogonally decomposed as:

$$\Phi = QA, \tag{2.11}$$

where $A \in \mathbb{R}^{m\times m}$ is an upper triangular matrix, as follows:

$$A = \begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1m} \\ 0 & 1 & a_{23} & \cdots & a_{2m} \\ 0 & 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & a_{m-1,m} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.12}$$

assuming $Q \in \mathbb{R}^{N\times m}$ is a matrix with orthogonal columns q_i, represented by:

$$Q = \begin{bmatrix} q_1 & q_2 & q_3 & \cdots & q_m \end{bmatrix}, \tag{2.13}$$

such that $Q^{\mathsf{T}}Q = D$, where D is a positive definite diagonal matrix:

$$d_i = q_i^{\mathsf{T}} q_i = \sum_{n}^{k=1} q_{ik} q_{ik}, \qquad 1 \le i \le m, \tag{2.14}$$

thus, the entries a_{ij} $(1 \le i < j \le m)$ of the matrix A are defined in matrix form as:

$$a_{rs} = \frac{q_r^{\mathsf{T}} \phi_s}{q_r^{\mathsf{T}} q_r}, \qquad 1 \le r \le s-1, \qquad 2 \le s \le m. \tag{2.15}$$

The space spanned by the orthogonal basis Q (2.13) is equivalent to that spanned by the basis set Φ (2.10). Consequently, (2.8) can be rewritten in the auxiliary model:

$$Y = \underbrace{(\Phi A^{-1})}_{Q} \underbrace{(A\theta)}_{g} + \xi = Qg + \xi, \tag{2.16}$$

where $g \in \mathbb{R}^m$ is an auxiliary parameter vector. The parameters of the model (2.16) are represented by:

$$g = (Q^{\mathsf{T}}Q)^{-1}Q^{\mathsf{T}}Y = D^{-1}Q^{\mathsf{T}}Y, \tag{2.17}$$

which can also be expressed as:

$$g_i = \frac{Y^\top q_i}{q_i^\top q_i}, \qquad 1 \leq i \leq m. \tag{2.18}$$

Once the parameters θ and g satisfy the triangular system $A\theta = g$, any orthogonalization method such as Householder, classical Gram-Schmidt, and modified Gram-Schmidt can be employed to solve the equation and estimate the original parameters (Aguirre, 2007). A more detailed discussion on these orthogonalization procedures can be found in Chen et al. (1989).

2.3.1.3 Error Reduction Ratio

Assuming $E[\Phi^\top \xi] = 0$, the output variance can be defined by multiplying (2.16) by itself and dividing by n, resulting in:

$$\frac{1}{n} Y^\top Y = \underbrace{\frac{1}{n} \sum_{m}^{i=1} g_i^2 q_i^\top q_i}_{\substack{\text{explained output} \\ \text{by the regressor}}} + \underbrace{\frac{1}{n} \xi^\top \xi}_{\substack{\text{unexplained} \\ \text{output}}} . \tag{2.19}$$

The interpretation of (2.19) is that the sum of the squared values of Y can be explained, using an orthogonal basis, as the sum of the squared values of each orthogonal regressor, respectively multiplied by their parameters. The portion unexplained by the regressors is equal to the sum of the squared values of the residual vector ξ. Therefore, (2.19) allows quantifying the importance of each individual regressor. In summary, the Rate of Error Reduction (ERR) due to the inclusion of the i-th regressor, expressed as a fraction of the sum of the squared values of the data, is given by:

$$\mathrm{ERR}_i = \frac{g_i^2 q_i^\top q_i}{Y^\top Y}, \qquad 1 \leq i \leq m. \tag{2.20}$$

There are several ways to establish the stopping criterion for structure determination algorithms (Akaike, 1974). A commonly used approach is to terminate the algorithm when the output variance of the model falls below a predetermined threshold ε:

$$1 - \sum_{i=1}^{m} \mathrm{ERR}_i \leq \varepsilon. \tag{2.21}$$

2.3.1.4 Implementation of the OFR algorithm

One possible implementation of the orthogonalization procedure is to use the classical Gram-Schmidt algorithm. The implementation can be carried out step by step, and along with other definitions presented in this section, we can construct the following sequence of steps for the classical OFR Algorithm 1:

Algorithm 1: Orthogonal Forward Regression

Input: $\{(y(k)), k = 1, \ldots, N\}$, $\mathcal{M} = \{\phi_i, i = 1, \ldots, m\}$, l, n_y, n_u, ε
Output: $\boldsymbol{\alpha} = \{\phi_i, i = 1, \ldots, m_0\}$, $\boldsymbol{\theta} = \{\theta_i, i = 1, \ldots, m_0\}$

$\Phi \leftarrow$ General regressor matrix based on n_y, n_u, l ▷ Equation (2.10)
for $i = 1 : m$ **do**
  $w_i \leftarrow \frac{\phi_i}{\|\phi_i\|_2}$
  $g_i \leftarrow \frac{Y^\top w_i}{w_i^\top w_i}$ ▷ Equation (2.18)
  $\mathrm{ERR}_i \leftarrow \frac{g_i^2 w_i^\top w_i}{Y^\top Y}$ ▷ Equation (2.20)
  $j \leftarrow \arg\max_{1 \le i \le m} \{\mathrm{ERR}_i\}$ ▷ Equation (2.22)
  $q_1 \leftarrow w_j$
  $a_{11} \leftarrow \|\phi_j\|_2$
  $g_1 \leftarrow g_j$
  $\mathrm{err}[1] \leftarrow \mathrm{ERR}_j$
  $y_n^{(1)} \leftarrow y - g_1 \phi$
  Remove ϕ_j from Φ
  $s \leftarrow 1$
end
repeat
  $s \leftarrow s + 1$
  for $i = 1 : m$ **do**
    Orthonomalize ϕ_i in relation $[q_1, q_2, \ldots, q_{(s-1)}]$
    $w_i^{(s)} \leftarrow \phi_i - \sum_{r=1}^{s-1} \frac{\phi_m^\top q_r}{q_r^\top q_r} q_r, \qquad \phi_i \in \Phi - \Phi_{(s-1)}$ ▷ Equation (2.23)
    if $w_i^\top w_i < 10^{-10}$ **then**
      Remove ϕ_i from Φ
      Next iteration
    end
    Calculate $\mathrm{ERR}_i(w_i, y_n^{(s-1)})$
  end
  $j \leftarrow \arg\max_{1 \le i \le m-s+1} \{\mathrm{ERR}_i\}$
  $q_s \leftarrow w_j$
  $a_{rs} \leftarrow \frac{q_r^\top \phi_j}{q_r^\top q_r}, \qquad 1 \le r \le s-1$ ▷ Equation (2.15)
  $a_{ss} \leftarrow \left\| \phi_j - \sum_{r=1}^{s-1} a_{rs} q_r \right\|_2$
  $g_s \leftarrow q_s^\top y_n^{(s-1)}$
  $\mathrm{err}[s] \leftarrow \mathrm{ERR}_j$
  $y_n^{(s)} \leftarrow y_n^{(s-1)} - g_s q_s$
  Remove ϕ_j from Φ
  $\mathrm{ESR} \leftarrow 1 - \sum_{s=1}^{m_0} \mathrm{err}(s)$ ▷ Equation (2.21)
until $ESR \le \varepsilon$
Estimate the coefficients $A\boldsymbol{\theta} = g$

- Define a search space $\mathcal{M} = \{\phi_i, i = 1, \ldots, m\}$ with all possible regressors and set a tolerance value ε;

- Construct a general regressor matrix Φ (2.10) with candidates based on the maximum delays (n_y, n_u) and nonlinearity degree l;

- **Step** ($s = 1$): Consider $i = 1, \ldots, m$, set $q_i = \phi_i$, and compute the auxiliary parameter vector g_i (2.18). Then, calculate the error reduction rate ERR_i (2.20) for each term in Φ, and select the term with the highest ERR:

$$j_1 = \arg\max_{1 \le i \le m} \{\mathrm{ERR}_i^{(s=1)}\}, \tag{2.22}$$

  then insert ϕ_j as the first term of the model and remove it from Φ. The first associated orthogonal vector can be chosen as $q_{_i} = \phi_j$, and the first term of the auxiliary vector as $g_{_1} = g_j^{(s=1)}$;

- **Step** ($s \ge 2$): Suppose a subset Φ_{s-1}, represented by $(s-1)$ significant terms of the model $(\phi_1, \phi_2, \ldots, \phi_{s-1})$, is selected in step $(s-1)$. At each step, the selected terms are transformed into a new set of orthogonal bases $(q_1, q_2, \ldots, q_{_{(s-1)}})$ through the Gram-Schmidt orthogonalization procedure. In the s-th step, orthogonalize each of the remaining terms $(i \ne j_1, i \ne j_2, \ldots, i \ne j_{s-1})$ with the $(s-1)$-th selected term:

$$q_i^{(s)} = \phi_i - \sum_{r=1}^{s-1} \frac{\phi_m^{\mathsf{T}} q_r}{q_r^{\mathsf{T}} q_r} q_r, \qquad \phi_i \in \Phi - \Phi_{(s-1)}. \tag{2.23}$$

- Calculate $g_{_i}$ and ERR_i for the regressor ϕ_i. Compare the significance of ERR_i for each remaining term and select the one with the highest ERR_i, denoted as j_s (2.22), among the remaining terms of the model;

- The significant regressor to compose the model can be chosen as $\alpha_s = \phi_{j_s}$, and the s-th associated orthogonal basis as $q_{_s} = q_{j_s}^{(s)}$. Define $g_{_s} = g_{_{j_s}}$ and calculate the entries a_{rs} of the matrix A, as in (2.15). The subsequent significant bases α can be selected in the same manner, step by step. At each step, the most significant terms to represent the output y are selected:

$$y = \theta_1 \alpha_1 + \cdots + \theta_m \alpha_m + e. \tag{2.24}$$

- Calculate the sum of ERR_i as per (2.21). If the condition that the output variance of the model falls below a predetermined limit ε is met, the algorithm stops. Otherwise, increment s by 1 and repeat the steps until the necessary condition is satisfied;

- The final model is the linear combination (2.2) of m_0 (usually $m_0 \ll m$) significant terms selected from m candidate terms:

$$\hat{y}(k) = \sum_{i=1}^{m_0} \theta_{j_i} \phi_{j_i}(k) + e(k), \tag{2.25}$$

where the coefficient vector $\boldsymbol{\theta} = [\theta_{j_1}, \theta_{j_2}, \ldots, \theta_{j_{m_0}}]^{\mathsf{T}}$ is estimated by solving the triangular system (2.16) $A\boldsymbol{\theta} = g$, with $\boldsymbol{g} = [g_1, g_2, \ldots, g_{m_0}]$, and the matrix A (2.12) calculated in the previous steps.

A more detailed discussion of the OFR algorithm, including methods that address any numerical ill-conditioning during the search process, can be found in Billings et al. (1988); Chen et al. (1989). The OFR algorithm is widely applied in the identification of nonlinear systems. Therefore, various variants of the algorithm have been developed to meet specific requirements or enhance model performance (Wei and Billings, 2008; Wei et al., 2004; Billings, 2013).

3 LOGISTIC-NARX MULTINOMIAL MODEL

This chapter presents a novel approach that integrates logistic regression with the NARX framework and the One-Versus-All (OVA) decomposition strategy, enabling its application to multiclass classification problems. The Orthogonal Forward Regression (OFR) algorithm is adapted to guide the selection of model terms, incorporating a k-fold cross-validation-based accuracy metric. This metric is used both to assess the generalization capability of the model and to rank the importance of candidate regressors. The proposed method not only extends the use of NARX models beyond traditional regression and binary classification but also introduces a transparent and interpretable structure for multiclass scenarios. Its effectiveness is demonstrated through classical benchmark datasets, highlighting the potential of the approach for accurate and explainable classification tasks.

3.1 INTRODUCTION

Classification in machine learning and statistics is a supervised learning approach that recognizes, comprehends, and assigns observations to predefined categories using labeled datasets (Bishop, 2006). This task is fundamental in various fields such as finance, healthcare, and engineering, where the objective is to build models capable of categorizing data into multiple classes. Common applications include medical diagnostics, credit scoring, handwritten character recognition, speech recognition, and biological classification (Theodoridis and Koutroumbas, 2006).

To address such problems, a wide range of algorithms has been developed, including logistic regression (Hosmer et al., 2013), random forest (Breiman, 2001), support vector machines (Cristianini and Shawe-Taylor, 2000), and k-nearest neighbors (Kuhn and Johnson, 2013). Although these methods often achieve high accuracy, they generally operate as black boxes, making it difficult to interpret how input variables influence the classification decisions. In contrast, the logistic-NARX approach proposed by Ayala Solares et al. (2019) offers a more transparent alternative, using a linear-in-the-parameters structure that provides interpretability and facilitates the analysis of variable contributions in binary classification problems.

This method has additional advantages, such as the direct inclusion of lag terms, which allows for modeling dynamic behavior more naturally. Moreover, the approach mitigates multicollinearity by orthogonalizing candidate terms during the model selection process, enhancing the robustness and interpretability of the resulting models. However, the original method is limited to binary classification, restricting its applicability to more complex scenarios with multiple output categories.

To overcome this limitation, and inspired by the logistic-NARX model for binary problems (Ayala Solares et al., 2019), this chapter proposes an extended formulation

aimed at solving multiclass classification tasks. The proposed method integrates machine learning techniques with system identification concepts, enabling the construction of interpretable models for multinomial problems. Rather than emphasizing dimensionality reduction or transformation of the input space, this approach focuses on understanding the role and relevance of each input variable within the model structure (Cai et al., 2018). This interpretative capability, often lacking in more complex classifiers, supports decision-making in real-world applications where both accuracy and transparency are essential.

3.2 LOGISTIC-NARX MULTINOMIAL MODEL APPROACH

This work proposes a hybrid multinomial classification framework that simultaneously performs feature extraction and selection. Unlike most NARX models applications, which focus on regression problems with continuous outputs, our method adapts the NARX paradigm to handle multiclass categorization. Specifically, we integrate the transparent and parsimonious structure of NARX with the probabilistic outputs of logistic regression. In doing so, we leverage the ability of NARX to mitigate multicollinearity and produce interpretable nonlinear regressors, while logistic regression transforms these regressors into class membership probabilities bounded in $[0, 1]$.

In logistic regression, each predicted value is constrained to the unit interval by the logistic (sigmoid) function:

$$f(x) \;=\; \frac{1}{1+\exp(-x)}, \quad x \in \mathbb{R}, \quad f(x) \in (0,1). \tag{3.1}$$

An inherent challenge in logistic modeling is *multicollinearity*, where two or more regressors are nearly linearly dependent. High correlation among predictors can make parameter estimates unstable and reduce the reliability of inference. Within the classical NARX identification pipeline, structure selection is based on computing the Error Reduction Ratio (ERR, see (2.20)) to rank candidate terms. However, when the target variable is categorical (multinomial), ERR is no longer directly applicable, since it presumes a continuous residual variance criterion.

To overcome this, we replace the ERR metric with a *classification-accuracy* criterion calculated using logistic regression. Concretely, each candidate regressor w_i (an orthonormalized version of a NARX term $\phi_i(\varphi(k))$) is tested as the sole predictor in a logistic model. We measure the cross-validated accuracy $r(w_i, y)$ of that one-term logistic model when predicting the class labels y. A high accuracy value indicates that w_i is strongly associated with the categorical outcome, analogous to the way in which a large ERR reflects a strong linear association in regression analysis. Formally, at each selection step, the variable is chosen according to:

$$j \;=\; \arg\max_{1\le i\le m}\Big\{\, r(w_i,\, y)\Big\}, \tag{3.2}$$

and include the corresponding original regressor $\phi_j(\varphi(k))$ in the NARX model. Here, m is the total number of remaining candidate terms, and $r(w_i, y)$ denotes the k-fold cross-validated accuracy of a logistic regression that uses only w_i to predict y.

After selecting a subset of $n_{\max}$ NARX terms $\{\phi_{j_1}, \ldots, \phi_{j_{n_{\max}}}\}$, we assemble the final probability model by plugging their linear combination into the logistic function:

$$p(k \mid x) \;=\; \frac{1}{1+\exp\Big(-\sum_{i=1}^{n_{\max}} \theta_{j_i}\, \phi_{j_i}\Big(\varphi(k)\Big)\Big)}, \tag{3.3}$$

here, $\varphi(k)$ is the original NARX regressor vector formed from past outputs and inputs (as in (2.2)), and $\{\theta_{j_i}\}$ are the parameters estimated by maximizing the multinomial logistic likelihood.

Since multiclass classification is more complex than a single binary split, we adopt a One-Versus-All (OVA) decomposition: for each class $v = 1, \ldots, C$, we fit a separate logistic-NARX model that discriminates "class v versus all other classes". Denote the probability output of the v-th binary logistic-NARX as $f_v(x)$. Then an input x is assigned to the class whose classifier yields the highest probability:

$$x \in w_v \iff v \;=\; \arg\max_{1\le r\le C}\; f_r(x). \tag{3.4}$$

Thus, each of the C binary logistic models selects its own most predictive NARX terms (as in (3.2) and the Orthogonal Forward Regression Algorithm 1), and at test time the final class label is determined by (3.4). In summary, our approach (Algorithm 2) combines:

- NARX structure search: orthogonalize all candidate regressors $\{\phi_i(\varphi(k))\}$ to obtain orthonormal bases $\{w_i\}$, then rank them by cross-validated logistic regression accuracy $r(w_i, y)$, which serves as a *relevance score* reflecting each term's contribution to classification performance;
- Logistic-NARX fitting: after selecting the top $n_{\max}$ regressors, estimate the parameters θ_{j_i} by fitting a logistic model on the combined linear predictor $\sum_i \theta_{j_i}\, \phi_{j_i}(\varphi(k))$;
- One-Versus-All decomposition: repeat the above for each class v to build C separate logistic-NARX classifiers, then use (3.4) to make a final multiclass decision.

By explicitly measuring *accuracy* at each term-selection step, the method both mitigates multicollinearity (via orthogonalization) and maintains interpretability: each

Algorithm 2: Logistic-NARX Multinomial Model

Input: $\{(y(k)), k = 1, \ldots, N\}$, $\mathcal{M} = \{\phi_i, i = 1, \ldots, m\}$, l, n_y, n_u, n_{max}
Output: $\boldsymbol{\alpha} = \{\alpha_i, i = 1, \ldots, n_{max}\}$, $\boldsymbol{\theta} = \{\theta_i, i = 1, \ldots, n_{max}\}$

1 **for** $i = 1 : m$ **do**
2 $\quad w_i \leftarrow \frac{\phi_i}{\|\phi_i\|_2}$
3 $\quad r_i \leftarrow$ Logistic regression accuracy in w_i and y
4 **end**
5 $j \leftarrow \arg\max_{1 \le i \le m} \{r(w_i, y)\}$ ▷ Equation (3.2)
6 $q_1 \leftarrow w_j$
7 $\alpha_1 \leftarrow \phi_j$
8 Train logistic model with α_1 and y using *One-Versus-All* ▷ Equation (3.4)
9 Compute cross-validation
10 Remove ϕ_j from $\mathcal{M}$
11 **for** $s = 2 : k$ **do**
12 $\quad$ **for** $i = 1 : m$ **do**
13 $\quad\quad w_i^{(s)} \leftarrow$ Orthogonalize ϕ_i in $[q_1, \ldots, q_{(s-1)}]$ ▷ Equation (2.23)
14 $\quad\quad$ **if** $w_i^{\mathsf{T}} w_i < 10^{-10}$ **then**
15 $\quad\quad\quad$ Remove ϕ_i from $\mathcal{M}$
16 $\quad\quad\quad$ Next iteration
17 $\quad\quad$ **end**
18 $\quad\quad r_i \leftarrow$ Logistic regression accuracy in w_i and y
19 $\quad$ **end**
20 $\quad j \leftarrow \max_{1 \le i \le m-s+1} \{r^{(i)}(w_i, y)\}$
21 $\quad q_s \leftarrow w_j$
22 $\quad \alpha_s \leftarrow \phi_j$
23 $\quad$ Remove ϕ_j from $\mathcal{M}$
24 $\quad \boldsymbol{\alpha} \leftarrow [\alpha_1, \ldots, \alpha_{(s)}]$
25 $\quad$ Train the logistics model with $\boldsymbol{\alpha}$ and y using *One-Versus-All*
26 $\quad$ Compute cross-validation
27 **end**
28 $\boldsymbol{\alpha} \leftarrow [\alpha_1, \ldots, \alpha_{(n_{max})}]$ ▷ matrix of selected terms
29 $\boldsymbol{\theta} \leftarrow [\theta_1, \ldots, \theta_{(n_{max})}]$ ▷ vector of estimated coefficients

selected ϕ_{j_i} enters the final logit as a transparent regressor, and its coefficient θ_{j_i} directly quantifies how that NARX term influences the log-odds of class membership. The implementation of the method is presented in Algorithm 2, and its main steps are described below:

- The inputs of Algorithm 2 include the vector $y(k)$ containing the class labels, the matrix $\mathcal{M}$ of candidate regressors ϕ_i (derived from combinations of input features), the maximum number of terms to be selected n_{max}, and the NARX model parameters (l, n_y, n_u) defined in Equation (2.1);

- Lines (1-5) identify the candidate terms ϕ_j with the highest discriminatory power,

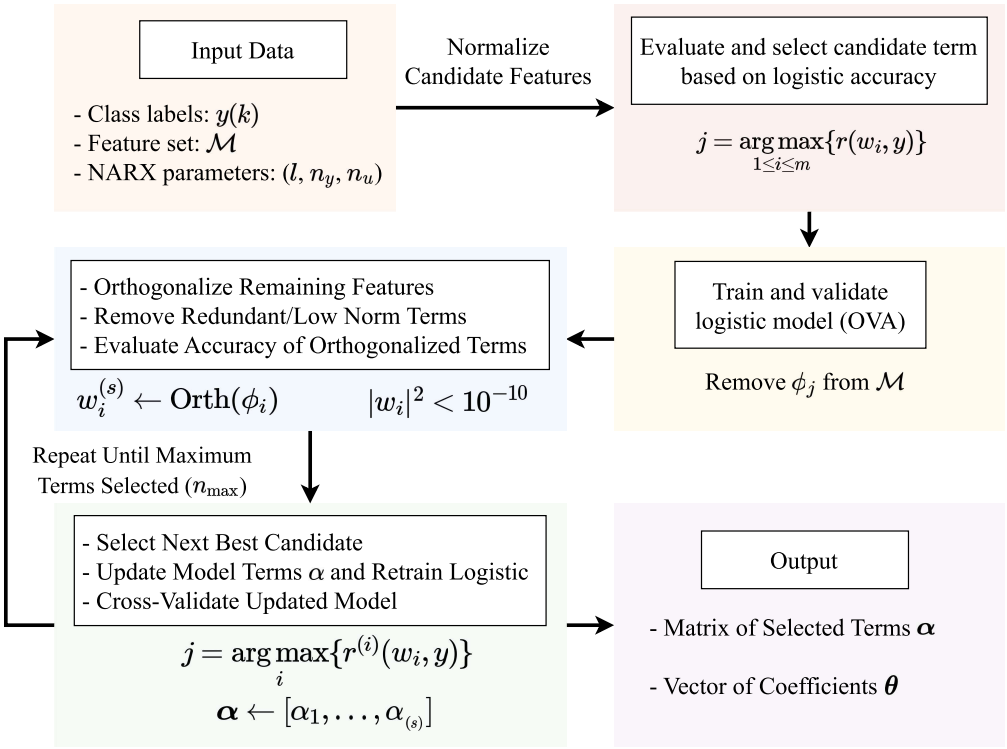
based on the logistic regression accuracy using each normalized candidate regressor;

- In lines (6-9), the logistic regression model is trained using the selected term α_1 and the label vector $y(k)$, applying the One-Versus-All (OVA) strategy and evaluating performance via cross-validation;
- From lines (12-27), the remaining candidate terms are orthogonalized with respect to the already selected ones using the modified Gram-Schmidt procedure. Each orthogonalized term is then evaluated using the logistic regression accuracy, and the most informative term is selected at each iteration. This process continues until n_{max} terms are selected;
- Lines (14-17) implement a filtering step based on the squared 2-norm (Euclidean norm) of each orthogonalized term (Wei and Billings, 2008). Terms with negligible norm are removed to avoid redundancy and prevent multicollinearity in the model;
- Finally, in lines (28-29), the matrix $\boldsymbol{\alpha}$ containing the selected regressors and the vector $\boldsymbol{\theta}$ of estimated coefficients are computed. Since the optimal number of terms is not known beforehand, the parameter n_{max} is usually chosen heuristically by analyzing the evolution of the model accuracy.

This algorithm tackles multiclass classification tasks by combining the interpretability of NARX models with the classification power of logistic regression. The NARX structure naturally incorporates lagged variables and their interactions, offering a transparent model structure that facilitates understanding of how input variables influence the classification decision. Figure 1 illustrates the main steps of the Logistic-NARX Multinomial algorithm. The method iteratively selects the most relevant terms based on logistic accuracy, applies orthogonalization to reduce redundancy, and builds a compact, interpretable model until the limit n_{max} is reached.

