IC-AISE'2024 PROCEEDINGS

Proceedings of the International Conference On AI In Systems Engineering (IC-AISE'2024)

Beni Mellal, Morocco
May 25-26 2024

Co-Editors
Said SAFI
Miloud FRIKEL

PARTNERS

International Conference on AI in Systems Engineering

(IC-AISE'2024)

Co-Editors

Said SAFI

Miloud FRIKEL

Beni Mellal, Morocco April 25-26, 2024

Prologue

The main objective of this International Conference is to allow interested parties to better discover current advances in fields related to Artificial Intelligence and its applications in different engineering domain. It will offer specialists the opportunity to establish privileged contacts in the scientific community, in particular with other laboratories and research teams. In addition, this international symposium will allow French-speaking researchers to discuss their work and initiatives; and will be able to give doctoral students the opportunity to gain a broad overview of their field of research and to benefit from a first contact that is both rigorous and benevolent with all the related activities. Finally, it will also be an opportunity to share on the perspectives and projects in progress.

Honor Committee

Mustapha Aboumaarouf: President of Sultan Moulay Slimane University, BeniMellal Morocco.

Abderrazak El Harti: Dean of the Polydisciplinary Faculty, Beni Mellal Morocco.

Belaid Bouikhalene: Head of LIMATI Laboratory

Chairman

Pr. Said SAFI
LIMATI Laboratory Polydisciplinary Faculty, Sultan Moulay Slimane University Beni Mellal, Morocco
safi.said@gmail.com

Keynote Speakers

Mohammed M'Saad : ENSICAEN, Caen University, Caen France.
Miloud Frikel: ENSICAEN, Caen University, Caen France.
Mathieu Poliquen : IUT of Caen University, Caen France.

Organizing Committee

Pr. Bahloul Rachid
Pr. Biniz Mohamed
Pr. Darif Anouar
Pr. Ellahiani Idriss
Pr. Falih Nouredine
Pr. Farchane Abderazak
Pr. Hakimi Said
Pr. Laaribi Aziz
Pr. Ouchitachen Hicham
Pr. Ousarhane Abdessamad
Pr. Sadqi Yassine
Pr. Manaut Bouzid

Scientific Committee

Pr. ANTARI Jilali (Ibn Zohr University, Morocco)

Pr. AYACHI Rachid (Sultan Moulay Slimane University, Morocco

Pr. AUHMANI Khalid (Cadi Ayyad University, Morocco)

Pr. BASLAM Mohamed (Sultan Moulay Slimane University, Morocco)

Pr. BELAID Bouikhalene (Sultan Moulay Slimane University, Morocco)

Pr. BINIZ Mohamed (Sultan Moulay Slimane University, Morocco)

Pr. BOUMEZZOUGH Ahmed (Sultan Moulay Slimane University, Morocco)

Pr. BOUIKHALENE Belaid (Sultan Moulay Slimane University, Morocco)

Pr. CHABAA Samira (Ibn Zohr University, Morocco)

Pr. ELBAAMRANI Khalid (Cadi Ayyad University, Morocco)

Pr. EL MOHADAB Mohamed (Chouaib Doukkali University, Morocco)

Pr. FALIH Noureddine (Sultan Moulay Slimane University, Morocco)

Pr. FARCHANE Abderrazak (Sultan Moulay Slimane University, Morocco)

Pr. FRIKEL Miloud (Caen Normandie University, France)

Pr. GEHAN Olivier (National Graduate School of Engineering and Research Center, France)

Pr. GÜNTHER Chantal (Caen Normandie University, France)

Pr. HAMZAOUI Mohammed (Picardie Jules Verne University, France)

Pr. MOUNCIF Hicham (Sultan Moulay Slimane University, Morocco)

Pr. JABRANE Youness (Université Cadi Ayyad, Maroc)

Pr. MALAOUI Abdessamad (Sultan Moulay Slimane University, Morocco)

Pr. NAIT-SIDI-MOH Ahmed (Picardie Jules Verne University, France)

Pr. PIGEON Eric (Caen Normandie University, France)

Pr. POULIQUEN Mathieu (Caen Normandie University, France)

Pr. REUTER Johannes (Berlin University, Germany)

Pr. SADQI Yassine (Sultan Moulay Slimane University, Morocco)

Pr. ZARAGOZA Carlos Manuel Astorgaa (Centro Nacional de Investigación y Desarrollo Tecnológico, Mexico)

Pr. ZEROUAL Abdelouhab (Cadi Ayyad University, Morocco)

Pr. ZIDANE Mohamed (Ibn Zohr University, Morocco)

ICAISE'2024 Chairman

Biography: Prof. Said Safi received the M.Sc. and the doctorate degree from Chouaib Doukkali University and Cadi Ayyad Univerrsity, in 1997 and 2002, respectively.
He has been a Professor of information theory and network at the National School for Applied Sciences, Tangier, Morocco, from 2003 to 2005. Since 2006, he is a Professor of applied mathematics and programming at Polydisciplinary Faculty, Sultan Moulay Slimane University, Beni Mellal, Morocco. In 2008 he received the Ph.D. degree in Telecommunication and Informatics from the Cadi Ayyad University.
In 2015 he received the degree of Professor in Sciences at Sultan Moulay Slimane University. His general interests span the areas of communications and signal processing, estimation, time-series analysis, and system identification subjects on which he has published more than 70 journal papers and more than 90 conference papers. Current research topics focus on transmitter and receiver diversity techniques for single and multi-user fading communication channels, and wide-band wireless communication systems and mobile network

Keynote speakers

● **Mohammed M'SAAD**
Title : Control System: An Engeneering Panoramaba

Biography : Prof. Mohamed M'SAAD He was educated at the Ecole Mohammadia d'Ingénieurs where he held an assistant professor position in September 1978. He started his research activities at the Laboratoire d'Electronique et d'Etude des Systèmes Automatiques where he prepared an engineering thesis of the Université de Mohammed V on the adaptive control of industrial processes. In November 1982, Mohammed M'SAAD joined the Laboratoire d'Automatique de Grenoble to prepare a PhD thesis of the Institut National Polytechnique de Grenoble, on the fundamental features of the adaptive control and its applicability, which he obtained in April 1987. In April 1988, he held a research position at the Centre National de Recherche Scientifique with an affectation in the Laboratoire d'Automatique de Grenoble. In September 1996, Mohammed M'SAAD held a professor position at the Ecole Nationale Supérieure d'Ingénieurs de Caen where he founded a control process laboratory in 1997 which became a control group at the GREYC UMR CNRS in January 2004. His main research activities are mainly devoted to the fundamental, methodological and applied features of the identification, observation and adaptive control of dynamical systems. He had several important scientific and collective responsibilities, namely the director of the GREYC UMR CNRS from January 2012 to Marsh 2016..

● **Miloud FRIKEL**
Title : Spatial Diversity For Parametric Separation Of Seismic Waves

Biography : Dr. Miloud FRIKEL is an Associate Professor at National Graduate School of Engineering and Research Center (ENSICAEN), and head of SATE's Department (Embedded Systems and Control), and he is the Deputy-Director of the Systems Engineering Lab of Normandy (LIS). He was with the R&T (Networks and Telecommunications) Department of the University of Caen (Normandy University). He received his Ph.D. degree from the Center of Mathematics and Scientific Computation CNRS URA 2053, France, in array signal processing. Dr. Frikel was with the Signal Processing Lab, Institut for Systems and Robotics, Institute Superior Tecnico, Lisbon, as a researcher in the field of wireless location and statistical array processing. And he worked in the Institute for Circuit and Signal Procesing of the Technical University of Munich, Germany. M. Frikel is member of German Foundation: Alexander von Humboldt Stiftung. His research interests span several areas, including statistical signal and array processing, cellular geolocation (wireless location), direction finding and source localization, blind channel identification for wireless communication systems and MC-CDMA systems.

● **Mathieu POULIQUEN**
Title : Identification Of Dynamic Systems From Binary Measurements, Some Solutions

Biography : Dr. Mathieu Pouliquen received his Engineering degree from the Ecole Nationale Superieure d'Ingenieurs de Caen, France, in 2000, then the PhD in Automatic Control from the University of Caen, in 2003. He was appointed Assistant Professor at the University of Caen in 2004, since 2021 he is Full Professor. He is a Researcher at the LIS laboratory and his research interests include linear and nonlinear system identification.

Contents

Keynote Talks

- **Control System: An engeneering Panorama**

Animated by: Mohammed M'Saad, Caen University, ENSICAEN, Caen France.

- **Spatial Diversity for Parametric Separation of Seismic Waves**

Animated by :Miloud Frikel, Caen University, ENSICAEN, Caen France..

- **Identification of dynamic systems from binary measurements, some solutions.** Animated by : Mathieu Poliquen : IUT of Caen University, Caen France.

.

Control System: An Engineering Panorama

Mohammed M'SAAD
ENSICAEN, Caen University
Caen, France
mohammed.msaad@ensicaen.fr

Abstract:

The main motivation of this conference consists in providing a comprehensive presentation of the control system development from an engineering perspective. A particular emphasis is put on the underlying engineering system development using my research activity real achievements

The mathematical concept behind deep learning

Miloud Frikel
LIS Laboratory, ENSICAEN School,
Caen University, Caen France
miloud.frikel@ensicaen.fr

Abstract :

This talk aims to use knowledge of signal processing for the interpretation of seismic data.

Indeed, the methods used for the analysis of seismic data are increasingly complex and require to the most recent developments in signal processing.

The objective of this talk is to provide an overview of the fundamental notions of signal processing that can be used in seismic.

And how spatial diversity or multi-sensor signal processing can help separate seismic wave fronts and precisely geolocate the epicenter of an earthquake. These analyzes will be illustrated on the seismic data collected from El Haouz, region of Marrakech, Morocco, in September 8, 2023.

Some elements on system identification from binary measurements of the output

Mathieu Pouliquen
UIT, Caen University,
Caen, France
mathieu.pouliquen@unicaen.fr

Abstract:

This talk deals with systems identification using binary-valued measurements. This identification context can occur for an economical reason (a low resolution sensor is less expensive than a high resolution sensor), for a technical reason related to the system to be identified (there is no high-resolution sensor available) or for a technical reason related to the implementation environment (only few binary data can be transmitted). Traditional identification methods cannot be successfully applied with such quantization constraints. In the past few years, a substantial effort has been devoted to the problem of identifying a system with binary-valued measurements. The objective here is to introduce to some identification algorithms dedicated to this particular context.

A Review of Traditional methods for Localizing Radiating Sources

Ilham Mahiri[1,2], Saïd Safi[1], and Miloud Frikel[2]

***Abstract*—The localization of radiating sources plays an essential role in ensuring reliable and efficient communication. Traditional methods, based on signal processing and antenna processing, have been widely studied and applied in various scenarios to estimate the direction of arrival (DOA) of signals emitted by these sources. This article provides an overview of advances in traditional source localization methods, which offer high-resolution DOA estimation and have found applications in sonar and wireless communication systems for target localization, interference suppression, and emergency response. Recent advances in adaptive antenna arrays are also discussed in the article. It delves into the theoretical foundations, practical implementations, strengths, limitations, and applications of these traditional methods, providing insight into their effectiveness and potential for real-world deployment.**

***Index Terms*—The localization of radiating sources, direction of arrival (DOA), traditional methods, wireless communication systems, signal processing, antenna techniques.**

I. INTRODUCTION

The localization of radiating sources can indeed be achieved using signal processing and antenna techniques. Antennas play a crucial role in collecting signals emitted by the sources, while signal processing is used to analyze these signals and estimate the parameters of the sources. Consequently, the localization of these sources is done by estimating the arrival angles [1]- [2], which allows determining the direction from which the signals are emitted. This process is integral to various fields such as radars, sonars, and wireless communication systems, where the accuracy of source localization is essential to ensure reliable and quality connectivity.

In this context, several traditional methods have been extensively studied and applied for source localization. Techniques such as MUSIC (MUltiple SIgnal Classification), Capon, beamforming, and ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) have been widely utilized due to their effectiveness in estimating the direction of arrival of signals. For example, Capon and beamforming minimize the variance of the estimated signal and provide high-resolution DOA estimation [3]. Similarly, MUSIC utilizes the eigenstructure of the received signal's covariance matrix to estimate the DOA with high accuracy [4]. ESPRIT, on the other hand, exploits rotational invariance properties of the signal to estimate DOA pairs directly from the received data [5]. These techniques have found applications in various scenarios, including radar and sonar systems for target localization, wireless communication systems for beamforming and interference suppression [6], and localization systems for emergency response and surveillance purposes [7]. Furthermore, recent advancements in adaptive antenna arrays and beamforming techniques have enhanced the performance of source localization in challenging environments. These techniques dynamically adjust the antenna array's radiation pattern to focus on the desired signal while suppressing interference and noise [8].

In the subsequent sections, we delve deeper into the theoretical foundations and practical implementations of these traditional methods, discussing their strengths, limitations, and applications in different scenarios of source localization.

II. SYSTEM MODEL

Consider a base station equipped with a linear array consisting of M omnidirectional sensors. This array is designed to capture $K(K < M)$ narrow-band plane wave signals arriving from various directions denoted by $\theta_1, \theta_2, \ldots, \theta_K$. Among these signals, one represents the direct signal while the remaining $K-1$ signals are reflections. The distance between the initial reference element and the i-th element within the array is represented by d_i. This array configuration ensures that there are at least two elements positioned at half of the wavelength of the radiation sources or even closer, as illustrated in Fig (1).

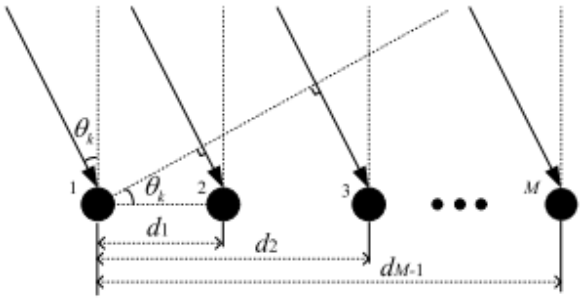


Figure 1. Linear network with N elements.

In general, array antenna received signal vector x(t) formed by the superposition of the multipath signals and noise, can be described as :

$$x(t) = \sum_{k=1}^{K} a(\theta_k) s_k(t) + n(t) \quad (1)$$

[1]Department of Mathematics and Informatics, University Sultan Moulay Slimane, Beni Mellal, Morocco
[2]Laboratoire d'Ingenierie des Systems - UR 7478 UNICAEN, ENSICAEN, Normandie Univ, LIS, Caen, France

where $x(t) = [x_1(t), x_2(t), \ldots, x_M(t)]^T$ is the M-dimensional snapshot data vector of the array, $v(t) = [n_1(t), n_2(t), \ldots, n_M(t)]^T$ is the M-dimensional noise data vector. The vector-valued function $a(\theta_k)$ is the array response vector (steering vector) for an array of M elements to the k-th source signal from the direction θ_k, expressed as follows:

$$a(\theta_k) = \begin{bmatrix} 1 \\ e^{-j\phi_{k1}} \\ \vdots \\ e^{-j\phi_{ki}} \\ \vdots \\ e^{-j\phi_{k(M-1)}} \end{bmatrix}^T \tag{2}$$

where the phase shift of the i-th element for each narrowband arrival signal can be defined as $\phi_{ki} = 2\pi f_c d_i \sin(\theta_k)/c$. f_c is the carrier frequency of the incident signals, c is the speed of light. It is assumed that the signals and noise are stationary, zero mean uncorrelated random processes. Further, the noise vector is Additive Gaussian White Noise (AGWN) with variance σ^2.

III. REVIEW OF SOME DOA ESTIMATION ALGORITHMS

The significance of traditional methods, recognized for their pivotal role in localization of radiating sources, underscores the necessity for a comprehensive review, focusing on their theoretical foundations.

A. *The conventional beamforming*

The conventional beamforming technique developed by Bartlett [9] is considered to be one of the oldest techniques for DOA estimation for signal sources. Here, the beamforming [10] steers the array in one direction at a time and measures the output power. The direction which gives maximum output power provides the true DOA for the incident sources.

The steering process is performed by linearly combining the sensor outputs. The principle is as follows:

- Create a weighting vector

$$\omega(\theta) = a(\theta) \tag{3}$$

- Calculate the mean squared magnitude of the output, $E[|y_{BF}|^2]$

$$y_{\text{BF}} = \omega(\theta)^H \cdot \mathbf{x}(t) \tag{4}$$

For each new angle of observation, the process is repeated until obtaining a set of observation angles within the desired range $\theta_{\min} \leq \theta \leq \theta_{\max}$. Subsequently, a pseudo-spectrum representing $\mathbb{E}[|y_{\text{BF}}|^2]$ is plotted as a function of θ. The angular positions of the sources are identified by locating the peaks of the pseudo-spectrum. This procedure verifies the operations can also be performed in matrix form.

$$\begin{aligned} P_{BF}(\theta) &= E[|y_{BF}|^2] = E[y_{BF} y_{BF}^H] \\ &= a^H(\theta) E[x(t) x^H(t)] a(\theta)] \end{aligned} \tag{5}$$

Then:

$$P_{BF}(\theta) = a^H(\theta) R_{xx} a(\theta) \tag{6}$$

B. *Capon's Method*

Capon's minimum variance method is an approach used for Direction of Arrival (DOA) estimation. It's designed as a beamformer to address the limitations of conventional beamformers, especially in scenarios with multiple narrowband sources coming from different directions. In such cases, the output power of the array includes signals from both desired and undesired sources, leading to reduced resolution with conventional beamformers [11]. Capon's method aims to mitigate the influence of undesired DOA by minimizing the total output power while preserving a constant gain along the desired direction.

We construct a filter that attenuates the signal more in a given direction as the received signal contains power in that direction. The sought-after filter is a solution of the following Hermitian form:

$$\min_{\omega} \mathbb{E}[|y(t)|^2] = \min_{\omega}(\omega^H R_{xx} \omega) \quad \text{subject to} \quad (\omega^H a(\theta)) = 1 \tag{7}$$

For a covariance matrix that is positive definite, the solution to equation (7). for the weight vector can be easily provided :

$$\omega = \frac{R_{xx}^{-1} a(\theta)}{a^H(\theta) R_{xx}^{-1} a(\theta)} \tag{8}$$

The weight obtained by equation (8) is referred to as the Capon method. Utilizing this weight vector from equation (8), the power of the array output signal adopts the following :

$$P_{\text{Capon}}(\theta) = \frac{1}{a^H(\theta) R_{xx}^{-1} a(\theta)} \tag{9}$$

In this scenario, DOA can be determined by identifying the K highest peaks in the spatial spectrum defined by equation (9). While Capon's method outperforms conventional beamformers, it remains [12] sensitive to factors such as the number of elements in the array and the Signal-to-Noise Ratio (SNR).

C. *The MUSIC method*

A simple and renowned technique, known as MUSIC, was independently introduced by Schmidt in 1979 [13], [14]and by Bienvenu and Kopp in 1980 under the name goniometer [15]. It allows determining the arrival angles using a pseudo-spectrum derived from the signal or noise subspace of the observation. MUSIC was specifically designed to find the DOA for multiple narrowband uncorrelated sources.

Consider the signal model in section 2.1 and assume that the covariance noise matrix has a uniform noise power on the diagonal as $\sigma^2 I$, where I is the identity matrix.

$$\begin{aligned} R_{xx} &= E[x(t) x^H(t)] \\ &= E[(As + n)(As + n)^H] \\ &= AE[ss^H]A^H + E[nn^H] \\ &= AR_{ss}A^H + \sigma^2 I \end{aligned} \tag{10}$$

The matrix R_{xx} being Hermitian and positive definite, its eigenvalues are real and positive. Its M non-zero eigenvalues are ordered in decreasing order:

$$[\mu_1 \geq \mu_2 \geq \ldots \geq \mu_M \tag{11}$$

The N eigenvalues of matrix R_{xx} can be written as follows:

$$\lambda_m = \mu_m + \sigma^2, \quad \text{for } m = 1, 2, \ldots, M \tag{12}$$

$$\lambda_m = \sigma^2, \quad \text{for } m = M+1, M+2, \ldots, N \tag{13}$$

The N eigenvectors associated with the N eigenvalues λ_m are: $(e_1, e_2, \ldots, e_M, \ldots, e_N)$. The spectral decomposition of the covariance matrix R_{xx} into eigenelements to separate the signal subspace from the noise subspace can be expressed as follows:

$$R_{xx} = \sum_{m=1}^{N} \lambda_m e_m e_m^H = E_s \Lambda_s E_s^H + E_n \Lambda_n E_n^H \tag{14}$$

With :

$$\Lambda_s = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M) \tag{15}$$

and

$$\Lambda_n = \text{diag}(\lambda_{M+1}, \lambda_{M+2}, \ldots, \lambda_N) = \text{diag}(\sigma^2, \ldots, \sigma^2) \tag{16}$$

The vector associated with the M most significant eigenvalues, containing the eigenvectors associated with the signal subspace, is represented as:

$$E_s = (e_1, e_2, \ldots, e_M) \tag{17}$$

Similarly, the vector of eigenvectors associated with the $(N - M)$ smallest eigenvalues, containing the eigenvectors associated with the noise subspace, is represented as:

$$E_n = (e_{M+1}, e_{M+2}, \ldots, e_N) \tag{18}$$

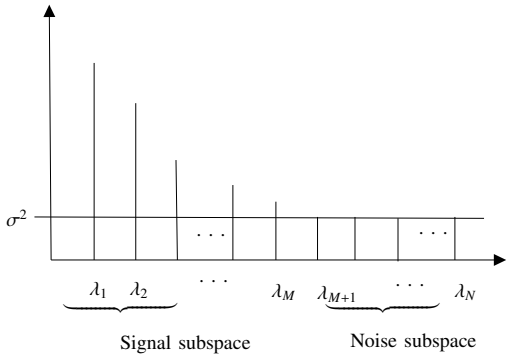


Figure 2. Representation of eigenvalues sorted in descending order.

The angular spectral function obtained by the MUSIC method allows determining the values for which this function is maximal and is defined as follows:

$$P_{\text{MUSIC}}(\theta) = \frac{1}{a^H(\theta) E_n E_n^H a(\theta)} \tag{19}$$

The pseudo-spectrum of MUSIC provides peaks corresponding to the exact arrival directions of the waves but does not inform us about the power of the sources.

*D. **Min-Norm method***

The Min-Norm algorithm, originally developed for frequency analysis and antenna processing, is used to estimate the arrival directions of signals based on measurements taken from an antenna array [16], [17]. It searches for a vector of minimum norm within the noise subspace to identify peaks in the pseudo-spectrum. Compared to the reference method MUSIC, the Min-Norm approach defines the pseudo-spectrum by searching for a vector x within the noise subspace, thereby minimizing its norm [18], [19]. The calculation of the minimum norm vector x satisfies three specific constraints [20]. Given that x belongs to the noise subspace E_n, it is automatically orthogonal to the projector V_s onto the signal subspace E_s, defined by:

$$V_s = \begin{pmatrix} v_1 & v_2 & \cdots & v_M \end{pmatrix} \quad V_s^H \, x = 0 \tag{20}$$

Moreover, the first element of x is equal to 1, and its Euclidean norm is minimal. Finally, since the Min-Norm method is based on spectral search, we can define the associated pseudo-spectrum formula as follows:

$$P_{\text{Min-Norm}}(\theta) = \frac{1}{|a^H(\theta) E_n E_n^H u_1|^2} \tag{21}$$

Where u_1 is the Cartesian basis vector (the first column of the identity matrix $N \times N$).

E. ESPRIT Algorithm

The ESPRIT is a fast, efficient, and robust method for estimating signal parameters. Introduced by Roy in 1989 [21], it determines the incidence directions of multiple sources in an antenna array by exploiting signal subspaces. Without requiring a search for maxima, it significantly reduces computations. Although it does not need criteria optimization or antenna calibration, it demands specific antenna geometries and perfectly identical sub-arrays. Its principle involves decomposing the initial array into two identical sub-arrays obtained by translation, as illustrated in Fig (3).

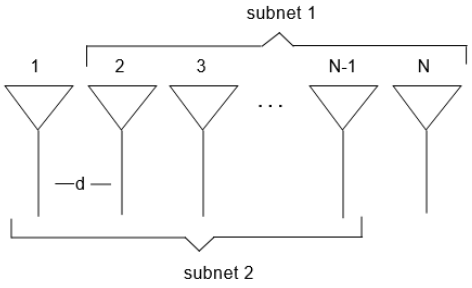


Figure 3. Principle of the ESPRIT method.

Arrival directions is contained within the passage matrix linking the observation vectors at the output of the sub-arrays. Designating respectively by $x_1(t)$ and $x_2(t)$ the observation vectors at the output of sub-array 1 and sub-array 2, the model of the received signal for the overall array is written as:

$$X(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} A \\ A\Phi \end{pmatrix} s(t) + n(t) \tag{22}$$

Where:

$$\Phi = \text{diag} \begin{pmatrix} \exp\left(j\frac{2\pi}{\lambda} d \sin\theta_1\right) \\ \exp\left(j\frac{2\pi}{\lambda} d \sin\theta_2\right) \\ \cdots \\ \exp\left(j\frac{2\pi}{\lambda} d \sin\theta_M\right) \end{pmatrix} \tag{23}$$

The steps of the ESPRIT algorithm for obtaining the estimation of incidence directions on the antenna array are:

1) Calculating the covariance matrix R_{xx} from the observation matrix $X(t)$ of the entire array:

$$R_{xx} = \begin{bmatrix} A \\ A\Phi \end{bmatrix} R_{ss} \begin{bmatrix} A^H \\ \Phi^H A^H \end{bmatrix} + \sigma^2 I \tag{24}$$

Where: R_{ss} is the correlation matrix of incident signals.

2) Eigenvalue decomposition of the covariance matrix or singular value decomposition of the observation matrix; The same procedure as in the case of the MUSIC algorithm:

$$R_{xx} = \sum_{m=1}^{N} \lambda_m e_m e_m^H = E_s \Lambda_s E_s^H + E_n \Lambda_n E_n^H \tag{25}$$

3) Selection of the signal subspace E_s from the eigenvector matrix; ESPRIT exploits not the noise subspace, as in the case of MUSIC, but the signal subspace $E_s = \{e_1, e_2, \ldots, e_M\}$.

4) Separation into translation-invariant subarrays of the signal subspace; The signal subspace E_s of the entire array can be decomposed into two subspaces E_1 and E_2 of dimension $(N-1)\times M$, which are the respective signal subspaces of subarrays 1 and 2. We can then write:

$$E_s = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} \tag{26}$$

These two matrices E_1 and E_2 are related by the following invertible linear transformation:

$$E_s = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = \begin{pmatrix} A^T \\ A\Phi^T \end{pmatrix} \tag{27}$$

With: $T = R_{21}R_{11}^{-1}$ where $R_{21} = \frac{1}{N} X_2 X_1^H$ and $R_{11} = \frac{1}{N} X_1 X_1^H$ are the covariance matrix between the two antenna subarrays. Using equation 3.24, we can write:

$$E_2 = A^T T^{-1} \Phi^T = E_1 \Psi \tag{28}$$

Where: $\Psi = T^{-1}\Phi^T$ of dimension $(M \times M)$.

5) Determination of eigenvalues; The eigenvalues of Ψ and Φ are common and are of the form:

$$\lambda_m = \exp\left(-j2\pi \frac{d}{\lambda} \sin\theta_m\right), \quad m = 1, \ldots, M \tag{29}$$

6) Estimation : The arrival angles are finally extracted from the eigenvalues and are expressed by the relation:

$$\hat{\theta}_m = \sin^{-1}\left(\frac{\lambda}{2\pi d} \arg(\lambda_m)\right) \tag{30}$$

We can also calculate the Root Mean Square Error (RMSE) between the theoretical angle and the angle estimated by a high-resolution method using the following formula:

$$\text{RMSE} = \sqrt{\left(\hat{\theta}_m - \theta_m\right)^2} \tag{31}$$

IV. SOURCE LOCALIZATION METHODS: ADVANTAGES AND CONSTRAINTS.

This section provides an overview of various source localization techniques, highlighting their respective strengths and limitations.

A. *Literature review of MUSIC, Min-Norm, and ESPRIT*

The MUSIC method, when adapted to the time domain, proposes a new approach to estimate propagation delays in locating the radiant source. In [22], pulsed radar signals are treated directly in the time domain, eliminating the need for frequency conversion via the Fourier transform by MUSIC. Instead, tailor-made filtering is used to select temporal data intervals likely to contain overlapping pulses. [23] strongly describes the dynamic nature of the resulting pseudo-spectra, highlighting its high resolution and remarkable ability to discern impulses even when they overlap. However, the method has limits, mainly its susceptibility to correlation, requiring an observation average of the covariance matrix, techniques available only in the frequency domain [25]. Although attempts have been made to adapt these techniques in the time domain [25], achieving complete optimization would require an integrated approach combining time and frequency techniques [24].

In this context, the literature highlights the Min-Norm approach, which shows increased sensitivity compared to the MUSIC algorithm concerning the "quality" of the covariance matrix. Marble [26] demonstrated that Min-Norm not only highlights separation capabilities superior to MUSIC but also involves higher estimate variance. This nuanced dynamic becomes evident in the operation of Min-Norm, particularly in the processing of frequency signals. For the specified simulation parameters, a particularly dynamic functionality is observed, thus emphasizing significant resolving power.

Similarly, for ESPRIT, the literature has highlighted its advantages in estimating parameters of noisy signals, particularly for estimating signal delays in radar signals. In [27], it was demonstrated that ESPRIT has lower variance than the MUSIC algorithm for noisy exponential signals.

Additionally, the total least squares (TLS) method, which minimizes the sum of the squares of errors in the dependent and independent variables, has been shown to provide better performance in some cases compared to the least squares (LS) [28]. Unlike the LS method, which only minimizes vertical errors, the TLS method also takes horizontal errors into account, making it more robust to measurement errors [28]. Overall, these results highlight the effectiveness of the ESPRIT algorithm in reducing measurement errors, especially in noisy conditions. Additionally, using the TLS method provides a more robust approach for parameter estimation by taking into account a wider variety of errors, making it an attractive option in various application contexts.

B. Literature review of beamforming and Capon

Beamforming, a cutting-edge technique in modern wireless networks, offers significant advantages in FD-MIMO networks [29]. By leveraging the dynamic control capabilities of amplitude and phase of individual antenna elements, electronic beamforming enables precise adaptation of radiation patterns in azimuthal and elevational planes [30]. This capability allows for more targeted and spatially separated transmission to a larger number of users simultaneously. By combining beamforming with MU-MIMO precoding techniques, FD-MIMO systems can significantly optimize network coverage and throughput [31]. Field trials in [32] confirm that this approach leads to tangible improvements in network performance. However, the application of beamforming also presents challenges. Signal processing and coordination requirements among antenna elements can be complex, necessitating additional resources in terms of computation and communication. Additionally, the accuracy of beam adaptation depends on the quality of available channel information, which may vary depending on network conditions. Despite these challenges, the potential benefits of beamforming in FD-MIMO systems fully justify its increasing adoption in wireless network architectures.

Moreover the Capon method offers several significant advantages for estimating the arrival direction of acoustic signals. By leveraging the directional sensitivity of vector sensors, it allows for improved spatial resolution and reduction of direction ambiguities [33]. This leads to better accuracy in estimating the azimuth and elevation angles of sound sources, even with simple array structures such as linear arrays. However, the Capon method can be sensitive to noise and signal conditions, which may affect its performance in complex acoustic environments. Additionally, accurately estimating covariance data for optimal performance can pose a practical challenge [34]. Despite these limitations, the Capon method remains a powerful approach for estimating the arrival direction of acoustic signals with vector sensors.

V. CONCLUSION

traditional methods for source localization, such as MUSIC [35], Capon, beamforming [36], and ESPRIT [37], have proven to be effective in estimating the direction of arrival (DOA) of signals in wireless communication systems. These methods have been widely applied in various scenarios, including radar, sonar, and wireless communication systems, for target localization, beamforming, interference suppression, and emergency response. However, these traditional methods are subject to certain constraints and limitations. One major constraint is their sensitivity to noise, interference, and multipath propagation, which can degrade the accuracy of DOA estimation, especially in complex and dynamic environments. Additionally, traditional methods often require a priori knowledge of the signal and environment parameters, which may not always be available or accurate in practical scenarios. Moreover, these methods may not scale well to large-scale or highly dynamic systems, where real-time processing and adaptation are crucial.

To address these challenges and optimize source localization, recent research efforts have focused on leveraging artificial intelligence (AI) techniques. Machine learning algorithms, particularly deep learning models [38], offer the potential to learn complex patterns and relationships from large datasets, enabling more robust and adaptive localization algorithms. For example, convolutional neural networks (CNN) can be trained to directly estimate DOA from raw signal data, bypassing the need for explicit feature extraction and parameter estimation. And in [39], a Transformer-based model is proposed for DOA estimation in wireless communication systems. This model outperforms traditional methods in terms of accuracy and robustness, especially in varying signal conditions and interference.

References

[1] A. Nehorai and E. Paldi, “Acoustic vector-sensor array processing,” in Proc. 26th Asilomar Conf. Signals, Syst., Comput., Pacific Grove, CA, Oct. 1992, pp. 192–198

[2] M. J. Berliner and J. F. Lindberg, ” Acoustic Particle Velocity Sensors: Design, Performance and Applications,” Woodbury, NY: AIP, 1996

[3] R. O. Schmidt, ”Multiple emitter location and signal parameter estimation,” IEEE Transactions on Antennas and Propagation, pp. 276-280, 1986.

[4] J. Capon, ”High-resolution frequency-wavenumber spectrum analysis,” Proceedings of the IEEE, pp.1408-1418, 1969.

[5] B. D. Van Veen, and K. M. Buckley, ”Beamforming: A versatile approach to spatial filtering,” IEEE ASSP Magazine, pp.4-24, 1988.

[6] P. Stoica, and R.L. Moses, ”Introduction to spectral analysis,” Prentice Hall, 1997.

[7] R. Roy, T. Kailath, and F. Ahmad, ”ESPRIT—a subspace rotation approach to estimation of parameters of cisoids in noise,” IEEE Transactions on Acoustics, Speech, and Signal Processing, pp.984-995, 1990.

[8] Li. J. Zhang, X. S. Han, ”An overview of precoding techniques for massive MIMO systems,” IEEE Wireless Communications, pp.74-81, 2014.

[9] M. S. Bartlett, “Periodogram analysis and continuous spectra,” Biometrica, Vol. 37, pp.1-16, 1950.

[10] G. Jankins and D. Watts, ”Spectral Analysis and its Applications,” San Francisco, CA: Holden-Day, 1968.

[11] J. Clerk Maxwell, ”A Treatise on Electricity and Magnetism,” 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[12] S. Petre, M. Randolph, ”Spectral analysis of signals,” Upper Saddle River, NJ : Pearson/Prentice Hall, (2005).

[13] R. O. Schmidt, ”Multiple emitter location and signal parameter estimation,” IEEE Transactions on Antennas and Propagation, vol. 34, no 3, pages 276-280, 1986.

[14] R. O. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation," Ph.D. thesis, Dept. Elect. Eng., Stanford university, Nov 1981.
[15] G. Bienvenu, L. Kopp, "Optimality of high resolution array processing using the eigen system approach," IEEE Trans. on Acoustic, Speech and Signal processing, vol. 31, no 5, pp. 1235-1248, Oct 1983.
[16] M. Young, "The Technical Writer's Handbook," Mill Valley, CA: University Science, 1989.
[17] D. ZHA, T. QIU, "Direction finding in non-Gaussian impulsive noise environments," Digital Signal Processing, vol. 17, no 2, pp.451–465, 2007.
[18] M. Hasan, A. Hasan , "Fast Approximated Sub-Space Algorithms," Tenth IEEE Workshop on Statistical Signal and Array Processing, pp.127–130, 2000.
[19] V. Guilhem, "Characterization of wideband sources in the time domain without constraints on the number of sensors," PhD thesis, École Centrale de Marseille, France, 2013.
[20] B. Cedric, "Contribution of super-resolution signal processing techniques to improving the performance of ground-penetrating radar," PhD thesis, Nantes, 2007.
[21] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," IEEE Transactions on Acoutics, Speech, and Signal Processing, vol. 37, no 7, pp.984–995, 1989.
[22] S. Marcos, "High-Resolution Methods, Antenna Processing and Spectral Analysis, Signal Processing," Hermès Ed., ISBN 2-86601-662-9, 1998.
[23] M. A. Pallas and G. Jourdain, "Delay Estimation and High-Resolution Methods," GRETSI Colloquium, pp. 97-100, Nice, 1987.
[24] S. Bozinoski, "Spatio-Temporal Analysis of Wideband Signals for Oceanic Acoustic Tomography," PhD thesis, INPG, 1996.
[25] A. El Hanafi, "Estimation of Radar Signal Propagation Delays in a Stratified Medium," PhD thesis, LCPC, 2005.
[26] S.L.Jr. Marple, "Digital Spectral Analysis with Applications," Prentice-Hall, Signal processing series, Alan V. Oppenheim, series editor, 1987.
[27] S.M. Kay, S.L.Jr. Marple, "Spectral Analysis - A Modern Perspective," Proceeding of IEEE, Vol. 69, No. 11, november 1981.
[28] G. Golub, C. Van Loan, "An Analysis of the Total Least Square Problem," SIAM Journal of numerical Analysis, Vol. 17, No. 6, 1980.
[29] Huawei," AAS possible application scenarios," 3GPP TSG-RAN WG4 60bis, R4-115012, Oct. 2011.
[30] W. Lee, S. R. Lee, H.-B. Kong, and I. Lee, "3D beamforming designs for single user MISO systems," in 2013 IEEE Global Communications Conference (GLOBECOM), Dec. 2013, pp. 3914–3919.
[31] J. Koppenborg, H. Halbauer, S. Saur, and C. Hoek, "3D beamforming trials with an active antenna array," in ITG Workshop on Smart Antennas, 2012, pp. 110–114.
[32] H. Halbauer, J. Koppenborg, J. Holfeld, M. Danneberg, M. Grieger, and G. Fettweis, "Field trial evaluation of 3D beamforming in a multicell scenario," in 17th International ITG Workshop on Smart Antennas (WSA), 2013, pp. 1–7.
[33] M. Hawkes, A. Nehorai, "Acoustic Vector-Sensor Beamforming and Capon Direction Estimation," IEEE Transactions on Signal Processing, vol. 46, no. 9, pp. 2291-2300, September 1998.
[34] M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and Capon direction estimation," in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Detroit, MI, May 1995, pp. 1673–1676.
[35] Y. Wang, Y. Huang, "Deep Learning for Direction-of-Arrival Estimation: A Survey", IEEE Communications Surveys and Tutorials, 22(4), 2403-2439, 2020.
[36] M. Al-Fayly, S. Goudos, Z. D. Zaharis," A Comprehensive Review of Deep Learning Models for DOA Estimation in MIMO Systems", IEEE Access, 8, 193289-193308, 2020.
[37] Y. Li, W. Zhang, and Z. Song, "Deep Learning in Wireless Communications", A Comprehensive Review. IEEE Internet of Things Journal, 8(2), 1236-1254, 2021.
[38] W. Li, C. Wang, and H. Dai, "Deep Reinforcement Learning for Wireless Communications", A Comprehensive Survey. IEEE Transactions on Cognitive Communications and Networking, 7(3), 834-854, 2021.
[39] Y. Liu, H. Chen and B. Wang, "A direction of arrival estimation method based on deep learning," Journal of Physics Conference Series 1550(3):032066, 2020.

A Smart University for a Smart City

Mohamed El Mohadab
LAROSERI
EL JADIDA, Morocco

Said Safi
Dept. MCS of SMSU
Beni Mellal, Morocco

Belaid Bouikhaleine
Dept. MCS of SMSU
Beni Mellal, Morocco

Hayat Jebbar
LAROSERI
EL JADIDA, Morocco

Nabil Ababou
Dept. MCS of SMSU
Beni Mellal, Morocco

ABSTRACT

Scientific research [1] represents a very important axis for the university, because it ensures its innovation and its productivity and develops the competencies of their researchers and the reputation and the glow of their research laboratory.

But the management and automation of this sector represents a great challenge for universities either for managers, directors, or the researchers from which comes the need of find relevant and effective solution.

To manage this sector, we have to study several computer solutions which ensure good management of the information system, in this direction we find that the Data Warehouses and the Data Mining are relatively well mastered when it comes to scientific research data because it is increasingly complex, diverse.

The Data Warehouse and the Data Mining is recognized as the core of the decision-making system: it integrates and stores data from the different functional areas of an organization for make it easily accessible to decision-making processes on order to ensure adaptation to the new change that can be brought to the system.

Keywords

Data warehouse, data mining, Data mart, OLAP, Modeling Process, Scientific Research, Research Management, Decision Support Systems.

1. INTRODUCTION

Lately the world has entered into an era that is called the societies and knowledge in which science, knowledge innovation its source of wealth. In this vision the university represents a pillar of development of each country.

In the last decade, the use of information communication technologies became important in all sectors, this technological revolution enhance establishments to integrate the information communication technologies in their information system in order to implement governance [2] and democratization of having information.

In this case, the Moroccan governments through the E-Governance strategy [3] move forward in the field of information technology and encourage all establishments to adopt new tools for developing their information system.

This new strategy [4] encourages public university to adopt information and communication technology tools like Data Warehouse to facilitate communicating information around the functional processes in the public university and to improve their performance especially in scientific research.

An information system, including a data warehouse system, is user-interfaced and designed to provide information useful to support strategy, operations, management analysis, and decision-making functions in an organization.

In this work, we contribute to the modeling and the implementation of management information decision support system in a public university especially in Scientific Research.

The paper has five parts. First section we propose a short recall of the scientific research in public university. Second section shows the relevant literature review of Data Warehouse. In the third section we introduce the context of study. The fourth section shows the case study. The fifth section of the paper shows the Pairing between Data Warehouse and Data Mining.

2. Scientific Research in Public University

Research is therefore not conducive to control and management. However, the rapid evolution of the highly competitive world of higher education today imposes constraints that require the establishment of a minimum management framework.

For the university nowadays to take risks is an essential aspect of the dynamism of the institution, but the risk must also be understood and managed.

Five ways in which management integrates culture into the university have been identified [5]:

- Strengthen the central steering structure.
- Develop peripheral activities.
- Diversify sources of funding.
- Mobilize academic skills.
- Integrate a corporate culture.

The study departments are the stones on which universities build their success, and the structures that link these departments directly to the center of the university, without intermediaries, shorten lines of communication and accelerate decision-making.

Good governance [6] contributes to the success of the institution when the external elements involved, the administrative staff and the scientific community work closely together. On the other hand, progress will be hampered if one of these elements takes over.

Research Management Studies [7]defines twelve characteristics of a productive research environment:

- Clear objectives that have a coordinating function.
- Emphasis on research.
- A culture of specific research.
- A positive group climate.
- Strong participatory governance.
- A decentralized organization.
- Frequent communication.
- Accessible resources (especially human resources).
- The size, age and diversity of the research group must be sufficient.
- Appropriate rewards.
- High emphasis on recruitment and selection.
- Management with the necessary experience and expertise in the field of research to put in place the appropriate organizational structures and apply participatory management methods.

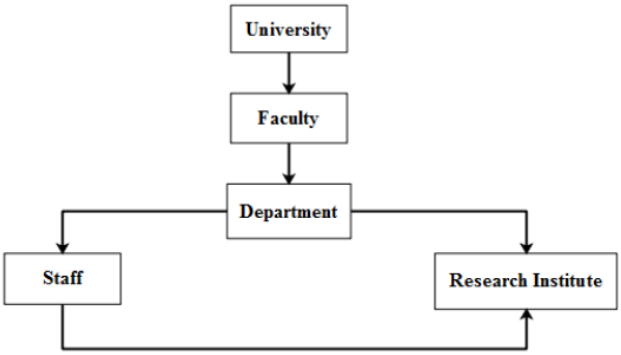


Fig. 1 Scientific Research Architecture

Among the main outcome indicators used in research universities were the following:

- Measures of resources: Financial resources.
- Number of researchers.
- Number and percentage of current researchers.
- Applications for research funding.
- Number of publications (by origin, for example, journals publishing articles submitted for examination by specialists).
- Quotes.
- Directed theses (completed and supported).
- Research applications (patents, licenses).
- Academic awards (editorial positions, Special rewards).

Scientific research competence today is important for many professions and activities: it is necessary not only to creatively apply the obtained knowledge but also to create new knowledge, to carry out the applied researches. Also, scientific research activity is the basic component of developing science education.

The purpose of the research is to describe the current situation of organization and realization of scientific research activity, to define essential factors promoting and hindering students' interest in scientific research activity, to determine lecturers' competence peculiarities in the sphere of organization and realization of scientific research.

In the first study years, the students should create reports, present works, raise problems and propose various problem solution variants; the most important criterion was the competence of experts and current research activities (scientific publications, participation in the national and international projects etc.).

The PhD students have possibilities to participate in seminars, projects, conferences, that lecturers willingly help the students to choose the research subject that interests them.

The participation in conference is great. On the other hand, it is obvious, that the main subject in scientific research activity is the student and sufficient lecturer's contribution promoting this activity; scientific research activity requires consistency, diligence, creativity.

Despite the financial difficulties, the institutions must find the possibilities and form conditions for professional improvement.

Between the most important recommendations:

- Apply modern study methods, promoting critical thinking and new subject search.
- Students should be more involved in the performance of the lecturers' research works, as assistants, putting data in order.
- Develop lecturers and students' team work.
- Include in study programs more subjects for education of research competences.
- Prepare more complex science projects, in which students could participate.

3. LITERATURE REVIEW OF DATA WARHOUSE

Nowadays, decision-makers need a synthetic and global view of the information circulating in their organization in order to guide and adapt their decision-making. To facilitate this process, they use decision support systems. These tools allow decision-makers to have quick and interactive access to a set of data to gain a global view of the activities of a given establishment [8]. The basic concept of a Data Warehouse evolved over the past 20 years.

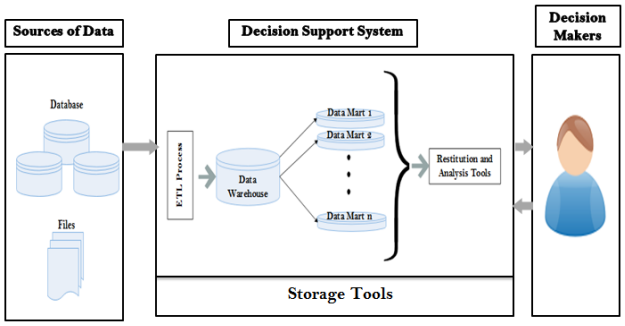


Fig. 2 Architecture of a decision support system

- **Data Warehouse:**

A Data Warehouse is the centralized storage space of an extract of relevant data sources for decision makers. Its organization must facilitate management data in the form of a unified vision and should allow for the conservation of necessary developments for decision-making [9].

- **Data Mart**

A data store is an extract from the warehouse suitable for a class of decision makers or for a particular purpose and organized according to a model adapted to decision-making processes [10].

- **Extraction, Transformation and Loading tools**[11]**:**

Defines a standardized procedure for each Data Warehouse; Extract signifies the connections and rules that the process has to follow to draw data from different sources; validation, completion, and standardization of data need to be done to transform into Data Warehouse compliant structure.

✓ Extraction

Extraction [12] is the first step in the process of providing data to the Data Warehouse. Extracting means reading and interpreting the source data and copying it to the preparation area for further manipulation.

✓ Transformation

It is a series of operations that aims to make the target data homogeneous and can be processed in a consistent way [13].

✓ Loading

This is the operation of loading the cleaned and prepared data into the Data Warehouse[14].

- **Analysis and Restitution of Data:**

Data from a Multi-Dimensional Database is queried using OLAP technology using tools graphs or in a textual language [15].

- **Online Analytical Processing:**

An OLAP system [16] is defined as a decision-making system in which data stores follow a multidimensional organization of data so provide effective support for OLAP analysis.

- **OLAP Data Modeling:**

The aim of multidimensional modeling is to organize data so that OLAP applications are effective and efficient [17], OLAP analyzes consist of following indicators considered as points observed in a space defined by different axes of analysis. Two approaches exist:

- **Modeling in cube:**

Among the first proposed models that speaks of cube[18], This vision manifests itself through a separation between the values and the structure of the elements[19], The data cube is formed by observation axes of indicators placed in the cells, for each observation axis, a graduation is chosen to observe the data at an adequate level of granularity, but among the negative point on this type of modeling, more than three axis of analysis meets representation problems multidimensional spaces.

- **Multidimensional modeling:**

In answering the problem manifested by cube modeling, other approaches try to overcome these limits based on a set of concepts such us: dimension, fact, hierarchy…; but this modeling suffers from the absence of a standardized consensus on this formalism [20], Axes of analysis represented by the dimensions the analyzed variables represented by the indicators.

- **OLAP manipulation operations**

You cannot find an operation set that provides all of the OLAP manipulation operations, but most of the proposals provide partial support for different categories of operations.

Amongst these different operations, the most important are the rotation and drilling operations which are based directly on the metaphor of the cube. Another operation can be found on a scientific literature and the many existing software.

✓ Drilling operations

Drilling operations [21] make it possible to analyze with more or less precision an indicator based on the hierarchical structure of the analysis axes.

The "roll-up" is to analyze the data according to a level of less detailed granularity as opposed to drilling down ("drill-down"), which allows analyzing the data with a finer level in granularity.

✓ Rotation operations

The rotation operations consist in most of the time to change the axis of analysis in use it is called dimension rotation.

The second case of rotation consists in changing the subject of the analysis into a constellation it is a rotation of the fact.

The third case of the rotation consists in maintaining the same axis of analysis but one change in the graduation it is a rotation in the hierarchy.

✓ Restriction operations

Restriction operations [22] can restrict the set analyzed data. ("slice") is to express a restriction on one of the data of one of the analysis axes by specifying a given cube slice. The specification of a sub-cube ("dice") is to express a restriction on data from an analysis indicator.

✓ Transformation operations

It adds a dimension attribute in as an indicator of analysis ("push") or it converted an indicator of analysis into parameter ("pull").

✓ Scheduling operations

"switch" [23] allows changing the position of the values of the parameters of the dimensions or to reorder the parameters of a hierarchy "nest" makes it possible to reorder the parameters of a hierarchy. We nest an attribute in another hierarchy.

4. THE CONTEXT OF STUDY

In this context, even Moroccan public universities are becoming more dependent on using information communication technologies tools for developing, communicating information around the functional processes, especially in the scientific research section. The aim of our research study is to develop a Data Warehouse related to scientific research.

Universities use Information and communications technology for academic purposes that differ from other organizations because they have different environments and circumstances. In order to make more informed decisions about a scientific research department, the ultimate goal for our system is

to facilitate the management of scientific research decision makers to get and find the relevant information they are looking for.