The computational complexity of the Logistic-NARX Multinomial algorithm is mainly determined by four components: (i) evaluation of feature relevance, (ii) training of the logistic regression model, (iii) orthogonalization of candidate terms, and (iv) the One-Versus-All (OVA) decomposition for multiclass problems. The feature relevance evaluation has linear complexity, $O(NM)$, where N is the number of samples and M is the number of features (Ayala Solares et al., 2019). Training the logistic regression model has a worst-case complexity of $O(M^3 + NM)$ (Komarek, 2004). Orthogonalization requires $O(N(M-1))$ (Senawi et al., 2017). The use of OVA adds an extra factor K (the number of classes), leading to an overall complexity of $O(K(M^3 + NM))$. By combining the interpretability of linear models with the accuracy of more advanced methods, such as Support Vector Machines (SVM), the proposed approach proves particularly valuable in contexts where both explainability and predictive performance are important, as illustrated in Figure 2.

Figure 1 – Flowchart illustrating the main stages of the Logistic-NARX Multinomial algorithm, including input processing, model structure identification, term selection, and final classification output.



Source: created by the author. (2024).

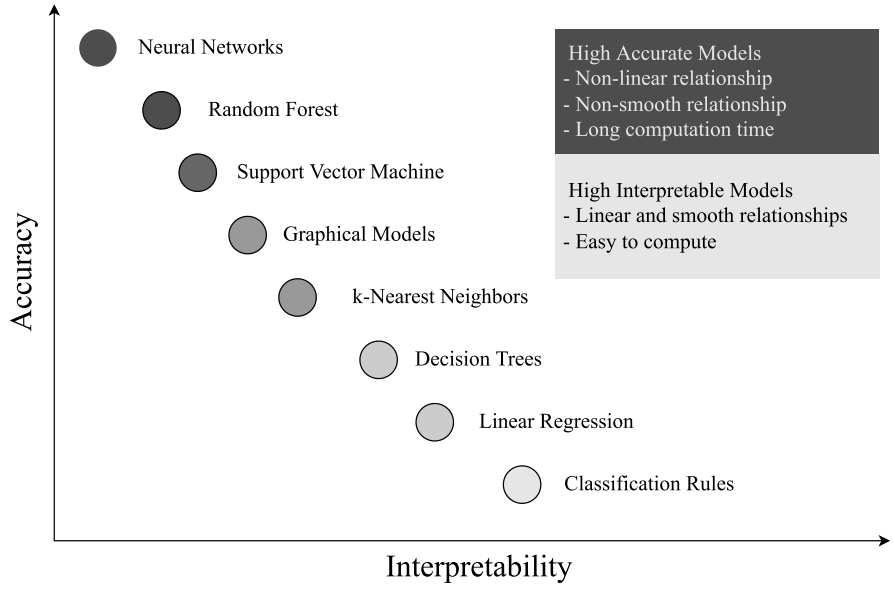
3.3 CASE STUDY SIMULATION

In this section, simulations are performed to evaluate the classification accuracy and model compactness achieved by the proposed methodology. The assessment is conducted using well-known benchmark datasets commonly employed in the machine learning literature. The objective is to analyze the effectiveness of the proposed Logistic-NARX Multinomial approach in comparison to widely used classification algorithms. Specifically, the method is tested on four multivariate datasets sourced from the UCI Machine Learning Repository[1], which represent diverse and realistic classification scenarios.

The datasets were selected to ensure a representative variety in terms of number of features, sample size, number of classes, and the presence of imbalanced or noisy data. The Iris dataset serves as a classic benchmark with relatively simple class boundaries. The Wine dataset introduces a greater number of features and an imbalanced class distribution. The Glass dataset includes six classes, with some minority classes and outliers, making it more challenging. Lastly, the Wave dataset features a higher-dimensional input space

[1] UCI Machine Learning Repository: https://archive.ics.uci.edu/ml.

Figure 2 – Comparative analysis of different classification techniques, highlighting their performance in terms of accuracy and model interpretability.



Source: created by the author. (2023).

and noisy observations. A summary of the characteristics of these datasets is provided in Table 1.

For each dataset, the performance of the proposed algorithm is compared against standard classification techniques, including Random Forest (RF) (Breiman, 2001), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), and k-Nearest Neighbors (KNN) (Cover and Hart, 1967). Evaluation is based on established performance metrics, enabling a comprehensive comparison of accuracy and model behavior.

Table 1 – Summary of the main characteristics of the selected datasets, including size, number of classes, and feature dimensionality.

Datasets	Classes	Features	Samples
Iris	3	4	150
Wine	3	13	178
Glass	6	9	214
Wave	3	40	5000

The algorithms evaluated in this section were implemented using the MATLAB programming environment and executed on a computer equipped with an Intel Core i5 processor (2.5 GHz) and 8 GB of RAM. To ensure consistency and fairness in the comparisons, all methods were evaluated using 5-fold cross-validation. The input features of each dataset were normalized, and identical preprocessing procedures were applied

across all classifiers.

The hyperparameters for each classifier were selected based on empirical testing with each dataset, choosing configurations that yielded the best cross-validation performance. The adopted settings for Random Forest (RF), Support Vector Machines (SVM), and k-Nearest Neighbors (KNN) for each dataset are as follows:

- *Iris* – RF with Gini's diversity index and a maximum of 5 splits; SVM with a polynomial kernel of degree 2; KNN using Minkowski distance (exponent 3) and $k = 10$ neighbors;
- *Wine* – RF with Gini's index and up to 20 splits; SVM with a polynomial kernel of degree 2; KNN with Euclidean distance and $k = 10$;
- *Glass* – RF with Gini's index and up to 100 splits; SVM with a polynomial kernel of degree 3; KNN with Euclidean distance and $k = 10$;
- *Wave* – RF with Gini's index and up to 20 splits; SVM with Gaussian (RBF) kernel; KNN with Euclidean distance and $k = 10$.

3.3.1 Model Selection with Cross-Validation and Statistical Tests

In the Logistic-NARX Multinomial framework, model complexity depends on the number of terms in the final structure. To balance accuracy and interpretability, this thesis adopts a clear and rigorous method to select the optimal number of terms, based on the following criteria:

- *Cross-validated accuracy*: The primary performance metric is the mean accuracy from k-fold cross-validation, ensuring generalizability to unseen data;
- *Statistical significance testing (paired t-tests)*: Differences in performance between models with varying numbers of terms are assessed using paired t-tests on fold-wise accuracy values. A significance level of $\alpha = 0.05$ is used (Fisher, 1992; Wasserman, 2013), representing the maximum acceptable probability of incorrectly rejecting the null hypothesis. P-values below this threshold indicate statistically significant differences; otherwise, models are considered statistically equivalent;
- *Error bar overlap*: Accuracy plots with standard deviation error bars are used to visually assess stability. Substantial overlap suggests non-significant performance differences that should be interpreted cautiously;
- *Model complexity*: When multiple configurations are statistically equivalent, the model with the fewest terms is preferred, following the principle of parsimony;

- *Stability (low variance)*: Among statistically equivalent models, the one with the smallest standard deviation across folds is favored, as it indicates more consistent performance.

The evaluation procedure consists of:

1. For each candidate number of terms, compute the cross-validated mean accuracy and standard deviation. A wide range of term counts is tested to ensure that adding more terms does not yield statistically significant improvements. The results tables in subsequent sections report only up to the optimal number of terms identified;
2. Visualize performance using accuracy curves with error bars;
3. Perform paired t-tests between the highest-performing configuration and all others;
4. In cases of statistical equivalence, select the model with:
   - Fewer terms (lower complexity);
   - Smaller standard deviation (greater stability).

This analysis framework is applied to all case studies in the following chapters. Each application includes a dedicated section presenting the accuracy results, statistical test outcomes, and the rationale for selecting the optimal number of terms for that dataset.

3.3.2 Static Multiclass System

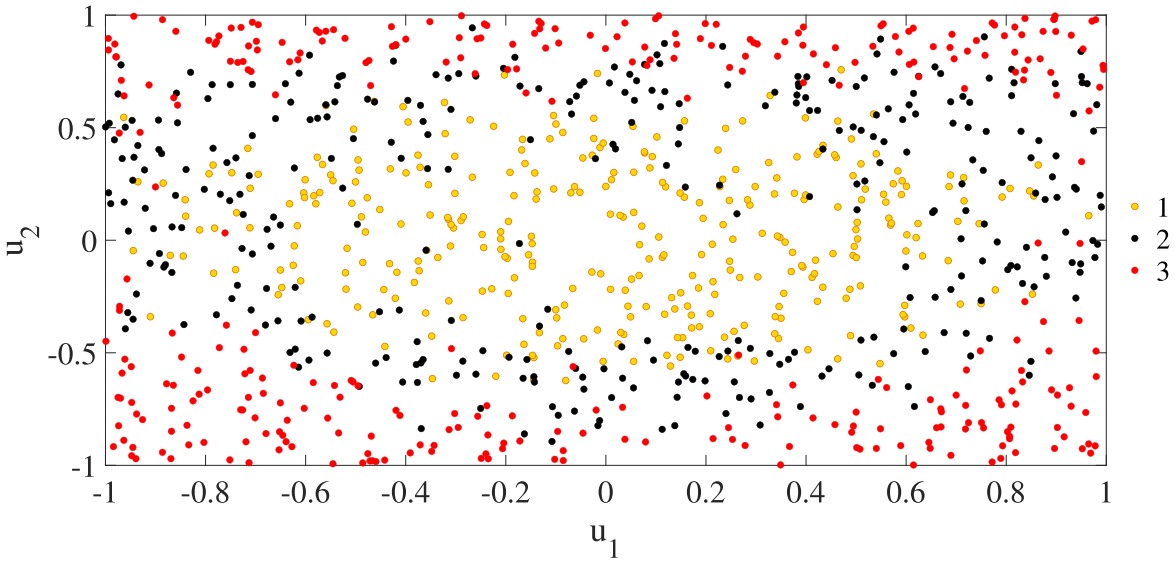
In this first example, the objective is to evaluate the algorithm's ability to correctly identify all model terms that define the decision boundary in a regression-like NARX framework. Consider the following input–output system: a latent score is computed from the inputs, and the class label is then obtained by applying a threshold to this common score:

$$\begin{aligned} s[k] &= u_1^2[k] + 2\,u_2^2[k] - 0.8\,u_1^2[k]\,u_2[k] + e[k], \\ y[k] &= \begin{cases} 1, & s[k] < \tau_1, \\ 2, & \tau_1 \le s[k] < \tau_2, \\ 3, & s[k] \ge \tau_2, \end{cases} \end{aligned} \tag{3.5}$$

inputs are independently and identically distributed (i.i.d.) following a uniform distribution over $[-1, 1]$, i.e., $u_i[k] \sim \mathcal{U}(-1, 1)$ for $i = 1, \ldots, 4$, and the measurement noise follows $e[k] \sim \mathcal{N}(0, , 0.3^2)$. Only u_1 and u_2 are relevant to the data-generating process, while u_3 and u_4 act as distractor variables. To obtain approximately balanced classes, (τ_1, τ_2) are set to the empirical quantiles $(Q_{1/3}, Q_{2/3})$ of $s[k]$ computed from the training set.

A total of 1,000 input–output samples are generated, as illustrated in Figure 3. Since this is a static problem, no time lags are considered, and the nonlinearity degree is fixed at $l = 3$, so that the candidate dictionary includes a constant term and all polynomial combinations of the four inputs at time k up to degree 3. The maximum model size is set to $n_{\max} = 10$, and performance is assessed using 5-fold cross-validation during the term selection phase.

Figure 3 – Scatter plot of the relevant input variables u_1 and u_2 for the generated static multiclass system dataset.



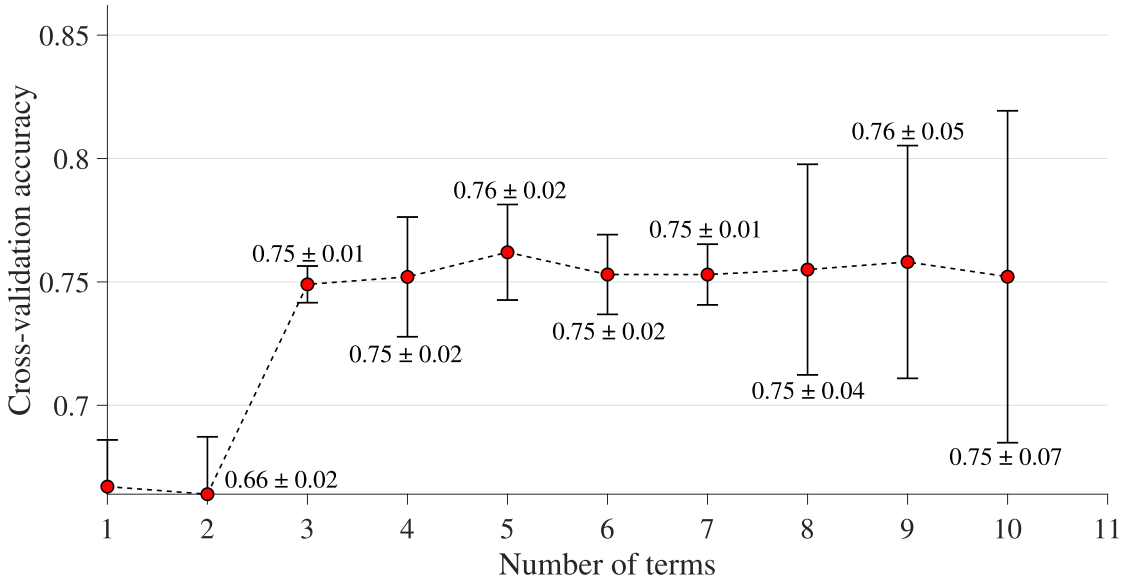
Source: created by the author. (2025).

Figure 4 shows that the mean cross-validation accuracy increased up to the 5-term model, which achieved the highest average performance (0.7620 ± 0.0193). Paired t-tests, reported in Table 2, comparing the 5-term model with the others revealed that only the 1-term ($p = 0.0025$) and 2-term ($p = 0.0014$) models exhibited statistically significant differences. All other configurations (3–10 terms) yielded p-values above 0.05, indicating statistical equivalence in performance. Considering parsimony and stability criteria, the 5-term model was selected as the optimal configuration, as it offers a balanced trade-off between high accuracy, low variability, and moderate complexity.

Table 3 presents the identified NARX model $\hat{y}(k)$ for the selected configuration, including the chosen terms, their estimated parameters, and relevance scores derived from cross-validated logistic regression accuracy. These scores quantify the individual contribution of each term to the overall model performance during the selection process. The results indicate that the algorithm correctly identified all model terms involved in the decision boundary of Equation (3.5). The estimated parameters correspond to log-odds

ratios and, therefore, are not expected to match the coefficients of the original decision boundary function directly.

Figure 4 – Average accuracy observed during the model term selection phase (static multiclass system).



Source: created by the author. (2023).

Table 2 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 5-term model (static multiclass system).

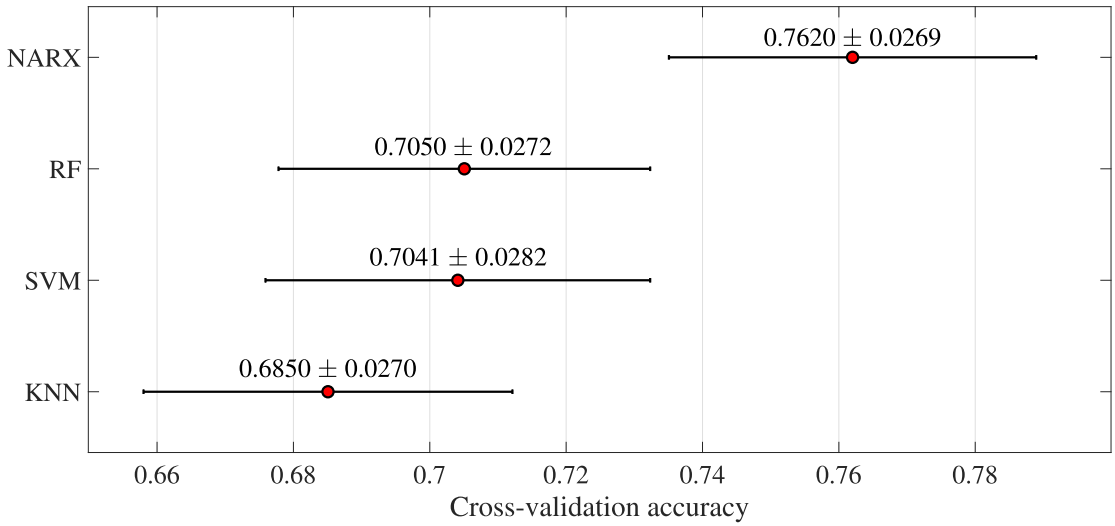
# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
1	0.6671	0.0189	0.0025	6	0.7530	0.0162	0.4866
2	0.6640	0.0233	0.0014	7	0.7530	0.0123	0.5555
3	0.7490	0.0074	0.2873	8	0.7550	0.0426	0.5914
4	0.7520	0.0243	0.5817	9	0.7581	0.0472	0.8837
5	0.7620	0.0193	—	10	0.7521	0.0672	0.7831

A comparative analysis was conducted to benchmark the proposed method against traditional classifiers. As shown in Figure 5, the Logistic-NARX Multinomial model achieved the highest average accuracy of 76.20 %, demonstrating superior cross-validation performance compared to standard techniques. To provide a more comprehensive evaluation, multiple metrics were considered. The results in Table 4 show that the proposed method consistently outperformed the compared techniques across all evaluated metrics.

Table 3 – Identified NARX model $\hat{y}(k)$ showing the top 5 selected terms, their estimated parameters, and corresponding relevance scores (correlation values) used during model construction (static multiclass system).

Model Terms	Parameter	Score	Model Terms	Parameter	Score
$u_2(k)u_2(k)$	0.2383	0.8134	constant	0.1690	0.6630
$u_1(k)u_1(k)$	0.0954	0.4550	$u_1(k)u_1(k)u_2(k)$	-0.0260	0.4491
$u_2(k)u_2(k)u_2(k)$	0.0205	0.4290			

Figure 5 – Comparative analysis of classification accuracy for the static multiclass system using different methods. Accuracy was evaluated through cross-validation with interval measurements and average performance indicated.



Source: created by the author. (2023).

Table 4 – Performance comparison of various classification methods based on multiple evaluation metrics (static multiclass system).

	L-NARX M	RF	SVM	KNN
Average Accuracy	0.7620	0.7050	0.7040	0.6850
Sensitivity	0.7575	0.7079	0.7109	0.6748
Specificity	0.8883	0.8537	0.8510	0.8494
Precision	0.7623	0.7052	0.7041	0.6853
F1 Score	0.7514	0.7043	0.7065	0.6738

3.3.3 Iris Dataset

The Fisher's Iris dataset, introduced by Ronald Fisher in 1936 (Fisher, 1936), is a classic multivariate dataset widely used for testing classification algorithms (Soni and Patel, 2017). It contains 150 balanced samples from three iris species (Setosa, Versicolor, and Virginica), with four numeric features per sample: sepal length, sepal width, petal length, and petal width, all measured in centimeters. Fisher originally used these variables to construct a linear discriminant model. The dataset is publicly available through the UCI Machine Learning Repository[2] and is summarized in Table 5.

Table 5 – Description of Fisher's Iris dataset, detailing the features (in centimeters) and class labels for each sample.

	Features (cm)				Classes
Samples	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...	...	...	...	...	...
50	6.4	3.5	4.5	1.2	Versicolor
...	...	...	...	...	...
150	5.9	3.0	5.0	1.8	Virginica

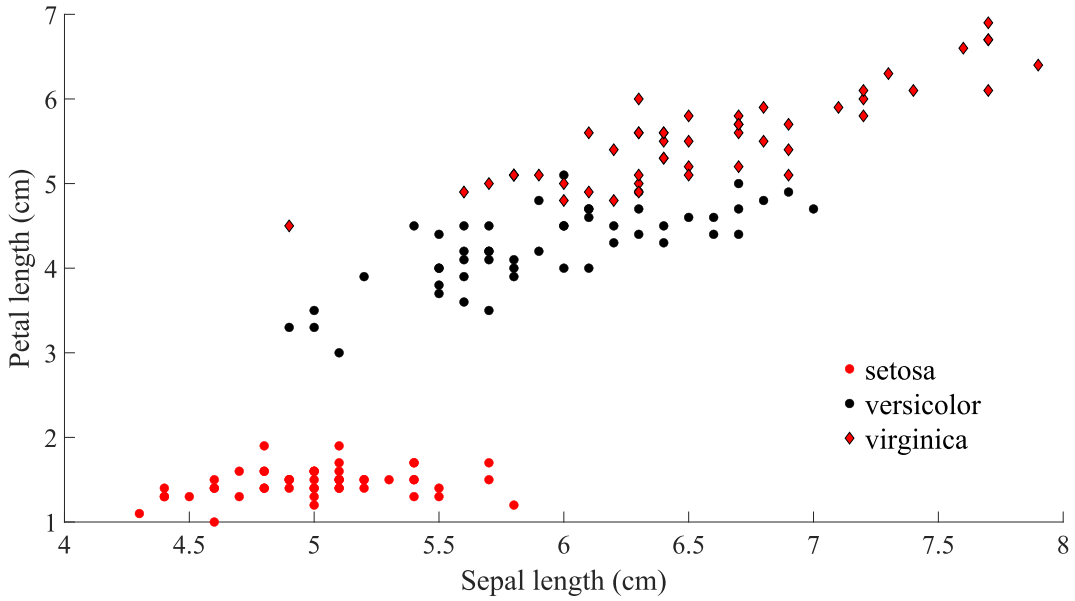
To evaluate the performance of the proposed classification method, experiments were conducted using Fisher's Iris dataset. This benchmark dataset allows for the analysis of model behavior in a controlled and well-understood setting. Figure 6 presents a bivariate comparison of petal and sepal lengths across the three iris species. From this visual analysis, it is evident that *Iris setosa* presents shorter petal and sepal lengths, *versicolor* shows intermediate values, and *virginica* is characterized by longer measurements.

The proposed algorithm was applied using a nonlinearity degree of $l = 2$ and a maximum of $n_{max} = 10$ selected terms, as defined in Equation (2.2). Since this dataset represents a static classification problem without temporal dependencies, time delays were not included in the NARX formulation. Consequently, the resulting search space comprised 15 candidate terms based solely on the original features.

Performance was evaluated using 5-fold cross-validation, ensuring consistent and robust estimation of the classifier's generalization capability. Figure 7 illustrates the accuracy obtained during the iterative term selection process. This step is essential

[2] Fisher's Iris dataset (UCI Machine Learning Repository): https://archive.ics.uci.edu/ml/datasets/iris

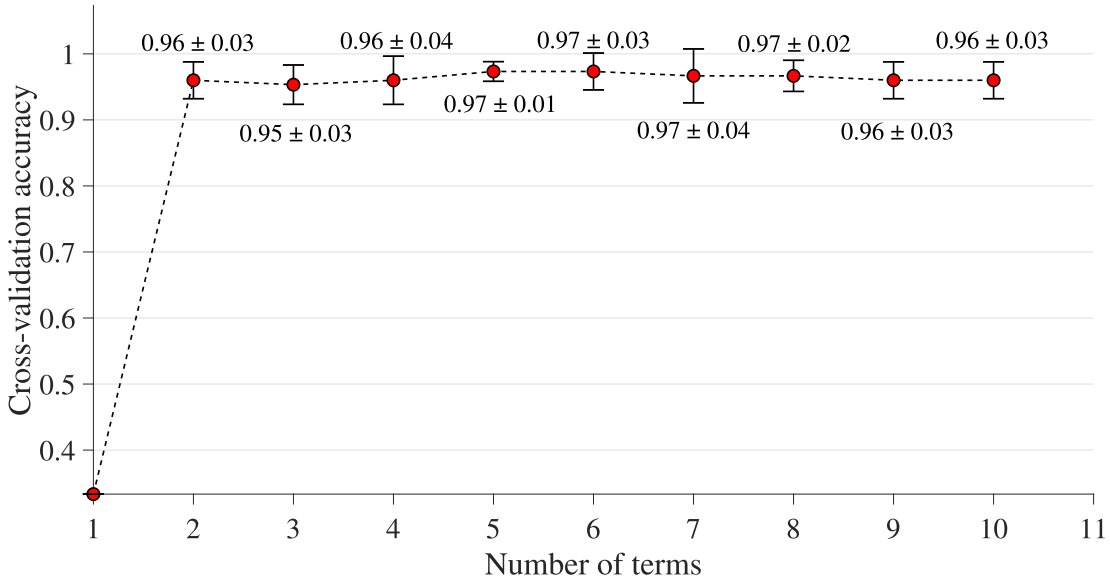
Figure 6 – Bivariate analysis comparing petal length and sepal length across the three Iris species, highlighting inter-class separability.