Although, universities are planning to renew their information systems in the future, this necessitates the call for more research efforts in this area. Because, the study concluded that Data Warehouse potentially improves services offered to decision makers and administrative staff to manage and take the right decision based on real and relevant data.

Change management [24] is a primary concern of many universities in terms of adopting a Data Warehouse, as activities, processes, and methodologies.

The aim is to develop a pertinent information system to support researchers and their scientific activities in the Moroccan public university.

- **Justification for the choice of Microsoft SQL Server**

SQL Server is an incredible database product that offers an excellent mix of performance, reliability, ease of administration, and new architectural options, yet enables the developer or DBA to control minute details. SQL Server is a dream system for a database developer.

The business intelligent SQL Server [25] give us the possibility to create a database that summarizes the OLTP data into new tables. Because it is still a refined set of data, it is often insufficient in performing analytics that are holistic to the entire organization.

The implementation of an automatic system for managing scientific research is based on SQL Server Software. The reasons for this choice are:

- Transform complex data.
- Modernize reporting.
- Flexible.
- Enable hybrid BI.

- **Restitution: Multidimensional OLAP Analysis**

Conceptually, Data Mart relies on multidimensional data modeling from the Data warehouse. The data is viewed in the form of dots in a multi-dimensional space with the data cube or hypercube metaphor. Each data represents a cell of the cube. The edges of the cube represent the axes of analysis of the data and include several graduations to allow observation of the data according to different levels of detail. This modeling allows the expression of online analyzes (OLAP) multidimensional.

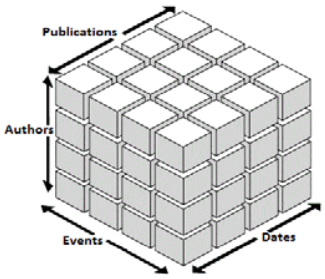


Fig. 3 Sample data cube

The interrogation of a multidimensional database is a succession of exploration operations. Following a multidimensional analysis of the technology used, or by OLAP analysis with reference to the type of data handled [26]. An OLAP analysis runs as follows:

- Selection of a first request
- Data Navigation by (drilling down, drilling up, data restriction, etc.)

By the notion of fact, we designate the object we want to analyze. In fact, one or more measures (indicators) are observed. A set of values is associated with each fact; these values are taken by the fact for each measure. These measurements are generally numerical.

A dimension offers the user points of view differentiated to analyze or observe facts and defines an axis of analysis. It consists of one or more attributes, called hierarchical levels. It corresponds to the levels of detail that can be observed on the facts. A level is composed of elements called members or modalities. The links between levels can have different cardinalities.

An OLAP cube represents the aggregated values of the measure in a multidimensional space defined by the user. he chooses the dimensions according to which he wants to analyze the facts and also for each dimensions the level the hierarchy on which he wants to work. The combination of the modalities of the selected dimensions designates the cells of the cube. The latter contains the value of the measure or measures corresponding to the combination modalities. This value of the measure can be either a detailed value or an aggregated value according to the hierarchical levels chosen.

5. CASE STUDY:

In this paper, we propose the study of publications of the research laboratory belonging to our public university for the purpose of analyzing scientific publications according to the main field of research.

The example in the figure 4 corresponds to the modeling of scientific research publication based on authors and keywords over time. The axes of the analysis are represented by the authors, the support, the keyword and the time dimensions. The dimension time is characterized by four parameters day, month, quarter, year organized hierarchically: the day parameter represents a finer graduation (to observe the precise date finer) than the graduation month, itself being a finer graduation of year. It is possible to group the month into quarter. The authors dimension contains two hierarchical levels in order to group the authors based on their status. The genre publication dimension contains the id of the work, the journal as well as the number of pages, the volume, … according to its national or international scope.

- Basic concepts:

- F all the facts,

- D all the dimensions,

- H all the hierarchies,

- M all measures,

- I the set of instances of dimensions,

- A the set of dimension attributes.

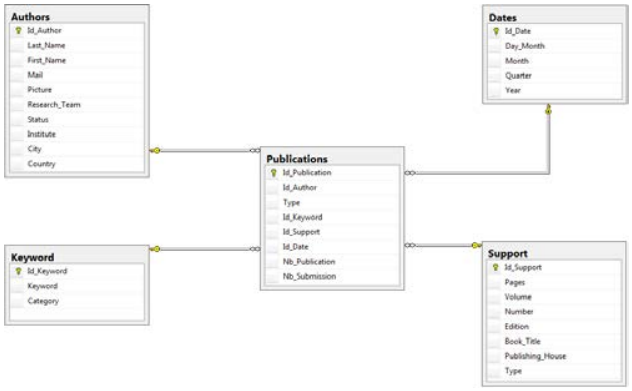


Fig. 4 Multidimensional modeling of publications

For example we can say that the dimension author D_1^{Authors} contains one level H_1^1Author, and the level H_1^1 contains several modalities between themes we find a_1^{11} Professor, a_2^{11} PhD student.

To count the number of publication in fact table, we use the aggregation functions in our case to count the number of publication related to Author x we use COUNT this function counts the number of instances in an aggregate.

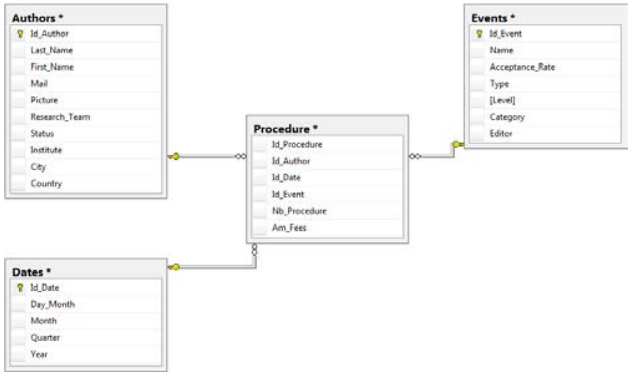


Fig. 5 Multidimensional modeling of procedures

For this case, we can say that the dimension D_3^{Events} contains one level H_3^1 Event; the level H_3^1 contains two modalities a_3^{11} national, a_3^{12} international.

To count the number of procedure in fact table, we use the Aggregation functions in our case to count the number of procedure related to laboratory **x** we use COUNT this function counts the number of instances in an aggregate.

A fact and its associated dimensions make up a schema star. One possible generalization is to describe a "constellation of stars" consisting of several facts and several dimensions possibly shared forming a constellation pattern.

A generalization from the two preceding figure consists in proposing a constellation model [27] which groups together a set of facts associated with dimensions which are provided with multiple hierarchies.

A constellation is defined by (N^{SR}, F^{SR}, D^{SR}, $Star^{SR}$) where:

- N^{SR} is the name of the constellation,

- $F^{SR} = \{F_1, \ldots\ldots\ldots\ldots\ldots\ldots F_n\}$ is a set of facts,

- $D^{SR} = \{D_1, \ldots\ldots\ldots\ldots\ldots\ldots D_m\}$ is a set of dimensions,

-$Star^{SR}: F^{SR} \rightarrow 2^{D^{SR}}$ is a function that combines facts with dimensions.

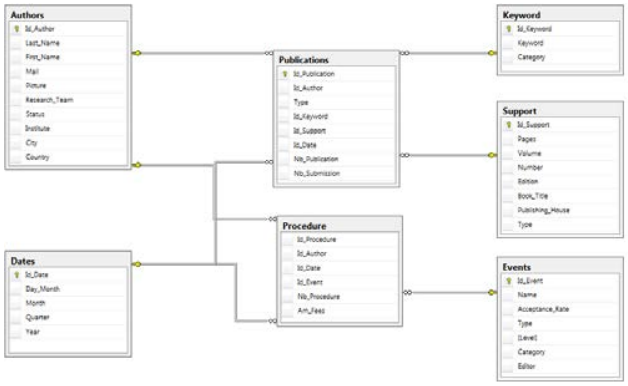


Fig. 6 Constellation schema of the multidimensional database

To simplify we designate the concept of fact F_i, respectively of dimension D_i and Hierarchy H_j, by its name N^{F_i}, respectively N^{D_i} and N^{H_j}, our formal description of the schema in constellation is as follows:

This constellation scientific research (SR) has two facts and five dimensions. It is defined by (N^{SR}, F^{SR}, D^{SR}, $Star^{SR}$) where:

N^{SR}= ' Scientific Research ';

$F^{SR} = \{F^{Publication}, F^{Procedure}\}$;

$D^{SR} = \{D^{Authors}, D^{Keyword}, D^{Dates}, D^{Support}, D^{Events}\}$;

$Star^{SR}(F^{Publication}) \rightarrow \{D^{Authors}, D^{Keyword}, D^{Dates}, D^{Support}\}$;

$Star^{SR}(F^{Procedure}) \rightarrow \{D^{Authors}, D^{Dates}, D^{Events}\}$;

The procedures carried out by researchers of the doctoral center can be studied according to the fact:

$F^{Procedure} = (N^{Procedure}, F^{Procedure}, D^{Procedure}, Star^{Procedure})$ where:

-$N^{F^{Procedure}} = ' Procedure'$;

-$M^{F^{Procedure}} = \{Nb\_Procedure, Am\_Fees\}$;

-$I^{F^{Procedure}} = \{i_1^{Procedure}, \ldots\ldots, i_n^{Procedure}\}$;

-$IStar^{F^{Procedure}} = \{i_k^{Procedure} \rightarrow \{i^{Authors}, i^{Events}, i^{Dates}\} | k \in [1..y], I_k{}^{Procedure} \in I^{F^{Procedure}} \wedge \exists i_{k_i}^{Authors} \in I^{D^{Authors}} \wedge \exists i_{k_j}^{Events} \in I^{D^{Events}} \wedge \exists i_{k_p}^{Dates} \in I^{D^{Dates}}\}$.

6. FROM DATA WAREHOUSE TO DATA MINING

In the previous section, we talked about the Data Warehouse its conception and its development, in this section we try to introduce some algorithms of the Data Mining [28] in order to better explore the existing data and to offer to the decision-maker a more realistic view in making predictions based on existent data in the Data Warehouse, for that we had to choose the datamining algorithm that produces results with the smallest possible error rate.

Data mining, contrary to the data warehouse, tries to find the independence and relationships that exist between the data based on a set of methods of adequate mathematical and statistical analysis.

The first example is evaluation of the number of endowment provided by the university council to the researchers based on the following criteria: Year Thesis, Supported this Year, Opinion of the Head of the Research Entity, Gender. We choose to work with Naïve Bayes algorithm, the reasons for this choice are:

1. For each hypothesis: We associate a probability observation of one or several instances may change this probability.

2. We can talk about the most hypotheses likely, based on the conditional probabilities and Bayes rule.

3. Forecasting the future from the past, while assume independence attributes.

4. Bayesian probability is the estimation of an event knowing a preliminary hypothesis is verified (knowledge).

Microsoft Naive Bayes [29], chooses Year Thesis as the most significant attribute for the first level. After that the algorithm chooses Opinion of the Head of the Research Entity and Supported this Year as the next attributes by importance and finally the Gender. At the end, there are evaluations of expected endowment depending on the values from analyzed attributes.

Input data will be randomly split into two sets, a training set and a testing set, with percentage of data for testing around 30%. The training set is used to create the mining model. The testing set is used to check model accuracy.

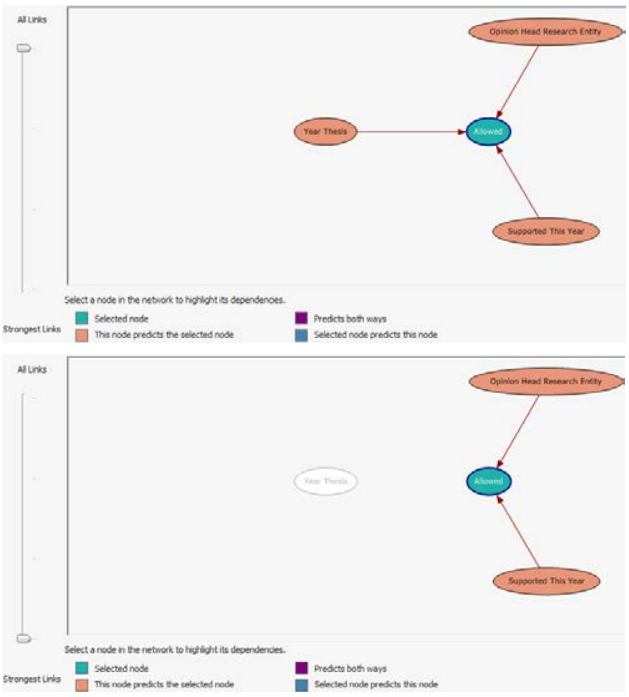


Fig. 7: Analysis of correlation between criterions for evaluation of the number of care provided by the university council.

Figure7 represents correlation between criteria for evaluation of the number of endowment provided by the university council. By moving cursor to the left-hand side of the Figures, the intensity of correlation between observed criterions is displayed. Therefore it is easily perceived that the obtained results depends on Year Thesis the least, the most on both Supported this Year, and Opinion of the Head of the Research Entity.

The second example discusses the grants giving to the laboratories based on the following criteria: Number Registered, Number Publication, and Number Event. Microsoft Naive Bayes chooses Number Publication as the most significant attribute for the first level. After that the algorithm chooses Number Registered as the next attributes by importance and finally the Number Event. At the end, there are evaluations of expected grants depending on the values from analyzed attributes.

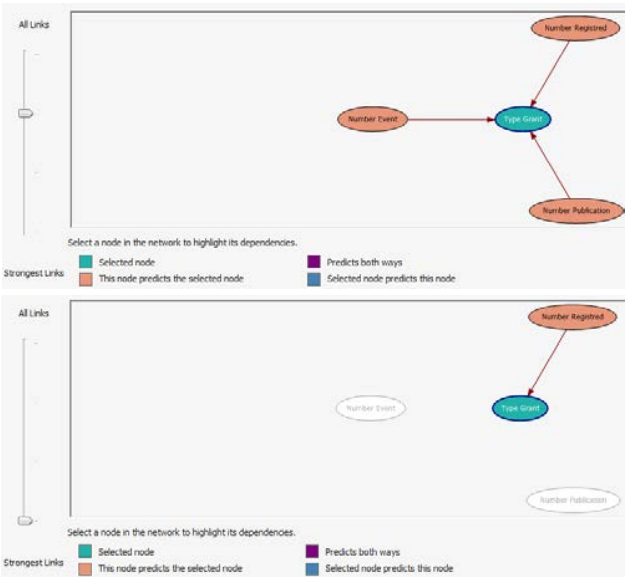


Fig. 8: Analysis of correlation between criterions for evaluation of the grants giving to the laboratories.

Figure8 represents correlation between criteria for evaluation of the grants giving to the laboratories. By moving cursor to the left-hand side of the figures, the intensity of correlation between observed criterions is displayed. Therefore it is easily perceived that the obtained results depends on Number Event the least, a bit more on Number Publication, the most on the Number Registered.

In our case, Now we want to decide situation when the Number of new Registered in the laboratory in the previous year is 5, the Number of Publications in the previous year is 14 and the Number of Events organized is 2 which is not available into our trained dataset.

Our example gets classified as 'Medium' about the type of grant giving to the laboratory.

To conclude, on the one hand the use of the data warehouse is a flexible solution for decision-makers in universities and on the other hand it is adapted to the administrative staff in the university because it can used daily tool for explore Data Warehouse like Microsoft Excel unlike other Olap tool.

Another advantage manifests through research on knowledge and information exciting in the Data Warehouse without having deep knowledge on the complex query languages.

Data Mining allows the user to analyze a large number of data by offering the possibility of the prediction at the requested time; the latter offers the decision-maker the opportunity to explore the current data and to have an idea about the future behavior of the institution. This marriage of existing data explorations and prediction of future behavior helps decision makers in the process of solving the problems.

7. Conclusion

The new orientation by the government by adopting the governance in all aspects of institutional Moroccan life leads to the change of the way of employing information system in this sector because the affect administrative staff, the quality of services, and the time of respond, etc.

Scientific research represents one of the axes that need modernization and automatization by adopting and implementing new solution like Data Warehouse. In this direction, this paper has examined the needs presented by decision-makers and the administrative staff and present as solution the implementation of Data Warehouse and Data Mining.

we demonstrated through the use of the Data Warehouse and Data Mining on a corpus of data relative to the university and especially on the scientific research without preliminary need of the deep knowledge on the algorithm of datamining nor on the language of request in a single and ultimate goal of optimizing the decision-making process for solving the problems encountered.

8. REFERENCES

[1] V. Lamanauskas and D. Augienė, "Development of Scientific Research Activity in University: A Position of the Experts," *Procedia - Soc. Behav. Sci.*, vol. 167, no. Supplement C, pp. 131–140, Jan. 2015.

[2] T. G. Weiss, "Governance, good governance and global governance: Conceptual and actual challenges," *Third World Q.*, vol. 21, no. 5, pp. 795–814, 2000.

[3] "Llorent-Bedmar - 2014 - Educational Reforms in Morocco Evolution and Curr.pdf." .

[4] "Morocco still aiming to boost education quality and access," *ICEF Monitor - Market intelligence for international student recruitment*, 09-Feb-2015. .

[5] D. A. Clark, "'The Two Joes Meet—Joe College, Joe Veteran': The G.I. Bill, College Education, and Postwar American Culture," *Hist. Educ. Q.*, vol. 38, no. 2, pp. 165–190, ed 1998.

[6] "M. Shattock - 2006 - Higher Education Management and Policy.pdf." .

[7] M. N. K. Saunders and F. Bezzina, "Reflections on conceptions of research methodology among management academics," *Eur. Manag. J.*, vol. 33, no. 5, pp. 297–304, Oct. 2015.

[8] G. Colliat, "OLAP, Relational, and Multidimensional Database Systems," *SIGMOD Rec*, vol. 25, no. 3, pp. 64–69, Sep. 1996.

[9] "Wiley: Building the Data Warehouse, 4th Edition - W. H. Inmon." [Online]. Available: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0764599445.html. [Accessed: 22-Nov-2017].

[10] D. L. Moody and M. A. Kortink, "From enterprise models to dimensional models: a methodology for data warehouse and data mart design.," in *DMDW*, 2000, p. 5.

[11] P. Vassiliadis, "A Survey of Extract–Transform–Load Technology," *Int. J. Data Warehous. Min. IJDWM*, vol. 5, no. 3, pp. 1–27, Jul. 2009.

[12] "Wiley: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data - Ralph Kimball, Joe Caserta." [Online]. Available: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0764567578.html. [Accessed: 22-Nov-2017].

[13] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 23, no. 2, pp. 91–104, Jul. 2011.

[14] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng Bull*, vol. 23, no. 4, pp. 3–13, 2000.

[15] "Data warehousing tool's architecture: from multidimensional analysis to data mining - IEEE Conference Publication." [Online]. Available: http://ieeexplore.ieee.org/abstract/document/617388/?reload=true. [Accessed: 22-Nov-2017].

[16] N. W. Alkharouf, D. C. Jamison, and B. F. Matthews, "Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases," *J. Biomed. Biotechnol.*, vol. 2005, no. 2, pp. 181–188, 2005.

[17] "Amazon.fr - The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses - Ralph Kimball - Livres." [Online]. Available: https://www.amazon.fr/Data-Warehouse-Toolkit-Techniques-Dimensional/dp/0471153370. [Accessed: 22-Nov-2017].

[18] A. Datta and T. Helen, "The cube data model: A conceptual model and algebra for online analytical processing in data warehouses," 1999. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.4722&rep=rep1&type=pdf. [Accessed: 22-Nov-2017].

[19] R. Torlone, "Multidimensional Databases," M. Rafanelli, Ed. Hershey, PA, USA: IGI Global, 2003, pp. 69–90.

[20] S. Rizzi, "OLAP Preferences: A Research Agenda," in *Proceedings of the ACM Tenth International Workshop on Data Warehousing and OLAP*, New York, NY, USA, 2007, pp. 99–100.

[21] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, "Algebraic and Graphic Languages for OLAP Manipulations," *Int. J. Data Warehous. Min. IJDWM*, vol. 4, no. 1, pp. 17–46, Jan. 2008.

[22] C. Ciferri, R. Ciferri, L. Gómez, M. Schneider, A. Vaisman, and E. Zimányi, "Cube algebra: A generic user-centric model and query language for OLAP cubes," *Int. J. Data Warehous. Min. IJDWM*, vol. 9, no. 2, pp. 39–65, 2013.

[23] J. Li and B. Xu, "ETL tool research and implementation based on drilling data warehouse," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 2010, vol. 6, pp. 2567–2569.

[24] B. H. Wixom and H. J. Watson, "An Empirical Investigation of the Factors Affecting Data Warehousing Success," *MIS Q.*, vol. 25, no. 1, pp. 17–41, 2001.

[25] "Wiley: Microsoft SQL Server 2012 Bible - Adam Jorgensen, Jorge Segarra, Patrick LeBlanc, et al." [Online]. Available:

http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118106873.html. [Accessed: 22-Nov-2017].

[26] N. Kumar, A. Gangopadhyay, G. Karabatis, S. Bapna, and Z. Chen, "Navigation Rules for Exploring Large Multidimensional Data Cubes," *Int. J. Data Warehous. Min. IJDWM*, vol. 2, no. 4, pp. 27–48, Oct. 2006.

[27] J. Pokorny, "Dealing with Dimensions in Data Warehousing," in *Knowledge Discovery for Business Information Systems*, Springer, Boston, MA, 2002, pp. 307–324.

[28] "Data Mining: Concepts and Techniques - 3rd Edition." [Online]. Available: https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1. [Accessed: 22-Nov-2017].

[29] C. L. Curotto and N. F. F. Ebecken, *Implementing Data Mining Algorithms in Microsoft SQL Server*. Southampton: WIT Press / Computational Mechanics, 2005.

Comparative analysis of HKNNRF and MLP for Block Cipher Algorithm identification

1nd Hamza Allaga
Laboratoire d'Innovation en Mathématiques et Applications
Théorie de l'Information - LIMATI.
Department of Mathematicsand and Informatics
Sultane Moulay Slimane University, Beni Mellal, Morocco.
hamzaallaga@gmail.com

2nd Khadija Lahrouni
Laboratory of Electrical Systems, Energy Efeciency and Telecommunications (LSEEET)
Department of Applied Physics, Faculty of Sciences and Technologies
Cadi Ayyad University, Marrakesh, Morocco

3rd Abderrazak Farchane
Laboratoire d'Innovation en Mathématiques et Applications
Théorie de l'Information - LIMATI
Department of Mathematicsand Informatics
Sultane Moulay Slimane University, Beni Mellal, Morocco.

4th Said Hakimi
Laboratoire d'Innovation en Mathématiques et Applications
Théorie de l'Information - LIMATI
Department of Mathematicsand and Informatics
Sultane Moulay Slimane University, Beni Mellal, Morocco.

***Abstract*—In this concise comparison, two studies were analyzed. These studies used machine learning techniques to improve the detection of cryptographic algorithms. Study A implemented a Hybrid K-Nearest Neighbor with Random Forest (HKNNRF) algorithm, whereas Study B utilized a Multi-Layer Perceptron (MLP) on block ciphers. For binary classification, HKNNRF and MLP had an acceptable accuracy, but their performance notably declined in multi-class scenarios, falling below 0.45 in Study B and to 0.36 in Study A. In the following examination, an assessment of the strengths and weaknesses of both studies is provided, underscoring the importance of combining various machine learning methodologies to fortify cryptanalysis. It is recommended that forthcoming studies focus on enhancing cryptographic security by creating sophisticated hybrids of deep learning and ensemble learning, rigorously evaluated using diverse and extensive datasets.**

***Index Terms*—Machine learning, cryptographic algorithm detection, deep learning, ensemble learning, Multi-Layer Perceptron, K-Nearest Neighbor, Random Forest, cybersecurity.**

I. Introduction

Cryptography holds a pivotal role in safeguarding privacy and ensuring secure communication, especially in the face of advancing computer technology [1]. The practice of cryptanalysis, involving the analysis and potential breaking of cryptographic systems, is crucial for upholding data transmission security [2]. The identification of block cipher algorithms refers to the process of analyzing and identifying the encryption algorithm used in a cryptographic system [3]. thus it stands as a cornerstone in this field.

Recently, Several papers proposed different methods and techniques for block cipher algorithm identification, such as [4] which proposed a method for identifying block cipher algorithms, specifically DES, 3DES, AES, and Blowfish, based on ciphertext sequences and NIST randomness test standard.

Identify applicable funding agency here. If none, delete this.

Other studies goes with extracting statistical features of ciphertext such as letter frequency information of ciphertext [4], [5]. meanwhile, [6] proposed a block cryptosystem recognition scheme based on Hamming weight distribution, which extracts cipher text features and uses an ensemble learning model to improve accuracy.

The primary technique for cryptosystem recognition is the classification approach based on machine learning [7]. Machine learning (ML) offers a compelling alternative. Due to its capacity to learn complex patterns within ciphertext data and effectively distinguishing between different block ciphers. In this regard we question the robustness of ML algorithms to handle cryptography problems. To what length its given accuracy can outperform classical methods? To our knowledge, two papers have proposed ML algorithms for Block Cipher Identification. The first one have used a combination of the K-Nearest Neighbor and Random Forest (HKNNRF), and the second is being based on Multi-Layer Perceptron (MLP). Thus, We aim in this paper to delve into a deep comparative analysis of these two studies in order to explore their pros and cons. for the rest or the article we're considering the following; Study A, represents the one with HKNNRF [8], in the other hand, [9] is always referd to as Study B.

Following the introduction, the paper is organized as follows: Section I provides an overview of the two Studies under consideration. Section II outlines the methodology adopted in both Studies. Section III explores the results with a focus on different metrics. Section IV offers a performance comparison of the findings. Finally, Section V provides suggestions for future research to improve accuracy within the same context.

II. Overview of Study A and B

A. Study A

Cryptographic algorithm identification plays a crucial role in cryptanalysis, but it becomes challenging when only ci-

phertexts are available. Traditional single-layer classifiers such as SVM, KNN, and RF face limitations in accuracy and efficiency. To address this challenge, authors of A proposed HKNNRF algorithm; a model that combines the strengths of KNN and RF algorithms. The model is inspired by the NIST randomness test methods and extracts ciphertext features to achieve superior accuracy compared to traditional classifiers.

Study A is authored by Ke Yuan and al. and was published in PeerJ Computer Science on 10th October 2022. the paper provides an in-depth overview of cryptographic algorithm identification principles, highlighting the transition from statistical methods to machine learning techniques. It also emphasizes the limitations of existing approaches and the necessity for ensemble learning models like HKNNRF.The model's workflow involves feature extraction using randomness tests, followed by training and testing stages, and providing a comprehensive and systematic approach to cryptographic algorithm identification.

Experimental results demonstrate the effectiveness of HKNNRF across different ciphertext file sizes and cryptographic algorithms. In binary-classification experiments, HKNNRF achieves impressive accuracy rates, outperforming baseline models such as SVM, KNN, and RF. The average binary-classification accuracy of HKNNRF is reported to be 69.5%, showcasing its ability to accurately identify block cipher algorithms.

B. Study B

The research of Ke Yuan and al. provided a new insight into the field of cryptanalysis through their Study B. This used the MLP algorithm to bridge the gap in accurately and consistently identifying block cipher algorithms. Authors detailed a methodology that derives feature sets from ciphertext using 15 different randomness tests proposed by the National Institute of Standards and Technology (NIST). They then distill these down to 10 salient features that are fed into the MLP for classification. The scope of the study included five popular block cipher algorithms, including AES, 3DES, Blowfish, CAST and RC2. The classifiers were rigorously evaluated in binary and multi-class settings against conventional machine learning benchmarks.

Intriguingly, the outcomes of this analysis cast the MLP classifier in a favorable light, evidencing a dominant average accuracy that eclipses the 75% threshold in binary classifications, a notable stride beyond its classical counterparts. The MLP's resilience against the variability of ciphertext lengths further solidifies its preferability. The research underscores the meticulous feature selection and the strategic implementation of the MLP framework, setting a precedent for integrating deep learning into the domain of cryptographic algorithm identification.

III. Methodology

In this section, methodologies of identifying cryptographic algorithms in each Study; A and B are presented. both approaches have incorporated diverse research methods and techniques to accurately distinguish cryptographic algorithms from ciphertext data. These works provided thorough frameworks for advancing the field through careful data preparation, feature extraction, classification algorithms, and rigorous experimental evaluation.

A. Research Methods in Study A

A comprehensive methodology for identifying block cipher algorithms based on ciphertext analysis is presented in Study A. its methodology encompasses several key components, including data preparation, feature extraction, classification algorithms, and experimental evaluation.

Data preparation phase involves generating ciphertext files using different cryptographic algorithms such as AES, 3DES, Blowfish, CAST, and RC2. These ciphertext files are created with fixed key parameters and in Electronic Code Book (ECB) mode, and are of varying sizes ranging from 1 KB to 512 KB. The study also applies 15 different NIST randomness test methods to extract features from the ciphertexts [10]. The chosen randomness test methods aim to capture statistical properties and randomness characteristics of the ciphertexts.

Feature extraction is a crucial step in the methodology, as it involves selecting 10 useful features from the extracted feature values to be used in classification. These features are derived from the results of the NIST randomness test methods and are intended to provide meaningful insights into the statistical and random properties of the ciphertexts.

The study employs four classification algorithms for comparison: Support Vector Machine (SVM), KNN, RF, and a hybrid model named HKNNRF. Each classification algorithm has its own unique approach to processing the extracted features. SVM uses the Gaussian kernel function, RF constructs multiple decision trees and integrates them, and KNN relies on proximity-based classification.finally, HKNNRF combines the advantages of both KNN and RF, utilizing the stacking technique in ensemble learning.

For experimental evaluation, binary and five classification experiments are performed on the cryptographic algorithms using the extracted features. The accuracy, precision and recall of each classification model are evaluated. In addition, the impact of file size on classification accuracy is evaluated. Confusion matrices are used to analyse the classification results and compare the performance of different algorithms.

A simplified summary of the whole methodology followed by authors of Study A for training , testing, features selecting, input and outupt the model KNNRT is illustrated in Figure 2.

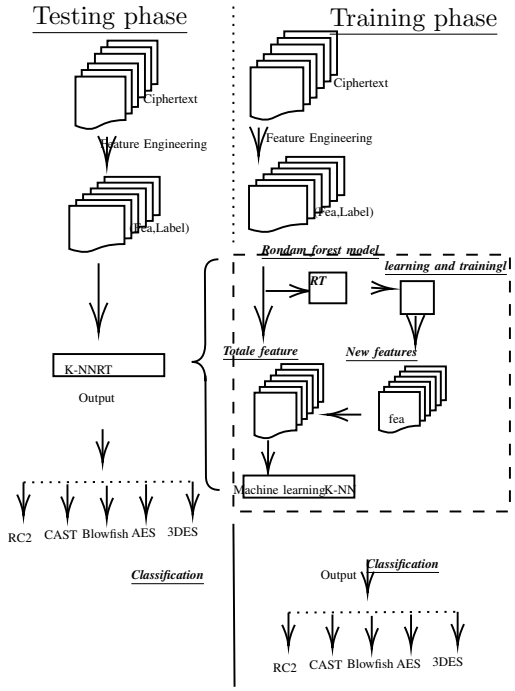


Fig. 1. The pattern of the identification mechanism for encryption algorithms, built around the HKNNRF model.

B. Research Methods in Study B

The study on cryptographic algorithm identification involves the use of various methods and techniques to accurately determine the specific cipher algorithm used to encrypt a given ciphertext file. This essay explores the use of cryptographic algorithms, feature extraction, the identification scheme, the operational workflow, the Multi-Layer Perceptron (MLP) algorithm, and the experimental environment in conducting this study.

First and foremost, the study involves the use of plaintexts generated using different cryptographic algorithms, which are represented as a set containing different cryptographic algorithms. These cryptographic algorithms are denoted as a_i, where 'i' ranges from 1 to k. The features extracted from the ciphertext files play a crucial role in identifying the cryptographic algorithm used, and these features are represented as a feature vector obtained through the extraction process, resulting in a set of d-dimensional features.

The identification scheme, represented as $\Delta = (A, I, h)$, comprises the set of cryptographic algorithms, the identification scheme, and the accuracy rate of the identification process. The identification scheme itself is represented by a triple $(P = (oper, fea, CA))$, which includes the operational workflow for identification, the features extracted from the ciphertext, and the classification model used, such as the MLP model.

The operational workflow of the identification scheme involves two stages: training and testing. During the training phase, a group of ciphertext files with known cryptosystems is collected, features are extracted using randomness tests, and a classification model is trained. In the testing phase, features are extracted from the ciphertext file to be identified, and these features are inputted into the trained classification model to predict the cryptographic algorithm used.

Randomness testing methods, including those defined in NIST FIPS140-2 [10], are employed to extract features from ciphertext files. These methods include tests for runs, binary matrix rank, template matching, and serial tests, among others. Features extracted from these tests are used to construct a feature vector representing the characteristics of the ciphertext.

The MLP algorithm, which is a type of artificial neural network used for classification tasks, is utilized in the study. It consists of input, hidden, and output layers, with fully connected nodes between layers. The activation function introduces non-linearity to the model, and during training, weights between nodes are adjusted iteratively using backpropagation to minimize prediction errors.

The experimental environment involves the use of machine learning models such as SVM, GNB, KNN, RF, and LR for data preparation, feature extraction, and evaluation. Data preparation involves generating ciphertexts using various encryption algorithms with specific parameters, and randomness tests are applied to extract features from ciphertext files. Evaluation is conducted using methods such as repeated random subsampling validation, where data is split into training and test sets for assessing classification accuracy.

A streamlined overview of the comprehensive methodology outlined by the authors of Study B is presented in Figure 2. This methodology encompasses the processes of training, testing, feature selection, as well as input and output considerations for the MLP model.

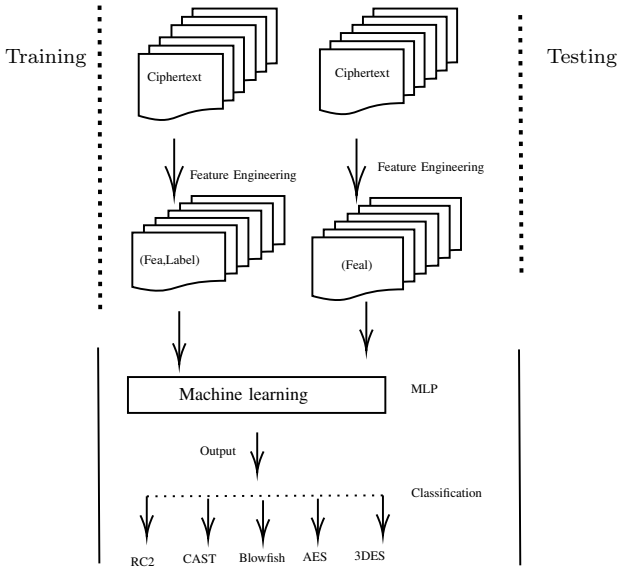


Fig. 2. Diagram for determining the block cipher algorithm by applying the MLP technique

IV. Findings

This section examines and discusses the findings and conclusions from both Studiess.

A. Outcomes of Study A

The findings from Study A present results from the use of two distinct methodologies to investigate the classification and recognition of encryption algorithms. The initial methodology concentrated on the binary classification of encryption algorithms, employing SVM, KNN, RF, and HKNNRF models to scrutinize 10 relevant features extracted for analysis. The results indicated that HKNNRF consistently displayed superior performance compared to other models, achieving the highest average accuracy of 72.5% for both AES and 3DES algorithms. Furthermore, it can be deduced that ensemble learning, as epitomized by HKNNRF, exhibited a competitive advantage over single-layer machine learning models : SVM, KNN, and RF. This is evident from the average identification accuracy of SVM, KNN, RF, and HKNNRF on varying sizes of ciphertext files, which averaged 0.565, 0.57, 0.595, and 0.695, respectively.

The second methodology focused on multiclassification recognition of cryptographic algorithms, utilizing a five-classification model based on 10 ciphertext features to identify the algorithms AES, 3DES, Blowfish, CAST, and RC2. Similar to the binary classification methodology, HKNNRF outperformed single-layer KNN, RF, and SVM models by significant margins, achieving the highest identification accuracy of 34%. The findings also indicated a correlation between file size and model performance, as well as the influence of different cryptographic algorithms on identification accuracy.

B. Outcomes of Study B

Using MLP algorithm by Study B has shown an improve in accuracy and stability in comparison to traditional machine learning models. By employing 15 randomness testing methods to extract ciphertext features, the Study ensures that relevant information is accurately captured for the identification task. The results of the study demonstrate the superiority of the MLP algorithm in binary classification, as well as its higher stability and robustness across different ciphertext file sizes.

In terms of binary classification results, the MLP model consistently outperforms other classical machine learning models across different ciphertext file sizes for encryption algorithms such as AES and 3DES. The average identification accuracy of the MLP model is notably higher, reaching up to 76.5% in comparison to the traditional models.

The average identification accuracy for five classification for Study B has shown a range of 34.2% to 41.6% of accuracy.

Furthermore, the effectiveness of feature extraction is demonstrated through the selection of 10 meaningful features out of 15 randomness testing methods. This approach ensures that relevant information is accurately captured for input to the MLP classifier, contributing to the overall superior performance of the algorithm in binary classification.

C. Performance Comparison

In the realm of data classification, both binary and multi-classification methods have been scrutinized and compared across various Studies. Figure 3 provides a comparative analysis of MLP and HKNNRF performance metrics across different file sizes in kilobytes (KB) for binary classification. Study B highlights the superior performance of the MLP model over traditional models, particularly in accurately identifying AES and 3DES encryption algorithms with an accuracy of up to 76.5%. Notably, the MLP model demonstrates a noticeable uptrend in performance with larger file sizes, with accuracy, precision, and recall all witnessing enhancements as file size expands, notably precision escalating from 0.79 to 0.867 as file sizes progress from 256 KB to 512 KB. Conversely, Study A designates the HKNNRF model as the most effective among the tested models, achieving the highest average accuracy of 72.5% for AES and 3DES. However, unlike the MLP model, the performance of HKNNRF does not display a consistent trend with changes in file size. While there is a slight rise in accuracy at the largest file size, both precision and recall fluctuate at lower file sizes, as depicted in Figure 3. The visual representation through line plots with distinct markers for each performance metric effectively underscores the contrast between MLP's consistent improvement with data size and HKNNRF's variable performance, thereby highlighting MLP's potential robustness in handling larger volumes of data.

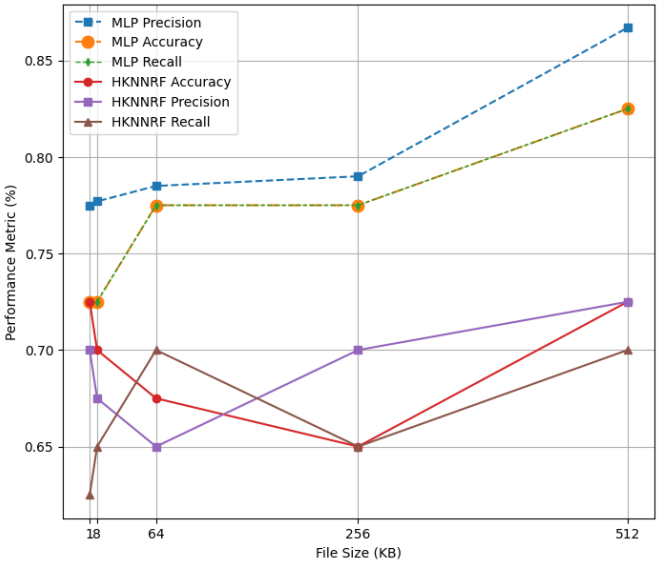


Fig. 3. Binary Classification Performance Comparison of MLP and HKNNRF Across File Sizes.

In multi-classification studies, Study B showcases notable success with MLP, indicating an average identification accuracy ranging from 34.2% to 46%, albeit without specifying exact figures in the comparison text. Conversely, Study A achieved an accuracy of 22% to 34% in multi-classification tasks. suggesting its performance HKNNRF model but may not rival MLP's effectiveness as presented in Study B. Figure 4

provides a comparative analysis of Multi-layer Perceptron and k-nearest neighbors with random forests performance across various file sizes in multi-classification scenarios. Both algorithms display differing levels of accuracy, precision, and recall across file sizes. MLP exhibits consistent performance, with accuracy ranging from 0.34 to 0.39 and precision from 0.34 to 0.42, while HKNNRF shows more fluctuating metrics, with accuracy ranging from 0.24 to 0.34 and precision from 0.23 to 0.37. This comparison underscores the distinct behaviors of the algorithms in multi- classification tasks, offering insights into their suitability for different file size scenarios.

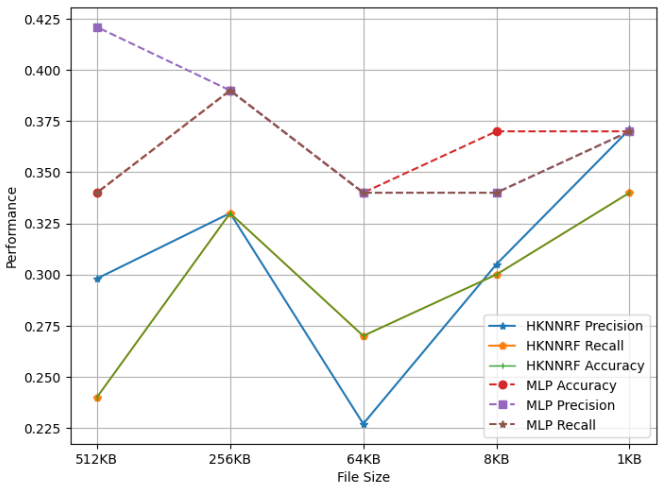


Fig. 4. Multi-Classification Performance Comparison of MLP and HKNNRF Across File Sizes.

V. Impact and Relevance

Both studies make significant contributions to the field of cryptographic algorithm identification by introducing innovative methodologies and achieving exceptional levels of accuracy. Study A underscores the promise of utilizing deep learning-based techniques for the identification of encryption algorithms, while Study B underscores the efficacy of ensemble learning models. The generalizability of the findings in both studies may be affected by external factors such as sample size and the specific algorithms selected. Additional research is warranted to corroborate and expand upon these findings, examining the integration of deep learning and ensemble learning methodologies to enhance the accuracy and reliability of cryptographic algorithm identification.

VI. Discussion

The following underscores the importance of various machine learning techniques in identifying cryptographic algorithms, comparing the approaches of Study A and Study B. Study B focuses on MLP in deep learning, while Study A utilizes HKNNRF model in ensemble learning. Both studies show promising findings in identifying block cipher schemes, indicating the potential for integrating diverse methods for improved precision. Nevertheless, there are limitations that necessitate future research to tackle these issues for further progress in cryptographic algorithm identification.

VII. Weakness

First and for most, The use of the same datasets in both studies A and B presents limitations in their practical application and distribution in real-world scenarios. Both studies focus on encryption algorithms and fixed parameters, neglecting important asymmetric algorithms such as RSA and ECC. This narrow focus restricts the potential usability of the models in real-world situations.

Furthermore, the sensitivity of both studies to variations in file sizes and dataset quality raises concerns about the reliability of these models across different data sizes. The lack of insight into feature extraction and selection processes in both studies complicates the optimization and understanding of models performance. These deficiencies highlight the necessity for more comprehensive approaches that encompass a wider range of algorithms, robust feature extraction methods, and thorough evaluation metrics to ensure the effectiveness and reliability of cryptographic identification models in real-world applications.

The inability of the models to handle multi-classification, with a high accuracy of only 46% for Study B and 36% for Study A, further highlights the limitations of the current approaches. The previous cryptosystem proved to be inadequate for handling multi-classification and large datasets, thus reducing its practical applicability.

there are numerous challenges that necessitate further investigation in order for us to resolve them. The current models' limitations in handling a wider range of algorithms, dataset variability, and multi-classification require more comprehensive approaches and thorough evaluation to ensure their practical applicability in real-world scenarios. Further research and development in this area are essential to overcome the current limitations and enhance the usability of cryptographic identification models.

VIII. Suggestions for Future Research

A variety of significant areas come to light when considering possible directions for further study in the field of cryptographic identification models. Initially, there is a need to expand algorithms, particularly by incorporating a wider variety of encryption algorithms, including asymmetric ones. This expansion would enhance the flexibility and applicability of models in real-world scenarios. Secondly, it is critical to broaden datasets to encompass a wider range of file sizes and types in order to assess the resilience of models under more realistic conditions. Furthermore, advancing feature engineering techniques tailored to cryptographic datasets could significantly enhance model performance and interpretability. Moreover, investigating algorithms for deep learning to achieve optimal results in multi-classification strategies and scalability solutions will be essential for handling complex classification tasks and large-scale datasets. Furthermore, focusing on post-detection measures to provide actionable insights after

encryption identification, conducting practical application trials.Through tackling these facets, scholars can propel the development and optimization of cryptographic identification models, ultimately amplifying their efficacy and pragmatic suitability.

IX. Conclusion

To sum up, both investigations greatly progress the identification of cryptographic algorithms. With the introduction of the HKNNRF algorithm in Study A, the potential of ensemble learning models in precisely identifying encryption schemes is demonstrated. Its exceptional results in binary and multiclassification tests demonstrate how useful it could be in practical settings. In contrast, Study B highlights the MLP algorithm's efficacy, especially in terms of enhancing accuracy and stability when contrasted with conventional models. The results highlight the significance of combining various machine learning approaches for reliable identification of cryptographic algorithms. The two research projects do, however, have certain drawbacks, such as dataset variability and a restricted focus on particular methods. These shortcomings should be addressed in future research by improving feature engineering methods, extending datasets, and covering more algorithmic ground. In general, these investigations establish the foundation for additional developments in cryptographic identification models, which will ultimately improve data security and privacy in the digital era.

References

[1] J-S Coron, "What is cryptography?", IEEE Security & Privacy, vol. 4, no. 1, pp. 70–73, 2006.

[2] B. Shavers, J. Bair v, "Chapter 6 - Cryptography and Encryption," in *Hiding Behind the Keyboard*, Brett Shavers and John Bair (eds.), Syngress, Boston, pp. 133-151, 2016, https://www.sciencedirect.com/science/article/pii/B9780128033401000069.

[3] A. Al-Sabaawi, "Cryptanalysis of Block Cipher: Method Implementation," International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-7, doi: 10.1109/ICONAT53423.2022.9726054, 2022.

[4] X. Yu, and K. Shi, "Block ciphers identification scheme based on randomness test." In 6th International Workshop on Advanced Algorithms and Control Engineering (IWAACE 2022), vol. 12350, pp. 375-380, 2022.

[5] W. Zhang, Y. Zhao, and S. Fan, "Cryptosystem Identification Scheme Based on ASCII Code Statistics", Security and Communication Networks (Hindawi), vol. 2020, pp. 1-10, Dec 15, 2020.

[6] L. Zhao, Y. Chi, Z. Xu, and Z. Yue, "Block Cipher Identification Scheme Based on Hamming Weight Distribution", IEEE Access, vol. 11, pp. 21364-21373, Jan 01, 2023.

[7] H. Liangwei, Z. Zhicheng, and Z. Yaqun , "Hierarchical identification scheme forcryptosystem based on random forest". Journal of Computers, 2018.

[8] K. Yuan, D. Yu, J. Feng, L. Yang, C. Jia, and Y. Huang, *A block cipher algorithm identification scheme based on hybrid k-nearest neighbor and random forest algorith*, PeerJ Computer Science, vol. 8, p. e1110, 2022.

[9] K. Yuan, D. Yu, W. Yang, Z. Li, L. Shen, Z. Du, W. Yang, "Identification of Block Cipher Algorithms Using Multi-Layer Perception Algorithm" Preprint, September 2023.

[10] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, N. Heckert, J. Dray, S. Vo, L. Bassham. " A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications" Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, 2001.

Diabetes Risk Factors Determination by extraction of Association Rules Mining algorithms

Youssef FAKIR
Laboratory of Information Processing and Decision Support , FST, Béni Mellal, Morocco
Email:info.dec07@yhoo.fr

Salim KHALIL
Laboratory of Information Processing and Decision Support, FST, Beni Mellal, Morocco,
Email:khalilsalim1@gmail.com

Ayoub MASHATE
FST, Beni Mellal, Morocco
Email:ayoub.mashate@usms.ac.ma

***Abstract*— The discovery of association rules, strong rules, correlations, sequential rules, episodes, multi-dimensional patterns, and a variety of other essential discovery tasks are among the application of datamining which can be used profitably in a wide range of areas, from network traffic data to medical records. In the healthcare industry, data mining plays an important role in disease prediction where diabetes is currently one of the most serious global health issues. The problem is presented as follows: "Find all frequent item sets in a large database of item transactions, where a frequent itemset occurs in at least a user-specified proportion of the database." Various scalable algorithms have been developed in response to this challenge and the growing volume of generated data. In this study, we attempted to improve the Eclat algorithm which one of Apriori's variants. Then apply it to a diabetes database to find different frequent patterns and association rules that will be useful in defining the relationship between the various diabetes measures. In addition, a comparison was performed to evaluate the performance of the implemented solution.**

***Keywords*—Data mining, Eclat, Apriori, Diffsets, Frequent Patterns, Diabete dataset, modified Eclat**

I. INTRODUCTION

Data mining has been used success-fully in a variety of fields to identify relevant information. Some of the primary industries where data mining is commonly employed are financial data analysis, telecommunications industry, biological data analysis, and intrusion detection [1]. It is mostly beneficial in the medical field for accurate diagnosis and selection of appropriate treatment modalities. The healthcare business generates a massive amount of intricate data regarding hospitals, patients, medical equipment, diseases, claims, treatment costs, and so on [2]. One of the approaches used in data mining is the search for a frequent itemset, which is any set of items that appears at least a certain number of times. The majority of the proposed pattern-mining algorithms are Apriori variants which use a bottom-up, breadth-first approach to find every single common itemset (horizontal mining algorithm) [3]. On the other hand, there is Eclat, which is an improvement on Apriori but with a new technique of searching that uses a vertical mining strategy to mimic depth-first search on a graph [4]. In most cases, vertical mining algorithms have proven to be more effective than horizontal mining algorithms in association mining. The main benefit of the vertical format is that it allows for fast frequency counting and automatic data reduction utilizing intersection operations on transaction ids (tids) [5]. The underlying difficulty with these approaches is that as the intermediate results of vertical tidlists become too large for memory, the algorithm's scalability reduces.

Vertical frequent item extraction algorithms, which extract sets of frequent items from a set of data by producing all combinations of row identifiers [4], are a good alternative to horizontal algorithms (Apriori) in the extraction of frequent elements from high dimensional datasets because high-dimensional datasets typically contain a large number of columns and a small number of rows.