Source: created by the author. (2023).

for identifying the most relevant terms and determining an appropriate model size that balances accuracy and interpretability.

Figure 7 – Average accuracy observed during the model term selection phase for the Fisher's Iris dataset, indicating the optimal number of terms.



Source: created by the author. (2023).

The average accuracy curve (Figure 7) shows that adding terms beyond a certain

point provides negligible improvements in predictive performance. The configuration with five terms achieved the highest mean cross-validation accuracy (0.9734 ± 0.0149) and the lowest variance among all tested models, indicating both strong performance and stability. To confirm these findings, paired t-tests were performed between the 5-term model and each alternative configuration (2–10 terms). As shown in Table 6, all p-values were greater than 0.05, confirming that the observed performance differences were not statistically significant.

Given the statistical equivalence among configurations, the final selection followed the principles of parsimony and performance stability. The 5-term model was chosen as the optimal configuration, offering a balance of high accuracy, low variance, and reduced complexity. Notably, even the more compact 2-term model (0.9600 ± 0.0279) delivered competitive results, underscoring the robustness and generalization capability of the modeling approach.

Table 7 presents the identified NARX model $\hat{y}(k)$ for the selected configuration, including the chosen terms, their estimated parameters, and relevance scores derived from cross-validated logistic regression accuracy as defined in Equation (3.2). These scores quantify the individual contribution of each term to the overall model performance during the selection process.

Table 6 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 5-term model.

# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
2	0.9600	0.0279	0.4759	7	0.9667	0.0365	0.6574
3	0.9534	0.0298	0.2081	8	0.9600	0.0279	0.4759
4	0.9600	0.0365	0.4764	9	0.9667	0.0279	0.6574
5	0.9734	0.0149	-	10	0.9533	0.0298	0.2081
6	0.9733	0.0279	0.9986				

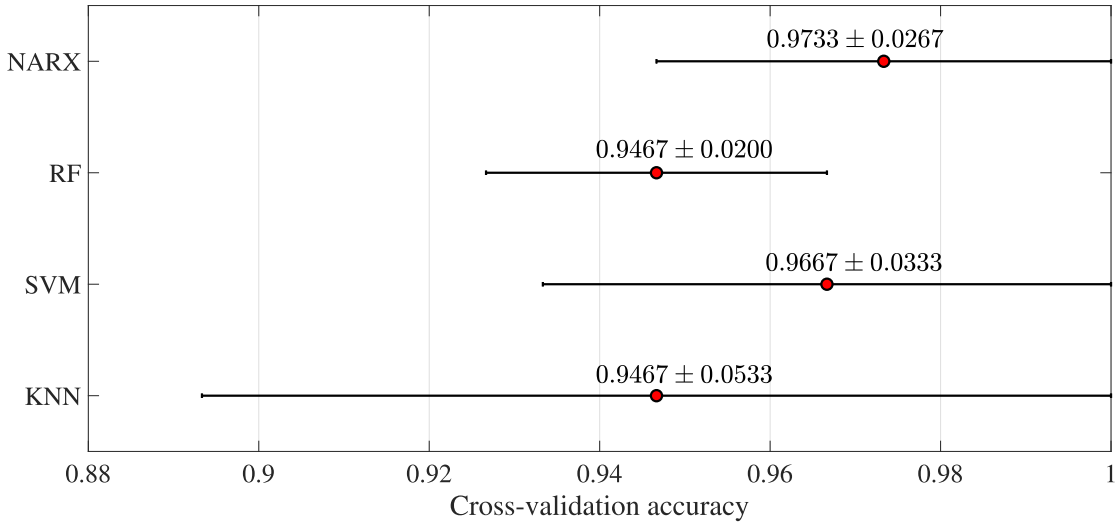
Table 7 – Identified NARX model $\hat{y}(k)$ for the static multiclass system, showing the selected terms, their estimated parameters, and corresponding relevance scores.

Model Terms	Parameter	Score	Model Terms	Parameter	Score
$u_3(k)$	0.1322	0.9533	Constant	0.4386	0.7714
$u_4(k)u_4(k)$	0.0218	0.5933	$u_3(k)u_3(k)$	0.0572	0.5067
$u_2(k)u_4(k)$	-0.0368	0.4667	$u_1(k)u_4(k)$	0.0035	0.4200

A comparative analysis was also conducted to benchmark the proposed method against traditional classifiers. As shown in Figure 8, the Logistic-NARX Multinomial model

achieved the highest average accuracy of 97.33 %, demonstrating superior performance under cross-validation when compared to standard techniques.

Figure 8 – Comparative analysis of classification accuracy for the Fisher's Iris dataset using different methods. Accuracy was evaluated through cross-validation with interval measurements and average performance indicated.



Source: created by the author. (2023).

Table 8 – The confusion matrix displays the performance of classification methods by aggregating partitions from validation sets created through cross-validation. It includes three classes: Iris Setosa (C1), Iris Virginica (C2), and Iris Versicolor (C3).

	C1	C2	C3
C1	50		
C2		48	2
C3		2	48

(a) Logist-NARX Multiclass.

	C1	C2	C3
C1	50		
C2		45	3
C3		5	47

(b) Random Forests.

	C1	C2	C3
C1	50		
C2		47	2
C3		3	48

(c) Support Vector Machine.

	C1	C2	C3
C1	50		
C2		48	6
C3		2	44

(d) K-Nearest Neighbors.

An effective way to visualize and assess classification performance is through the confusion matrix, which summarizes the number of correct and incorrect predictions by

comparing predicted and actual class labels. Based on the confusion matrix (see Table 8), several evaluation metrics can be derived, as presented in Table 9. Although accuracy is a commonly used metric, it may be insufficient in scenarios with imbalanced classes. For this reason, multiple metrics are considered to provide a more comprehensive evaluation. The results in Table 9 show that the proposed method consistently outperforms the compared techniques across all evaluated metrics.

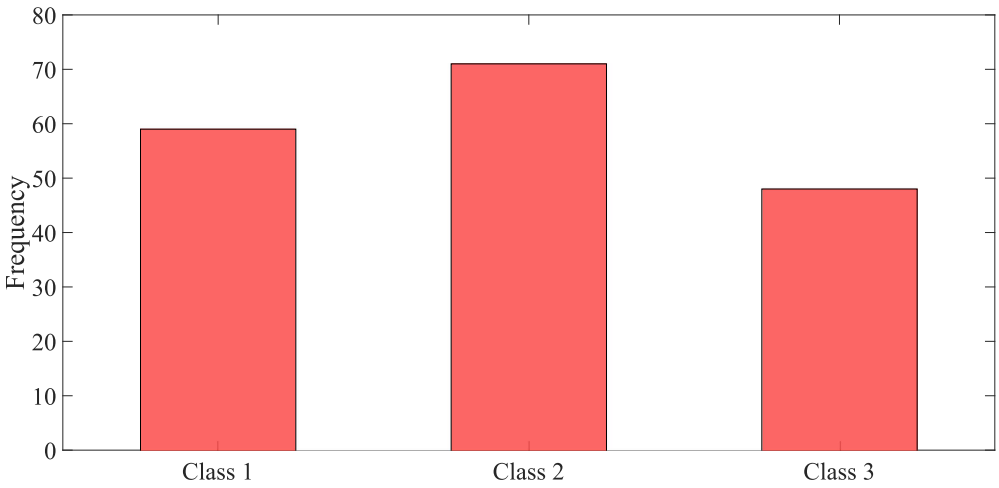
Table 9 – Performance comparison of various classification methods applied to the Fisher's Iris dataset, based on multiple evaluation metrics derived from the respective confusion matrices.

	L-NARX M	RF	SVM	KNN
Average Accuracy	0.9733	0.9467	0.9667	0.9467
Sensitivity	0.9735	0.9471	0.9668	0.9485
Specificity	0.9867	0.9735	0.9834	0.9738
Precision	0.9732	0.9467	0.9665	0.9467
F1 Score	0.9731	0.9466	0.9661	0.9466

3.3.4 Wine Dataset

The Wine dataset, originally compiled by Forina et al. (1990), contains chemical measurements of 178 wine samples from three different cultivars grown in the same region of Italy. It includes 13 features such as alcohol content and color intensity, which serve to classify the wines by cultivar. The objective is to predict the wine class based on these physicochemical attributes. This benchmark dataset is widely used in classification studies and is publicly available in the *UCI Machine Learning Repository*[3].

Figure 9 – Class frequency distribution of the Wine dataset, illustrating the number of samples per class.



Source: created by the author. (2023).

The class distribution of the Wine dataset (see Figure 9) shows a moderate imbalance among categories. In this experiment, the proposed method was configured with a nonlinearity degree of $l = 2$ and a maximum of $n_{\max} = 10$ selected terms, producing a search space of 105 candidates. Since the dataset represents a static classification problem, where each sample is described by fixed chemical descriptors, no time delays were included in the NARX formulation.

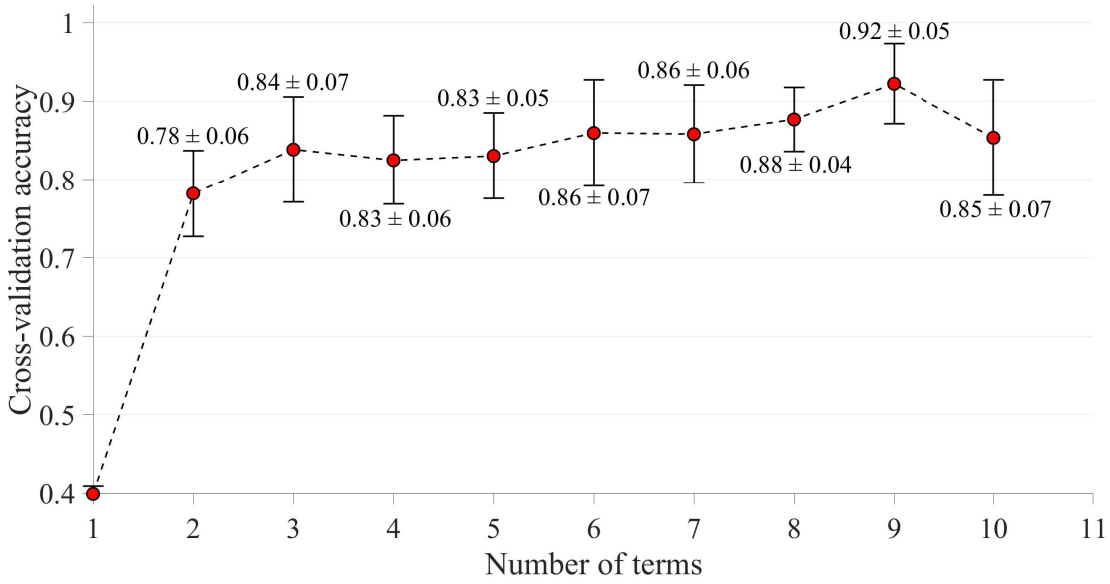
Results of cross-validation for models containing 1 to 10 terms are presented in Figure 10. The highest mean accuracy (0.9205 ± 0.0456) was obtained with 9 terms (see table 10). Statistical analysis using paired two-sided t-tests ($\alpha = 0.05$) indicated significant differences for most simpler configurations ($p < 0.05$ for 2 and 3 terms), while models with 4, 5, and 6 terms performed similarly to the 9-term model ($p > 0.05$).

Overall, although the 9-term configuration achieved the best accuracy, some lower-complexity models also delivered competitive results. For instance, the 6-term model

[3] Wine dataset in the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/wine.

(0.8598 ± 0.0676) provides comparable performance with advantages in interpretability and computational efficiency.

Figure 10 – Average accuracy computed during the model's term selection phase for the Wine dataset, used to determine the optimal number of terms for constructing the final model with maximum classification performance.



Source: created by the author. (2023).

Table 10 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 9-term model.

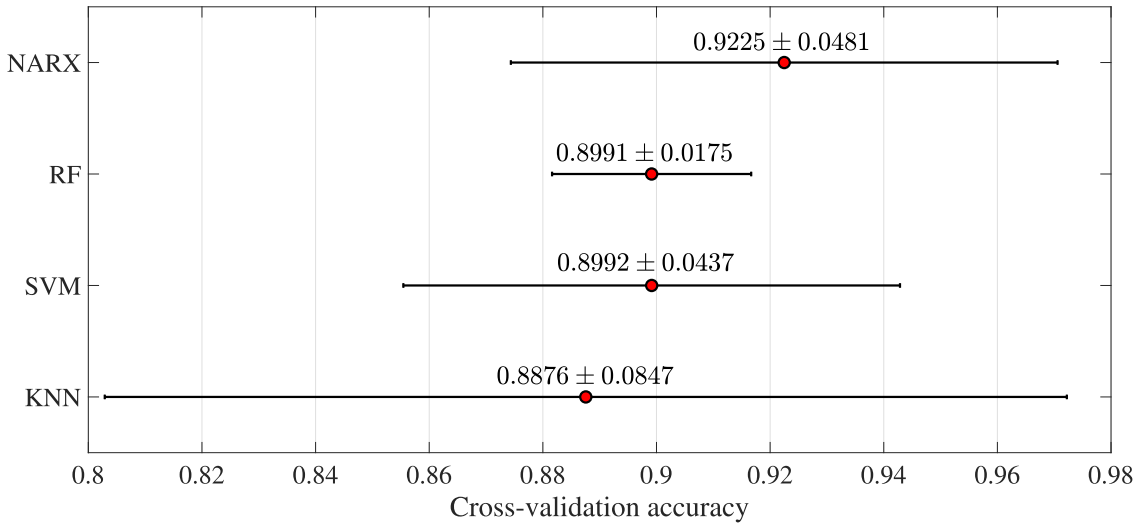
# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
2	0.7821	0.0551	0.0020	7	0.8716	0.0565	0.3568
3	0.8386	0.0673	0.0004	8	0.8892	0.0456	0.6059
4	0.8252	0.0566	0.0807	9	0.9205	0.0456	—
5	0.8305	0.0547	0.1091	10	0.8738	0.0747	0.4923
6	0.8598	0.0676	0.2268				

Figure 11 displays the classification accuracy of the proposed method in comparison to traditional techniques. The Logistic-NARX Multinomial approach achieved an average accuracy of 92 %, outperforming competing models. The narrow confidence intervals suggest strong generalization capabilities, even under class imbalance conditions. The confusion matrix analysis (see Table 12) further confirms the competitive performance of the proposed method, with Support Vector Machines (SVM) yielding similar results but slightly more classification errors.

Table 11 – Identified NARX model $\hat{y}(k)$ for the Wine dataset, presenting the selected terms, their estimated parameters, and corresponding relevance scores used during the model selection process.

Model Terms	Parameter	Score	Model Terms	Parameter	Score
Constant	0.5333	0.7759	$u_7(k)$	-0.1446	0.7754
$u_{10}(k)u_{13}(k)$	0.7491	0.6630	$u_1(k)u_{10}(k)$	0.2371	0.6458
$u_1(k)u_1(k)$	-0.6990	0.6569	$u_2(k)u_7(k)$	-0.5193	0.4948
$u_3(k)u_7(k)$	0.5661	0.4492	$u_7(k)u_{11}(k)$	0.3936	0.4652
$u_6(k)u_{13}(k)$	0.2870	0.4436			

Figure 11 – Accuracy outcomes for various classification methods applied to the Wine dataset, evaluated through cross-validation. Results are presented at regular intervals along with their respective average accuracy values.



Source: created by the author. (2023).

Table 13 offers a comprehensive overview of overall performance using various metrics. Both the L-NARX M and SVM methods produced favorable results, particularly in terms of the F1 Score, which represents the harmonic mean of precision and recall. This highlights the effectiveness of these methods in achieving a balanced performance across different evaluation criteria.

Table 12 – Confusion matrix illustrating the performance of the classification methods, obtained by aggregating the validation set predictions from cross-validation partitions. The classes represent the three distinct cultivars present in the Wine dataset.

	C1	C2	C3
C1	57	4	
C2	2	65	3
C3		2	45

(a) Logist-NARX Multiclass.

	C1	C2	C3
C1	52	3	3
C2	2	60	7
C3	5	8	38

(b) Random Forests.

	C1	C2	C3
C1	55	4	
C2	2	63	2
C3	2	4	46

(c) Support Vector Machine.

	C1	C2	C3
C1	54	8	1
C2	4	62	9
C3	1	1	38

(d) K-Nearest Neighbors.

Table 13 – Performance comparison among different classification methods applied to the Wine dataset, using multiple evaluation metrics derived from the corresponding confusion matrices.

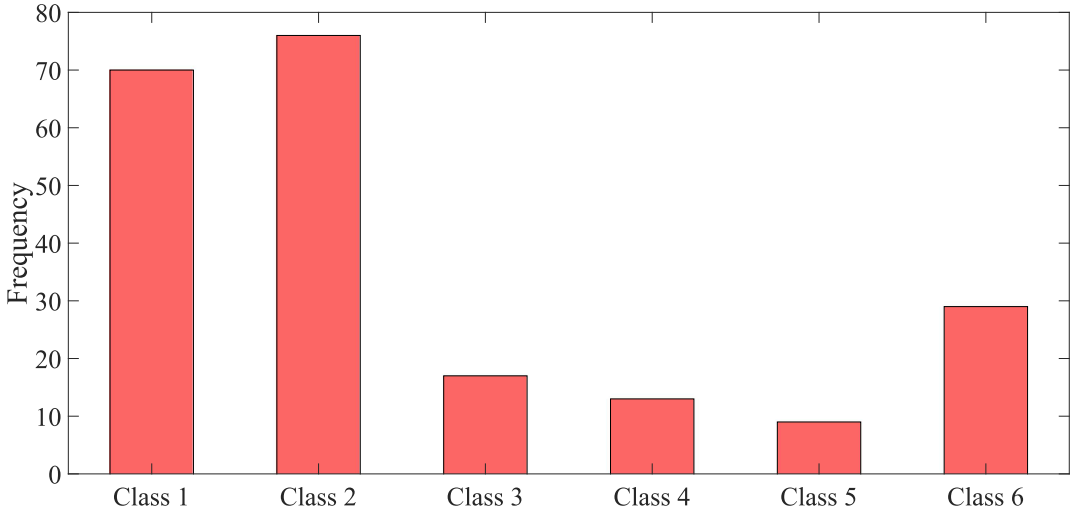
	L-NARX M	RF	SVM	KNN
Average Accuracy	0.9225	0.8991	0.8992	0.8876
Sensitivity	0.9246	0.8962	0.8998	0.8900
Specificity	0.9603	0.9483	0.9477	0.9435
Precision	0.9192	0.9027	0.9030	0.8875
F1 Score	0.9215	0.8987	0.9013	0.8873

3.3.5 Glass Dataset

The Glass dataset was developed by the Centre for Forensic Science Research to support forensic investigations involving the classification of glass types (Zhong and Fukushima, 2007; Denoeux, 2000). It is publicly available through the *UCI Machine Learning Repository*[4]. The dataset contains 214 instances and includes chemical composition data (e.g., Na, Fe, K oxides) used to classify samples into seven distinct glass categories. Each sample is described by nine numerical features derived from chemical analysis.

Figure 12 presents the class distribution, revealing a significant imbalance across categories. Although some classes are underrepresented, all categories are considered equally important in this classification task. The dataset is considered challenging due to the presence of minority classes and potential outliers.

Figure 12 – Class frequency distribution in the Glass dataset, highlighting the presence of significant class imbalance due to the uneven number of samples across categories.



Source: created by the author. (2023).

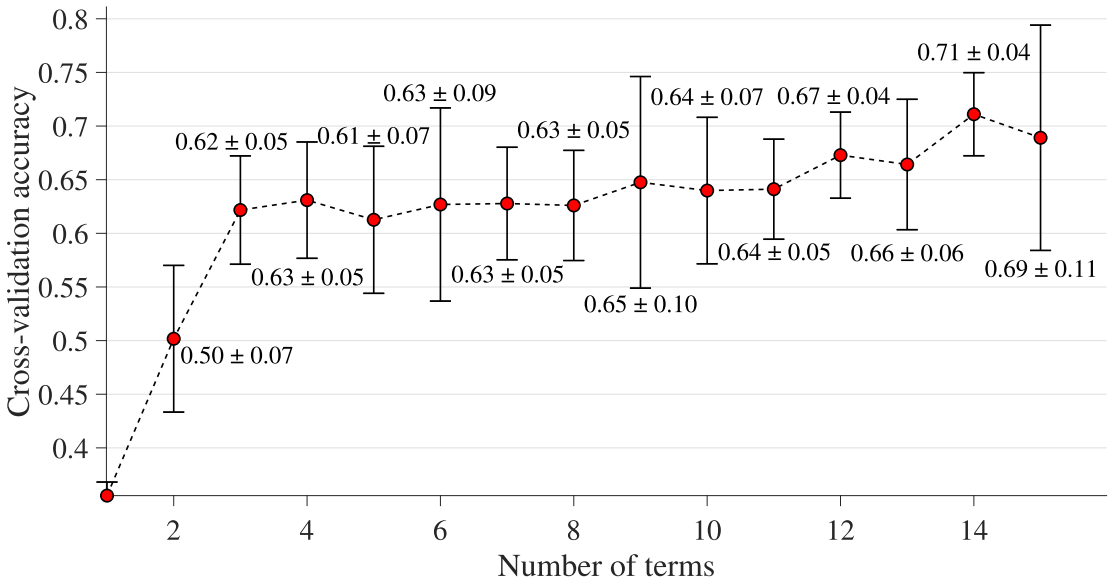
For the application of the proposed method (2.2), the parameters were set to a nonlinearity degree of $l = 2$ and a maximum of $n_{max} = 10$ selected terms, yielding a search space of 55 candidate terms. As this is a static classification problem—where each instance represents an independent chemical analysis of a glass sample—no time delays were incorporated into the NARX structure.

The cross-validation results for this case study (see Figure 13), summarized in Table 14, reveal a gradual increase in accuracy as the number of terms grows, with the best average performance achieved for the 14-term model (0.6943 ± 0.1072). Paired t-tests

[4] Creator: B. German – Central Research Establishment Science Service Aldermaston, Berkshire. Donor: Vina Spiehler, Diagnostic Products Corporation (213) 776-0180, 1987. Glass dataset (*UCI Machine Learning Repository*): https://archive.ics.uci.edu/ml/datasets/glass+identification.

comparing each configuration to the 14-term model show that most models with fewer terms present statistically lower accuracy ($p < 0.05$), particularly for the configurations with 2, 4, 7, and 11 terms. However, models with 13 terms ($p = 0.6847$) and 12 terms ($p = 0.3543$) exhibit no statistically significant difference in accuracy when compared to the 14-term model, suggesting potential for complexity reduction without significant performance loss. Nevertheless, considering the higher variance observed in some simpler configurations and the more consistent accuracy at higher complexities, the 14-term model is retained as the preferred configuration for this case study.

Figure 13 – Average accuracy calculated during the model term selection phase for the Glass dataset, used to identify the optimal number of terms for constructing the final classification model with maximum performance.



Source: created by the author. (2023).

Table 15 presents the identified NARX model $\hat{y}(k)$, listing the selected terms, their estimated parameters, and relevance scores. Due to the inherent complexity of the dataset, including class imbalance and outliers, all classification methods yielded modest performance levels. Nevertheless, the Logistic-NARX model achieved results comparable to the best-performing technique, reinforcing its potential as a competitive and interpretable alternative, as illustrated in Figure 14.

Table 16 presents a comparative evaluation of the tested classification methods using standard performance metrics derived from confusion matrices. The proposed NARX model outperformed all other methods in terms of overall accuracy (71.1 %), F1 Score (70.0 %), and precision (63.6 %). Although Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) also demonstrated competitive results, especially in sensitivity and

Table 14 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 14-term model.

# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
2	0.5017	0.0684	0.0054	9	0.6448	0.0455	0.1907
3	0.6217	0.0505	0.0549	10	0.6640	0.0571	0.3080
4	0.6309	0.0542	0.0222	11	0.6400	0.0490	0.1108
5	0.6126	0.0686	0.0702	12	0.6663	0.0563	0.3543
6	0.6269	0.0900	0.1655	13	0.7094	0.0419	0.6847
7	0.6309	0.0505	0.0465	14	0.6943	0.1072	—
8	0.6537	0.0784	0.4388				

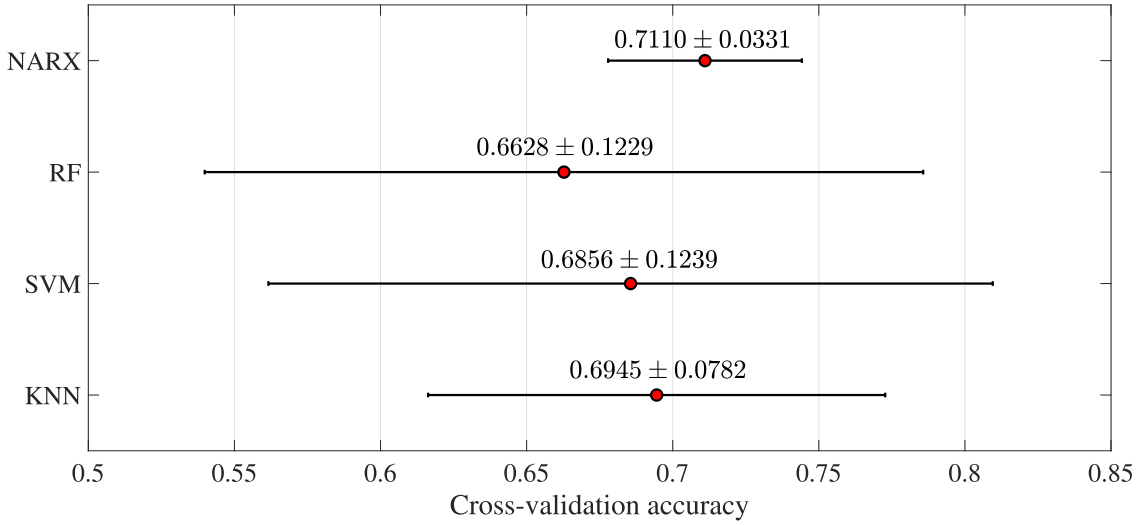
Table 15 – Identified NARX model $\hat{y}(k)$ for the Glass dataset, presenting the selected model terms along with their estimated parameters and associated relevance scores.

Model Terms	Parameter	Score	Model Terms	Parameter	Score
Constant	0.4211	0.7287	$u_3(k)u_7(k)$	0.0151	0.5198
$u_8(k)$	0.2210	0.5403	$u_6(k)u_6(k)$	0.0336	0.4754
$u_4(k)u_6(k)$	−0.0682	0.4808	$u_8(k)u_8(k)$	−0.0344	0.4398
$u_4(k)$	0.0993	0.4530	$u_3(k)u_9(k)$	0.0353	0.4375
$u_4(k)u_8(k)$	−0.0147	0.4727	$u_6(k)$	−0.0174	0.4257
$u_2(k)u_4(k)$	−0.0496	0.4254	$u_4(k)u_7(k)$	−0.0475	0.4171
$u_7(k)$	0.0660	0.4075	$u_9(k)u_9(k)$	−0.0031	0.3887
$u_1(k)u_6(k)$	0.0579	0.3888			

specificity, they fell slightly behind in balanced performance metrics.

Given the class imbalance inherent in the Glass dataset, these results may still be improved through techniques designed to address data imbalance. Common approaches include *undersampling*, which removes instances from majority classes, and *oversampling*, which synthetically augments minority class instances. Such strategies could further enhance classifier robustness across all classes.

Figure 14 – A comparative analysis of accuracy results for various classification methods applied to the Glass dataset is showcased. The accuracy, evaluated through cross-validation, is depicted at intervals, accompanied by their respective average values.



Source: created by the author. (2023).

Table 16 – Performance comparison among different classification methods applied to the Glass dataset, based on multiple evaluation metrics derived from the corresponding confusion matrices.

	NARX	RF	SVM	KNN
Accuracy	0.7110	0.6628	0.6856	0.6945
Sensitivity	0.6600	0.6488	0.6751	0.6633
Specificity	0.9342	0.9229	0.9281	0.9316
Precision	0.6356	0.5933	0.6385	0.6370
F1 Score	0.7005	0.6164	0.6529	0.6484

3.3.6 Wave Dataset

To examine the effectiveness of the parameter extraction and dimensionality reduction-based method, several experiments were conducted by applying the method to the Wave dataset (Valentini, 2004). The samples are generated by the waveform database generator presented in Breiman et al. (1984), which is available on the UCI Machine Learning Repository[5]. The dataset comprises 5000 samples and 40 attributes (features) describing 3 classes as waveforms, all of which include noise.

Each class in the Wave dataset is generated by combining two of three basic waveforms: $h_1(t)$, $h_2(t)$, and $h_3(t)$. To create an instance $\mathbf{x}_i$, a uniformly distributed random variable $u \sim \mathcal{U}(0,1)$ and a noise vector $\mathbf{e} = [e_t]^\mathsf{T}$, where $e_t \sim \mathcal{N}(0,\sigma^2)$ for $t = 1,\ldots,40$, are used. The input vector is then defined as:

$$\mathbf{x}_i = u\mathbf{h}_1 + (1-u)\mathbf{h}_2 + \mathbf{e}, \tag{3.6}$$

where the waveform pairs vary by class: class 1 uses (h_1, h_2), class 2 uses (h_1, h_3), and class 3 uses (h_2, h_3). Due to the synthetic nature of the dataset, nearly half of the 40 features are irrelevant. Thus, identifying the most informative variables is crucial not only to enhance classification accuracy but also to reduce computational cost and improve model generalization.

The Wave dataset exhibits a temporal structure, as its attributes represent sequential values with correlations across time steps. Given that waveforms evolve over time, past observations carry useful information for distinguishing between classes. By incorporating delays, the NARX model captures these temporal dependencies, improving classification performance by leveraging the underlying dynamics of the data.

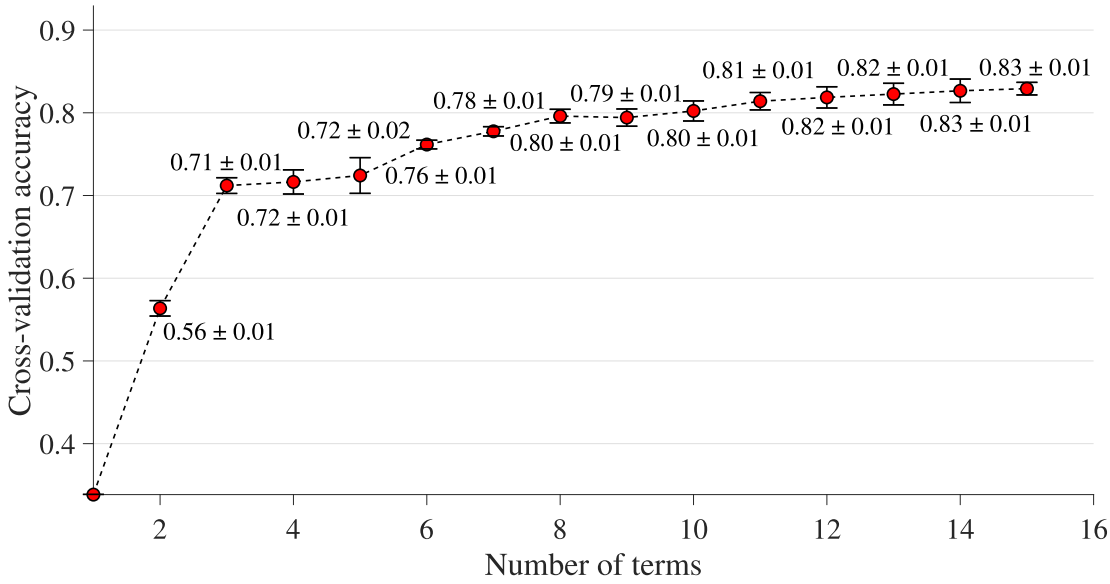
In applying the proposed method (2.2) to the Wave dataset, the following parameters were adopted: nonlinearity degree $l = 2$, maximum number of selected terms $n_{max} = 10$, and input delay $n_u = 10$, resulting in a search space of 80601 candidate terms. This large number of features and temporal dependencies contributes to the increased complexity of the problem.

The cross-validation results for this case study, shown in Figure 15, indicate a gradual improvement in model performance as the number of terms increases, reaching a maximum mean accuracy of 0.8422 ($\pm$ standard deviation) for the configuration with 15 terms. This configuration was identified as the best-performing model among all those tested.

Table 17 shows that, when comparing each configuration to the 15-term model

[5] Creator: Wadsworth International: California. Donor: David Aha. Waveform Database Generator (Version 2) (*UCI Machine Learning Repository*): http://archive.ics.uci.edu/ml/datasets/waveform+database+generator+(version+2).

Figure 15 – Average accuracy calculated during the term selection phase of the model for the Wave dataset, used to determine the optimal number of terms for constructing the final model with maximum classification performance.



Source: created by the author. (2023).

using paired t-tests, models with fewer terms (e.g., 2 to 6) exhibit statistically significant performance differences, with very low p-values close to zero. This confirms that these simpler configurations underperform relative to the best model.

From approximately 10 terms onward, the mean accuracy approaches the maximum value and the standard deviations remain relatively small, indicating greater stability. Nonetheless, the statistical results reveal that some models with slightly fewer terms, such as those with 12 or 13, achieve accuracy statistically equivalent to the 15-term configuration. This suggests the possibility of selecting more parsimonious models when the aim is to reduce complexity without compromising predictive performance (see Table 18).

The classification performance for the Wave dataset is summarized in Figure 16 and Table 19. The proposed method outperformed Random Forest (RF) and k-Nearest Neighbors (KNN), achieving an average accuracy of 82.9 % and F1 score of 82.9 %. Although Support Vector Machines (SVM) showed slightly higher metrics overall, the Logistic-NARX model demonstrated competitive performance with the added advantage of interpretability and dimensionality reduction. These results confirm the method's ability to effectively address complex, high-dimensional classification tasks.

Table 17 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 15-term model.

# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
2	0.5636	0.0084	0.0000	9	0.7944	0.0088	0.0003
3	0.7112	0.0087	0.0000	10	0.8022	0.0122	0.0010
4	0.7133	0.0193	0.0000	11	0.8138	0.0095	0.0168
5	0.7206	0.0212	0.0000	12	0.8232	0.0118	0.0861
6	0.7616	0.0054	0.0000	13	0.8226	0.0118	0.0699
7	0.7776	0.0054	0.0000	14	0.8266	0.0118	0.1465
8	0.7962	0.0090	0.0005	15	0.8292	0.0103	-

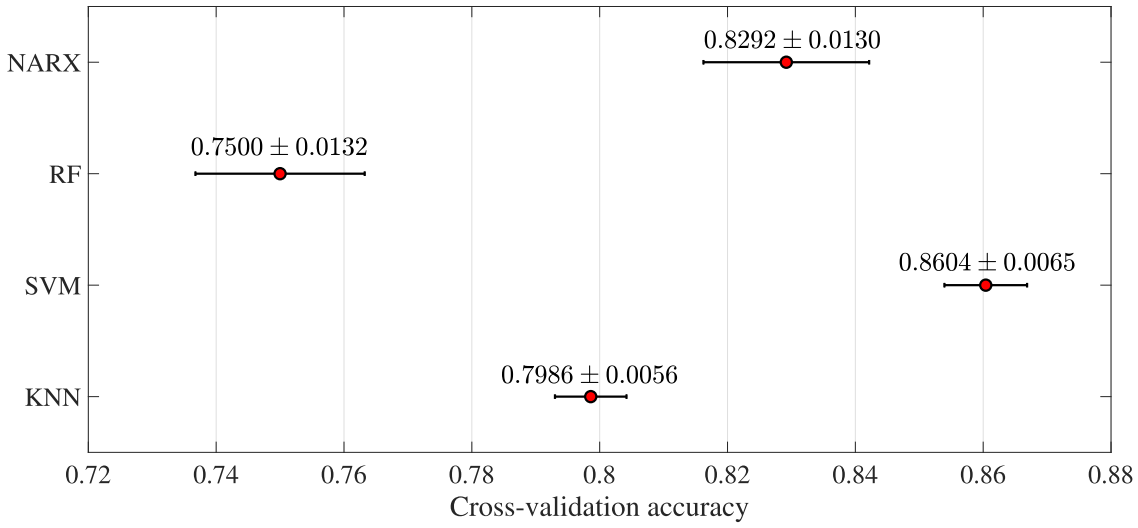
Table 18 – Identified NARX model $\hat{y}(k)$ for the Wave dataset, including the selected model terms along with their estimated parameters and associated relevance scores.

Model Terms	Parameter	Score	Model Terms	Parameter	Score
Constant	4.4978	0.7690	$u_7(k)$	-0.8246	0.5638
$u_{11}(k)$	0.9394	0.5388	$u_8(k)u_{11}(k)$	-0.2979	0.4542
$u_{12}(k)u_{16}(k)$	0.4512	0.4278	$u_{10}(k)$	0.5870	0.4278
$u_{15}(k)u_{16}(k)$	0.1612	0.4300	$u_8(k)$	-0.2781	0.4164
$u_9(k)u_{10}(k)$	-0.2658	0.4116	$u_6(k)u_7(k)$	0.3556	0.4028
$u_{13}(k)$	0.7814	0.4032	$u_6(k)u_{17}(k)$	-0.2372	0.4026
$u_7(k)u_9(k)$	0.2290	0.3964	$u_9(k)$	0.2863	0.3982
$u_{11}(k)u_{13}(k)$	-0.0849	0.3978			

Table 19 – Performance comparison among different classification methods applied to the Wave dataset, based on multiple evaluation metrics derived from the corresponding confusion matrices.

	NARX	RF	SVM	KNN
Accuracy	0.8292	0.7500	0.8604	0.7986
Sensitivity	0.8292	0.7531	0.8619	0.7986
Specificity	0.9146	0.8777	0.9310	0.8993
Precision	0.8294	0.7510	0.8609	0.7989
F1 Score	0.8292	0.7474	0.8600	0.7986

Figure 16 – Examination of accuracy results for different classification methods applied to the Wave dataset. Cross-validated accuracy is presented at regular intervals, accompanied by their corresponding average values.



Source: created by the author. (2023).

3.4 MODEL EVALUATION AND INTERPRETABILITY

This section discusses the key findings from applying the Logistic-NARX Multinomial model to four benchmark datasets, emphasizing classification accuracy, variable importance, and interpretability of the resulting models. To better illustrate the concept of variable importance and model interpretability, let us first consider a simplified version of the model built using the Fisher's Iris dataset, composed of only two terms:

$$\hat{y}(k) = 0.4386 + 0.1322\, u_3(k) + 0.0218\, u_4(k)u_4(k), \tag{3.7}$$

here, $u_3(k)$ corresponds to Petal Length and $u_4(k)$ to Petal Width. The coefficient of $u_3(k)$ is positive, indicating that an increase in Petal Length is associated with a higher probability of classifying a sample into a specific class. Meanwhile, the quadratic term $u_4(k)^2$ (Petal Width squared) introduces a nonlinear effect, reflecting how larger petal widths influence the classification in a non-proportional way. Even with only two terms, the model can already provide interpretable insights into how individual and nonlinear effects of variables contribute to decision-making. Expanding to the complete identified model for the Iris dataset, we have:

$$\begin{aligned}\hat{y}(k) = {} & 0.4386 + 0.1322\, u_3(k) + 0.0218\, u_4(k)u_4(k) + 0.0572\, u_3(k)u_3(k) \\ & - 0.0368\, u_2(k)u_4(k) + 0.0035\, u_1(k)u_4(k),\end{aligned} \tag{3.8}$$

with input variables corresponding to: $u_1(k)$: Sepal Length, $u_2(k)$: Sepal Width, $u_3(k)$: Petal Length, $u_4(k)$: Petal Width. This equation makes clear:

- Petal Length ($u_3(k)$) is strongly positively associated with the output class, both linearly and quadratically;
- Petal Width contributes nonlinearly via its squared term and in combination with Sepal Width;
- Interaction terms such as $u_2(k)u_4(k)$ introduce complex dependencies, which are still fully interpretable due to the explicit form of the equation.

Thus, each term's coefficient quantifies its relative influence on the classification decision, offering a transparent framework for variable importance analysis. The Logistic-NARX Multinomial model demonstrated strong classification performance across the datasets tested. Table 20 shows the accuracy comparison with classical classifiers like Random Forest, Support Vector Machines, and K-Nearest Neighbors. For instance, in

Table 20 – Comparison of average accuracy obtained through cross-validation.

	Iris	*Wine*	*Glass*	*Wave*
NARX	0.9733 ± 0.026	0.9225 ± 0.048	0.7110 ± 0.033	0.8292 ± 0.013
RF	0.9467 ± 0.020	0.8991 ± 0.017	0.6628 ± 0.122	0.7500 ± 0.013
SVM	0.9667 ± 0.033	0.8992 ± 0.043	0.6856 ± 0.123	0.8604 ± 0.006
KNN	0.9467 ± 0.053	0.8876 ± 0.084	0.6945 ± 0.078	0.7986 ± 0.005

the Iris dataset, the proposed model achieved 97.33% accuracy, outperforming the other models while maintaining an interpretable and compact representation.

In terms of variable relevance, Table 21 summarizes how many variables and corresponding model terms were retained. For the Iris dataset, from four original variables, only one was necessary to derive two meaningful model terms, achieving a 75% reduction. This underscores the model's efficiency in isolating the most relevant variables and their interactions without sacrificing performance.

Table 21 – Performance of the proposed method in feature extraction, leading to the selection of a reduced-dimensionality model while preserving classification effectiveness.

Datasets	Total	Model Terms	Select Features	Reduction (%)	Accuracy	
					$\overline{x}$	max
Iris	4	2	1	75.00	0.9667	0.9997
Wine	13	9	8	38.46	0.9225	0.9706
Glass	9	14	8	11.11	0.7110	0.7441
Wave	40	14	11	72.50	0.8292	0.8422

3.5 DISCUSSION

The analysis of the Logistic-NARX Multinomial model reveals several key findings and areas for future development. A notable observation is the consistent inclusion of the constant term across all experiments, underscoring its importance in capturing baseline behavior and supporting decision threshold calibration, particularly under normalized conditions.

Despite demonstrating competitive accuracy and interpretability in various benchmark datasets, the model presents limitations that merit consideration. One primary constraint lies in its reliance on polynomial basis functions. While these offer simplicity and analytical clarity, they may fall short in capturing intricate nonlinearities inherent in more complex classification tasks. Future research could explore alternative basis expansions,

such as radial basis functions or wavelets, to enhance model flexibility and approximation capability.

Another technical challenge involves the selection of lag values for input and output variables (n_u and n_y). This step remains an open issue in the literature, as increasing these values significantly enlarges the search space of possible terms. This exponential growth can render the term selection phase computationally intensive, both in processing time and memory usage (Wei et al., 2004). Efficient strategies for lag estimation or adaptive selection methods could address this bottleneck.

The presence of multicollinearity among input features also poses a concern. Highly correlated variables can distort coefficient estimation and reduce model robustness. Approaches such as iterative Orthogonal Forward Regression (OFR) (Guo et al., 2015) or ultra-OFR (Guo et al., 2016) offer promising avenues to mitigate this issue by ensuring orthogonalization during the selection of regressors.

Moreover, class imbalance remains a significant obstacle. As is typical in real-world classification tasks, especially in safety-critical applications, some categories are underrepresented. This imbalance can hinder the model's ability to correctly classify rare but important cases. To overcome this, strategies like data resampling, cost-sensitive learning, or ensemble methods could be employed to improve minority class detection without compromising overall accuracy (Kubat and Matwin, 1997; Batista et al., 2000).

3.6 SUMMARY

This chapter provided an evaluation of the proposed Logistic-NARX Multinomial model using four benchmark datasets: *Iris*, *Wine*, *Glass*, and *Wave*—selected for their varying levels of dimensionality, noise, and class imbalance. Results showed that the proposed model consistently achieved competitive or superior classification accuracy when compared to traditional methods like Random Forests, Support Vector Machines, and k-Nearest Neighbors. Additionally, the model demonstrated effective dimensionality reduction, maintaining interpretability through explicit functional expressions. Notably, high performance was achieved even in imbalanced and noisy scenarios, reinforcing the method's applicability to complex classification tasks with enhanced transparency and reduced computational burden.

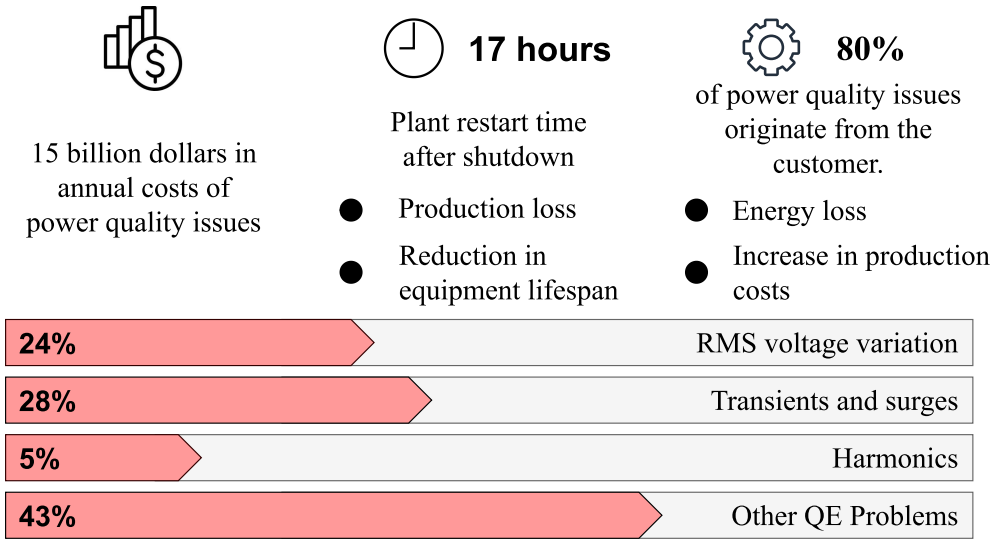
4 CLASSIFICATION OF PQ DISTURBANCE

The next chapter presents the application of the NARX classification method to Power Quality Event Classification. It covers key steps, including event simulation, signal preprocessing, and parameter extraction using Higher Order Statistics (HOS), followed by selection via the Fisher Criterion. The chapter also demonstrates the discriminative power of the extracted features and compares the proposed approach with traditional methods, highlighting its effectiveness in this specific domain.

4.1 INTRODUCTION

Power Quality has become a prominent area of research in recent years, attracting growing interest from both academia and industry. This trend is largely driven by the increasing presence of non-linear loads, the evolution of power electronics, the widespread adoption of microprocessed systems in the electrical grid, and the expansion of renewable and distributed generation sources (Bollen et al., 2017). Non-linear loads contribute to distortions in voltage and current waveforms, while microprocessed equipment tends to be highly sensitive to such disturbances (Mishra, 2019). Any deviation from the nominal characteristics of electrical signals is classified as a power quality disturbance, which can cause equipment malfunctions and disrupt industrial processes—often leading to substantial financial losses (Figure 17). In this context, the following section explores the development and improvement of monitoring systems for Power Quality, with particular emphasis on the classification and detection of disturbances (Nagata et al., 2018).

Figure 17 – Infographic illustrating the impact of different events and conditions on Power Quality, highlighting key sources of disturbance and their effects on electrical systems.



Source: created by the author. (2023).

In general, the classification and detection of Power Quality disturbances involve pre-processing steps that include parameter extraction and dimensionality reduction. Among the parameter extraction techniques, Higher Order Statistics (HOS) have been widely investigated as a promising approach for this application (Mendel, 1991; Ferreira et al., 2009). HOS is favored due to its robustness to Gaussian noise, efficiency in capturing signal characteristics from voltage waveforms, and relatively low computational cost. To reduce the dimensionality of the extracted parameter space, the Fisher linear discriminant is commonly applied. This technique selects the most representative features that maximize class separability (Duda et al., 2012).

Several classification algorithms have been proposed for addressing the Power Quality problem, including Support Vector Machines (Nagata et al., 2020), Neural Networks (Naik and Kundu, 2014), Random Forests (Zhang et al., 2003), and k-Nearest Neighbors (Pan et al., 2017). Although these methods often achieve high accuracy, they typically suffer from limited interpretability, making it difficult to understand the relationships between input variables and model predictions.

To overcome this limitation, the present work applies the Logistic-NARX Multinomial model to the Power Quality classification task. This approach aims to combine competitive predictive performance with a transparent and interpretable model structure.