The Bottom-up Lattice Traversal and Equivalence Class Clustering algorithm, or Eclat, is one of these algorithms. This is one of the most commonly used approaches for obtaining association rules and it is an improved and more scalable version of the Apriori algorithm. While Apriori works in a horizontal direction, similar to a graph's breadth search, the Eclat method works in a vertical direction, similar to a graph's depth search, making it a faster algorithm than Apriori [5]. The main idea is to

calculate a candidate's support value using transaction Id Sets (tidsets) to avoid the production of subsets that do not exist in the prefix tree. When the function is initially invoked, all of the individual elements and their tidsets are used. The function is then executed recursively, and each recursive call checks and combines each item-tidset pair with other item-tidset pairs. This procedure is repeated until no more candidate element-tidset pairs can be combined as visualised in Figure 1.

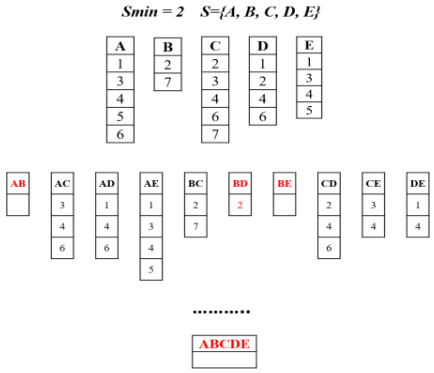


Fig.1: The Eclat algorithm

Most algorithms include a lot of critical metrics on their rules, however Eclat does not. ECLAT [6], for example, does not provide us with the Confidence and Lift metrics that are required for interpretation in the alternative models. On the other hand, it allows the model to be faster: the user can choose between having more metrics and being faster.

II. IMPROVING ECLAT USING DIFFSETS

A. *Benefits of using diffsets*

DEclat conducts a deep search in the subset tree. The Diffset (the difference of two sets) form (Figure 2), demonstrated by Zaki and Gouda in 2003 [4], enables them to exploiton supports that are substantially lower than those of the standard Eclat approach. Consider **t(A)** to be the tidset of an element **A** and **t(AB)** to be the tidset of **A** and **B**. The set of transaction identifiers present in **t(A)** and not in **t(AB)** is known as the diffset of **A** and **B** [7] and is given by $d(AB) = t(A) - t(AB)$ and the support of A and B is determined as $\sigma(AB) = \sigma(A) - |d(AB)|$.

Tidsets					Diffsets				
A	C	D	T	W	A	C	D	T	W
1	1	2	1	1	2		1	2	6
3	2	4	3	2	6		3	4	
4	3	5	5	3					
5	4	6	6	4					
	5			5					
	6								

Fig.2: Difference between tidsets & diffsets

As a result, diffset is at least two or three times smaller than tidset [4], making it more memory efficient and faster to compute. However, we should note that it is suitable for a dense database, but degrades with a base spare data, and that for a sparse database, it should switch between tidset and diffset.

B. *dEclat's pseudo code and implementation*

The algorithm's implementation is based on the presented pseudo-code [4] show bellow, where all data was supplied as metrics, which were then reformed into transactions using Python's built-in capabilities. Each transaction consists of a collection of keywords that describe the metrics. The transactions were then stored in a transactional database. The algorithm's input parameters are this transactional database instance and the minimum support parameter (minSupp), which defines an acceptance threshold. It iterates through all pairs of elements (i,j) in [P] where i < j. For each pair, it creates a new item set R as the union of i and j, and calculates the difference between the support of i and the support of j, denoted as d(R). If the support of R is greater than or equal to minsSupp, R is added to a set of frequent itemsets Ti. The algorithm then recursively applies the algorithm to each Ti, with the same minsSupp threshold, in order to find all frequent itemsets.

dEclat's pseudo-code

```
DiffEclat ([P], minsSupp):
For all element i in [P], do
    For all element j in [P], with j > i do:
        R = Union (element i, element j);
        d(R) = d(element i) - d(element j) ;
    If support(R) >= minSupp , then:
    Ti = Ti⋃{R};  //Ti initially empty
For all Ti,  do: DiffEclat (Ti, minsSupp);
```

III. DATA PRE-PROCESSING

The dataset used is "Pima Indians Diabetes Database" extracted from Kaggle. The National Institute of Diabetes and Digestive and Kidney Diseases created this database with the goal of predicting whether a patient has diabetes based on diagnostic measures and analytical dimensions put into the database [8]. By selecting the phenomenon from the wider dataset, some constraints were added. By analyzing the dataset, it has 9 attributes, 768 female patient records, 500 negative cases (65.1%), and 268 positive instances (34.9%) with no missing values. We can see that for all attributes, all values are present, but there are null values (zeros) and because the "SkinThickness" and "Insulin" attributes contain too many missing data (very high percentage), we will eliminate them from this analysis [9]. Based on the database's findings, it is unclear how well the property "DiabetesPedigreeFunction" predicts the beginning of diabetes [9], which will also be eliminated.

As seen in Table 1, the attributes employed are: 'Glucose,' 'Blood Pressure,' 'BMI,' 'Age,' 'Pregnancies' (and 'Outcome'), where age is divided into seven categories, glucose measurements are divided into three classes, blood pressure measurements are divided into three classes, BMI measurements are divided into four classes, and the number of pregnancies is divided into four groups [10]. Then, our database will be changed into a transactional format, with each transaction containing 6 items (for a total of 23 items) [11,12,13,14,15,16,17,18].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Performance evaluation

Algorithms performance is valued using computing time, memory consumption and the number items generated. Experimental were done on a personal PC with 8Gb of RAM and an Intel i5 2.5GHz by varying minsupp. Experimental results show that DEclat takes more time than the other algorithms when the minimum support level is low (e.g., 20%). According to Table 2 and Fig.3, Eclat takes less time than the other algorithms to find frequent itemsets when the minimum support level is high (e.g., 70%).

Table 3 shows that the memory consumption of Apriori is high compared to the other algorithms, especially when the minimum support level is low. By analyzing the experimental results, FP-Growth has a consistent memory consumption for different minimum support levels.

TABLE 1: Binning dataset

BMI (Body Mass Index):	**Age:**
BMI < 18.5 : Underweight	**0 - 9 :** Childhood
18.5 - 24.9 : Healthy weight	**10 - 19 :** Adolesence
25 - 26.9 : Over weight	**20 - 29 :** Early adulthood
BMI >= 30 : Obesity	**30 - 39 :** Adulthood
	40 - 59 : Middle Age
	60 - 79 : Early elder
	80 <= : Late elder
Glucose test:	**Blood pressure:**
<= 140 mg/dl : Normal Glucose	**<= 80 mmHg :** Normal
140 - 199 mg/dl : Prediabete Glucose	**80 - 89 mmHg :** Hyper tension 1
200 mg/dl <= : Diabete Glucoe	**90 mmHg <= :** Hyper tension 2
Pregnancies:	**Outcome (Class attribute)**
0 : NoPreg	**0 :** No
1 - 4 : MinusFivePreg	**1 :** Yes
5 - 9 : FiveToTenPreg	
10 <= : PlusTenPreg	

TABLE 2: Computing time (s) VS minsupp

	Minimum supports (%)					
Algorithms	**20**	**30**	**40**	**50**	**60**	**70**
Eclat	2,488	0,679	0,488	0,351	0,165	0,045
DEclat	7,167	1,384	0,464	0,124	0,034	0,019
Apriori	0,0178	0,0163	0,0129	0,0046	0,0090	0,0061
FP-Growth	0,0231	0,0249	0,0149	0,0199	0,0149	0,0149

TABLE 3: Memory consumption variation in KB

	Minimum supports (%)					
Algorithms	**20**	**30**	**40**	**50**	**60**	**70**
Eclat	798,49	375,75	282,16	195,09	155,00	127,38
DEclat	33,48	19,98	9,85	4,83	2,47	1,96
Apriori	909,92	418,20	356,79	153,39	102,34	66,19
FP-Growth	169,71	67,22	66,81	66,81	66,81	66,81

A small experiment is carried out to evaluate the minimum support's impact on the algorithm's performance by generating frequent items and varying the minimal support. As we can see in the Fig.3, Fig.4 and Fig.5, the algorithm's performance increase as the minimum support decrease.

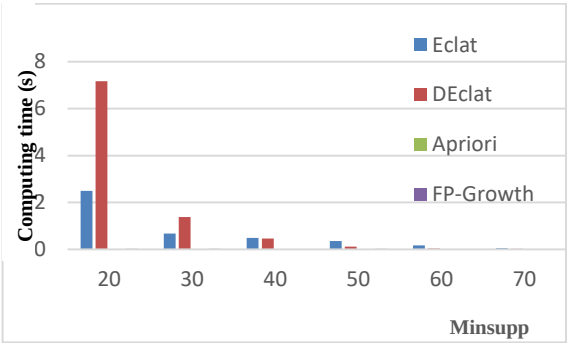


Fig.3: Computing time vs Minsupp

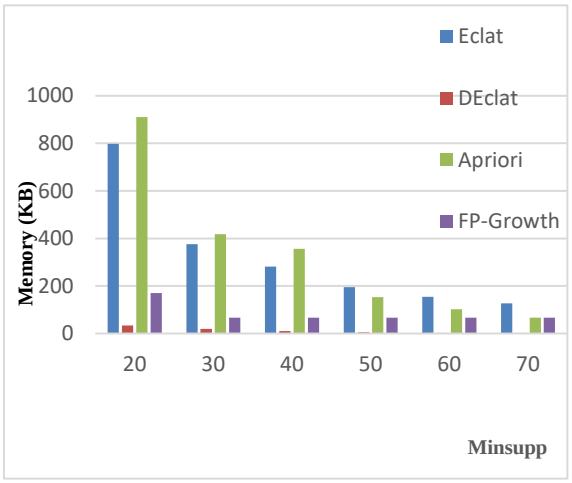


Fig.4: Memory space vs Minsupp

The tables 3, 4 and 5 show the performance results of the four association rule mining algorithms on the diabetes dataset. The tables presents the algorithms's performance based on minimum support, which is the minimum frequency of itemsets that the algorithms find as frequent.

Table 4 shows that the number of generated itemsets decreases as the minimum support level increases for all algorithms.

The memory consumption of DEclat is not necessarily better or worse than the other algorithms, as it varies depending on the minimum support percentage and can be higher or lower than the other algorithms for different values of minimum support. However, the presented results show that DEclat is the best of the four algorithms in terms of memory consumption. Therefore, it's important to consider the trade-off between memory consumption and other factors when choosing the best algorithm for a particular task. Figure 3, 4 and 5 illustrate respectively the computing time, the memory space occupied and the generated items by varying the minsupp.

TABLE 4: Generated items

	Minimum supports (%)					
Algorithms	**20**	**30**	**40**	**50**	**60**	**70**
Eclat	47	31	16	8	4	2
DEclat	54	32	16	8	4	2
Apriori	54	32	16	8	4	2
FP-Growth	54	32	16	8	4	2

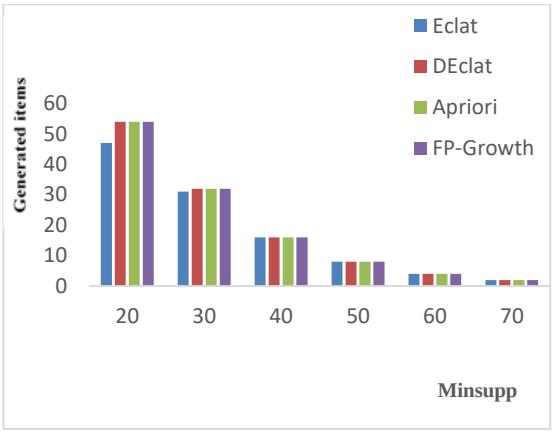


Fig.5: Generated items vs Minsupp

B. *Association rules's generation*

Table 5 summarizes the results of an association rule mining process performed on the diabetes dataset. The minimum support and a confidence values were set to 30% and 80% respectively. Each row lists an antecedent (a set of conditions), a consequent (the result), the support (the number of instances the rule occurs in the dataset), the confidence (the proportion of times the consequent occurs given the antecedent), and lift (the ratio of the observed support to that expected if the antecedent and consequent were independent).

For example, the first rule, **"OutcomeNo -> NormalGlucose"**, has a support of 440, a confidence of 0.88, and a lift of 1.527. This means that there were 440 transactions in the dataset where both OutcomeNo and NormalGlucose were present, and in 88% of those transactions, NormalGlucose appeared when OutcomeNo was present. The lift of 1.527 indicates that the likelihood of **NormalGlucose** appearing in transactions that include OutcomeNo is 1.527 times higher than the likelihood of

NormalGlucose appearing in transactions that do not include OutcomeNo. The process did find 21 association rules in 2.068 seconds. Overall, these results suggest that there are significant associations between various factors related to diabetes and health outcomes. High lift values indicate that the factors listed in the antecedent and consequent are more likely to occur together than would be expected by chance, while high confidence values indicate that the presence of the antecedent is a strong predictor of the presence of the consequent. For example, one of the strongest rules in the Table 6is **"OutcomeNo, EarlyAdulthood -> NormalGlucose"** with a confidence of 0.9166 and lift of 1.591. This means that when both "OutcomeNo" and "EarlyAdulthood" are present, there is a 91.66% probability of the presence of "NormalGlucose". The lift of 1.591 indicates that the presence of "OutcomeNo" and "EarlyAdulthood" is 1.591 times more likely to lead to "NormalGlucose" than if the variables were independent. Another strong rule is **"NormalBloodPressure, EarlyAdulthood -> OutcomeNo"** with a confidence of 0.8056 and lift of 1.611. This means that when both "NormalBloodPressure" and "EarlyAdulthood" are present, there is an 80.56% probability of the presence of "OutcomeNo". The lift of 1.611 indicates that the presence of "NormalBloodPressure" and "EarlyAdulthood" is 1.611 times more likely to lead to"OutcomeNo" than if the variables were independent. From the provided tables of extracted association rules we can conlude that **age**, **blood pressure** and **glucose levels** are the strongest factors of being diabetic or not.

V. CONCLUSION

In this paper, an analysis was performed to compare various frequent pattern mining algorithms, including Apriori, Eclat, FP-Growth, and DEclat, which proved that DEclat help optimize the memory usage during association rules mining. The same diabetes database transactions were used to compare these algorithms and retrieve the association rules proved that age, blood pressure and glucose are the strongest factors to predict where, or not, a patient is diabetic.

References

[1] Mohamed Azalmad & Youssef Fakir:, Data Mining Approach for Intrusion Detection. CBI 2021: 201-219, DOI: 10.1007/978-3-030-76508-8_15, In book: Business Intelligence

[2] Lakshmi KS & Vadivu G. Extracting Association Rules from Medical Health Records Using multi- Criteria Decision Analysis. Procedia Computer Science. 2017 Jan;115:290-5.

[3] Guns T, Nijssen S & De Raedt L. Itemset mining:A constraint programming perspective. Artificial Intelligence. 2011 Aug;175(12):1951-83.

[4] Zaki MJ, & Gouda K. Fast vertical mining using diffsets. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '03. New York, NY, USA: Association for computing machinery; 2003. p. 326- 35.

[5] Darrab S & Ergenc B. Vertical Pattern Mining Algorithm for multiple Support Thresholds. Procedia Computer Science. 2017 Jan;112:417-26.

[6] Al-Banam R, Farhan MS & Othman NA. An Efficient Spark-Based Hybrid Frequent Itemset mining Algorithm for Big Data. Data. 2022 Jan; 7(1):11.

[7] Mahadi Man & Abdul Jalil, Frequent itemset mining: technique to improve eclat based algorithm, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 6, December 2019, pp. 5471~5478 ISSN: 2088-8708, DOI: 10.11591/ijece.v9i6.pp5471-5478

[8] Wu H, Yang S, Huang Z, He J &Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics in medicine Unlocked. 2018 Jan; 10:100-7.

[9] Tiwari P, Singh V. Diabetes disease prediction using significant attribute selection and classification approach. Journal of Physics: Conference Series. 2021 jan;1714(1):012013.

[10] meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of medical Sciences. 2013 Feb; 29(2):93-9.

[11] Kavakiotis I, Tsave O, Salifoglou A, maglaveras N, Vlahavas I & Chouvarda I. Machine Learning and Data mining methods in Diabetes Research. Computational and Structural Biotechnology Journal. 2017 Jan;15:104-16.

[12] Youssef Fakir, Abdelfatah Maarouf & Rachid El Ayachi, Mining Frequents Itemset and Association Rules in Diabetic Dataset. CBI 2022: 146-157

[13] Youssef Fakir, Diabetes Prediction by Machine Learning Algorithms and Risks Factors. CBI 2023: 44-56

[14] Youssef Fakir, & Rachid Elayachi, Closed frequent itemsets mining based on It-Tree. Global Journal of Computer Sciences: Theory and Research, 11(1), 01–11., 2021, https://doi.org/10.18844/gjcs.v11i1.4912

[15] Youssef Fakir & Amina Bouchehait, Diabetes prediction by machine learning algorithms after association rules extraction on SPARK using massive dataset, Journal of Harbin Engineering University, (accepted and to be published)

[16] Xiong, P. Tan & V. Kumar, "Mining strong affinity association patterns in data sets with skewed support distribution", in: Proceedings of the Third IEEE International Conference on Data Mining, ICDM, 2003, pp. 387-394

[17] Ya-Han Hu & Yen-Liang Chen, "Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism", Decision Support Systems, 2006, 42, pp. 1-24

[18] Ding, "Efficient association rule mining among infrequent items", Ph.D. Thesis, University of Illinois at Chicago, 2005.

[19] Fakir et al, Extraction of itemsets frequents, International Journal of Mathematics Research. ISSN 0976-5840 Volume 12, Number 1 (2020), pp. 23-32

TABLE 5: Extracted association rules with 80% and a minimum support of 30%

Antecedent	Consequent	Support	Confidence	Lift
OutcomeNo	NormalGlucose	440	0,88	1,527
EarlyAdulthood	NormalGlucose	331	0,8358	1,451
EarlyAdulthood	NormalBloodPressure	319	0,8055	1,430
MinusFivePreg	NormalBloodPressure	308	0,8083	1,435
OutcomeNo , NormalBloodPressure	NormalGlucose	348	0,9038	1,569
NormalBloodPressure , EarlyAdulthood	NormalGlucose	275	0,8620	1,496
NormalGlucose , EarlyAdulthood	NormalBloodPressure	275	0,8308	1,475
NormalBloodPressure , MinusFivePreg	NormalGlucose	250	0,8116	1,409
NormalGlucose , MinusFivePreg	NormalBloodPressure	250	0,8333	1,480
OutcomeNo , EarlyAdulthood	NormalGlucose	286	0,9166	1,591
NormalGlucose , EarlyAdulthood	OutcomeNo	286	0,8640	1,728
OutcomeNo , MinusFivePreg	NormalGlucose	253	0,8939	1,552
NormalGlucose , MinusFivePreg	OutcomeNo	253	0,8433	1,686
EarlyAdulthood , MinusFivePreg	NormalGlucose	234	0,8509	1,477
NormalBloodPressure , EarlyAdulthood	OutcomeNo	257	0,8056	1,611
OutcomeNo , EarlyAdulthood	NormalBloodPressure	257	0,8237	1,463
OutcomeNo , MinusFivePreg	NormalBloodPressure	237	0,8374	1,487
EarlyAdulthood , MinusFivePreg	NormalBloodPressure	233	0,8472	1,504
NormalBloodPressure , OutcomeNo , EarlyAdult-hood	NormalGlucose	242	0,9416	1,634
NormalGlucose,NormalBloodPressure,	OutcomeNo	242	0,88	1,76
NormalGlucose , OutcomeNo , EarlyAdulthood	NormalBloodPressure	242	0,8461	1,502

Comparison of some Ellipsoidal Outer Bounding Identification Algorithms for Output Error systems

Hasna El Maizi[1,2], Mathieu Pouliquen[1], Saïd Safi[2] and Miloud Frikel[1]

***Abstract*—This paper is concerned with the identification of Output Error (OE) models in the presence of unknown but bounded noise. Three Ellipsoidal Outer Bounding (EOB) algorithms are compared. The first algorithm aims to estimate the parameters of the system such that the output error is bounded by the noise bound. The second algorithm is a variant of the first algorithm. It is based on the minimization of the volume of an ellipsoid enclosing the real parameter vector. The third algorithm focuses also on the size of the bounding ellipsoid but using another optimization strategy. The algorithms are described, commented and results of some numerical simulations are provided to compare them.**

***Index Terms*—Ellipsoidal Outer Bounding (EOB) Algorithm, Output Error (OE) Model, Bounded Noise.**

I. Introduction

The identification of dynamic systems, also called data based modelling, is a focal point over the years in engineering science with a multitude of works dedicated to this subject. Different methods to identify dynamic systems are now available (see for instance [1] and [2]), main differences between methods lying in the structure of the model, the quality of available data and the implementation framework. In this paper we focus on the identification of discrete time linear systems considering the bounded noise assumption.

Set-membership identification methods are methods dedicated to such a context. These methods produce a set of estimates consistent with the measurement data, the model structure and the bound on the noise. Among these identification methods, one class of Set-membership methods is the Ellipsoidal Outer Bounding Algorithms class. These algorithms are particularly interesting for their low computational load and their ability to be adapted to different model structures. From the fact that the center of the ellipsoid and its shape can be computed from different manners, different contributions are available, see for instance [3], [4], [5], [6], [7], [8], [9], [10], [11], [12],and [13]. About the structure of the model we are interested in Output Error model. One of the benefits of this model structure is its separation of the disturbance model from the process model by directly introducing noise on the output. Our objective here is to compare performance of three ellipsoidal algorithms dedicated to such structure model, these three algorithms are based on three different strategies in the design of the ellipsoidal.

[1]Laboratoire d'Ingenierie des Systemes - UR 7478 Normandie Univ, UNICAEN, ENSICAEN, LIS, Caen, France

[2]Department of Mathematics and Informatics, University Sultan Moulay Slimane, Beni Mellal, Morocco

The first algorithm is the one proposed in [14]. It aims at estimate a model such that the magnitude of the output error is bounded by the upper bound on the noise. The second algorithm is introduced in [15]. It is an alternative to the first one. Its aim is to compute the minimal volume ellipsoid that encloses the real parameter vector while constraining the magnitude of the output error to be bounded by the upper bound on the noise. This is achieved through the minimization of a criterion based on the determinant of a matrix characterizing the shape of the ellipsoid. The third algorithm shares the same objective as the second one but uses another criterion in order to reduce the size of the ellipsoid, the minimized criterion is here based on the trace of a matrix characterizing the shape of the ellipsoid.

The remainder of the paper is structured as follows: the considered identification problem is first formulated in Section II, the three considered identification algorithms are next presented in Section III, these algorithms are then evaluated and compared using numerical simulations in Section IV and finally Section V concludes the paper.

II. PROBLEM FORMULATION

The considered discrete time system, depicted in Fig.1, is parameterized by an Output Error (OE) model described as:

$$y_t = G^*(q)u_t + v_t \tag{1}$$

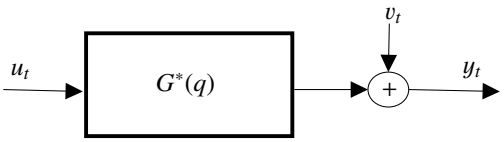


Figure 1. The considered system .

where u_t represents the input of the system, y_t the output of the system and v_t is an unknown bounded noise. We assume knowledge of an upper bound δ_v on its magnitude:

$$|v_t| \leq \delta_v. \tag{2}$$

$G^*(q)$ is the transfer function of the considered system described as follows

$$G^*(q) = \frac{B^*(q)}{A^*(q)} \tag{3}$$

where

$$\begin{cases} A^*(q) = 1 + a_1^* q^{-1} + \ldots + a_{n_a}^* q^{-n_a} \\ B^*(q) = b_0^* + b_1^* q^{-1} + \ldots + b_{n_b}^* q^{-n_b} \end{cases} \tag{4}$$

and q^{-1} is the delay operator such that $q^{-1}x_t = x_{t-1}$. Degrees n_a and n_b are assumed to be known. Let us define the parameter vector θ^* as

$$\theta^{*T} = \begin{pmatrix} a_1^* & \cdots & a_{n_a}^* & b_1^* & \cdots & b_{n_b}^* \end{pmatrix} \tag{5}$$

and denote $\hat{\theta}_t$ the estimate of θ^* at time t.

At the current time t we also define the a priori and a posteriori estimates of y_t by

$$\begin{cases} \hat{y}_{t/\hat{\theta}_{t-1}} = \hat{\phi}_t^T \hat{\theta}_{t-1} \\ \hat{y}_{t/\hat{\theta}_t} = \hat{\phi}_t^T \hat{\theta}_t \end{cases}. \tag{6}$$

where $\hat{\phi}_t$ is the following regression vector

$$\hat{\phi}_t^T = \begin{pmatrix} -\hat{y}_{t-1/\hat{\theta}_{t-1}} & \cdots & -\hat{y}_{t-n_a/\hat{\theta}_{t-na}} & u_t & \cdots & u_{t-n_b} \end{pmatrix} \tag{7}$$

The **objective** of the paper is to propose a comparison of three EOB identification algorithms. Theses algorithms use $\left\{u_t, y_t\right\}_{t=1}^{N}$, with N the amount of available data, and the knowledge of δ_v. They allow the estimation of a model $\widehat{G}(q)$ parameterized as $G^*(q)$ and such that

$$|y_t - \widehat{G}(q)u_t| \leq \delta_v. \tag{8}$$

III. IDENTIFICATION ALGORITHM

In this section we present the three EOB algorithms for the identification of $G^*(q)$. These algorithms aim to find a parameter vector $\hat{\theta}_t$ center of an ellipsoid ξ_t defined by

$$\xi_t = \left\{\theta \in \mathbb{R}^n, (\theta - \hat{\theta}_t)^T P_t^{-1} \ (\theta - \hat{\theta}_t) \leq \rho_t\right\}. \tag{9}$$

Note that if $\theta^* \in \xi_t$, then the smaller the ellipsoid is, the more precisely we know the true parameter.