4.2 PQ EVENT MODELING AND FEATURE ENGINEERING

4.2.1 Simulation of Power Quality Events

A system developed for the detection and classification of electrical disturbances in Power Quality (PQ) should correctly identify the occurrence of each abnormality in a discrete-time power system voltage signal $v[n]$ composed of N samples, which can be expressed as a sum of contributions from various types of phenomena (Diego Ferreira, 2010; Ribeiro and Pereira, 2007):

$$v[n] = v(t)\Big|_{t=\frac{n}{f_s}} := f[n] + h[n] + i[n] + t[n] + r[n], \qquad 0 \leq n \leq N-1, \tag{4.1}$$

where f_s is the sampling frequency, and the sequences $f[n]$, $h[n]$, $i[n]$, $t[n]$, and $r[n]$ represent the fundamental component, harmonics, interharmonics, transients, and noise, respectively. Each of these signals can be defined as:

$$f[n] := A_0[n] \cos\left(2\pi \frac{f_0[n]}{f_s} n + \theta_0[n]\right), \tag{4.2}$$

in (4.2), the term $A_0[n]$ represents the amplitude of the voltage signal, $f_0[n]$ is the fundamental frequency, and $\theta_0[n]$ is the phase of the voltage signal.

$$h[n] := \sum_{m=1}^{M} h_m[n], \tag{4.3}$$

$$i[n] \coloneqq \sum_{j=1}^{J} i_j[n], \tag{4.4}$$

where $h_m[n]$ and $i_j[n]$ are the m-th harmonic and the j-th interharmonic, respectively.

$$t[n] \coloneqq t_{\text{imp}}[n] + t_{\text{not}}[n] + t_{\text{osc}}[n], \tag{4.5}$$

where $r[n]$ is noise with normal distribution $\mathcal{N}(0, \sigma_r^2)$ and independent of $f[n]$, $h[n]$, $i[n]$, and $t[n]$. In Equations (4.3) and (4.4), the terms $h_m[n]$ and $i_j[n]$ can be defined as:

$$h_m[n] \coloneqq A_m[n] \cos\left[2\pi m \frac{f_0[n]}{f_s} n + \theta_m[n]\right] \left[u\left[n - n_{h_{m,i}}\right] - u\left[n - n_{h_{m,f}}\right]\right], \tag{4.6}$$

$$i_j[n] \coloneqq A_{I,j}[n] \cos\left[2\pi \frac{f_{I,j}[n]}{f_s} n + \theta_{I,j}[n]\right] \left[u\left[n - n_{i_{j,i}}\right] - u\left[n - n_{i_{j,f}}\right]\right], \tag{4.7}$$

where $u[n]$ denotes the unit step sequence, $n_{h_{m,i}}$ and $n_{h_{m,f}}$ represent the start and end samples of the harmonics, respectively. Similarly, $n_{i_{j,i}}$ and $n_{i_{j,f}}$ represent the start and end samples of the interharmonics, respectively. In Equation (4.6), $A_m[n]$ is the amplitude, and $\theta_m[n]$ is the phase of the m-th harmonic. Similarly, in (4.7), $A_{I,j}[n]$, $f_{I,j}[n]$, and $\theta_{I,j}[n]$ are the amplitude, frequency, and phase of the j-th interharmonic, respectively. Then, in (4.5), $t_{\text{imp}}[n]$, $t_{\text{not}}[n]$, and $t_{\text{osc}}[n]$ represent impulsive transients, notches, and oscillatory transients, respectively, and are expressed as Ribeiro and Pereira (2007):

$$t_{\text{imp}}[n] \coloneqq \sum_{i=1}^{N_{\text{imp}}} t_{\text{imp},i}[n] \left[u\left[n - n_{t_{\text{imp},i}}\right] - u\left[n - n_{t_{\text{imp},f}}\right]\right], \tag{4.8}$$

$$t_{\text{not}}[n] \coloneqq \sum_{i=1}^{N_{\text{not}}} t_{\text{not},i}[n] \left[u\left[n - n_{t_{\text{not},i}}\right] - u\left[n - n_{t_{\text{not},f}}\right]\right], \tag{4.9}$$

$$t_{\text{osc}}[n] \coloneqq \sum_{i=1}^{N_{\text{osc}}} A_{\text{osc},i}[n] \exp\left[-\alpha_{\text{osc},i}\left(n - n_{\text{osc},i}\right)\right] \left[u\left[n - n_{t_{\text{osc},i}}\right] - u\left[n - n_{t_{\text{osc},f}}\right]\right], \tag{4.10}$$

where $t_{\text{imp},i}[n]$ and $t_{\text{not},i}[n]$ are the n-th samples of the i-th impulsive transient and the i-th notch, respectively. $n_{t_{\text{imp},i}}, n_{t_{\text{not},i}}$, and $n_{t_{\text{osc},i}}$ represent the start samples of each impulsive transient, and $n_{t_{\text{imp},f}}, n_{t_{\text{not},f}}$, and $n_{t_{\text{osc},f}}$ represent the samples marking the end. Finally, (4.10) defines the exponential decay components, as well as the direct current components ($\alpha = 0$) generated by geomagnetic disturbances.

The Power Quality events (PQ) addressed are isolated disturbances reflected in the voltage waveform. The disturbances were synthetically generated following the standards in the IEEE-1159 standard (IEEE, 2019), belonging to five different classes: Harmonic (C1), Sags/Swells (C2), Spikes (C3), Notches (C4), and Pure (C5), detailed in Table 22. The events were simulated with a sampling frequency $f_s = 15.36\,\text{kHz}$, i.e., 256 samples per cycle of the fundamental component of 60 Hz. A total of 250 events or signal windows were generated for each class, where the size of each event is equal to 4 cycles of the fundamental component with $N = 1024$ samples. Additionally, a Signal-to-Noise Ratio

Table 22 – A brief description of Power Quality events, representing the different classes considered in the proposed classification problem.

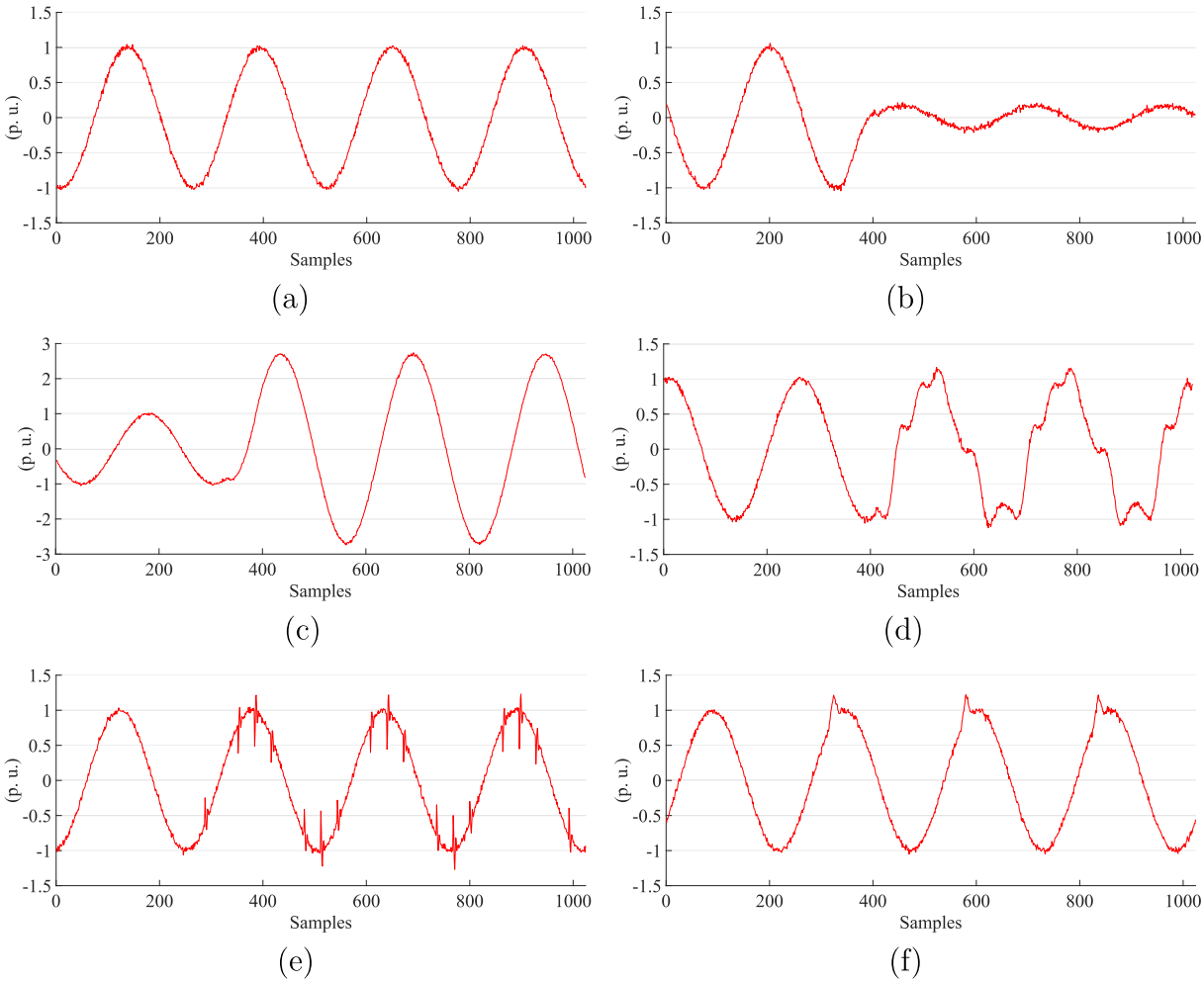
Event	Description
Harmonics	Sinusoidal voltages or currents with frequencies that are integer multiples of the fundamental frequency, caused by rectified inputs and power supplies used in electronic systems.
Sag	Decrease in RMS voltage, associated with faults, load switching, and motor starting.
Swell	Increase in RMS voltage, caused by faults in the power system.
Spike	Sudden change in the nominal voltage condition, caused by electrical discharges, load switching, and short circuits.
Notch	Periodic voltage disturbance caused by the operation of power devices during the switching from one phase to another.

(SNR) of 30 dB was considered. The modeling of events is represented in Table 23, where the parameters adopted in the simulation were randomly chosen within predefined intervals. The parameters have a uniform statistical distribution, achieving a high level of generalization in the classification process. The power quality events were generated using the MATLAB programming language, resulting in the waveforms depicted in Figure 18.

Table 23 – Mathematical models and associated parameters representing the main types of Power Quality events considered in this study.

Classes	Event	Equation	Parameters
C1	Harmonic	$\text{sen}(\omega t) + \sum_{n=3}^{7} \alpha_n \text{sen}(n\omega t)$	$0.01 \leq \alpha_n \leq 0.2$
C2	Sag/Swell	$(1 - \alpha(u(t - t_1) -$	$0.1 \leq \alpha \leq 0.8$
		$u(t - t_2)))\text{sen}(\omega t)$	$T \leq (t_2 - t_1) \leq 9T$
C3	Spike	$\text{sen}(\omega t) + \alpha \exp(-$	$0.1 \leq \alpha \leq 0.8$
		$(t - t_1)/\tau)u(t - t_1)$	$8\text{ms} \leq \tau \leq 40\text{ms}$
C4	Notch	$\text{sen}(\omega t) - \text{sign}(\text{sen}(\omega \text{t}))$	$0.1 \leq K \leq 0.4$
		$\sum_{n=0}^{4} K(u(t - (t_1 + 0.02n)) -$	$0.1T \leq (t_2 - t_1)$
		$u(t - (t_2 + 0.02n)))$	$(t_2 - t_1) \leq 0.05T$
C5	Pure	$A\text{sen}(\omega t), \quad A = 1(\text{p.u.})$	$\omega = 2\pi 60$ rad/s

Figure 18 – Power Quality events isolated and reflected in synthetically generated voltage waveforms: (a) Pure, (b) Sag, (c) Swell, (d) Harmonics, (e) Notch, and (f) Spike.



Source: created by the author. (2023).

4.2.2 Parameter Extraction: Higher Order Statistics

Parameter extraction is a fundamental step for ensuring reliable signal processing and for developing effective disturbance classifiers. In the context of Power Quality analysis, this process typically involves preprocessing stages that include the extraction and selection of representative features. Among the various techniques explored in the literature, Higher Order Statistics (HOS) has emerged as a promising approach for parameter extraction in Power Quality classification tasks (Mendel, 1991; Ferreira et al., 2009).

HOS involves the computation of statistical measures known as cumulants, which can effectively capture non-linear and non-Gaussian properties of signals. These cumulants have been widely used in signal analysis due to their robustness to Gaussian noise and their ability to reveal hidden structures within the signal (Nikias and Mendel, 1993; Nikias and Petropulu, 1993). Several studies have successfully applied cumulant-based feature

extraction to Power Quality disturbance classification, including the works of (Ferreira et al., 2009; Gerek and Ece, 2006; Ribeiro et al., 2006), demonstrating notable effectiveness in distinguishing different types of disturbances.

Conceptually, HOS extends the idea of correlation by introducing higher-order dependencies in the data (Mendel, 1991). One of its key advantages is the ability to project the signal into a new parameter space where the separability of different classes is often enhanced compared to the original domain (Nagata et al., 2020; Ferreira et al., 2009). HOS is particularly suited for analyzing non-Gaussian processes and non-linear systems, making it well-suited for the characteristics typically found in Power Quality disturbances.

Higher Order Statistics can be expressed through moments or cumulants. While moments are more suitable for analyzing deterministic signals, cumulants are preferred for stochastic signal modeling, which is generally the case in Power Quality analysis. Given the inherent non-linearity of many power disturbances, cumulant-based parameter extraction stands out as a powerful tool for enhancing classification performance.

The vectors extracted from the voltage signal using HOS-based techniques are capable of providing very well-defined information for each voltage event class ($\omega_i, i = 1, \ldots, C$). The expressions for the second and fourth-order cumulants of a random signal $x[n]$, when $E\{\mathrm{x}[n]\} = 0$, are represented as:

$$c_{2,x}[i] = E\{x[n]x[n+i]\}, \tag{4.11}$$

$$c_{4,x}[i] = E\left\{x[n]x^3[n+i]\right\} - 3c_{2,x}[i]c_{2,x}[0], \tag{4.12}$$

assuming i is the i-th lag, and $x[n]$ is the n-th element of the vector $\mathbf{x}$. Higher Order Statistics can be defined in terms of cumulants, assuming a periodic and finite-length vector of length N and a signal $x[n]$ with zero mean $E\{x[n]\} = 0$. The calculations for second and fourth-order cumulants can be expressed as:

$$\hat{c}_{2,x}[i] := \frac{2}{N}\sum_{n=0}^{N/2-1} x[n]x[n+i], \tag{4.13}$$

$$\hat{c}_{4,x}[i] := \frac{2}{N}\sum_{n=0}^{N/2-1} x[n]x^3[n+i] - \frac{2}{N^2}\sum_{n=0}^{N/2-1} x[n]x[n+i]\sum_{n=0}^{N/2-1} x^2[n], \tag{4.14}$$

where $i = 0, 1, \cdots, N/2-1$, Equations (4.13) and (4.14) cannot be used if $i > N/2+1$ since $n+i$ will be greater than N, thus losing some information in the cumulant calculations. In Ribeiro et al. (2006), the author proposes an alternative approach where each cumulant is calculated using all available N signal samples. Essentially, a circular buffer is formed in the signals, such that if the value of $n+1$ exceeds N by k units, this value is replaced

by k, thereby considering the preceding samples to the value i that were previously not used. In fact, it is as if there is a continuity from the last sample to the first, making the signal virtually circular. Thus, the replacement of $n+1$ when $n+1 > N$ can be given by:

$$\mathrm{mod}(n+i, N) = [n+i] - bN, \tag{4.15}$$

assuming b is the integer obtained when disregarding the decimal places of the division of $n+i$ by N, which actually results in the remainder of the division of $n+i$ by N. Thus, (4.11) and (4.12) can be estimated, for finite N, as:

$$\hat{c}_{2,x}[i] := \frac{1}{N}\sum_{n=0}^{N-1} x[n]x[\mathrm{mod}[n+i, N]], \tag{4.16}$$

$$\hat{c}_{4,x}[i] := \frac{1}{N}\sum_{n=0}^{N-1} x[n]x^3[\mathrm{mod}[n+i, N]] - \frac{1}{N^2}\sum_{n=0}^{N-1} x[n]x[\mathrm{mod}[n+i, N]]\sum_{n=0}^{N-1} x^2[n], \tag{4.17}$$

where $\mathrm{mod}[n+i, N]$ is the integer remainder of the division of $n+i$ by N. The approximations presented in (4.16) and (4.17) lead to a good simplification for problems where a finite-length vector is used. The approximations are more suitable when the signal is periodic. Therefore, considering the periodic nature of voltage signals in power systems, this is a good approximation of the HOS. It can be observed that for a signal with N samples, there are N cumulants for each order of HOS. For classification and detection purposes, the combination of a few of these cumulants is sufficient to achieve good performance. In summary, the equations can be used to extract parameters from voltage signals for power quality analysis.

4.2.3 Feature Selection using Fisher Criterion

Fisher's Discriminant Analysis (FDA), also known as Fisher's Linear Discriminant, is a statistical technique used to find a linear combination of variables that best separates two or more data classes (Theodoridis and Koutroumbas, 2006). While FDA is primarily used for classification, its underlying mathematical formulation provides an effective criterion to evaluate the discriminative power of individual features. This criterion, known as the Fisher Score, is widely used in filter-based feature selection methods, especially in high-dimensional datasets where selecting the most relevant attributes can significantly improve classification performance.

The goal of FDA is to find a projection vector that maximizes the separation between classes, by maximizing the ratio of variance between classes to the variance within classes. Considering a dataset with N training examples, each with d attributes and two distinct classes, let X be a data vector with dimension $(d \times 1)$, and Y a binary class variable taking values -1 or 1. FDA seeks a vector w with dimension $(d \times 1)$ that maximizes the class separation criterion $J(w)$:

$$J(w) = \frac{(w^T S_B w)}{(w^T S_W w)}, \tag{4.18}$$

where S_B is the between-class scatter matrix, and S_W is the within-class scatter matrix. The between-class scatter matrix S_B is defined as:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T, \tag{4.19}$$

where μ_1 and μ_2 are the means of the training examples in class 1 and 2, respectively. The matrix S_B measures the spread between the class means. The within-class scatter matrix S_W is defined as:

$$S_W = \sum_{i=1}^{N} (X_i - \mu_{Y_i})(X_i - \mu_{Y_i})^T, \tag{4.20}$$

where X_i is the i-th training example, μ_{Y_i} is the mean of the training examples in the class of X_i, and the sum is over all training examples. The matrix S_W measures the spread within each class. The solution for the vector w is obtained by maximizing the criterion $J(w)$ shown in (4.18), which can be rewritten as:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} = \frac{w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w}{w^T S_W w}, \tag{4.21}$$

To find the vector w that maximizes $J(w)$, we can use the method of Lagrange multipliers. The solution is given by:

$$w = S_W^{-1}(\mu_1 - \mu_2), \tag{4.22}$$

Once the vector w is determined, the classification of a new test example is performed by projecting it onto the direction of w and checking on which side of the threshold w_0 the projection lies. The threshold w_0 can be chosen to maximize the classification accuracy on a validation set.

Beyond classification, the Fisher criterion can also be extended to evaluate the relevance of individual features for feature selection. In this context, the separability of classes for each feature can be computed using the following criterion:

$$J_c = (m_1 - m_2)^2 \circ \frac{1}{D_1^2 + D_2^2}, \tag{4.23}$$

where $J_c = [J_1 \dots J_{L_t}]^T$, L_t is the total number of features, m_1 and m_2 are the mean vectors of each class, and D_1^2 and D_2^2 are the corresponding within-class variance vectors. The symbol ($\circ$) denotes the Hadamard (element-wise) product. This score ranks each

feature according to how well it separates the two classes: features with higher J_c values provide greater discriminative power.

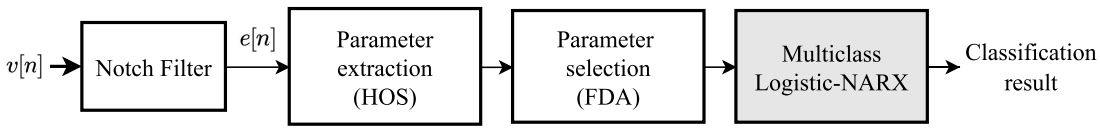
In summary, although Fisher's Linear Discriminant is traditionally employed for classification, its mathematical foundation supports its use as a powerful criterion for feature selection. This approach, commonly known as the Fisher Score, allows the selection of the most relevant features by quantifying class separability for each attribute individually, making it especially useful in preprocessing steps for high-dimensional datasets.

4.3 NARX-BASED PQ CLASSIFICATION

One of the main objectives in Power Quality monitoring is the accurate analysis of electrical disturbances. To achieve this, the voltage signals from the monitored system must first be recorded. However, continuously storing raw voltage waveforms leads to the generation of large data volumes. As a result, it is essential to implement detection mechanisms that selectively store only those signal segments containing relevant disturbances. Once these significant events are captured, classification techniques can be applied to analyze the stored data, allowing for the identification and localization of the disturbance sources. It is important to note that this analysis is typically performed offline, using previously recorded signals.

The classification of Power Quality (PQ) disturbances involves three main stages: application of a notch filter, parameter extraction through signal processing, and event classification using the Multinomial Logistic-NARX model (Section 3). These components are integrated into the system illustrated in Figure 19. The process begins with the simulation of voltage disturbances, modeled according to the procedures described in earlier sections. The resulting signals, denoted as $v[n]$, undergo preprocessing using a second-order Infinite Impulse Response (IIR) notch filter. This filter is centered at the fundamental frequency of 60 Hz and employs a notch factor of $\rho_0 = 0.97$. The filtered output, $e[n]$, effectively isolates the disturbance components by removing the fundamental frequency, thus enhancing the visibility of events superimposed on the original waveform.

Figure 19 – Proposed system architecture for the classification of Power Quality events, detailing the processing flow from signal input to final classification output.



Source: created by the author. (2023).

In the parameter extraction step, the voltage signal $e[n]$, composed of events or

windows of $N = 1024$ samples (equivalent to 4 cycles of the fundamental component), is processed by applying (4.16) and (4.17), extracting vectors or cumulants using Higher Order Statistics (HOS). The second and fourth-order cumulants provide a parameter vector $\boldsymbol{p} = [\hat{c}_{4,x}[i], \hat{c}_{2,x}[i]]$, totaling $2 \times N$ parameters obtained for each signal window. The next step involves parameter selection through Fisher Discriminant Analysis, aiming to choose a reduced set of data composed of the most representative parameters obtained during extraction. The method reduces the dimensionality of the parameter space and maximizes separability between disturbance classes, thereby reducing computational complexity and processing time. The implemented Fisher Discriminant Analysis selects 2 parameters for each class by considering the highest values of J_c (4.23) related to each cumulant, resulting in 10 parameters per event. Thus, the original dimension of each event has been reduced from 1024 to 10 samples.

The final stage consists of event classification using the Multinomial Logistic-NARX model. In this step, selected cumulant samples serve as input features for training the predictor. The dataset comprises 250 samples per class, totaling 1,250 balanced samples. To evaluate the model's performance, k-fold cross-validation with $k = 5$ was applied. The method parameters include a nonlinearity degree of $l = 2$, a maximum of $n_{max} = 10$ selected terms (as detailed in Algorithm 2), and a search space comprising 66 candidate model terms.

The inclusion of delayed inputs in the NARX structure is particularly relevant given the transient nature of Power Quality Disturbances (PQDs), which often occur over short time spans—typically within one or a few cycles of the 60 Hz fundamental frequency. To effectively capture these temporal dependencies, delays ranging from 1 to 10 samples were considered (input delay $n_u = 10$). This range reflects a practical compromise: it enables the model to retain short-term memory of the signal while avoiding unnecessary complexity. The use of delays is consistent with principles from system identification, where incorporating past information improves dynamic modeling and classification performance. Although alternative delay selection strategies exist—such as data-driven optimization and heuristic methods—the approach adopted here proved adequate for characterizing common PQDs, including sags, swells, spikes, and notches.

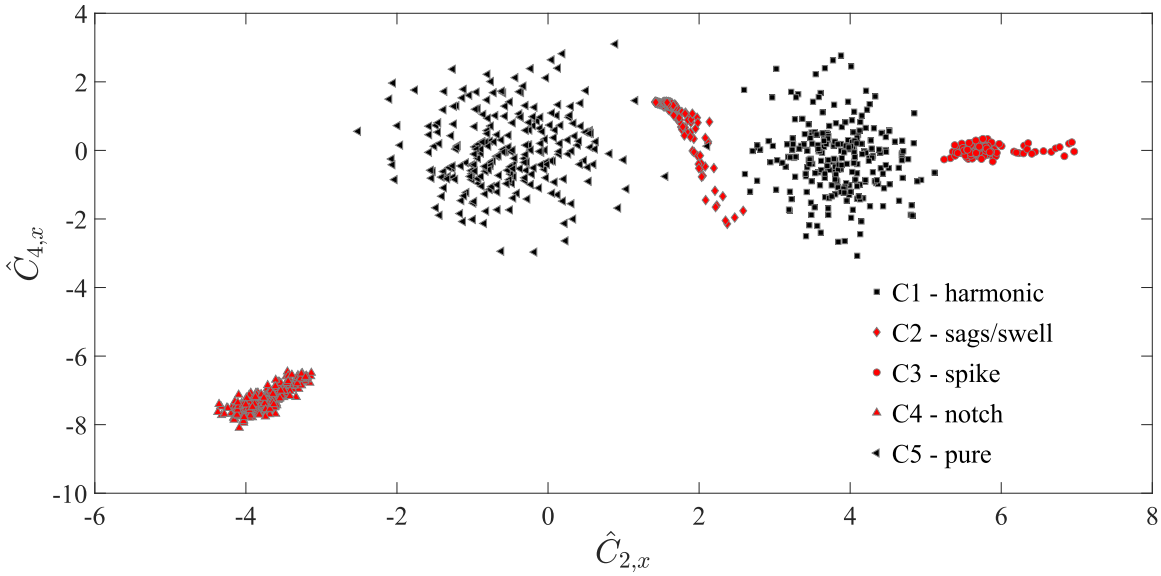
4.4 RESULTS

This section presents the results obtained using the proposed classification system for identifying power quality disturbances. The evaluation focuses on the model's accuracy and robustness, comparing its performance against several widely adopted classification algorithms. This comparative analysis allows for a comprehensive assessment of the proposed approach's effectiveness in real-world scenarios.