The three algorithms are very similar. The main differences between them lie in the design of the center $\hat{\theta}_t$ of the ellipsoid and in the design of the shape of the ellipsoid through the computation of P_t and ρ_t.

A. A first proposed identification Algorithm

The first algorithm is the one proposed in [14]. It is called the Filtered-Output Error-Ellipsoidal Outer Bounding (F-OE-EOB) algorithm. It uses an adaptation filter $F(q)$ chosen by the user. The filtered version of y_t and u_t are defined by

$$\begin{cases} y_t^F = \dfrac{1}{F(q)} y_t \\ u_t^F = \dfrac{1}{F(q)} u_t \end{cases}. \tag{10}$$

We define the filtered a priori prediction error and the filtered a posteriori prediction error by

$$\begin{cases} \varepsilon^F_{t/\hat{\theta}_{t-1}} = y_t^F - \hat{\phi}_t^{F^T} \hat{\theta}_{t-1} \\ \varepsilon^F_{t/\hat{\theta}_t} = y_t^F - \hat{\phi}_t^{F^T} \hat{\theta}_t \end{cases} \tag{11}$$

and we also define the a priori and a posteriori adaptation errors by

$$\begin{cases} \eta_{t/\hat{\theta}_{t-1}} = \varepsilon^F_{t/\hat{\theta}_{t-1}} + \ (F(q) - 1)\varepsilon^F_{t/\hat{\theta}_t} \\ \eta_{t/\hat{\theta}_t} = \ F(q)\varepsilon^F_{t/\hat{\theta}_t} \end{cases}. \tag{12}$$

The algorithm is given in Tab. I. Note that this algorithm is an iterative algorithm, n_g being the number of iterations. The iterative implementation is due to the fact that, as shown in [14], the ideal adaptation filter (i.e. the one ensuring the convergence of $\hat{\theta}_t$ towards θ^*) is $F(q) = A^*(q)$, the adaptation filter is then updated at each iteration.

By shortly analyzing this algorithm it appears that its aim is to provide a parameter vector such that the a posteriori adaptation error is bounded by δ_v. This is realized through the term σ_t in (16) and (17). This property is summarized below.

Property 1. *[14]*

Consider the F-OE-EOB identification algorithm in Tab. I. If $\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F > 0$, then $\eta_{t/\hat{\theta}_t}$ is such that

$$|\eta_{t/\hat{\theta}_t}| \leq \delta_v \tag{13}$$

■

It is shown in [14] that if $F(q) = A^*(q)$, then the estimated model $\widehat{G}(q)$ is such that

$$|y_t - \widehat{G}(q)u_t| \leq \delta_v. \tag{14}$$

which is coherent with the objective.

TABLE I. The F-OE-EOB identification algorithm

Filtred-OE-EOB Algorithm

Data: $\left\{u_t, y_t\right\}_{t=1}^{N}$, n_g, $\hat{\theta}_{t=0}$, $P_{t=0}$, δ_v

Result: $\hat{\theta}_t$

for $t = 1 : n_g$ **do**

– **Design the filter** $F(q)$ **such that** $F(q) = \hat{A}(q)$

for $t = n+1 : N$ **do**

– **Calculation of** $\left\{\hat{y}^F_{t/\hat{\theta}_{t-1}}, \ \varepsilon^F_{t/\hat{\theta}_{t-1}}, \ \eta_{t/\hat{\theta}_{t-1}}\right\}$

$$\begin{cases} \hat{y}^F_{t/\hat{\theta}_{t-1}} = \hat{\phi}_t^{F^T} \hat{\theta}_{t-1} \\ \varepsilon^F_{t/\hat{\theta}_{t-1}} = y_t^F - \hat{\phi}_t^{F^T} \hat{\theta}_{t-1} \\ \eta_{t/\hat{\theta}_{t-1}} = \varepsilon^F_{t/\hat{\theta}_{t-1}} + \ (F(q) - 1)\varepsilon^F_{t/\hat{\theta}_t} \end{cases} \tag{15}$$

– **Update of the estimated parameter vector** $\hat{\theta}_t$

if $|\eta_{t/\hat{\theta}_{t-1}}| > \delta_v$ ***and*** $\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F > 0$ **then**

$$\sigma_t = \frac{1}{\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F} \left(\left| \frac{\eta_{t/\hat{\theta}_{t-1}}}{\delta_v} \right| - 1 \right) \tag{16}$$

else

$$\sigma_t = 0 \tag{17}$$

$$\begin{cases} \Gamma_t = \dfrac{P_{t-1} \ \hat{\phi}_t^F \ \sigma_t}{1 + \hat{\phi}_t^{F^T} P_{t-1} \ \hat{\phi}_t^F \sigma_t} \\ P_t = \left(I_n \ - \Gamma_t \ \hat{\phi}_t^{F^T}\right) P_{t-1} \\ \hat{\theta}_t = \hat{\theta}_{t-1} \ + \ \Gamma_t \ \eta_{t/\hat{\theta}_{t-1}} \\ \varepsilon^F_{t/\hat{\theta}_t} = y_t^F - \hat{\phi}_t^{F^T} \hat{\theta}_t \\ \eta_{t/\hat{\theta}_t} = F(q)\varepsilon^F_{t/\hat{\theta}_t} \\ \rho_t = \rho_{t-1} \ + \sigma_t \delta_v^2 - \sigma_t \dfrac{\eta^2_{t/\hat{\theta}_{t-1}}}{1 + \hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F \sigma_t}. \end{cases} \tag{18}$$

B. A second identification algorithm

The second algorithm is a variant of the F-OE-EOB algorithm. It might be noticed that the F-OE-EOB algorithm presented above doesn't focus on the shape of the ellipsoid. The second algorithm presented in this section has been first proposed in [15]. It aims to minimize the size of the ellipsoid by minimizing $\mathcal{V}_{\xi_t} \stackrel{\text{def}}{=} \det(\rho_t P_t)$. $\mathcal{V}_{\xi_t}$ represents the product of the squares of the lengths of the half-axes of the ellipsoid, it is therefore proportional to the volume of the ellipsoid.

The algorithm is given in Tab. II. Here again it is an iterative algorithm with an iterated design of the adaptation filter, the ideal filter being $F(q) = A^*(q)$. The difference with the first algorithm lies in the synthesis of the term σ_t in (23), (26) and (27). By shortly analyzing this algorithm it appears that its aim is to minimize $\mathcal{V}_{\xi_t}$ while keeping the magnitude of the a posteriori adaptation error $\eta_{t/\hat{\theta}_{t-1}}$ under δ_v. This yields to the following property.

Property 2. *[15]*

Consider the F-MD-OE-EOB identification algorithm in Tab. II. If $\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F > 0$*, then* $\eta_{t/\hat{\theta}_t}$ *is such that*

$$|\eta_{t/\hat{\theta}_t}| \leq \delta_v \tag{19}$$

and $\mathcal{V}_{\xi_t}$ *satisfies*

$$\mathcal{V}_{\xi_t} \leq \mathcal{V}_{\xi_{t-1}}. \tag{20}$$

Moreover, if $|\eta_{t/\hat{\theta}_{t-1}}| \leq \delta_v$*, then*

$$\sigma_t = argmin_\sigma \ \mathcal{V}_{\xi_t} \tag{21}$$

■

It is shown in [15] that this algorithm provides also $\widehat{G}(q)$ such that

$$|y_t - \widehat{G}(q) u_t| \leq \delta_v \tag{22}$$

and when $|\eta_{t/\hat{\theta}_{t-1}}| \leq \delta_v$, then the algorithm provides the minimum volume ellipsoid.

C. A third identification algorithm

The third algorithm proposed in this section focuses also on the shape of the ellipsoid. While the second algorithm aims to minimize $\mathcal{V}_{\xi_t} = \det(\rho_t P_t)$, the third algorithm aims to minimize $\mathcal{U}_{\xi_t} \stackrel{\text{def}}{=} \text{tr}(\rho_t P_t)$. This alternative strategy to reduce the size of the ellipsoid is justified by the fact that $\mathcal{U}_{\xi_t}$ represents the sum of the squares of the lengths of the half-axes of the ellipsoid, hence it is another way to characterize the shape of ellipsoid. Such a similar strategy is proposed in [3] in the case of an ARX model.

The third algorithm is presented in Tab. III. Here again, the difference with the two previous algorithms is the design of σ_t. σ_t is computed here so as to minimize $\text{tr}(\rho_t P_t)$ while keeping the magnitude of the a posteriori adaptation error $\eta_{t/\hat{\theta}_t}$ below δ_v. This is summarized in the following property.

Property 3.

Consider the F-MT-OE-EOB identification algorithm in Tab. III. If $\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F > 0$*, then* $\eta_{t/\hat{\theta}_t}$ *is such that*

$$|\eta_{t/\hat{\theta}_t}| \leq \delta_v \tag{28}$$

TABLE II. The Filtered-Minimal Determinant-OE-EOB identification algorithm

Filtered-Minimal Determinant-OE-EOB identification algorithm

Data: $\{u_t, y_t\}_{t=1}^N$, n_g, $\hat{\theta}_{t=0}$, $P_{t=0}$, δ_v

Result: $\hat{\theta}_t$

for $t = 1 : n_g$ **do**

– **Design the filter** $F(q)$ **such that** $F(q) = \hat{A}(q)$

for $t = n+1 : N$ **do**

– **Calculation of** $\left\{\hat{y}^F_{t/\hat{\theta}_{t-1}},\ \varepsilon^F_{t/\hat{\theta}_{t-1}},\ \eta_{t/\hat{\theta}_{t-1}}\right\}$ **as shown in Tab.I.**

– **Update of the estimated parameter vector** $\hat{\theta}_t$ **using the following design of** σ_t**.**

if $|\eta_{t/\hat{\theta}_{t-1}}| > \delta_v$ ***and*** $\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F > 0$ **then**

$$\sigma_t = \frac{1}{\hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F} \left(\left| \frac{\eta_{t/\hat{\theta}_{t-1}}}{\delta_v} \right| - 1 \right) \tag{23}$$

else

$$G_t = \hat{\phi}_t^{F^T} P_{t-1} \hat{\phi}_t^F \tag{24}$$

$$\begin{cases} a_1 = \delta^2 (n-1) G_t^2 \\ a_2 = \left((2n-1)\delta_v^2 - \rho_{t-1} G_t + \eta^2_{t/\hat{\theta}_{t-1}} \right) G_t \\ a_3 = n\left(\delta_v^2 - \eta^2_{t/\hat{\theta}_{t-1}} \right) - \rho_{t-1} G_t \end{cases} \tag{25}$$

if $a_3 \geq 0$ **then**

$$\sigma_t = 0 \tag{26}$$

else

$$\sigma_t = \frac{-a_2 + \sqrt{a_2^2 - 4a_1 a_3}}{2a_1} \tag{27}$$

and $\mathcal{U}_{\xi_t}$ *satisfies*

$$\mathcal{U}_{\xi_t} \leq \mathcal{U}_{\xi_{t-1}}. \tag{29}$$

Moreover, if $|\eta_{t/\hat{\theta}_{t-1}}| \leq \delta_v$*, then*

$$\sigma_t = argmin_\sigma \ \mathcal{U}_{\xi_t} \tag{30}$$

■

The proof of this property follows the same steps than the proof of property 2 (see [15]). Currently no additional property has been established on this third algorithm, nevertheless some numerical results are presented in the next section so as to characterize performance of this algorithm and to compare it with the two previous algorithms.

IV. NUMERICAL SIMULATION

Some simulation results are reported in this section in order to compare various aspects of the three previous algorithms. Data are generated according to (1) and (2). The considered system is as follows:

$$\begin{cases} B^*(q) &= 0.173 \\ A^*(q) &= 1 - 1.425q^{-1} + 0.496q^{-2} \end{cases} \tag{36}$$

TABLE III. The Filtred-Minimal Trace-OE-EOB identification algorithm

Filtred-Minimal Trace-OE-EOB Algorithm

Data: $\{y_t\}_{t=1}^{N}$, n_c, n_d, $\hat{\theta}_{t=0}$, $P_{t=0}$, δ_v

Result: $\hat{\theta}_t$

for $t = 1 : n_g$ **do**

– **Design the filter** $F(q)$ **such that** $F(q) = \hat{A}(q)$

for $t = n+1 : N$ **do**

– **Calculation of** $\left\{\hat{y}^F_{t/\hat{\theta}_{t-1}},\ \varepsilon^F_{t/\hat{\theta}_{t-1}},\ \eta_{t/\hat{\theta}_{t-1}}\right\}$ **as shown in Tab.I**

– **Update the estimated parameter vector** $\hat{\theta}_t$ **using the following design of** σ_t

if $|\eta_{t/\hat{\theta}_{t-1}}| > \delta_v$ ***and*** $\hat{\phi}_t^{F\,T} P_{t-1} \hat{\phi}_t^F > 0$ **then**

$$\sigma_t = \frac{1}{\hat{\phi}_t^{F\,T} P_{t-1} \hat{\phi}_t^F} \left(\left| \frac{\eta_{t/\hat{\theta}_{t-1}}}{\delta_v} \right| - 1 \right) \tag{31}$$

else

$$\begin{cases} G_t = \hat{\phi}_t^{F\,T} P_{t-1}^T \hat{\phi}_t^F \\ B_t = \hat{\phi}_t^{F\,T} P_{t-1}^2 \hat{\phi}_t^F \\ \mathcal{U}_{\xi_{t-1}} = det(\rho_{t-1} P_{t-1}) \end{cases} \tag{32}$$

$$\begin{cases} a_1 = 1 \\ a_2 = \dfrac{3}{G_t} \\ a_3 = \dfrac{G_t \left((3\ \delta_v^2 - \eta^2_{t/\hat{\theta}_{t-1}}) \mathcal{U}_{\xi_{t-1}} - B_t \rho^2_{t-1} \right) + 2 \left((\eta^2_{t/\hat{\theta}_{t-1}} - \delta_v^2) B_t \rho_{t-1} \right)}{\delta_v^2 G_t^2 (G_t \mathcal{U}_{\xi_{t-1}} - B_t \rho_{t-1})} \\ a_4 = \dfrac{\left(\delta_v^2 - \eta^2_{t/\hat{\theta}_{t-1}} \right) \mathcal{U}_{\xi_{t-1}} - B_t \rho^2_{t-1}}{\delta_v^2 G_t^2 (G_t \mathcal{U}_{\xi_{t-1}} - B_t \rho_{t-1})} \end{cases} \tag{33}$$

if $a_4 \geq 0$ **then**

$$\sigma_t = 0 \tag{34}$$

else

$$\sigma_t = \text{real root} \left\{ a_1 \sigma^3 + a_2 \sigma^2 + a_3 \sigma + a_4 = 0 \right\} \tag{35}$$

The input sequence $\{u_t\}$ is a random sequence with a uniform distribution over the interval $[-2\ ;\ 2]$. The noise is also a random sequence with a uniform distribution over the interval $[-\delta_v; \delta_v]$ with $\delta_v = 0.5$ (this corresponds to a Signal to Noise Ratio not far from 7dB). The three algorithms are implemented with a maximum number of iterations $n_g = 20$ (the default value in MATLAB for similar iterative identification algorithms) and the number of available data is $N = 500$.

Fig.2 presents the a posteriori adaptation error and $\pm\delta_v$ as a function of t for the three identification algorithms. These results are consistent with properties 1, 2 and 3.

The convergence of the estimated parameter vectors $\hat{\theta}_t$ towards the true vector θ^* as a function of t is shown in Fig. 3. One can observe that the estimated parameters gradually converge to the true parameters. Furthermore, it is noticeable that the speed of convergence of the estimated parameters towards the true parameters is faster with the second algorithm compared to the first and third algorithms.

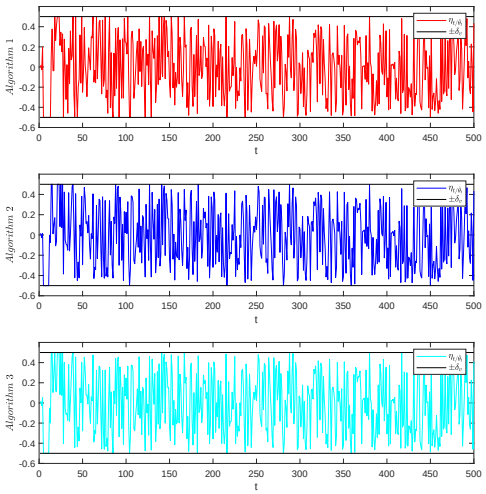


Figure 2. A posteriori adaptation error $\eta_{t/\hat{\theta}_t}$ and the bound $\pm\delta_v$ for each algorithm.

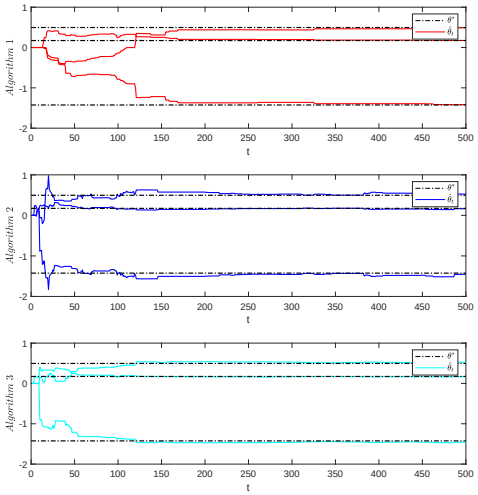


Figure 3. Convergence of estimated parameters $\hat{\theta}_t$ towards the true parameters θ^*.

Fig.4 presents $y_t - \widehat{G}(q)u_t$ and $\pm\delta_v$ as a function of t for the three presented identification algorithms. It appears that the three algorithms allow the estimation of model such that the output error is bounded by δ_v as stated in the objective in section II.

Ellipsoids $\xi_{t=N}$, representing the uncertainty zone on θ^* (Dot in black), computed using the three algorithms are shown

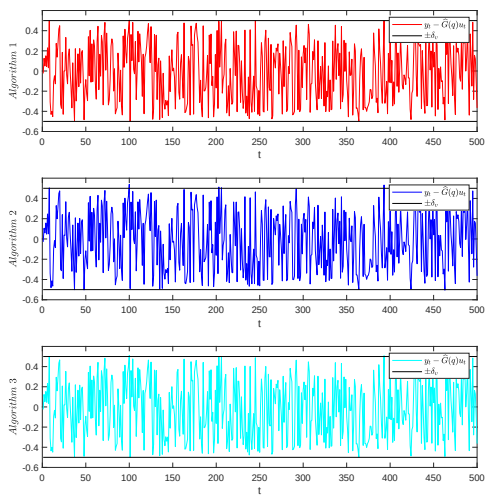


Figure 4. Output error $y_t - \widehat{G}(q)u_t$ and the bound $\pm\delta_v$.

in Fig. 5, 6, and 7. The ellipsoid obtained by the second algorithm (Fig. 6) is smaller than ellipsoids obtained by the first algorithm (Fig. 5) and third algorithm (Fig. 7), indicating that the second algorithm generally performs better than the two others.

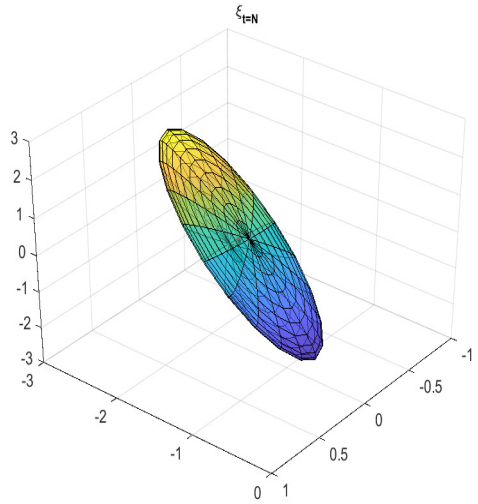


Figure 5. Algorithm 1: Ellipsoid $\xi_{t=N}$, θ^* (dot in red) and $\hat{\theta}_t$ (dot in black).

In order to compare the evolution of the size of the ellipsoids, a first index presented on Fig.8 is the evolution of the ratio between the volume of the ellipsoid at time t and that at time $t = 0$. We notice that this volume ratio decreases progressively. This reduction in volume indicates a significant reduction in the uncertainty zone on the parameters. Furthermore, it is noticeable that the volume ratio

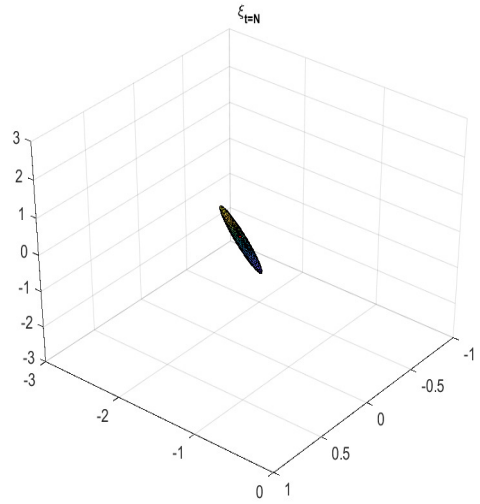


Figure 6. Algorithm 2: Ellipsoid $\xi_{t=N}$, θ^* (dot in red) and $\hat{\theta}_t$ (dot in black).

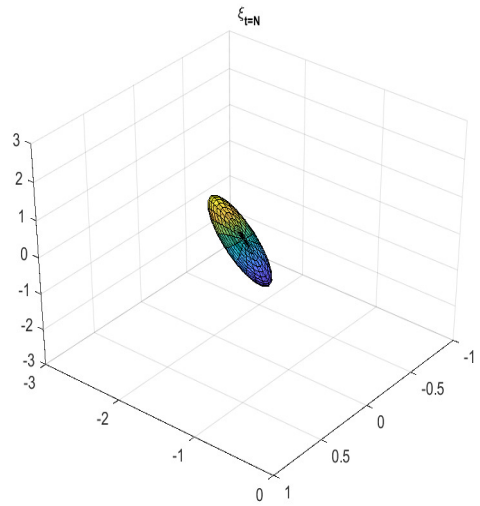


Figure 7. Algorithm 3: Ellipsoid $\xi_{t=N}$, θ^* (dot in red) and $\hat{\theta}_t$ (dot in black).

characterizing the second proposed algorithm is always lower than those characterizing the first and third algorithms. This indicates that the volume of the ellipsoid obtained by the second proposed algorithm is smaller than the volume of the ellipsoid obtained by the first and third algorithms. This is confirmed by Fig. 5, Fig. 6 and Fig. 7 on which the ellipsoids are represented.

A second index is the evolution of the ratio between $\mathcal{U}_{\xi_t}$ and $\mathcal{U}_{\xi_{t=0}}$. This index is depicted on Fig. 9 for the three presented algorithms. We notice that this ratio globally decreases during the adaptation process for the three algorithms, indicating a

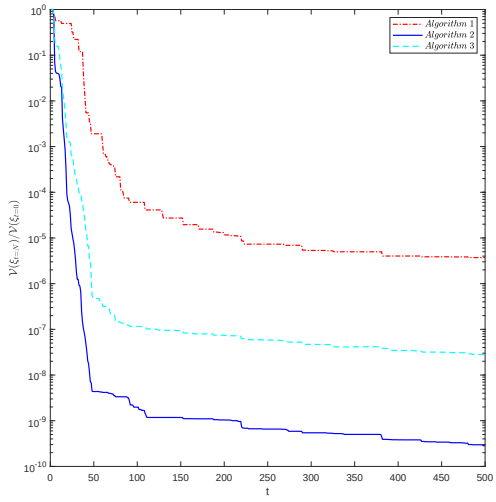


Figure 8. The evolution of the ratio between the volume of the ellipsoid at time t and that at time $t = 0$.

reduction in the size of the ellipsoids. Furthermore, we observe that the ratio characterizing the second proposed algorithm is lower than those characterizing the first and third algorithms.

From these results, it appears that the second and third algorithms allow to reduce the size of the ellipsoid with respect to the first algorithm. Moreover, event if the objective of the third algorithm is to minimize locally for each time t the trace ratio (property 3), then this is not sufficient to reduce the size of the ellipsoid with respect to the strategy of the second algorithm which minimizes locally for each time t the volume ratio (property 2).

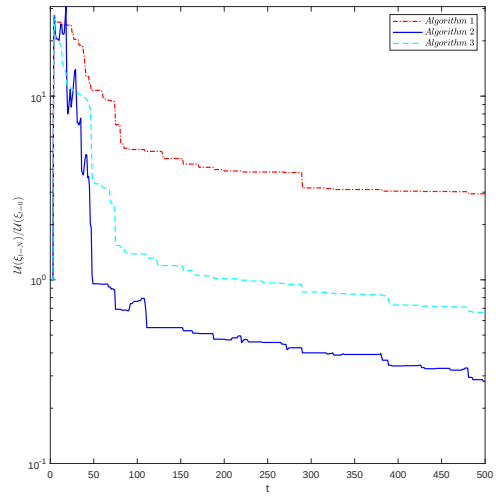


Figure 9. The evolution of the ratio between the trace of the ellipsoid at time t and that at time $t = 0$.

V. CONCLUSIONS

In this paper, three identification EOB algorithms for identifying OE model with bounded noise are compared. The second algorithm and the third algorithm are alternatives to the first algorithm. The main interest of these alternatives is to reduce the size of the ellipsoid that encloses the real parameter vector while ensuring that the magnitude of the output error is bounded by the noise bound. The well behavior of each algorithm is illustrated trough some numerical simulations. The interest of the second and third algorithms with respect to the size of the ellipsoid is confirmed. It also appears that the second algorithm provides the minimal volume ellipsoid and allows a faster convergence of the estimates.

REFERENCES

[1] L. Ljung, "System identification theory for the user," *Prentice Hall*, 1999.

[2] R. Pintelon and J. Schoukens, "System identification, a frequency domain approach," *Wiley*, 2012.

[3] E. Fogel and Y.-F. Huang, "On the value of informaton in system identification - bounded noise case," *Automatica*, vol. 18, no. 2, pp. 229–238, 1982.

[4] S. Dasgupta and Y.-F. Huang, "Asymptotically convergent modified recursive least-squares with data-dependent updating and forgetting factor for systems with bounded noise," *IEEE Transactions on information theory*, vol. 33, no. 3, pp. 383–392, 1987.

[5] R. Lozano-Leal and R. Ortega, "Reformulation of the parameter identification problem for systems with bounded disturbances," *Automatica*, vol. 23, no. 2, pp. 247–251, 1987.

[6] J. Deller Jr, M. Nayeri, and M. Liu, "Unifying the landmark developments in optimal bounding ellipsoid identification," *International journal of adaptive control and signal processing*, vol. 8, no. 1, pp. 43–60, 1994.

[7] T. Lin, M. Nayeri, and J. Deller Jr, "A consistently convergent obe algorithm with automatic estimation of error bounds," *International Journal of Adaptive Control and Signal Processing*, vol. 12, no. 4, pp. 305–324, 1998.

[8] D. Joachim and J. R. Deller, "Multiweight optimization in optimal bounding ellipsoid algorithms," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 679–690, 2006.

[9] M. Pouliquen and M. M'Saad, "Further stability and convergence analysis of a set membership identification," in *Proceedings of the 2nd International Symposium on Control, Communication and Signal Processing*, 2006.

[10] M. Pouliquen, O. Gehan, E. Pigeon, and M. Frikel, "Closed loop output error identification with bounded disturbances." *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 870–875, 2012.

[11] M. Pouliquen, O. Gehan, and E. Pigeon, "Bounded-error identification for closed-loop systems," *Automatica*, vol. 50, no. 7, pp. 1884–1890, 2014.

[12] E. Mousavinejad, X. Ge, Q.-L. Han, T. J. Lim, and L. Vlacic, "An ellipsoidal set-membership approach to distributed joint state and sensor fault estimation of autonomous ground vehicles," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 6, pp. 1107–1118, 2021.

[13] Q. Jin, R. Xiong, H. Mu, and J. Wang, "A set-membership algorithm based parameter identification method for lithium-ion batteries," *Energy Procedia*, vol. 152, pp. 580–585, 2018.

[14] M. Pouliquen, E. Pigeon, and O. Gehan, "Output error identification for multi-input multi-output systems with bounded disturbances," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, 2011, pp. 7200–7205.

[15] H. El Maizi, M. Pouliquen, S. Safi, and M. Frikel, "Identification of output error model with bounded disturbances," *11th IEEE International Conference on Systems and Control, Sousse, Accepted*, 2023.

Evaluating the Influence of PSO Inertia Weight on Bit Flipping Decoding Performance in QC-MDPC McEliece Cryptosystems

Abdellatif Kichna
LIMATI Laboratory
Polydisciplinary Faculty
Sultan Moulay Slimane University
Beni Mellal, Morocco
abdellatif.kichna@usms.ac.ma

Abderrazak Farchane
LIMATI Laboratory
Polydisciplinary Faculty
Sultan Moulay Slimane University
Beni Mellal, Morocco
a.farchane@gmail.com

Said Hakimi
LIMATI Laboratory
Polydisciplinary Faculty
Sultan Moulay Slimane University
Beni Mellal, Morocco
h_saidhakimi@yahoo.fr

***Abstract*—QC-MDPC McEliece cryptosystems represent a significant advancement in post-quantum cryptography, offering both security and efficiency in encryption. In the process of decryption, bit flipping plays a crucial role in error correction, ensuring the integrity of the decrypted message. However, optimizing the efficiency of bit flipping decoding remains a challenge. In this study, we pioneer the application of Particle Swarm Optimization (PSO) in conjunction with QC-MDPC McEliece cryptosystems to enhance the efficiency of bit flipping decoding. Specifically, we focus on investigating the influence of PSO inertia weight, a key parameter controlling the exploration and exploitation trade-off, on the performance of bit flipping decoding. Through empirical experiments and performance evaluations, we systematically analyze the effects of different inertia weight settings on the decryption capabilities of QC-MDPC McEliece cryptosystems. Our research not only sheds light on the novel application of PSO in cryptographic decryption but also provides valuable insights into the optimal tuning of PSO inertia weight parameters for improving the reliability and security of post-quantum cryptographic systems.**

***Index Terms*—bit-flipping decoder, Code-based cryptography, Particle Swarm Optimization, post-quantum cryptography, QC-MDPC codes**

I. Introduction

In recent years, the field of cryptography has been driven by the need for encryption methods that can withstand the growing threat posed by quantum computing. This has led to the development of post-quantum cryptographic systems, which offer potential solutions to these challenges. Among these systems, QC-MDPC (Quasi-Cyclic Moderate Density Parity-Check) McEliece cryptosystems have garnered attention for their ability to provide secure and efficient encryption in the post-quantum era. At the heart of QC-MDPC McEliece cryptosystems lies the use of bit flipping, a fundamental technique employed for error correction during decryption to ensure the accuracy and integrity of decrypted messages.

However, while QC-MDPC McEliece cryptosystems show promise for post-quantum security, optimizing the efficiency of the bit flipping decoding process remains a significant challenge. Traditionally, achieving optimal decoding efficiency requires the integration of advanced optimization techniques. In this context, Particle Swarm Optimization (PSO) has emerged as a powerful optimization algorithm, drawing inspiration from the social behavior of swarms, capable of efficiently navigating complex solution spaces and finding near-optimal solutions.

The integration of PSO with QC-MDPC McEliece cryptosystems offers a novel approach to enhancing decoding efficiency and improving the overall performance of post-quantum cryptographic systems. By leveraging PSO, the optimization process aims to strike a balance between exploration and exploitation, ultimately improving the decoding process and enhancing the reliability and security of cryptographic systems.

In this study, we embark on exploring the integration of PSO with QC-MDPC McEliece cryptosystems, with a specific focus on the influence of PSO inertia weight on decoding performance. Through a series of rigorous empirical experiments and performance evaluations, our objective is to uncover the intricate relationship between inertia weight configurations and decoding efficiency. By providing insights into the optimization of PSO parameters, our research aims to contribute to the advancement of post-quantum cryptographic systems and reinforce their resilience against emerging threats posed by quantum computing.

II. Background and Related Work

A. QC-MDPC-McEliece Cryptosystem

The QC-MDPC-based McEliece cryptosystem is an adaptation of the traditional McEliece cryptosystem, employing quasi-cyclic moderate density parity-check (QC-MDPC) codes [7]. It was developed to address the challenge posed by large key sizes in the original McEliece scheme, all while maintaining robust resistance against quantum attacks. The QC-MDPC variant introduces efficient key generation and encryption/decryption procedures, positioning it as a competitive contender in the realm of post-quantum cryptography.

1) Generation of Keys: To initiate the key generation process for a QC-MDPC McEliece system, we aim for a predetermined security level denoted as λ. Key parameters, including n_0, n, r, and w, are chosen to align with the desired security level λ.

Initially, we construct a random QC-MDPC code. This is accomplished by randomly selecting a vector h in $\mathbb{F}_2^n$ domain, which maintains a weight of w. We partition h into $n_0 = n/r$ equal segments, represented as $h = [h_0|h_1|...|h_{n_0-1}]$. The parity check matrix H is constructed subsequently as follows:

$$H = [H_0|H_1|...|H_{n_0-1}] \tag{1}$$

Here, each H_i constitutes a cyclic matrix, with h_i as the first row. A cyclic matrix is a square matrix where each row (or column) is a cyclic permutation of the previous row (or column). This means that each subsequent row (or column) is obtained by shifting the elements of the previous row (or column) cyclically to the right (or left). To ensure the accuracy of the process, the block H_{n_0-1} must be invertible. If this condition is not met, the procedure is restarted with a new randomly chosen h. H is kept as the private key.

The public key, denoted as G, is derived from H as:

$$G = (I|Q) \tag{2}$$

where I is the identity matrix and Q is:

$$Q = \begin{pmatrix} (H_{n_0-1}^{-1}H_0) \\ (H_{n_0-1}^{-1}H_1) \\ \vdots \\ (H_{n_0-1}^{-1}H_{n_0-2}) \end{pmatrix}. \tag{3}$$

2) Process of Encryption: Encryption is a fundamental step in the QC-MDPC McEliece Cryptosystem that allows secure communication of information. The process begins with a message vector m that resides in the finite field $\mathbb{F}_q^k$. This vector holds the actual information to be securely transmitted.

To obfuscate this information and protect it from potential eavesdroppers, a unique random error vector e is generated, residing in the finite field $\mathbb{F}_q^n$. This error vector serves a crucial purpose in introducing 'noise' into the system. Importantly, this vector is characterized by a weight $\leq t$, where t denotes the maximum number of errors that the QC-MDPC code can handle and correct.

The error vector e and message vector m are then used in conjunction to generate the ciphertext - the scrambled, unintelligible version of the message that is safe to transmit over the public channel. This process is carried out using the public key matrix G, a significant component of the McEliece Cryptosystem. The relationship is expressed succinctly by the equation $c = mG + e$, where the ciphertext c is derived from the multiplication of the message vector m and the public key matrix G, with the error vector e added.

The resulting ciphertext vector c now effectively disguises the original message m, making it secure for transmission. It's worth emphasizing that this encryption process leverages the mathematical properties of finite fields and the structure of the QC-MDPC McEliece Cryptosystem to ensure strong cryptographic security.

3) Decryption process: The decryption process, also known as decoding, is the mirror image of the encryption process - it is here that the original message m is recovered from the transmitted ciphertext $c \in \mathbb{F}_2^n$. This delicate procedure of reversing the encryption process is crucial for any cryptosystem, and the QC-MDPC McEliece Cryptosystem is no exception.

To correctly unravel the ciphertext and reveal the original message, a decoding algorithm is essential. A prominent method suggested in [7] is the bit flipping algorithm, which is a technique originally proposed by Robert Gallager in 1963 for decoding Low-Density Parity-Check (LDPC) codes, which are a type of error correcting codes [**?**].

At the heart of this algorithm lies a basic but powerful observation: each bit in the syndrome, a mathematical expression that captures the relationship between the received and transmitted codewords, reveals whether its corresponding equation is satisfied or not. If an equation is not satisfied, it signifies an error in the transmitted code [**?**].

In practice, positions in the received code that are implicated in a high number of unsatisfied equations are likely to be errors. Recognizing this, the bit-flipping algorithm iteratively flips these suspect bits based on a predefined threshold. The threshold is a predetermined value used to determine whether a bit in the received vector should be flipped during the decoding process. The algorithm compares the magnitude of the error associated with each bit to this threshold. If the error magnitude exceeds the threshold, the algorithm flips the corresponding bit in an effort to reduce the overall number of unsatisfied equations. This iterative process aims to home in on the transmitted code and, ultimately, the original message.

4) The Role of Threshold in the Bit Flipping decoder: In [7], a heuristic rule was introduced to define the threshold, primarily based on the maximum number of unsatisfied parity-check equations (UPC) at a given iteration, denoted as max(UPC). A constant value δ is subtracted from max(UPC) to set the bit-flipping threshold. The choice of δ is a crucial aspect of this method, with the suggested value of 5 obtained empirically for their specific system configuration and conditions as determined by [7].

In effect, this method introduces an adaptive, dynamic thresholding strategy that adjusts the threshold depending on the present condition of the decoding process [**?**]. At the start of the decoding, when the number of errors and unsatisfied equations is likely to be high, the threshold will also be relatively high. This prevents the algorithm from flipping bits too aggressively, reducing the risk of making premature or incorrect corrections.

As the decoding process progresses and the number of unsatisfied equations decreases, max(UPC) and thus the threshold will also decrease. This allows the algorithm to become more aggressive in flipping bits, helping to clear up residual errors.

The advantage of this method is its adaptability to varying decoding conditions. By making the threshold proportional to max(UPC), the decoder's behavior is made flexible. When

there are many errors, it is cautious; when there are few errors, it is assertive. This makes the decoder robust to a variety of error conditions and helps maintain reliable performance.

However, despite its efficiency, the bit-flipping decoding method is not without its drawbacks. One potential issue with this approach lies in its deterministic nature. In certain instances, the algorithm might reach a state of deadlock, where no bit satisfies the flipping condition, yet the syndrome is not zero. This situation represents a decoding failure.

Another concern involves the choice of the δ value. The proposed δ of 5, however, is largely heuristic and empirical. Although this value may work well for the specific system configuration and conditions that Misoczki et al. considered, it might not be universally applicable or optimal for other conditions or variations in the system. In other words, different QC-MDPC codes, operating conditions, or noise levels might necessitate different δ values for optimal decoding performance.

Given these inherent limitations of the heuristic thresholding strategy in the bit-flipping decoding process, there is a significant need to explore more flexible and robust solutions. One such promising approach is the implementation of optimization algorithms that can dynamically adjust the threshold according to the evolving decoding conditions. In this context, Particle Swarm Optimization (PSO) emerges as an attractive proposition.

III. Bit Flipping decoding algorithm

According to the definition of the parity-check matrix, each codeword can be considered as a solution to a linear equation system defined by the parity-check matrix. During encryption, an error-vector is added to the codeword, which means that with high probability, the encrypted ciphertext vector is not a codeword. As a result, the syndrome of the ciphertext becomes a non-zero vector. Each bit of the syndrome is calculated using a parity-check equation, and the i-th bit in the syndrome is generally calculated by multiplying the i-th row of the parity-check matrix with the ciphertext vector. If that bit is non-zero, it means that the parity-check is unsatisfied for the bits located at the positions of non-zero elements in the i-th row of the parity-check matrix.

In [7], in order to decode their MDPC code, the authors used the original Gallager's Bit Flipping algorithm [8] that was first introduced for LDPC codes. The bit-flipping algorithm takes advantage of counting the number of unsatisfied parity-checks for each ciphertext bit. This information is then used to decide whether a bit should be flipped or not, by using the exclusive OR operation (XOR). The algorithm works as follows: first, it starts by calculating the syndrome of the received codeword. Then it checks the number of unsatisfied parity check equations associated with each codeword bit and flips each bit that violates more than b unsatisfied equations. These steps are repeated until the syndrome becomes zero or the maximum number of iterations is reached. In this case, a decoding error is said to be occurred.

Algorithm 1 Bit flipping decoding algorithm

Require: $H \in \mathbb{F}_2^{r \times n}$ (parity check matrix), $y \in \mathbb{F}_2^n$ (received vector)
Ensure: $y \in \mathbb{F}_2^n$

for $iter = 1 \dots MaxIter$ **do**
  $s = yH^T$
  if $s = 0^r$ **then**
    break
  end if
  $upc = compute(H, s)$
  $th = threshold(context)$
  for $i = 1 \dots n$ **do**
    if $upc[i] \geq th$ **then**
      $y[i] = 1 - y[i]$
    end if
  end for
end for
return y

IV. Particle Swarm Optimization: An Overview

Particle Swarm Optimization (PSO) is a heuristic global optimization method introduced by Eberhart and Kennedy in 1995 [11], inspired by the collective dynamics observed in bird migration or fish schooling. The PSO algorithm operates based on the cooperation and intelligence inherent in swarm behavior.

Within the PSO framework, individual solutions are metaphorically visualized as 'birds' navigating the search field, known as 'particles.' Each particle, denoted by i, explores the solution space within d dimensions. The position of particle i is represented as x_{id}, and its velocity as v_{id}. Notations such as the inertia weight (w), cognitive parameter (c_1), social parameter (c_2), and random numbers (r_1 and r_2) ranging from 0 to 1 are integral to the algorithm. Personal best (p_{id}) and global best (p_{gd}) positions, as well as the fitness function ($f(x_{id})$) at the current position of a particle, further contribute to the PSO dynamics. Additionally, the termination condition, often defined in terms of a specific number of iterations or a fitness threshold, shapes the orchestration of the cooperative exploration of the solution space by the swarm of particles.

The basic PSO algorithm includes the following steps:

1) **Initialization:** Commence with a cluster of particles, each endowed with random placements and velocities across d dimensions within the problem scope. The location of each particle is characterized by a vector, symbolizing a probable solution. Likewise, each particle's velocity is signified by a vector, indicating the trajectory of the particle's motion. Moreover, each particle is equipped with a memory function, preserving its optimal achieved position, along with the globally optimal position attained by any particle.
2) **Fitness Evaluation:** For every particle, compute the fitness value tied to its present location. If this position surpasses its past optimum, the best position

gets updated. Analogously, should the present location outperform the global best, the global best position is refreshed.

3) **Velocity and Position Update:** Revise the velocity and location of every particle according to the subsequent formulas:

$$v_{id} = w \cdot v_{id} + c_1 \cdot r_1 \cdot (p_{id} - x_{id}) + c_2 \cdot r_2 \cdot (p_{gd} - x_{id}) \quad (4)$$

$$x_{id} = x_{id} + v_{id} \quad (5)$$

where v_{id} represents the particle's velocity, w stands for the inertia weight, c_1 and c_2 denote cognitive and social parameters, r_1 and r_2 are random numbers ranging from 0 to 1, p_{id} and p_{gd} correspond to the personal best and global best positions, respectively, and x_{id} indicates the particle's current location.

4) **Termination:** Continue executing steps 2 and 3 until a predetermined stopping condition is achieved. This could be the completion of a specific number of iterations or the attainment of a fitness value that meets or exceeds a particular threshold.

The advantage of PSO is that it has a fast convergence rate and it does not require the gradient of the problem, making it suitable for non-differentiable optimization problems [**?**]. Furthermore, PSO has been applied in various fields, such as neural network training, fuzzy system control, and other areas of engineering.

V. Experimental Setup

Our experimental setup aimed to investigate the influence of PSO inertia weight on bit flipping decoding performance in QC-MDPC McEliece cryptosystems. Initially, we designed a series of experiments to assess decoding efficiency under varying inertia weight configurations. We utilized a standard QC-MDPC McEliece cryptosystem implementation, integrating PSO for threshold optimization during the decoding process.

The experimental procedures involved:

- Parameter Initialization: We initialized the PSO algorithm with a range of inertia weight values, encompassing both low and high values to explore a broad spectrum of configurations.
- Decoding Simulations: Using a predetermined set of encoded ciphertexts, we simulated the decoding process with different inertia weight settings. Each simulation aimed to decode the ciphertext using the bit flipping algorithm, adjusting the threshold based on the specified inertia weight.
- Performance Evaluation: After each decoding simulation, we evaluated the performance metrics, including decoding accuracy, convergence speed, and computational efficiency. These metrics provided insights into the impact of inertia weight on decoding performance.

Across the experiments, we meticulously recorded several performance metrics to assess the efficacy of PSO inertia weight optimization. Key metrics included decoding accuracy, convergence speed, and computational efficiency. These metrics served as quantitative indicators of the algorithm's performance under different inertia weight configurations.

VI. Results and Discussion

The results obtained from our experiments, as depicted in figure 1, provide valuable insights into the relationship between PSO inertia weight w and decoding performance metrics, particularly the fitness value represented by the syndrome weight. Our findings reveal a compelling trend wherein the convergence behavior of the PSO algorithm exhibits a clear dependence on the inertia weight parameter. Specifically, as illustrated in the experimental data, when w is small, the PSO algorithm struggles to converge efficiently, resulting in a lower success rate and prolonged convergence time. This phenomenon can be attributed to the reduced influence of particle velocity updates on the exploration and exploitation processes within the swarm. However, as w increases, we observe a notable improvement in the convergence behavior of the PSO algorithm, accompanied by a corresponding enhancement in the success rate of decoding. The heightened w values facilitate more pronounced exploration and exploitation of the solution space by individual particles, leading to a more effective search for optimal solutions. Consequently, decoding performance, as measured by the fitness value or syndrome weight, demonstrates improvement with increasing w values, indicating a positive correlation between inertia weight and PSO algorithm convergence.

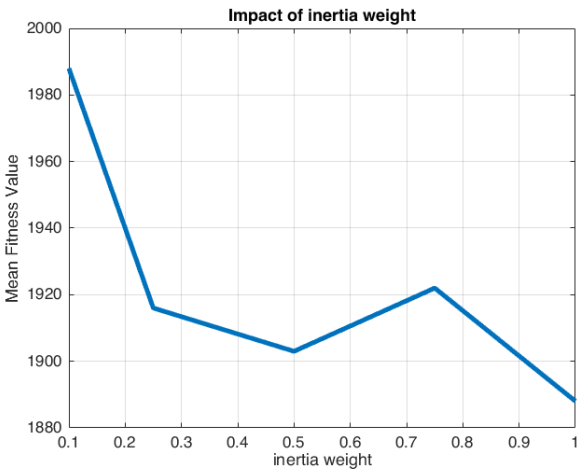


Fig. 1. Impact of Inertia Weight on PSO Fitness Value

Table I provides a summary of the corrected bits and wrong flips observed for each inertia weight value across all three iterations, facilitating a comprehensive comparison of the algorithm's performance under different parameter settings. The results of our experiments, which investigated the influence of inertia weight values ranging from 0.1 to 1 on the performance of the particle swarm optimization (PSO) algorithm, provide valuable insights into the dynamics of the decoding process in QC-MDPC McEliece cryptosystems.

We examine the behavior of the PSO algorithm concerning the number of corrected bits in each iteration. Across all three iterations (denoted as Iteration 1, Iteration 2, and Iteration 3), we observed variations in the number of corrected bits corresponding to different inertia weight values. For example, in Iteration 1, the corrected bits ranged from 32.5 to 37.5, with higher inertia weight values generally associated with higher numbers of corrected bits. This trend suggests that increasing the inertia weight promotes exploration within the solution space, enabling the PSO algorithm to traverse a broader range of potential solutions and consequently correct more errors in the decoding process.

TABLE I
PERFORMANCE METRICS FOR DIFFERENT INERTIA WEIGHT VALUES

w	Corrected Bits			Wrong Flips		
	Iter 1	Iter 2	Iter 3	Iter 1	Iter 2	Iter 3
0.1	34.5	39.7	10.0	12.0	15.9	8.1
0.2	35.7	36.5	12.0	12.8	13.9	6.7
0.3	37.5	39.5	7.8	11.9	13.4	5.7
0.4	34.4	39.7	10.1	9.8	14.1	7.1
0.5	35.5	40.4	13.7	10.4	13.7	5.9
0.6	33.3	40.4	10.9	9.7	10.6	4.9
0.7	32.5	40.1	12.0	9.9	13.4	5.9
0.8	35.8	38.5	10.3	12.5	12.7	6.8
0.9	36.5	39.3	8.6	9.8	11.6	4.2
1.0	35.1	40.8	8.5	9.1	11.9	6.4

Similarly, in Iterations 2 and 3, we observed comparable trends, with higher inertia weight values consistently yielding higher numbers of corrected bits. This consistency across iterations underscores the robustness of the relationship between inertia weight and the efficiency of error correction in QC-MDPC McEliece cryptosystems. Moreover, the gradual increase in the number of corrected bits as the inertia weight value rises indicates a progressive improvement in the algorithm's ability to converge towards optimal solutions, thereby enhancing the overall decoding performance.

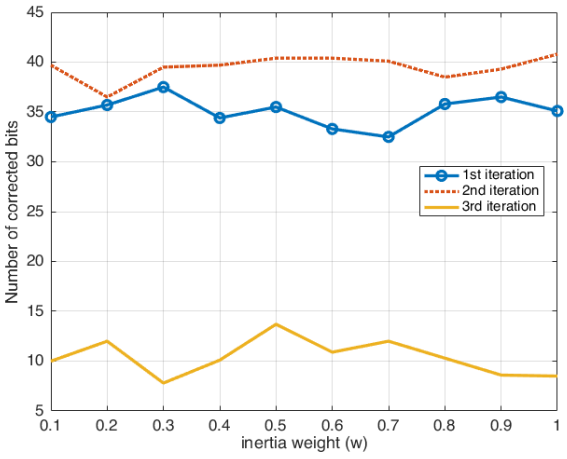


Fig. 2. Impact of Inertia Weight on the number of corrected bits

We also consider the occurrence of wrong flips in each iteration, which serves as a complementary metric for assessing the reliability of the decoding process. Interestingly, we observed a contrasting trend compared to the number of corrected bits. Specifically, lower inertia weight values tended to result in fewer wrong flips, indicating a more cautious approach to error correction. In contrast, higher inertia weight values were associated with an increased incidence of wrong flips, suggesting a greater tendency towards aggressive exploration within the solution space.