A key step in the classification process involves feature selection based on the

Fisher Discriminant criterion, which prioritizes variables with the highest discriminative power. Beyond numerical validation, the discriminative capability of the selected features can also be visually interpreted. Figure 20 highlights the separability achieved among the different classes, illustrating how the extracted features—particularly the second- and fourth-order cumulants—contribute significantly to distinguishing between event types.

Figure 20 – Selection of parameters in Power Quality event classification using second- and fourth-order cumulants within the parameter space.

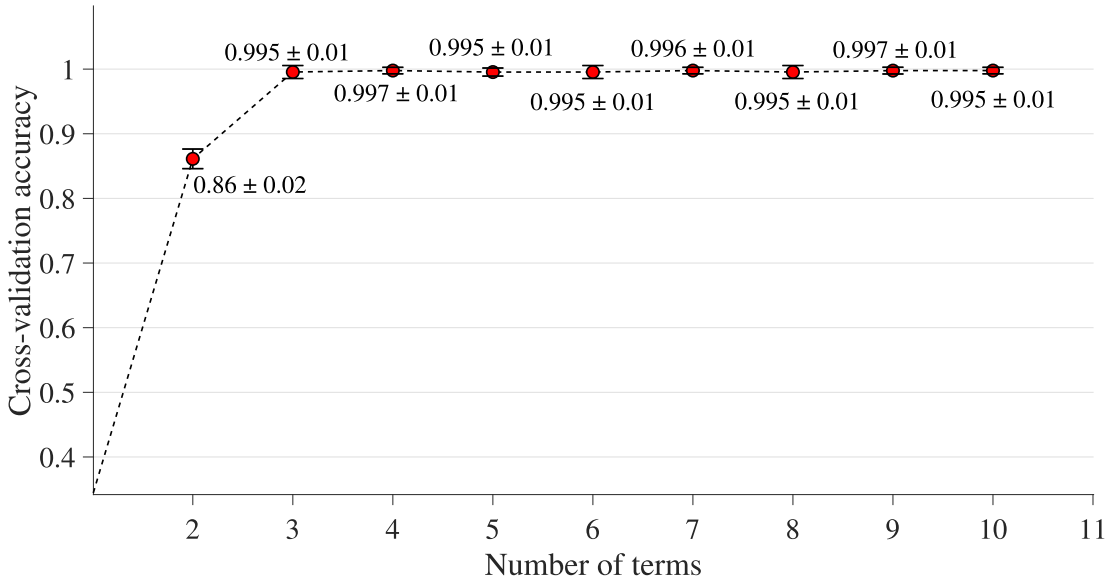


Source: created by the author. (2023).

Figure 21 shows the cross-validation accuracy as a function of the number of terms. A marked improvement is observed when moving from two to three terms (from 0.860±0.020 to 0.995 ± 0.009). Paired t-tests, shown in Table 24, using the three-term configuration as the reference, confirmed that this gain is statistically significant ($p = 3.68 \times 10^{-4}$). For configurations with 4–10 terms, mean accuracies lie in the narrow range 0.995–0.997 with standard deviations of approximately 0.01. Paired comparisons against the three-term model yielded no statistically significant differences e.g., $p \approx 0.374$ for 4, 7, and 9 terms), indicating statistical equivalence. Considering parsimony and stability, the three-term model was selected as the preferred configuration, since adding further terms does not provide statistically reliable performance gains while increasing model complexity.

Table 25 summarizes the selected terms of the identified NARX model $\hat{y}(k)$, along with their estimated parameters and relevance scores, following the methodology described in Section 3.2 and implemented via Algorithm 2. These terms represent the most influential regressors selected for the final model, whose output is given by Equation 4.24. The model's structure reflects a sparse and interpretable representation, where each term captures essential nonlinear interactions among delayed input signals:

Figure 21 – Average accuracy calculated during the model's term selection phase for the Wave dataset, enabling the identification of the optimal number of terms for constructing the final model with maximum performance.



Source: created by the author. (2023).

Table 24 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 3-term model.

# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
2	0.8600	0.0200	3.68e-4	7	0.9960	0.0067	0.3739
3	0.9950	0.0089	—	8	0.9950	0.0089	1.0000
4	0.9970	0.0045	0.3739	9	0.9970	0.0045	0.3739
5	0.9950	0.0089	1.0000	10	0.9950	0.0089	1.0000
6	0.9950	0.0089	1.0000				

$$\begin{aligned}\hat{y}(k) = &0.3529 - 0.2652\,u_5(k-7) + 0.2781\,u_1(k-2)u_4(k-4) \\ &+ 0.1412\,u_4(k-5)u_9(k-6).\end{aligned} \tag{4.24}$$

To better interpret these terms, Table 26 maps each component of the model to its corresponding cumulant order and delay. This relationship reinforces the contribution of higher-order statistics, particularly second and fourth-order cumulants, in distinguishing power quality events. The identified delays also suggest that both short-term and long-range dependencies play a role in shaping the model's decision function.

Table 25 – Identified NARX model $\hat{y}(k)$, including the selected model terms, their estimated parameters, and corresponding relevance scores based on correlation analysis.

Model Terms	Parameter	Score	Model Terms	Parameter	Score
constant	0.3529	0.8501	$u_1(k-2)u_4(k-4)$	0.2781	0.8658
$u_5(k-7)$	-0.2652	0.8314	$u_4(k-5)u_9(k-6)$	0.1412	0.5854

Table 26 – Mapping of selected model terms to their corresponding cumulant orders and time delays, providing insight into the temporal structure and statistical characteristics of the input features.

Model Term	Cumulant Type(s)	Cumulant Delay(s)
$u_5(k-7)$	Second-order	4
$u_1(k-2) \cdot u_4(k-4)$	Second-order $\times$ Second-order	3 and 40
$u_4(k-5) \cdot u_9(k-6)$	Second-order $\times$ Fourth-order	40 and 33

The effectiveness of the proposed algorithm was assessed through a comparative analysis using 5-fold cross-validation on a validation dataset of $k = 1250$ samples. The benchmark models include Random Forest (RF) with 500 trees and the Gini index, Support Vector Machines (SVM) with a radial basis function kernel, and K-Nearest Neighbors (KNN).

Table 27 presents the performance metrics for each method. The proposed model achieves results comparable to SVM, particularly in terms of accuracy, sensitivity, and F1 score, while outperforming RF and KNN across all evaluated metrics. Although SVM slightly surpasses the proposed method in some measures, the Logistic-NARX model offers the added benefit of interpretability, allowing insights into the influence of individual regressors, a distinct advantage in applications requiring model transparency.

Table 27 – Performance comparison among different classification methods, based on evaluation metrics derived from confusion matrices.

	Proposed	RF	SVM	KNN
Accuracy	0.9976	0.9936	0.9984	0.9944
Sensitivity	0.9976	0.9936	0.9984	0.9944
Specificity	0.9994	0.9984	0.9996	0.9986
Precision	0.9976	0.9936	0.9984	0.9944
F1 Score	0.9976	0.9936	0.9984	0.9944

4.5 DISCUSSION

The application of the Logistic-NARX Multinomial model to the Power Quality (PQ) event classification problem confirmed the method's capacity to achieve high classification accuracy while maintaining a transparent and parsimonious model structure. The method effectively leveraged nonlinear and time-delayed interactions among selected input features, with only a small number of terms required to reach competitive performance.

The integration of Higher-Order Statistics (HOS) for feature extraction, followed by Fisher-based feature selection, proved essential in reducing dimensionality and enhancing class separability. As visualized in Figure 20, the selected second and fourth-order cumulants allowed for clear discrimination between different PQ events, supporting the construction of a model that balances accuracy and interpretability.

The comparative results presented in Table 27 indicate that while the SVM slightly outperformed the proposed model in some metrics, the Logistic-NARX method remained highly competitive and offered the added benefit of interpretability. The sparse model structure presented in Equation 4.24 enables domain experts to understand which signal characteristics and delays contribute most significantly to classification decisions—an essential feature in power system diagnostics.

Despite the favorable results, it is important to recognize the controlled nature of the study. The dataset was balanced, noise-free, and based on simulated events, which simplifies the classification task. Real-world applications may introduce challenges such as overlapping disturbances, noisy signals, and class imbalance (Chawla et al., 2002; Kuhn and Johnson, 2013). Therefore, future work should consider expanding the method's application to more complex and imbalanced scenarios, possibly integrating cost-sensitive learning or ensemble techniques to enhance robustness.

4.6 SUMMARY

This chapter presented the application of the Logistic-NARX Multinomial model to the classification of Power Quality (PQ) disturbances. By combining signal preprocessing, higher-order statistical feature extraction, and term selection via the Fisher criterion, the model was able to achieve high accuracy with minimal complexity. The results demonstrated that the proposed approach not only competes with established classifiers such as SVM and Random Forest but also provides the advantage of interpretability, essential for engineering applications where understanding model decisions is crucial. The findings reinforce the suitability of the NARX-based framework for dynamic signal classification tasks in power systems.

5 RAILWAY TRACK RISK ASSESSMENT

This section applies the proposed Logistic-NARX Multinomial method to assess operational risks in railway tracks by classifying critical conditions using dynamic data from rail vehicles. The approach integrates multibody dynamic simulation, feature extraction, and machine learning to detect track defects indirectly, offering a novel interpretable alternative to traditional inspection methods. The results highlight the model's predictive accuracy and potential to enhance georeferenced maintenance planning in railway infrastructure.

5.1 INTRODUCTION

Railways remain a critical component of modern transportation infrastructure, offering resilience and efficiency over long distances. Despite their inherent reliability, rail networks face increasing operational demands that accelerate track degradation, especially under higher speeds and axle loads. Track geometry faults, which contribute to a significant share of rail accidents, necessitate frequent and accurate monitoring to ensure safety and performance (Lasisi and Attoh-Okine, 2018; Koohmishi et al., 2024).

Traditional inspection methods rely heavily on specialized track geometry vehicles, which are costly, require trained personnel, and limit inspection frequency, particularly for privately operated or low-budget railways (Wang et al., 2021). In response to these constraints, recent research has explored the use of onboard sensors and inverse dynamic models to infer track conditions indirectly. These techniques leverage vehicle-mounted accelerometers and numerical models to estimate critical interaction forces between wheels and rails, as outlined by Sun et al. (2015); Barbosa (2016); Karis et al. (2018).

Machine learning has emerged as a powerful ally in this context, enabling the extraction of meaningful insights from high-frequency sensor data. Studies have demonstrated the effectiveness of data-driven methods, such as regression models, neural networks, and deep learning, for estimating wheel–rail forces and detecting track defects (Malekjafarian et al., 2023; Mosleh et al., 2023; Bhat et al., 2023; Sun et al., 2024; Marasco et al., 2024). When combined with multibody simulations, these techniques can generate synthetic datasets that mirror realistic track-vehicle interactions under diverse scenarios.

This chapter contributes to the thesis by presenting an integrated methodology that combines dynamic multibody simulation with machine learning to assess track quality. Introduces a novel classification-based approach for estimating operational risks using acceleration data from instrumented rail vehicles (IRVs). Beyond showcasing the effectiveness of the Logistic-NARX model in this application, this section offers an innovative framework for the railway sector, supporting indirect georeferenced monitoring and enabling more proactive infrastructure management.

5.2 RAILWAY TRACK SAFETY AND STABILITY

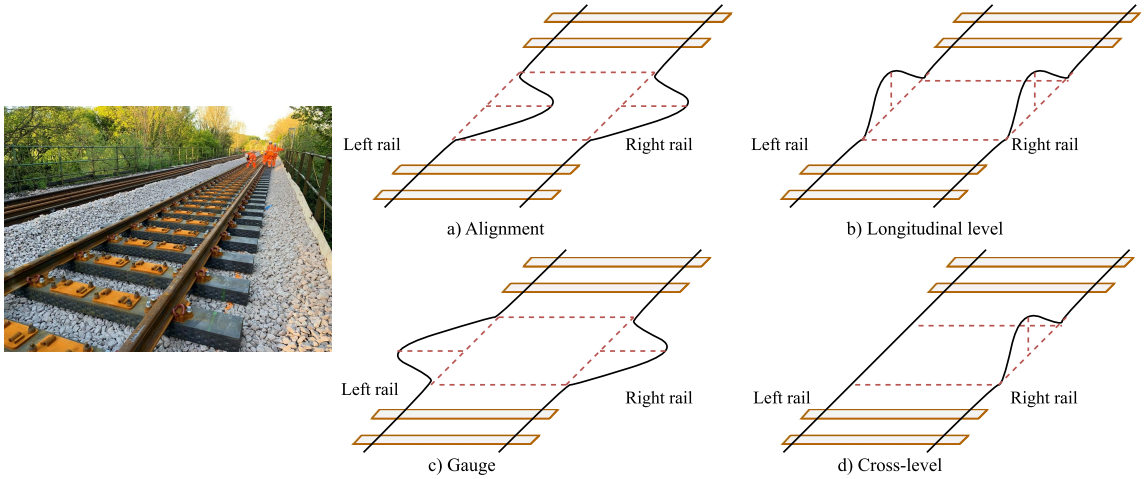
5.2.1 Track Geometry Safety Standards

Track geometry refers to the spatial configuration and alignment of railroad tracks, serving as a fundamental parameter for ensuring safe and efficient train operations. For decades, it has been the primary indicator used to monitor infrastructure condition and maintain operational reliability. Over time, however, track geometry deteriorates due to environmental factors, cyclic loading from train passages, and substructure settlement. As the severity of geometric deviations increases, so does the risk of operational issues such as derailments, excessive vibrations, and uneven wear of vehicle components.

Track irregularities emerge as physical manifestations of substructure and superstructure degradation, leading to measurable deviations in parameters such as alignment, longitudinal level, gauge, and cross-level. These irregularities can negatively impact vehicle stability and ride quality. Figure 22 illustrates typical examples of such defects.

Vehicle dynamics, being highly sensitive to track condition, offer an indirect yet informative means of evaluating infrastructure integrity. Two key indicators frequently used in this context are wheel unloading and the lateral-to-vertical force ratio (L/V). Monitoring these dynamic responses enables the identification of critical zones where elevated accelerations suggest structural weaknesses, guiding targeted inspections and preventive maintenance actions to mitigate safety risks and ensure operational continuity.

Figure 22 – Illustration of track irregularities, including: (a) alignment, (b) longitudinal level, (c) gauge, and (d) cross-level.

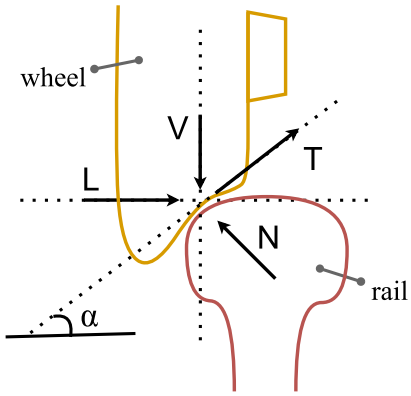


Source: created by the author. (2024).

5.2.2 Lateral-to-Vertical Force Ratio (L/V)

One of the most critical parameters for ensuring rail vehicle safety and preventing derailments is the ratio between lateral and vertical forces at the wheel–rail contact interface. This ratio is commonly assessed through the contact plane angle, which delineates the stability threshold of the interface forces, as illustrated in Figure 23.

Figure 23 – Representation of the forces acting on the wheel-rail interface, illustrating the geometric relationships of force components projected onto the contact plane.



Source: created by the author. (2023).

This stability boundary is defined by the geometric relationship between the forces projected onto the wheel–rail contact plane, as formulated by Nadal (1908):

$$\frac{\mathrm{L}}{\mathrm{V}} = \frac{\tan(\alpha) - \mu}{1 + \mu\tan(\mu)}, \qquad \mathrm{T} = \mu\mathrm{N}, \tag{5.1}$$

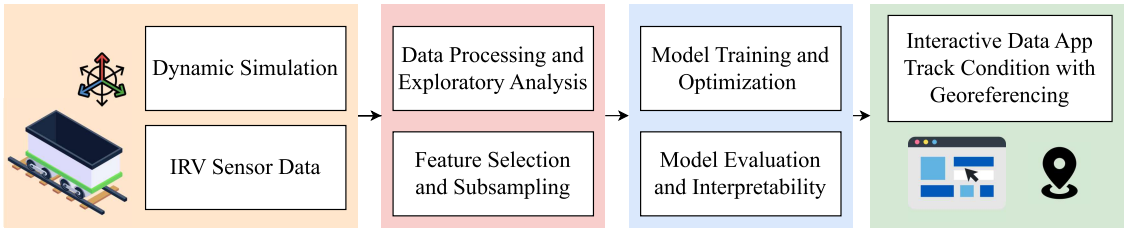
where L and V represent the lateral and vertical forces at the contact interface, α is the angle of the contact plane, μ is the friction coefficient between the contacting surfaces, T denotes the tangential force, and N is the normal force. This expression is one of the most widely adopted in the railway sector, serving as a reference for defining safety thresholds and critical values of the derailment coefficient.

Another common derailment mechanism, particularly on straight track sections—is associated with vertical force reduction, known as wheel unloading. In such cases, a drop in vertical load V can cause the L/V ratio to rise sharply, even if lateral forces remain constant. Wheel unloading is typically caused by rigid body dynamics, which alter the distribution of vertical loads across the wagon. When the unloading rate becomes significant, the wheel may lift off the rail, leading to an immediate loss of contact and an increased risk of derailment.

5.3 CLASSIFICATION-BASED TRACK CONDITION

This study proposes a data-driven methodology for assessing railway track conditions using acceleration and angular velocity measurements. An overview of the track condition monitoring framework is presented in Figure 24. The core of the approach lies in the development of a classification model, trained to distinguish between predefined safety-critical classes based on established operational limits. The dataset used to train and validate the model was constructed by integrating dynamic multibody simulations with real-world measurements collected from an instrumented rail vehicle (IRV), enabling the generation of realistic and representative scenarios for model learning.

Figure 24 – Overview of the track condition monitoring system, illustrating the structure and workflow of the data-driven classification model.



Source: created by the author. (2023).

5.3.1 Multibody Dynamic Simulation

The raw dataset was generated using VAMPIRE, a widely adopted software in the railway industry for dynamic multibody simulations. Accurate modeling of vehicle–track interaction requires several parameters, including wheel and rail profiles, speed profiles, vehicle dynamics, and geometric track irregularities. In this study, real track geometry data was integrated into the simulation to replicate the dynamic behavior of an Instrumented Railway Vehicle (IRV).

Table 28 summarizes the generated dataset, which includes input features such as accelerations and carbody modes computed at key vehicle components: the carbody, leading bogie, and trailing bogie. The output variables correspond to normalized amplitudes of the Lateral-to-Vertical Force Ratio (L/V) and wheel unloading, both critical indicators of operational risk. The simulated responses were validated against real-world measurements, confirming the model's accuracy. This digital twin of the IRV is well-established and currently utilized in industry for condition monitoring and predictive maintenance planning.

Table 29 describes the parameters considered in the creation of the dynamic

Table 28 – Dataset generated by the model, comprising input features and corresponding target response values.

Input Features	
Timestamp	Data time in seconds (s).
Distance_km	Distance kilometers (Km).
Type_load	Wagon empty or loaded.
Section	Locations railway network.
Speed	Speed scales in (Km/h).
Acel_lat_cbd, Acel_vert_cbd, Roll_cbd, Yaw_cbd, Pitch_cbd	Lateral e Vertical acceleration; roll, pitch, yaw motion; positioned in the carbody.
Acel_lat_lead, Acel_vert_lead, Roll_lead, Yaw_lead, Pitch_lead	Lateral e vertical acceleration; roll, pitch, yaw motion in leading bogies.
Acel_lat_trail, Acel_vert_trail, Roll_trail, Yaw_trail, Pitch_trail	Lateral e Vertical acceleration; roll, pitch, yaw motion in trailing bogies.
Target(labels)	
unloading_ratio, lv_ratio	Unloading and lateral-to-vertical force Ratio.

multibody model, with the current methodology applied to HPD vehicle[1]. The generated model is calibrated using real data obtained from an instrumented HPD vehicle (see figure 25).

Table 29 – Parameters considered in the creation of the dynamic multibody model.

Parameter	**Description**
Track gauge	1000 mm
Speed range	25 km/h to 70 km/h
Train composition	90 wagons, distributed power: (2 locomotives, 45 wagons, 2 locomotives, 45 wagons, 1 locomotive)
Gross tonnage per train	7000 gross tons
Curve radius	Ranging from 75 meters (23.5 degrees) to 2000 meters (0.89 degrees). Predominantly sharp curves with radius less than 150 meters (11.7 degrees) in 35% of the corridor length
Load per axle	25 tons per axle (metric gauge)
Vehicle	HPD modeled and instrumented. Load per axle of 20 tons. Old Ride Control bogie with constant damping. Lateral block bearings. Center of gravity at 2000 mm. Wheel profile type (AAR 1B NF)

[1] HPD vehicle are recognized worldwide for the transport of grain, sugar, corn, soybeans and other commodities.

Figure 25 – Railcar of the HPD (Hoppers) type.

Source: created by (Smit, 2023).

5.3.2 Instrumented Railway Vehicle

The dataset used in this study was acquired from the instrumentation of the HPD rail vehicle, which is equipped with accelerometers and gyroscopes to measure linear acceleration and angular velocity (see Figure 26). The system also records GPS coordinates, vehicle speed, and timestamps. Sensor modules were strategically installed on three key locations of the vehicle: the carbody, the leading bogie, and the trailing bogie.

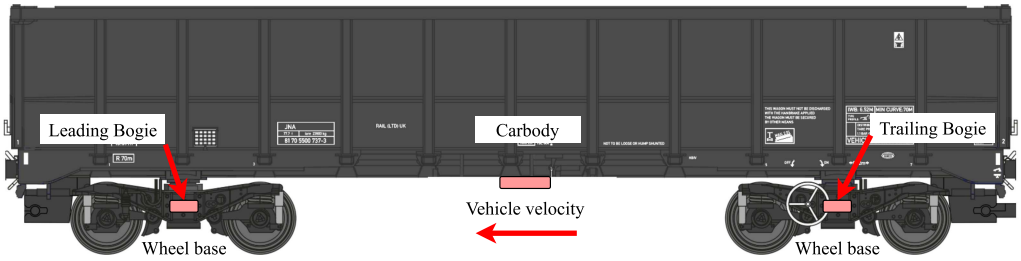
Figure 26 – Inertial sensor modules installed on the railway vehicle, including placements on the carbody, leading bogie, and trailing bogie.

Source: Created by the author (2023).

Sensor placement was determined based on dynamic principles to ensure comprehensive capture of the vehicle's operational behavior. The carbody provides integrated responses from both primary and secondary suspensions, enabling global analysis of ride

quality and structural dynamics. The bogies, on the other hand, are essential for detecting localized effects such as lateral dynamics during curve negotiation and identifying asymmetries caused by wheel or suspension defects. Figure 27 illustrates the sensor layout across the vehicle structure.

Figure 27 – Sensor module placement on the rail vehicle, capturing dynamic responses from the carbody and bogies.



Source: created by the author. (2024).

Data acquisition and processing followed the standards defined by the Federal Railroad Administration (FRA) (Transportation Research Board, 2020), which recommend a minimum sampling rate of 100 Hz for acceleration data. In this study, a sampling rate of 200 Hz was adopted for both simulated and real measurements to ensure accurate frequency resolution.

To enhance signal integrity and suppress noise, a preprocessing stage was applied. Moving average filters were used on angular motion data (roll, pitch, and yaw), as well as on critical output indicators such as wheel unloading and Lateral-to-Vertical Force Ratio (L/V), improving the quality of criticality mapping. Furthermore, because vibration components in the 2 Hz to 10 Hz range are closely associated with instability phenomena, a second-order Butterworth low-pass filter with a 10 Hz cutoff frequency was applied to acceleration signals. Forward-backward filtering was used to minimize phase distortion and temporal lag (Gustafsson, 1996).

5.3.3 Analysis of Critical Safety Limits

Railway safety standards rely on key indicators such as the Lateral-to-Vertical Force Ratio (L/V) and wheel unloading to assess operational risk and track condition. These metrics are fundamental for identifying dynamic instabilities and ensuring compliance with established safety protocols, including those set by the Federal Railroad Administration (FRA) (Transportation Research Board, 2020).

In this study, safety criteria defined by engineers from a Brazilian railway operator were adopted. While the critical thresholds match those established by the FRA, additional risk bands were introduced to enhance predictive capabilities. These extended classifications

incorporate more conservative thresholds for L/V and wheel unloading, enabling the early identification of potentially hazardous track sections.

Based on these criteria, the dataset was labeled according to four severity levels: Normal (no risk), P2 (low risk), P1 (moderate risk), and P0 (high risk). Table 30 details these thresholds, illustrating how this classification framework supports proactive maintenance planning and a more granular evaluation of track health.

Table 30 – Analysis of critical safety limits. This table categorizes track conditions into ascending levels of severity based on predefined safety standards, relating to wheel unloading and the lateral-to-vertical force ratio (L/V) amplitude (%) for both empty and loaded vehicle conditions.

		L/V	
	Unloading	loaded	empty
N	$x \leq 0.5$	$x \leq 0.6$	$x \leq 0.6$
P2	$0.5 \leq x < 0.6$	–	$0.6 \leq x < 0.8$
P1	$0.6 \leq x < 0.85$	$0.6 \leq x < 0.8$	$0.8 \leq x < 1.0$
P0	$x \geq 0.85$	$0.8 \leq x < 1.0$	$x \geq 1.0$

The analysis incorporates distance-based windows to evaluate safety-critical thresholds more conservatively. Specifically, a sliding window of 1.5 m is applied, and when multiple criticality levels are detected within this span, the most severe class is assigned, provided it occupies at least 25 % of the window length.

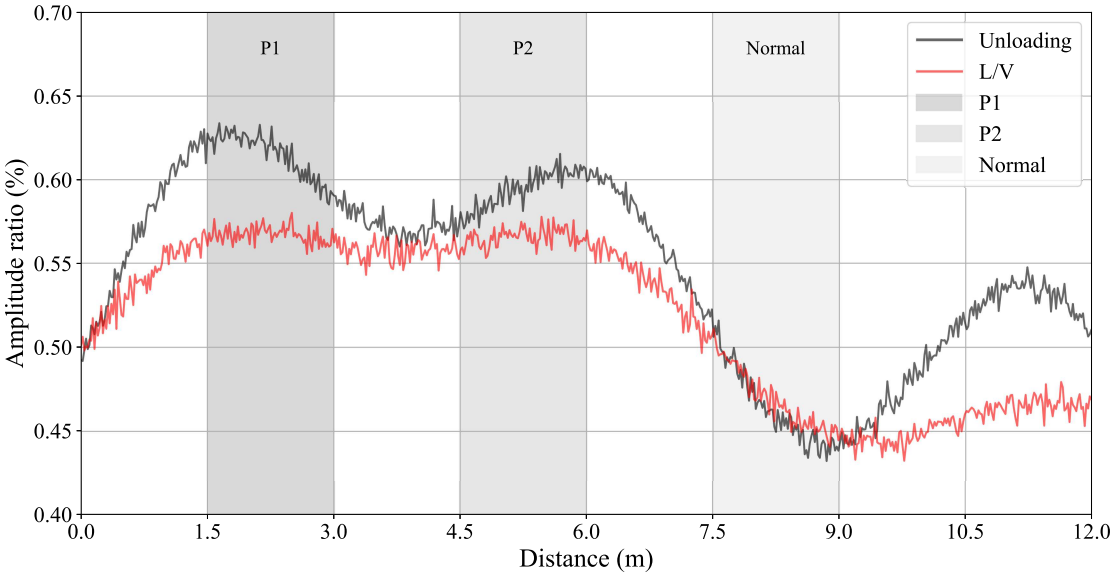
Figure 28 illustrates the amplitude signals of the Lateral-to-Vertical Force Ratio (L/V) and wheel unloading, which serve as target labels. It also highlights how critical events are detected within the specified spatial window. This strategy ensures that transient peaks or brief threshold crossings are not overlooked, reinforcing the reliability of class labeling for the supervised learning framework. Ultimately, these parameters enable accurate annotation of track conditions across the four severity classes used in the classification model.

5.3.4 Feature Selection and Subsampling

This study employed a structured feature selection strategy to enhance both the interpretability and performance of the classification model by reducing dimensionality and eliminating irrelevant or redundant variables. The following methods were applied to identify the most informative features while minimizing noise and redundancy:

- Pearson correlation matrix: used to detect and eliminate highly correlated features, reducing redundancy without significant information loss (Guyon and Elisseeff, 2003);
- Constant feature removal: features with low variance were excluded, as they contribute minimally to predictive performance;

Figure 28 – Amplitude of the Lateral to Vertical Force Ratio (L/V) and Wheel Unloading Signals. The figure displays the amplitudes of the L/V force ratio and wheel unloading signals, which serve as targets or labels. It also highlights critical situations occurring within the specified windows at a distance of 1.5 m.



Source: created by the author. (2024).

- Boruta algorithm: a robust wrapper method that compares real features to randomized shadow features, ensuring the selection of truly relevant variables (Kursa and Rudnicki, 2010);
- Recursive Feature Elimination (RFE): iteratively removes less important features and uses cross-validation to determine the optimal subset that balances accuracy and model simplicity (Guyon et al., 2002).

Table 31 summarizes the features retained and discarded during this process. Notably, due to strong correlations among dynamic signals, only acceleration features measured on the carbody and bogies of the trailing vehicle were retained. This outcome highlights the importance of sensor placement in capturing representative dynamic responses for effective classification.

Table 31 – Selected and removed features during the feature selection process.

Removed Features	Selected Features
Roll_lead, Roll_trail, Roll_cbd, Yaw_lead, Yaw_trail, Acel_lat_lead, Acel_vert_lead, Pitch_lead, Pitch_trail, Pitch_cbd	Acel_lat_cbd, Acel_lat_trail, Acel_vert_cbd, Acel_vert_trail, Yaw_cbd, Type_load, Section, Speed

Due to the nature of railway track monitoring, the collected dataset exhibits a significant class imbalance: normal track conditions (majority class) are substantially

overrepresented compared to samples indicating track geometry defects (minority class). This imbalance can negatively impact model performance, as standard machine learning algorithms tend to favor the majority class, often resulting in poor detection of rare but critical defect instances.

To mitigate this issue, we applied the One-Sided Selection (OSS) technique (Kubat and Matwin, 2000), a data-level undersampling method that reduces the size of the majority class by removing redundant or non-informative instances. Crucially, OSS preserves minority class samples, which are essential for learning to recognize rare fault conditions. This method is particularly suited to the application at hand, as it improves the classifier's ability to detect significant anomalies while minimizing bias toward normal operating conditions. Table 32 presents the post-processing class distribution, demonstrating a more balanced dataset that maintains representativeness without excessively reducing the majority class.

Table 32 – Class distribution before and after applying the One-Sided Selection undersampling technique, showing the frequency (%) and number of samples per class.

	Before		After	
	(%)	Samples	(%)	Samples
N	97.2	5.96×10^6	54.3	198953
P2	1.44	88427	23.9	88427
P1	1.05	64478	17.4	64478
P0	0.25	15352	4.19	15352

5.4 LOGISTIC-NARX MULTINOMIAL MODEL APPROACH

To evaluate the applicability and effectiveness of the proposed Logistic-NARX Multinomial classification method in a real-world engineering context, the model is applied to the problem of railway track condition assessment based on dynamic responses from rail vehicles. This case study not only tests the model's predictive performance but also highlights its capacity to produce interpretable results in a complex operational environment.

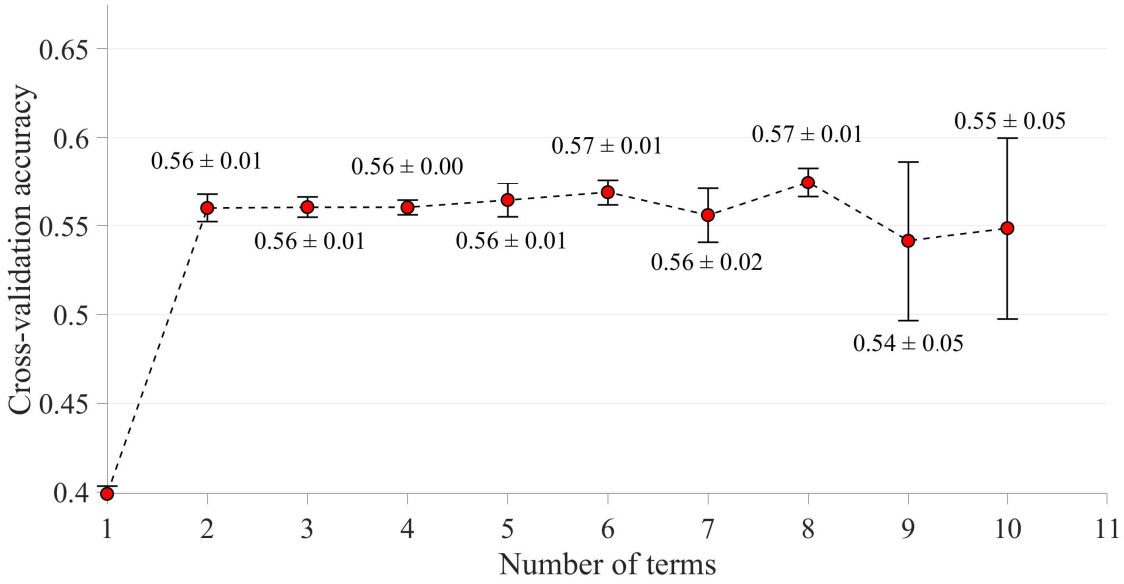
The NARX-based method is benchmarked against conventional classification algorithms, with all models implemented in MATLAB to ensure consistency. Differing from the earlier Power Quality application, the current implementation adopts parameters tailored to the railway domain, including a nonlinearity degree of $l = 2$ and a large candidate pool of 13041 regressors (2.2). Model evaluation is conducted using 5-fold cross-validation to ensure reliable and generalizable performance estimates. This analysis provides a robust validation of the proposed method's potential for supporting data-driven decision-making in railway maintenance and safety management.

The cross-validation accuracy curve for this case study (see Figure 29) shows a steep improvement from one to two terms, followed by relatively stable performance across subsequent model complexities. The maximum average accuracy was obtained with eight terms (0.5746 ± 0.0084), although several configurations achieved comparable results within the margin of error.

Paired t-tests were conducted to compare each configuration against the six-term model (see Table 33). The results indicate that most differences were not statistically significant at the 5% threshold, suggesting that increasing the number of terms beyond six yields only marginal benefits in predictive performance. Furthermore, configurations with fewer than six terms, while simpler, generally displayed slightly reduced accuracy.

Considering the trade-off between model complexity, statistical equivalence, and predictive accuracy, the six-term model was selected as the preferred configuration for this case study. This choice balances parsimony with robust performance, avoiding unnecessary complexity while maintaining accuracy within the range of the best-performing models. The details of the identified NARX model, represented by $\hat{y}(k)$, including the terms, correlation values, and corresponding parameters, are comprehensively presented in Table 34.

Figure 29 – The average accuracy calculated during the term selection phase of the model for the railway dataset classification, enabling identification of the optimal number of terms for constructing the final model with the highest performance.



Source: created by the author. (2024).

Table 35 presents the selected model terms alongside their corresponding input features, highlighting key interactions identified by the Logistic-NARX Multinomial

Table 33 – Cross-validation mean accuracy, standard deviation, and p-values from paired t-tests comparing each configuration to the 6-term model.

# Terms	Mean	STD	p-value	# Terms	Mean	STD	p-value
1	0.3989	0.0044	2×10^{-6}	6	0.5645	0.0095	-
2	0.5600	0.0077	0.1873	7	0.5558	0.0155	0.0935
3	0.5605	0.0057	0.1715	8	0.5746	0.0084	0.1621
4	0.5603	0.0041	0.0369	9	0.5215	0.0507	0.1517
5	0.5645	0.0095	0.4017	10	0.5485	0.0501	0.3627

model. This mapping reinforces the interpretability of the model by showing how specific variables—such as lateral accelerations, yaw, and track sections—interact to influence the output classification.

Table 34 – Identified NARX model $\hat{y}(k)$, including the selected model terms, their estimated parameters, and corresponding correlation coefficients.

Model Terms	Parameter	Score	Model Terms	Parameter	Score
$u_6(k-1)u_6(k-1)$	0.6082	0.8043	$u_5(k-8)u_5(k-9)$	-0.1219	0.4560
$u_6(k-5)u_{11}(k-2)$	-0.0025	0.4010	$u_6(k-2)u_{15}(k-5)$	-0.0028	0.4035
$u_7(k-9)u_{12}(k-6)$	0.0334	0.3835	$u_6(k-4)u_{13}(k-1)$	-0.0179	0.3805
$u_1(k-5)u_2(k-7)$	-0.0014	0.3956	$u_2(k-9)u_5(k-9)$	0.0263	0.3750

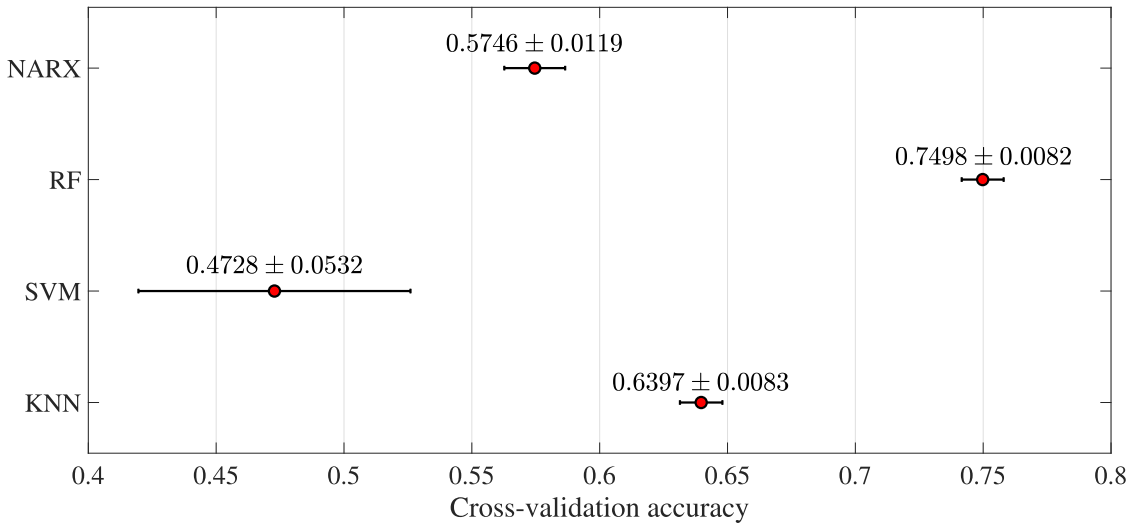
Table 35 – Identified NARX model $\hat{y}(k)$, including the selected model terms along with the corresponding input feature names for interpretability.

Model Terms	Feature
$u_6(k-1)u_6(k-1)$	Type_load × Type_load
$u_5(k-8)u_5(k-9)$	Yaw_cbd × Yaw_cbd
$u_6(k-5)u_{11}(k-2)$	Type_load × Section
$u_6(k-2)u_{15}(k-5)$	Type_load × Section
$u_7(k-9)u_{12}(k-6)$	Section × Section
$u_6(k-4)u_{13}(k-1)$	Type_load × Section
$u_1(k-5)u_2(k-7)$	Accel_lat_cbd × Accel_lat_trail
$u_2(k-9)u_5(k-9)$	Accel_lat_trail × Yaw_cbd

The performance of the proposed method was benchmarked against other commonly used classification techniques. Figure 30 displays the accuracy results as intervals obtained through cross-validation. While the proposed Logistic-NARX Multinomial model achieved

an average accuracy of 57 %, outperforming the SVM approach, it remained below the performance of the Random Forest (RF) model, which reached an accuracy of 75 %.

Figure 30 – An examination of accuracy results for different classification methods applied to the railway dataset. Cross-validated accuracy is presented at regular intervals, along with their corresponding average values.



Source: created by the author. (2023).

The classification performance of the proposed Logistic-NARX Multinomial method was compared against Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), as summarized in the confusion matrices (Table 36) and performance metrics (Table 37). Among the evaluated methods, Random Forest consistently outperformed the others across all key metrics, achieving the highest accuracy (74.98%), sensitivity (75.29%), precision (74.98%), and F1 score (75.08%). In contrast, the proposed NARX-based model showed limited effectiveness in this application, with lower performance particularly in distinguishing between intermediate classes. Given these results, Random Forest was selected for further refinement through hyperparameter optimization, as it demonstrated the best overall balance between predictive power and class discrimination in the railway track condition assessment task.

5.5 MODEL SEARCH AND OPTIMIZATION

In this section, a machine learning model is developed to classify safety-critical railway track conditions using features derived from Multibody Dynamic Simulations and data from an Instrumented Railway Vehicle (IRV). The objective is to estimate the probability of exceeding operational safety thresholds.

Table 36 – Confusion matrix representing the performance of the classification methods. The matrices were generated by summing each partition of the validation sets generated in cross-validation. The indicated classes are P0 (C1), P1 (C2), P2 (C3), Normal (C4).

	C1	C2	C3	C4
C1	49.6%	14.4%	30.6%	5.4%
C2	29.8%	49.9%	15.9%	4.4%
C3	22.1%	10.0%	61.2%	6.7%
C4	14.0%	11.7%	8.8%	65.5%

(a) Logist-NARX Multiclass.

	C1	C2	C3	C4
C1	73.5%	14.3%	8.9%	3.3%
C2	13.2%	69.1%	14.5%	3.2%
C3	8.5%	14.2%	68.5%	8.8%
C4	2.0%	3.9%	4.1%	90.0%

(b) Random Forests.

	C1	C2	C3	C4
C1	82.1%	8.1%	5.7%	4.1%
C2	44.1%	36.7%	15.2%	4.1%
C3	28.5%	30.8%	36.8%	3.9%
C4	6.5%	10.8%	16.4%	66.3%

(c) Support Vector Machine.

	C1	C2	C3	C4
C1	69.8%	14.0%	10.8%	5.4%
C2	20.6%	58.2%	16.2%	5.0%
C3	19.0%	12.8%	56.8%	11.4%
C4	5.8%	11.5%	9.0%	73.7%

(d) K-Nearest Neighbors.

Table 37 – A performance comparison among different classification methods using various evaluation metrics derived from confusion matrices for the railway dataset.

	NARX	RF	SVM	KNN
Accuracy	0.5746	0.7498	0.4728	0.6397
Sensitivity	0.5656	0.7529	0.5546	0.6465
Specificity	0.8653	0.9165	0.8354	0.8808
Precision	0.5746	0.7498	0.4728	0.6397
F1 Score	0.5468	0.7508	0.4105	0.6378

To identify the most effective approach, Random Forest, XGBoost, and LightGBM (Hastie et al., 2009) were evaluated using 5-fold Stratified Cross-Validation. XGBoost and LightGBM were included for their strong performance in scenarios with class imbalance and structured inputs, such as time-series signals and categorical features representing sensor positions or event types. The top-performing model was further optimized using the GridSearchCV method (Pedregosa et al., 2011), which systematically searches for the best combination of hyperparameters to enhance classification accuracy and generalization.

5.5.1 Model Evaluation

This section presents the performance evaluation of classification models based on multiple metrics. As shown in Table 38, the results demonstrate that Random Forest (RF) outperforms the other models across all evaluation criteria. Specifically, it achieved the highest accuracy (0.9311), precision (0.9428), recall (0.9012), and F1-score (0.9231), indicating superior performance in predicting the likelihood of exceeding critical safety thresholds. Based on these results, Random Forest was selected as the optimal model for hyperparameter tuning.

Table 38 – Comparison of classification methods based on average metrics and standard deviations obtained using Stratified K-Fold Cross-Validation.

	Accuracy	Precision	Recall	F1-Score
KNN	0.8331± 0.0007	0.8461± 0.0006	0.8442± 0.0007	0.8140± 0.0007
RF	0.9311± 0.0008	0.9428± 0.0007	0.9012± 0.0008	0.9231± 0.0008
XGBoost	0.8063± 0.0010	0.8263± 0.0012	0.8563± 0.0010	0.8132± 0.0010
LightGBM	0.7595± 0.0008	0.7788± 0.0008	0.7295± 0.0008	0.7957± 0.0008

5.5.2 Random Forest with Hyperparameter Tuning

The Random Forest (RF) algorithm was selected not only for its superior performance in initial evaluations but also for its robustness in handling high-dimensional, heterogeneous datasets, characteristics typical in railway track quality assessment. Its ability to model non-linear relationships and effectively process both numerical and categorical inputs makes it well-suited for integrating multisensor data and operational conditions in this domain.

To further enhance its performance, hyperparameter optimization was conducted using the GridSearchCV method. This technique systematically evaluates combinations of key parameters, such as the number of estimators (trees), maximum tree depth, and minimum samples required to split a node, to identify the configuration that maximizes model accuracy and generalization.

5.5.3 Results

Table 39 presents the performance metrics, precision, recall (sensitivity), and F1-score, for each class in the track condition classification. The classes are ordered by increasing severity, from Normal (non-critical) to P0 (most critical). In this context, sensitivity is the most relevant metric, as the primary objective is to detect all critical conditions. From a conservative safety standpoint, it is preferable to identify as many critical cases (P0, P1, P2) as possible, even if this leads to some false positives.

The model demonstrates high sensitivity for the critical classes: 0.7428 for P0, 0.9074 for P1, and 0.9317 for P2. In particular, correctly detecting P0 conditions, those representing the highest risk, is essential, and a sensitivity near 75% indicates the model captures most severe faults. However, approximately 12% of P0 cases are still misclassified as less severe, which may underestimate the risk in some instances.

Table 39 – General report of performance metrics for the validation of different classes, highlighting classification effectiveness across categories.

	Precision	Recall	F1-Score
P0	0.9281	0.7428	0.8251
P1	0.6097	0.9074	0.7293
P2	0.2471	0.9317	0.3906
Normal	0.9985	0.9525	0.9750
Accuracy	-	-	0.9512

Despite the model's high sensitivity, the precision for the intermediate classes P1 and P2 remains relatively low, 0.6097 and 0.2471, respectively, indicating a notable incidence of false positives. This suggests a tendency to overestimate the severity of certain track segments, potentially leading to unnecessary maintenance interventions. However, in safety-critical contexts like railway infrastructure, such conservatism may be preferable to the risk of undetected defects.

The F1-score, which harmonizes precision and recall, is highest for the P0 class (0.8251), reinforcing the model's reliability in identifying the most hazardous scenarios. The Normal class also performs exceptionally well across all evaluated metrics, effectively minimizing false alarms in healthy track segments.

As illustrated by the confusion matrices in Table 40, the model shows strong classification capability, with high values along the diagonal indicating accurate predictions. Misclassifications primarily occur between P1 and P2, highlighting potential ambiguity between these intermediate conditions. This suggests that further refinement in feature selection or threshold calibration could enhance class separation.

Complementing this analysis, the feature importance results shown in Figure 31, derived from the Random Forest model, reveal the most relevant variables contributing to the prediction of critical track states. These insights provide valuable guidance for future model enhancements and targeted monitoring strategies.

5.5.4 Data App Track Condition with Georeferencing

The interactive application developed in this study provides a practical decision-support tool for railroad specialists engaged in assessing track conditions and operational safety. As illustrated in Figure 32, the application enables dynamic visualization of geo-

Table 40 – Confusion matrix based on the metrics of sensitivity and precision using validation data. The indicated classes are P0 (C1), P1 (C2), P2 (C3), Normal (C4).

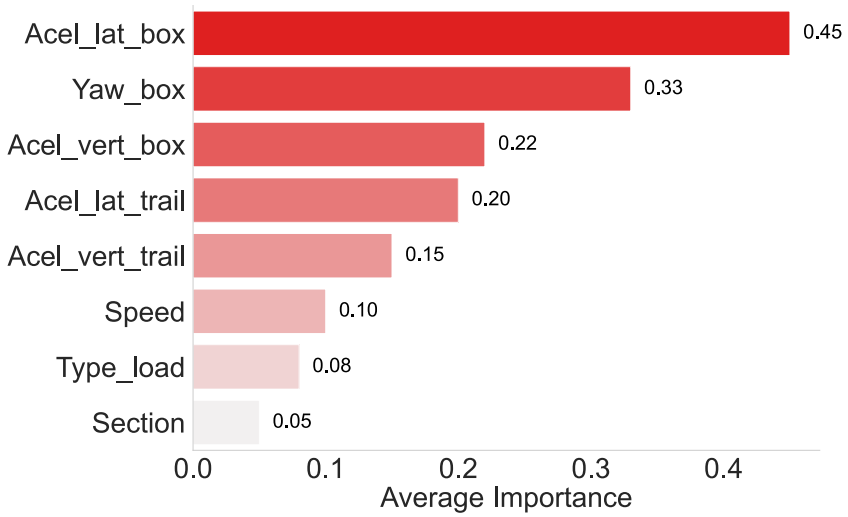
	Recall			
	P0	P1	P2	N
P0	0.75	0.085	0.043	0.12
P1	0.0031	0.91	0.053	0.035
P2	0.001	0.024	0.93	0.042
N	–	0.005	0.042	0.95
	Prediction label			

(a) Sensitivity.

	Precision			
	P0	P1	P2	N
P0	0.93	0.014	0.002	–
P1	0.016	0.61	0.01	–
P2	0.007	0.023	0.25	–
N	0.047	0.35	0.74	0.99
	Prediction label			

(b) Precision.

Figure 31 – Feature importance analysis illustrating the relative significance of each feature in the model's prediction process.

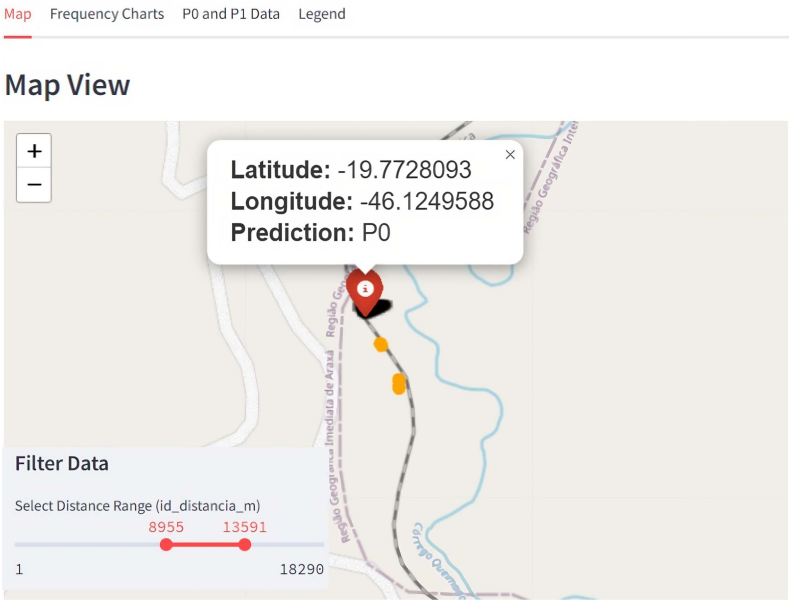


Source: created by the author. (2024).

referenced safety indices, allowing users to identify and monitor critical points along the railway network through an intuitive map interface.

Users can apply distance-based filters to focus on specific track segments and analyze detailed information through frequency charts, which reveal temporal trends in quality indicators. Additionally, the platform provides tabular summaries with the predicted probability of each criticality class and enables data export in CSV format for further offline analysis. This tool enhances the usability of the classification results derived from the methodology presented in this thesis, supporting proactive maintenance planning and informed decision-making in railway infrastructure management.

Figure 32 – Interactive map interface of the application, displaying georeferenced safety indices alongside filtering options for user-defined distance ranges.



Source: created by the author. (2024).

5.6 DISCUSSION

The application of the Logistic-NARX Multinomial model to the railway track condition assessment task highlighted both the potential and the limitations of interpretable nonlinear modeling in complex operational scenarios. Despite achieving only moderate predictive accuracy in its initial implementation, outperformed by more established classifiers such as Random Forest, the NARX-based method provided valuable insights into the dynamics of critical variables, reinforcing its suitability for exploratory modeling and feature interaction analysis.

Among these, lateral acceleration stands out due to its direct link with wheel unloading, a key factor in derailment scenarios. Given the vehicle's geometry, with a center of gravity approximately 2 meters above the rail and a nominal gauge of 1 meter, it is particularly prone to generating large overturning moments during curve negotiation. This effect is exacerbated under abrupt lateral accelerations, which can destabilize the vehicle. In such cases, the lateral-to-vertical force ratio (L/V) becomes critical. When this ratio nears or surpasses established safety limits (typically in the range of 0.8 to 1.0, depending on the standard), the risk of derailment increases significantly. This is especially true on curves with small radii, such as 100 meters, where lateral forces are inherently higher. The model's emphasis on lateral acceleration corroborates these dynamics, reinforcing its relevance as a primary indicator of operational safety in railway applications.

From an engineering perspective, the integration of multibody simulations and

machine learning represents a methodological advancement with tangible implications. The use of simulation-generated data allows for a realistic representation of operational dynamics without requiring direct measurements of wheel–rail forces, which are typically inaccessible in regular operation. This approach not only enhances the interpretability of the model but also facilitates risk assessment in contexts where direct instrumentation is impractical or cost-prohibitive.

5.7 SUMMARY

This chapter introduced a novel methodology for railway track condition assessment by integrating multibody dynamic simulations with machine learning models, specifically focusing on interpretable classification using acceleration and angular motion data. The proposed approach enables the estimation of critical operational risks, such as wheel unloading and L/V thresholds, through indirect measurements obtained from onboard sensors. While the Logistic-NARX model demonstrated limited accuracy compared to ensemble methods like Random Forest, it contributed valuable interpretability and insight into feature relevance. The methodology also culminated in the development of a geo-referenced application, providing real-time visualization and decision support for maintenance planning. Together, these contributions reinforce the practical viability and strategic benefits of combining physical modeling with data-driven approaches in railway infrastructure monitoring.

6 CONCLUSIONS

6.1 SUMMARY AND CONCLUSIONS

This thesis introduced a novel classification framework that integrates system identification principles with machine learning techniques, culminating in the development of the Logistic-NARX Multinomial model. This approach expands the classical use of NARX models beyond regression and binary classification by enabling their application to multiclass problems, while maintaining a sparse, interpretable, and transparent structure.

The proposed methodology combines the flexibility of logistic regression with the structural rigor of NARX modeling, incorporating an accuracy-driven term selection mechanism based on cross-validation. The result is a classification model capable of capturing nonlinear dependencies, highlighting meaningful variable interactions, and offering interpretable mathematical expressions that aid in understanding the underlying processes.

Experimental results on benchmark datasets confirmed that the model achieves competitive classification performance with significantly fewer terms, demonstrating its efficiency in dimensionality reduction and robustness in scenarios involving noise and class imbalance. In particular, the method excelled in providing interpretability and model transparency, features often lacking in black-box classifiers.

The methodology was also validated in two real-world engineering applications. In the case of power quality disturbance classification, the model successfully identified discriminative descriptors derived from higher-order statistics, effectively categorizing electrical events. This demonstrated the model's capacity to extract and prioritize relevant features even in highly nonlinear and transient contexts.

The second and more comprehensive application addressed railway track condition assessment. In this domain, the Logistic-NARX Multinomial model was embedded in a broader methodological pipeline involving multibody dynamics simulations, sensor-based data collection, feature selection, and georeferenced safety visualization. This application not only served as a validation case for the proposed model but also represented an innovative contribution to railway engineering: enabling the indirect estimation of critical safety indicators, such as lateral-to-vertical force ratio and wheel unloading, through an interpretable and scalable classification system. The methodology demonstrated potential for integration into intelligent maintenance strategies and decision-support tools for railway infrastructure monitoring.

Nonetheless, the research also revealed important limitations and opportunities for future improvement. Challenges such as basis function flexibility, lag selection, multicollinearity, and class imbalance emerged as critical factors influencing model performance. Ad-

dressing these issues—through alternative function bases, adaptive lag optimization, robust orthogonalization procedures, and advanced imbalance mitigation techniques—constitutes a promising direction for further refinement of the approach.

In conclusion, this work makes both methodological and applied contributions. It advances the field of interpretable classification by proposing a hybrid, sparse, and transparent model applicable to a wide range of engineering challenges. Simultaneously, it introduces an innovative methodology for railway safety assessment, bridging theoretical modeling with practical deployment in critical infrastructure. The proposed framework lays the groundwork for future applications where clarity, trust, and domain-aligned reasoning are essential to support real-time, data-driven decision-making.

6.2 FUTURE WORKS

This work opens several avenues for future research aimed at enhancing the performance, scalability, and applicability of the Logistic-NARX Multinomial model, both from a methodological and application-driven perspective.

One key direction concerns the refinement of the model's functional basis. The current use of polynomial basis functions, while offering analytical clarity and ease of interpretation, may limit the ability to capture more intricate nonlinearities present in complex classification tasks. Future research could explore the incorporation of alternative basis expansions—such as radial basis functions, wavelets, or kernel-based approaches—to increase the flexibility and representational power of the model without compromising its interpretability.

Another technical challenge involves the selection of input and output lags (n_u and n_y). This step significantly influences the search space of candidate terms. As lag values increase, the number of possible combinations grows exponentially, leading to a sharp increase in computational complexity (Wei et al., 2004). Developing adaptive or automated strategies for lag selection, possibly guided by relevance measures or mutual information, would enhance the scalability of the method, especially in high-dimensional scenarios.

The issue of multicollinearity among regressors also remains critical. Highly correlated variables can adversely affect coefficient estimation and compromise model robustness. Future work may investigate the adoption of enhanced orthogonalization techniques, such as iterative Orthogonal Forward Regression (OFR) (Guo et al., 2015) or ultra-OFR (Guo et al., 2016), to mitigate redundancy during term selection and improve model stability.

From an application standpoint, Power Quality scenarios could benefit from extending the model to handle concurrent disturbances and more realistic operational conditions,

including imbalanced and overlapping classes. Further, integrating higher-order feature engineering and resampling techniques may improve the robustness and generalization of the model in practical environments.

In the railway track monitoring context, future work could focus on improving classification accuracy for critical safety-related classes, such as Class P2, by refining labeling strategies and exploring data segmentation methods. Expanding the validation framework to include diverse operational conditions, such as varying train configurations, speeds, and track geometries, would increase the model's robustness and reliability. Additionally, incorporating geometric descriptors like curvature, cross-level, gauge, and alignment into the feature space may provide deeper insights into degradation mechanisms.

A promising direction also involves the use of frequency-domain analysis, such as Wavelet Transforms, to extract features that capture localized changes in acceleration and vibration patterns (Lupea and Lupea, 2025). These frequency-sensitive features may reveal subtle anomalies or degradation levels not visible in the time domain alone, improving both detection accuracy and interpretability.

Finally, the Data App Track Condition platform developed in this work demonstrates the potential for real-time, interactive visualization of predictive safety indices. Future enhancements could integrate the methodological improvements described above, transforming the tool into a powerful decision-support system for railway operators engaged in preventive maintenance and infrastructure risk management.

BIBLIOGRAPHY

Abu-Mostafa, Y. S. (2012). *Learning from data*, volume 4. AML Book, New York.

Aguirre, L. (2007). *Introdução à identificação de sistemas - Técnicas lineares e não-lineares aplicadas a sistemas reais*. Editora UFMG.

Aguirre, L. and Jácome, C. (1998). Cluster analysis of NARMAX models for signal-dependent systems. *IEE Proceedings - Control Theory and Applications*, 145(4):409–414.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ayala Solares, J. R. and Wei, H.-L. (2015). Nonlinear model structure detection and parameter estimation using a novel bagging method based on distance correlation metric. *Nonlinear Dynamics*, 82(1-2):201–215.

Ayala Solares, J. R., Wei, H.-L., and Billings, S. A. (2019). A novel logistic-NARX model as a classifier for dynamic binary classification. *Neural Computing and Applications*, 31(1):11–25.

Baldacchino, T., Anderson, S. R., and Kadirkamanathan, V. (2012). Structure detection and parameter estimation for NARX models in a unified EM framework. *Automatica*, 48(5):857–865.

Barbosa, R. S. (2016). New method for railway track quality identification through the safety dynamic performance of instrumented railway vehicle. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 38(8).

Batista, G. E. A. P. A., Carvalho, A. C. P. L. F., and Monard, M. C. (2000). Applying One-Sided Selection to Unbalanced Datasets. In *Advances in Artificial Intelligence: Lecture Notes in Computer Science*, pages 315–325. Springer Berlin Heidelberg.

Bhat, S., Karegowda, A. G., and A, L. R. (2023). Multiclass Classification of Rail Track Defects Using Deep Learning Techniques. In *2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, pages 1–6. IEEE.

Billings, S. and Wei, H.-L. (2019). NARMAX Model as a Sparse, Interpretable and Transparent Machine Learning Approach for Big Medical and Healthcare Data Analysis. In *IEEE 5th International Conference on Data Science and Systems*, pages 2743–2750.

Billings, S. A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons.

Billings, S. A., Korenberg, M. J., and Chen, S. (1988). Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, 19(8):1559–1568.

Billings, S. A. and Wei, H. L. (2005). A new class of wavelet networks for nonlinear system identification. *IEEE Transactions on Neural Networks*, 16(4):862–874.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bollen, M. H. J., Das, R., Djokic, S., Ciufo, P., Meyer, J., Ronnberg, S. K., and Zavodam, F. (2017). Power Quality Concerns in Implementing Smart Distribution-Grid Applications. *IEEE Transactions on Smart Grid*, 8(1):391–399.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Routledge, New York, USA.

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(2):321–357.

Chen, S., Billings, S. A., and Luo, W. (1989). Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896.

Corrêa, M. V. and Aguirre, L. A. (2004). Identificação não-linear caixa-cinza: uma revisão e novos resultados. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, 15(2):109–126.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Denoeux, T. (2000). A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(2):131–150.

Diego Ferreira, D. (2010). *Análise de Distúrbios Elétricos em Sistemas de Potência*. PhD thesis, Universidade Federal do Rio de Janeiro.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Ferreira, D., Cerqueira, A., Duque, C., and Ribeiro, M. (2009). HOS-based method for classification of power quality disturbances. *Electronics Letters*, 45(3):183.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Fisher, R. A. (1992). Statistical Methods for Research Workers. pages 66–70.

Forina, M., Leardi, R., Armanino, C., and Lanteri, S. (1990). PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. *Journal of Chemometrics*, 4(2):191–193.

Gerek, O. and Ece, D. (2006). Power-Quality Event Analysis Using Higher Order Cumulants and Quadratic Classifiers. *IEEE Transactions on Power Delivery*, 21(2):883–889.

Gu, Y. and Wei, H.-L. (2018). A robust model structure selection method for small sample size and multiple datasets problems. *Information Sciences*, 451-452:195–209.

Guo, Y., Guo, L., Billings, S., and Wei, H.-L. (2015). An iterative orthogonal forward regression algorithm. *International Journal of Systems Science*, 46(5):776–789.

Guo, Y., Guo, L., Billings, S., and Wei, H.-L. (2016). Ultra-Orthogonal Forward Regression Algorithms for the Identification of Non-Linear Dynamic Systems. *Neurocomputing*, 173:715–723.

Gustafsson, F. (1996). Determining the initial states in forward-backward filtering. *IEEE Transactions on Signal Processing*, 44(4):988–992.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null):1157–1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1/3):389–422.

Haber, R. and Keviczky, L. (1999). *Nonlinear System Identification - Input-Output Modeling Approach*. Springer Netherlands.

Haber, R. and Unbehauen, H. (1990). Structure identification of nonlinear dynamic systems—A survey on input/output approaches. *Automatica*, 26(4):651–677.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York, New York, NY.

Hosmer, D. W., Lemeshow, S., and Sturdivant, X. R. (2013). *Applied Logistic Regression*. John Wiley & Sons.

IEEE (2019). *IEEE Std 1159-2019 - Recommended Practice for Monitoring Electric Power Quality*. IEEE.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3).

Jayaprakash, K. and Balamurugan, S. P. (2021). Analysis of Plant Disease Detection and Classification Models: A Computer Vision Perspective. *Journal of Computational and Theoretical Nanoscience*, 17(12).

John Walker, S. (2014). Big Data: A Revolution That Will Transform How We Live, Work, and Think. *International Journal of Advertising*, 33(1).

Karis, T., Berg, M., Stichel, S., Li, M., Thomas, D., and Dirks, B. (2018). Correlation of track irregularities and vehicle responses based on measured data. *Vehicle System Dynamics*, 56(6):967–981.

Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification*. PhD thesis, Carnegie Mellon University.

Koohmishi, M., Kaewunruen, S., Chang, L., and Guo, Y. (2024). Advancing railway track health monitoring: Integrating GPR, InSAR and machine learning for enhanced asset management. *Automation in Construction*, 162:105378.

Kubat, M. and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann.

Kubat, M. and Matwin, S. (2000). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Fourteenth International Conference on Machine Learning*, 1.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kukreja, S. L., Löfberg, J., and Brenner, M. J. (2006). A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes*, 39(1):814–819.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11):1–13.

Lasisi, A. and Attoh-Okine, N. (2018). Principal components analysis and track quality index: A machine learning approach. *Transportation Research Part C: Emerging Technologies*, 91:230–248.

Lee, Y. W. and Schetzen, M. (1965). Measurement of the wiener kernels of a non-linear system by cross-correlation. *International Journal of Control*, 2(3):237–254.

Liu, L., Jones, B. F., Uzzi, B., and Wang, D. (2023). Data, measurement and empirical methods in the science of science.

Lupea, I. and Lupea, M. (2025). Continuous Wavelet Transform and CNN for Fault Detection in a Helical Gearbox. *Applied Sciences*, 15(2):950.

Malekjafarian, A., OBrien, E., Quirke, P., and Bowe, C. (2019). Railway track monitoring using train measurements: An experimental case study. *Applied Sciences*, 9(22).

Malekjafarian, A., Sarrabezolles, C.-A., Khan, M. A., and Golpayegani, F. (2023). A Machine-Learning-Based Approach for Railway Track Monitoring Using Acceleration Measured on an In-Service Train. *Sensors*, 23(17):7568.

Manyika, J., Chui Brown, M., Bughin, B. J., Dobbs, R., Roxburgh, C., and Hung Byers, A. (2011). Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute*.

Marasco, G., Oldani, F., Chiaia, B., Ventura, G., Dominici, F., Rossi, C., Iacobini, F., and Vecchi, A. (2024). Machine learning approach to the safety assessment of a prestressed concrete railway bridge. *Structure and Infrastructure Engineering*, 20(4):566–580.

Mendel, J. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proceedings of the IEEE*, 79(3):278–305.

Mishra, M. (2019). Power quality disturbance detection and classification using signal processing and soft computing techniques: A comprehensive review. *International Transactions on Electrical Energy Systems*, 29(8).

Mosleh, A., Meixedo, A., Ribeiro, D., Montenegro, P., and Calçada, R. (2023). Machine learning approach for wheel flat detection of railway train wheels. *Transportation Research Procedia*, 72:4199–4206.

Nadal, M. J. (1908). Locomotives á Vapeur. *Collection Encyclopédie Scientifique, Bibliotéque de Mécanique Appliquée et Génie*, 186.

Nagata, E. A., Ferreira, D. D., Bollen, M. H., Barbosa, B. H., Ribeiro, E. G., Duque, C. A., and Ribeiro, P. F. (2020). Real-time voltage sag detection and classification for power quality diagnostics. *Measurement*, 164.

Nagata, E. A., Ferreira, D. D., Duque, C. A., and Cequeira, A. S. (2018). Voltage sag and swell detection and segmentation based on Independent Component Analysis. *Electric Power Systems Research*, 155:274–280.

Naik, C. A. and Kundu, P. (2014). Power quality disturbance classification employing S-transform and three-module artificial neural network. *International Transactions on Electrical Energy Systems*, 24(9):1301–1322.

Nelles, O. (2020). *Nonlinear System Identification From Classical Approaches to Neural Networks and Fuzzy Models*. Springer International Publishing.

Nikias, C. and Mendel, J. (1993). Signal processing with higher-order spectra. *IEEE Signal Processing Magazine*, 10(3):10–37.

Nikias, C. L. and Petropulu, A. P. (1993). *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Prentice Hall.

Oppenheim, A. V., Willsky, A. S., and Nawab, S. H. (1997). *Signals and systems*, volume 2. Prentice hall, NJ.

Pan, D., Zhao, Z., Zhang, L., and Tang, C. (2017). Recursive clustering K-nearest neighbors algorithm and the application in the classification of power quality disturbances. In *Conference on Energy Internet and Energy System Integration*, pages 1–5.

Pazzani, M. (2000). Knowledge discovery from data. *IEEE Intelligent Systems*, 15(2):10–12.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pintelon, R. and Schoukens, J. (2012). *System Identification: A Frequency Domain Approach*. John Wiley & Sons.

Ribeiro, M. V., Marques, C. A., Duque, C. A., Cerqueira, A. S., and Pereira, J. L. (2006). Power quality disturbances detection using HOS. *IEEE Power Engineering Society General Meeting*, pages 1–6.

Ribeiro, M. V. and Pereira, J. L. R. (2007). Classification of Single and Multiple Disturbances in Electric Signals. *EURASIP Journal on Advances in Signal Processing*, 2007(1):18.

Rogers, S. and Girolami, M. (2011). *A first course in machine learning*. Chapman and Hall/CRC, New York.

Rugh, W. J. (1981). *Nonlinear system theory*. Johns Hopkins University Press.

Schetzen, M. (1980). *The Volterra and Wiener Theories of Nonlinear Systems*. John Wiley & Sons.

Senawi, A., Wei, H.-L., and Billings, S. A. (2017). A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recognition*, 67:47–61.

Sette, S. and Boullart, L. (2001). Genetic programming: principles and applications. *Engineering Applications of Artificial Intelligence*, 14(6):727–736.

Shu, X. and Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, 110.

Silva, P. H. O., Cerqueira, A. S., and Nepomuceno, E. G. (2021). Hybrid Method Based on NARX models and Machine Learning for Pattern Recognition. In *Anais do XV Simpósio Brasileiro de Automação Inteligente (SBAI 2021)*. SBA.

Silva, P. H. O., Cerqueira, A. S., and Nepomuceno, E. G. (2024). Insightful Railway Track Evaluation: Leveraging NARX Feature Interpretation. In *Anais do Congresso Brasileiro de Automática (CBA)*. SBA.

Silva, P. H. O., Cerqueira, A. S., Nepomuceno, E. G., and Oliveira, A. F. (2020). Classificação de Distúrbios na Qualidade de Energia Usando Modelagem Logística-NARX Multinomial. In *Anais do Congresso Brasileiro de Automática*.

Silva, P. H. O., Marotta, R. D., Cerqueira, A. S., Nepomuceno, E. G., and Lopes, L. A. S. (2022). Avaliação da Condição da Via Permanente usando Dados de Dinâmica de Veículos Ferroviários: Uma Abordagem de Aprendizado de Máquina. In *Anais do Congresso Brasileiro de Automática (CBA)*.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P. Y., Hjalmarsson, H., and Juditsky, A. (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724.

Smit, J. (2023). HPD 604780-7 (com logomarca da VLI).

Soni, K. G. and Patel, D. A. (2017). Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data. *International Journal of Computational Intelligence Research*, 13(5):889–906.

Sun, L., Liang, J., Zhang, C., Wu, D., and Zhang, Y. (2024). Meta-Transfer Metric Learning for Time Series Classification in 6G-Supported Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11.

Sun, Y. Q., Cole, C., and Spiryagin, M. (2015). Monitoring vertical wheel-rail contact forces based on freight wagon inverse modelling. *Advances in Mechanical Engineering*, 7(5):1–11.

Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition*. Academic Press.

Transportation Research Board (2020). *Review of the Federal Railroad Administrations Research and Development Program*. DC: The National Academies Press, Washington, D.C.

Valentini, G. (2004). Random Aggregated and Bagged Ensembles of SVMs: An Empirical Bias–Variance Analysis. In *Lecture Notes in Computer Science*, chapter Multiple C, pages 263–272. International Workshop on Multiple Classifier Systems.

Vidyalashmi, K., Chandana L, M., Nandana, J., Azhikodan, G., Priya, K. L., Yokoyama, K., and Paramasivam, S. K. (2024). Analysing the performance of the NARX model for forecasting the water level in the Chikugo River estuary, Japan. *Environmental Research*, 251:118531.

Wang, H., Berkers, J., van den Hurk, N., and Layegh, N. F. (2021). Study of loaded versus unloaded measurements in railway track inspection. *Measurement*, 169:108556.

Wasserman, L. (2013). *All of Statistics : a Concise Course in Statistical Inference*. Springer.

Wei, H. L. and Billings, S. A. (2008). Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information. *International Journal of Modelling, Identification and Control*, 3(4):341–356.

Wei, H. L., Billings, S. A., and Liu, J. (2004). Term and variable selection for non-linear system identification. *International Journal of Control*, 77(1):86–110.

Weierstrass, K. (1885). Über die analytische darstellbarkeit sogenannter willkürlicher functio- nen einer reellen veränderlichen. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 2:633–639.

Wiener, N. (1958). Nonlinear problems in random theory. *Massachusetts Institute of Technology*.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Wu, X., Zhu, X., Wu, G. Q., and Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1).

Zhang, H., Liu, P., and Malik, O. (2003). Detection and classification of power quality disturbances in noisy conditions. *IEE Proceedings - Generation, Transmission and Distribution*, 150(5):567.

Zhang, S. and Lang, Z.-Q. (2022). Orthogonal least squares based fast feature selection for linear classification. *Pattern Recognition*, 123:108419.

Zhong, P. and Fukushima, M. (2007). Regularized nonsmooth Newton method for multi-class support vector machines. *Optimization Methods and Software*, 22(1):225–236.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.