The contrast between the trends in the number of corrected bits and wrong flips underscores a fundamental dilemma present in optimization algorithms such as PSO. Higher inertia weight values encourage broader exploration, possibly resulting in more corrections, yet also raise the chance of incorrect adjustments or wrong flips. On the other hand, lower inertia weight values emphasize exploitation and cautious decision-making, decreasing the risk of wrong flips but potentially constraining the algorithm's capacity to explore a wide range of solutions.

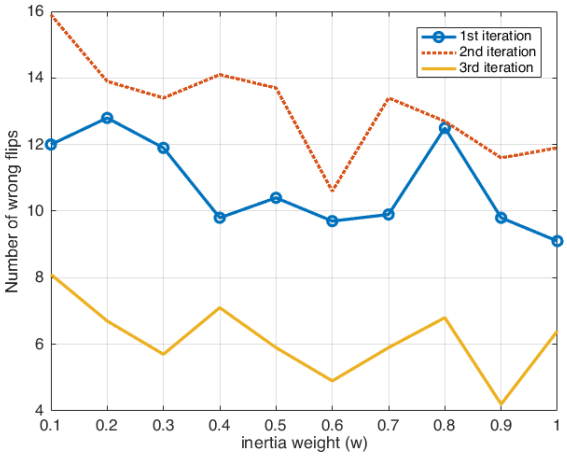


Fig. 3. Impact of Inertia Weight on the number of wrong flips

VII. CONCLUSION AND FUTURE WORKS

In conclusion, our investigation highlights the significance of PSO inertia weight optimization in enhancing the efficiency and reliability of bit flipping decoding in QC-MDPC McEliece cryptosystems. By leveraging PSO to fine-tune inertia weight parameters, cryptographic practitioners can bolster the security posture of post-quantum cryptographic systems, paving the way for robust protection against emerging quantum threats.

While our study provides valuable insights into the role of PSO inertia weight in bit flipping decoding, it is not without limitations. Future research endeavors could explore additional factors, such as population size and cognitive parameters, to further refine the optimization process. Additionally, conducting experiments with real-world cryptographic datasets could offer a more comprehensive understanding of PSO's applicability in practical decryption scenarios.

REFERENCES

[1] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring." Proceedings 35th Annual Symposium on Foundations of Computer Science, pp. 124-134. IEEE, 1994.

[2] R. J. McEliece, "A public-key cryptosystem based on algebraic," Coding Thv, 4244, 114-116, 1978.

[3] C. Monico, J. Rosenthal, and A. Shokrollahi, "Using low density parity check codes in the McEliece cryptosystem," 2000 IEEE International Symposium on Information Theory (Cat. No.00CH37060), p. 215. IEEE, 2000.

[4] P. Gaborit, "Shorter keys for code based cryptography," In Proceedings of the 2005 International Workshop on Coding and Cryptography (WCC 2005), pp. 81-91. 2005.

[5] M. Baldi and F. Chiaraluce, "Cryptanalysis of a new instance of McEliece cryptosystem based on QC-LDPC codes," 2007 IEEE International Symposium on Information Theory, 2007.

[6] T. Fabšič, V. Hromada, P. Stankovski, P. Zajac, Q. Guo, and T. Johansson, "A reaction attack on the QC-LDPC mceliece cryptosystem," Post-Quantum Cryptography, pp. 51–68, 2017.

[7] R. Misoczki, J.-P. Tillich, N. Sendrier, and P. S. Barreto, "MDPC-McEliece: New mceliece variants from moderate density parity-check codes," 2013 IEEE International Symposium on Information Theory, 2013.

[8] R. G. Gallager, "Low-density parity-check codes," 1963.

[9] N. Drucker, S. Gueron, K. Dusan, "Additional implementation of BIKE," https://bikesuite.org/additional.html, 2019.

[10] N. Drucker, S. Gueron, and D. Kostic, "QC-MDPC decoders with several shades of Gray," Post-Quantum Cryptography, pp. 35–50, 2020.

[11] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of ICNN'95-International Conference on Neural Networks, vol. 4, pp. 1942-1948, 1995.

[12] N. Sendrier and V. Vasseur, "On the decoding failure rate of QC-MDPC bit-flipping decoders," Post-Quantum Cryptography, pp. 404–416, 2019.

[13] N. Aragon, P.S. Barreto, S. Bettaieb, L. Bidoux, O. Blazy, J.C. Deneuville, P. Gaborit, S. Gueron, T. Guneysu, C.A. Melchor, and R. Misoczki, "BIKE: bit flipping key encapsulation," 2017.

[14] N. Aragon, P. S. L. M. Barreto, S. Bettaieb, L. Bidoux, O. Blazy, J.-C. Deneuville, P. Gaborit, S. Gueron, T. Güneysu, C. A. Melchor, R. Misoczki, E. Persichetti, N. Sendrier, J.-P. Tillich, V. Vasseur, and G. Zémor. "BIKE: Bit Flipping Key Encapsulation," Round 2 Submission. 3.0. 2019.

[15] A. Nilsson, I. E. Bocharova, B. D. Kudryashov, and T. Johansson, "A Weighted Bit Flipping Decoder for QC-MDPC-based Cryptosystems," in 2021 IEEE International Symposium on Information Theory (ISIT), IEEE Press, pp. 1266-1271, 2021.

[16] B. Imine, N. Hadj-Said, and A. Ali-Pacha, "McEliece cryptosystem based on Plotkin construction with QC-MDPC and QC-LDPC codes," arXiv:2211.14206 [cs], Dec. 2022.

[17] A. Janoska, "MDPC decoding algorithms and their impact on the McEliece cryptosystem," Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, 2018.

[18] J. Liu, X. Tong, Z. Wang, M. Zhang, and J. Ma, "An improved McEliece cryptosystem based on QC-MDPC code with Compact Key Size," Telecommunication Systems, vol. 80, no. 1, pp. 17–32, 2022.

[19] G. Wu, R. Yang, and Z. Dai, "Design of McEliece cryptosystem based on QC-MDPC codes," 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology (ICEICT), 2020.

Fuzzy Logic for Energy Efficiency Enhancing in Internet of Things (IoT)

1st LAGNFDI Oussama
LIMATI LABORATORY
Sultan Moulay Slimane University
Beni mellal, Morocco
lagnfdi.o@gmail.com

2nd MYYARA Marouane
LIMATI LABORATORY
Sultan Moulay Slimane University
Beni mellal, Morocco
mar.myyara@gmail.com

3rd DARIF Anouar
LIMATI LABORATORY
Sultan Moulay Slimane University
Beni mellal, Morocco
anouar.darif@gmail.com

***Abstract*—As the Internet of Things (IoT) continues to proliferate, managing energy consumption becomes a critical concern.Our study focuses on optimizing energy consumption within IoT environments using fuzzy logic, aiming to develop efficient models for enhanced performance. With the rapid evolution of IoT and increasing device connectivity, effective energy management is paramount. We conducted comprehensive research on IoT concepts and devised systems based on fuzzy logic to significantly improve energy efficiency. Our findings highlight the potential of fuzzy logic in optimizing energy consumption within IoT networks. The proposed models signify a substantial stride towards achieving optimal energy utilization in IoT systems. They hold considerable promise for driving future advancements in IoT technology and contributing to sustainability efforts. Our study contributes valuable insights into enhancing energy efficiency in IoT, paving the way for more environmentally friendly and resource-efficient IoT implementations.**

***Index Terms*—Fuzzy logic, energie consmpution, IoT, nergy Efficiency Enhancing, Bnadwidth, Time transsmisison**

I. Introduction

The projected quantity of substantial Internet of Things (IoT) devices in the foreseeable future is estimated to reach billions and is consistently escalating [1] [2]. These devices are expected to maintain connectivity with highly dependable and low-latency networks, continuously transmitting data throughout their operational lifespan [3] [4].The fundamental purpose of IoT devices is to consistently gather and transmit perceived data to interact with the physical environment. The hardware components of an IoT device typically include a sensor powered by a battery, an actuator, and a communication system.The Internet of Things (IoT) has a profound impact on industries and daily life [5] [6]. It enables the interconnectivity of physical devices, vehicles, and appliances embedded with sensors, software, and network connectivity. This connectivity allows for data collection and exchange, aiming to create a seamless connection between the physical and digital world [9].The growing worldwide need for energy, which is expected to rise by more than 25% over the next two decades, poses a crucial challenge in today's society [10].Addressing this difficulty efficiently while minimizing waste has become a top priority. As the network's capacity to accommodate IoT devices is expanding with the advent of 5G and comparable technologies, managing data complexity and optimizing energy usage for IoT devices is evolving into a challenging endeavor for researchers [11]. The implementation of IoT into energy management systems provides numerous benefits, altering how firms manage energy usage and conservation.Better energy monitoring IoT devices provide exceptional precision in energy monitoring, allowing organizations to track real-time energy use patterns. This increased visibility into usage patterns allows for more informed decision-making and proactive efforts to improve energy use. Leveraging IoT-generated data enables firms to assess energy consumption trends and discover inefficiencies, establishing the framework for focused strategies to optimise energy usage, reduce waste, and improve overall operational efficiency [12] [13].

IoT integration promotes the development of dynamic smart energy management systems, which use IoT-enabled devices to automate operations, regulate energy usage in real-time, and intelligently respond to changing demands, resulting in more efficient resource utilization.By using the power of IoT,businesses unlock unprecedented opportunities for precise energy management, reducing costs, mitigating environmental impact, and fostering sustainable practices [14]. To overcome these challenges, scientists have adopted fuzzy logic, a mathematical framework designed to deal with uncertainties. Fuzzy logic is very useful in the field of IoT due to its ability to manage complex models, handle imprecise input, and provide understandable reasoning. In the field of IoT, researchers use fuzzy logic to improve performance, evaluate efficiency, and increase energy consumption [8] [9].

In this paper, our main objective is to optimize energy consumption in IoT networks using Fuzzy logic to enhance efficiency. Our research methodology prioritizes energy efficiency as a primary concern, addressing identified issues through Fuzzy logic implementation. We demonstrate the effectiveness of our approach in improving IoT network energy efficiency. The paper is structured with related works in section 2, IoT

Identify applicable funding agency here. If none, delete this.

introduction in section 3, fuzzy logic reviews in section 4, problem scenario and our model in Section 5, and finally a conclusion in Section 6.

II. Relted works

Limited research has been conducted on optimizing the energy consumption of IoT devices. Nonetheless, substantial advancements have been made in optimizing energy usage within systems that leverage IoT devices, such as (smart grids, industry 4.0, and smart cities).
The [15]research focuses on energy optimization in smart cities with interconnected devices. It proposes a model for optimizing energy in smart homes and cities, using IoT, 5G, and cloud tech. The model targets key smart city features like lighting, billboards, homes, and parking. Mathematical modeling evaluation suggests it could enhance energy efficiency in smart cities.In [16] assesses LoRa network energy usage under the LoRaWAN protocol, comparing a reference scenario with three variants considering post-packet transmission end-device states. Simulation reveals a notable rise in network energy consumption post-uplink packet transmission without acknowledgment. The conclusion underscores the necessity of factoring in end-device states and acknowledgments for precise network energy consumption evaluation.
A study by [17] delves into the transformative capabilities of IoT in optimizing energy utilization, highlighting how IoT-powered smart energy management systems are reshaping the landscape of energy resource monitoring, control, and conservation. It discusses the impact of IoT on energy efficiency and conservation efforts, exploring how IoT is revolutionizing energy resource management.A research was undertaken in [18] focuses on developing an energy-efficient framework to balance energy consumption in IoT devices. It utilizes an Energy Harvesting MAC protocol and applies Reinforcement Learning for node modeling. The results demonstrate an impressive 80% reduction in energy usage by IoT devices, significantly enhancing overall performance compared to current energy harvesting solutions for sustainable IoT systems.

III. Internet of things (IoT)

A. Definition

The Internet of Things (IoT) is simply a network of physical gadgets, automobiles, appliances, and other objects that are integrated with sensors, software, and network connectivity, enabling these objects to collect and exchange data.IoT devices form interconnected networks, facilitating communication and data exchange among themselves and other internet-enabled devices. They autonomously monitor environmental conditions, manage traffic, control machinery, and track inventory and shipments.

B. IoT Archeticture

The classic three-layer design of IoT, which was used in its early phases, comprises of Perception, Network and Application layer

- **The Perception Layer :** gathers data via sensors and embedded tech for spatial and object detection.
- **The Network Layer :** manages data distribution, storage, transfer, and connectivity for IoT devices
- **The application layer :** lets users interact in smart home apps to activate equipment, supporting IoT applications like smart homes, and healthcare.

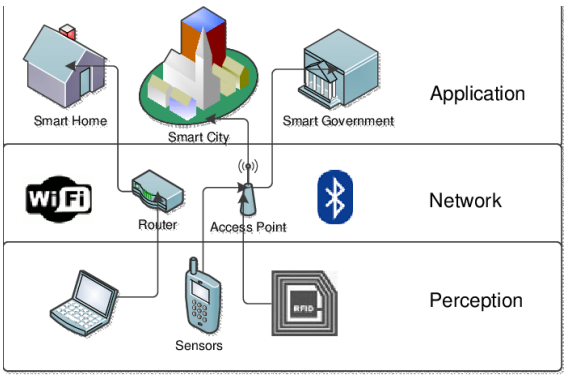


Fig. 1. IoT Three-Layer-Architecture

IV. Fuzzy system

Fuzzy logic is a mathematical framework for handling uncertainty by allowing degrees of truth between true and false. It extends classical binary logic to represent and manipulate imprecise information more effectively [10]. Fuzzy logic is widely used in various fields such as engineering, artificial intelligence, and control systems. It enables more nuanced decision-making in situations where precise rules are difficult to define or where systems operate in uncertain environments.

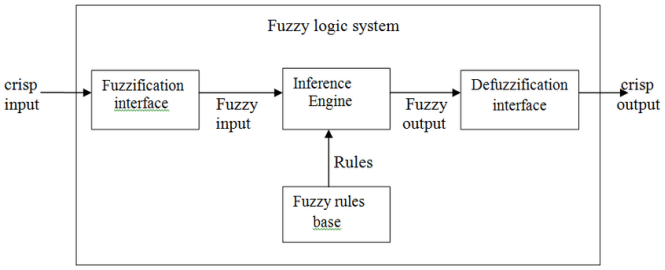


Fig. 2. Caption

Knowledge Base: It houses IF-THEN rules from experts. Inference Engine: Simulates human reasoning through fuzzy inference on inputs and IF-THEN rules. Defuzzification Module: Converts fuzzy sets from the inference engine into precise values.

V. Assessment Energy consumption Model

The Internet of Things (IoT) is a vast network of devices that are interconnected, sharing and receiving data. This

continuous data flow, however, can result in substantial energy usage. As such, the design and operation of IoT systems must prioritize energy efficiency to ensure sustainability.

In our cas, We will determine the energy consumption associated with the transmission time and bandwidth of networks in the Internet of Things (IoT). This entails measuring the energy consumed during data transmission and reception, taking into account signal strength, data rate, and network circumstances.

- **Energy consumption (mW) ;** refers to the amount of electrical power used by devices for sensing, processing, and transmitting data. Optimizing energy consumption is vital for extending battery life, reducing costs, and enhancing sustainability in IoT deployments.
- **Transmission time (ms) :** refers to the duration it takes for data to be sent from one point to another in a network or communication system.
- **Bandwidth (Kpbs) :** is the measure of how much data can be transmitted in a given amount of time through a network or communication channel.

Furthermore,We employ the Mean of Maxima (MOM) defuzzification method to interpret fuzzy sets on energy consumption in IoT networks, enhancing understanding. This aids in designing energy-efficient IoT systems.

$$Energy = \frac{\sum_{i=1}^{n} y_i u_i}{\sum_{i=1}^{n} u_i}$$

Where : y_i represents the data points, u_i represents the membership values corresponding to the data points, and n is the total number of data points.
Also In this research, our energy consumption evaluation model is based on a new technique. We utilized fuzzy logic to evaluate energy consumption.The assessment of eenrgie consmpution perfomed shows in figure 3.

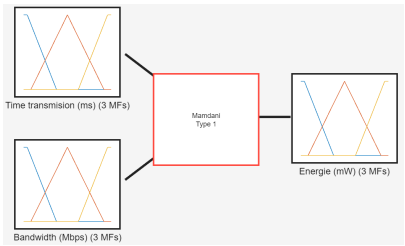


Fig. 3. Assessment energie consmpution model

We use fuzzy variables and linguistic phrases to characterize input parameters such as ”transmission time” (short, average, and long) and ”bandwidth” (narrow, moderate, and wide).Using fuzzy variables, we create an output variable that categorizes energy consumption as Low, Medium, or High.
The inputs and output parameters have specified ranges as follows: Transmission time ranges from 0 to 100 milliseconds, Bandwidth ranges from 0 to 100 kilobits per second (kbps), and Energy consumption ranges from 0 to 100.Also, two membership functions, trapezoidal and triangular, have been identified for the parameters.
The membership function graph can be drawn in various shapes, such as triangles, trapezoids, and bell curves, to appropriately depict the distribution of data within the system.A triangular fuzzy number is denoted by a triplet (a, b, c), illustrated in Fig. 3. Its membership function is computed using equation under.

$$\mu(y) = \begin{cases} \frac{y-a}{b-a}, & a \le y \le b \\ \frac{c-y}{c-b}, & b \le y \le c \\ 0, & Further \end{cases}$$

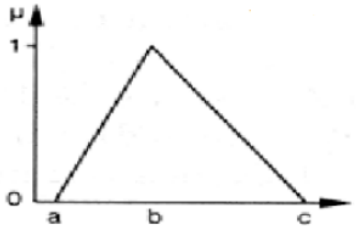


Fig. 4. Triangular Fuzzy Shape

A trapezoid’s fuzzy number can be described using four variables (a, b, c, d), illustrated in Figure 4. The membership function is determined using the equation under.

$$\mu(y) = \begin{cases} \frac{y-a}{b-a}, & a \le y \le b \\ 1, & b \le y \le c \\ \frac{d-y}{d-c}, & c \le y \le d \end{cases}$$

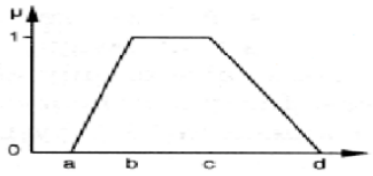


Fig. 5. Trapezoid Fuzzy Shape

VI. IMPLEMENTATION AND RESULTS

We developed the proposed model using MATLAB’s Fuzzy Logic Designer, following the methods outlined above for creating FCS. Our research aims to evaluate energy consumption in the Internet of Things.We used triangle and trapezoid shapes to represent the fuzzy set for all inputs and outputs in three levels.

A. Result

We tested the new model by creating a data set of 15 conditions using MATLAB’s random function. The model successfully calculates energy consumption for all 15 input parameters.Table 1 displays the results achieved in this model.

TABLE I
ENERGIE CONSMPUTION

Condition	Bandwidth(kbps)	Time trms(ms)	Energie cnsmpt(mW)
1	5	24	12
2	12	24	10.5
3	21	24	50
4	50	24	6
5	60	24	96,5
6	13	24	50
7	90	24	4
8	54	24	3.5
9	12	37	7
10	21	24	12
11	31	51	9.5
12	83	10	91.5
13	41	22	50
14	88	55	94
15	77	11	88.5

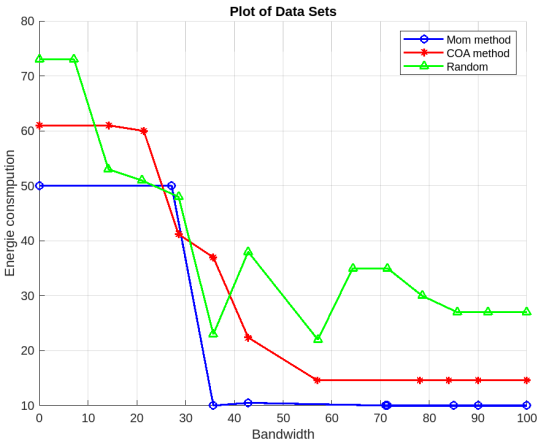


Fig. 6. Energy consumption as a function of bandwidth

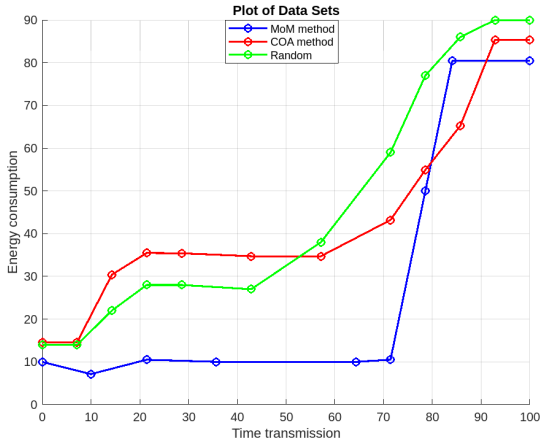


Fig. 7. Energy consumption as a function of time transmission

B. Discussion

The study thoroughly examined energy use under 15 different parameter circumstances, including variations in bandwidth and time transmission. Surprisingly, the mean of maxima technique consistently beat both the random function and fuzzy logic center of gravity methods, demonstrating its strength in maximizing energy efficiency.This superiority can be attributed to the mean of maxima method's inherent capacity to prioritize high values, resulting in good energy management under changing situations. The study's findings underline the importance of algorithmic optimization in reducing energy usage, particularly in dynamic contexts.The mean of maxima method demonstrated versatility and scalability, making it an attractive choice for energy-efficient applications.

Algorithmic complexities, such as weighing maximum values, have a considerable impact on energy consumption outcomes, emphasizing the importance of algorithm design in energy optimization tactics. While the random function and fuzzy logic center of gravity approaches are useful in some situations, they performed poorly in compared to the mean of maxima method across the various parameter circumstances studied. Furthermore, the study emphasizes the importance of personalized techniques based on unique parameter conditions, as energy optimization tactics can differ depending on the system's requirements and limits. Real-world applications are required to validate these discoveries and transform them into viable energy-saving solutions for a variety of industries.

Factors such as processing overhead, algorithmic complexity, and flexibility to changing contexts all have an impact on these technologies' practical application. Future research efforts could include investigating other parameters, delving deeper into computational optimizations, and conducting large-scale field testing to evaluate and develop energy optimization methodologies. Overall, the study provides useful insights into increasing energy efficiency and lays the framework for future advances in energy-conscious system design and optimization.

VII. CONCLUSION

Our comparative examination of energy consumption models, which used a random function, fuzzy logic center of gravity, and mean of maxima methods, yielded important insights under 15 different parameter settings. The mean of maximum approach regularly outperformed random function and fuzzy logic center of gravity methods in terms of energy efficiency. This demonstrates the method's efficacy in optimizing energy use, especially in circumstances including bandwidth and time transmission changes.The study underscores the importance of algorithmic optimization in reducing energy usage and the need for personalized techniques depending on individual parameter conditions. Informed decision-making in energy-efficient system design and optimization methodologies are some of the practical ramifications. Future research could investigate other parameters, test findings in real-world implementations, and address factors like processing overhead and scalability.Overall, the mean of maxima technique appears to

be a potential solution for lowering energy consumption in a variety of settings, thereby contributing to sustainable and energy-efficient practices in numerous industries.

REFERENCES

[1] Raval, Maulin, Shubhendu Bhardwaj, Aparna Aravelli, Jaya Dofe, and Hardik Gohel. "Smart energy optimization for massive IoT using artificial intelligence." Internet of Things 13 (2021): 100354.

[2] Stankovic, John A., Tu Le, Abdeltawab Hendawi, and Yuan Tian. "Hardware/software security patches for internet of trillions of things." arXiv preprint arXiv:1903.05266 (2019).

[3] Shabana Anjum, Shaik, Rafidah Md Noor, Ismail Ahmedy, and Mohammad Hossein Anisi. "Energy optimization of sustainable Internet of Things (IoT) systems using an energy harvesting medium access protocol." In IOP Conference Series: Earth and Environmental Science, vol. 268, no. 1, p. 012094. IOP Publishing, 2019.

[4] Mujawar, Mubarak A., Hardik Gohel, Sheetal Kaushik Bhardwaj, Sesha Srinivasan, Nicolerta Hickman, and Ajeet Kaushik. "Nano-enabled biosensing systems for intelligent healthcare: towards COVID-19 management." Materials Today Chemistry 17 (2020): 100306.

[5] Faye, Ibrahima, Pape Abdoulaye Fam, and Mamadou Lamine Ndiaye. "Energy consumption of IoT devices: An accurate evaluation to better predict battery lifetime." Radio Science 57, no. 12 (2022): 1-10.

[6] Gohel, Hardik A., Himanshu Upadhyay, Leonel Lagos, Kevin Cooper, and Andrew Sanzetenea. "Predictive maintenance architecture development for nuclear infrastructure using machine learning." Nuclear Engineering and Technology 52, no. 7 (2020): 1436-1442.

[7] Krishnan, R. Santhana, E. Golden Julie, Y. Harold Robinson, S. Raja, Raghvendra Kumar, and Pham Huy Thong. "Fuzzy logic based smart irrigation system using internet of things." Journal of cleaner production 252 (2020): 119902.

[8] Lin, Yu-Hsien, Chao-Ming Yu, and Chia-Yu Wu. "Towards the design and implementation of an image-based navigation system of an autonomous underwater vehicle combining a color recognition technique and a fuzzy logic controller." Sensors 21, no. 12 (2021): 4053.

[9] Hasan, Nadine, Ayaskanta Mishra, and Arun Kumar Ray. "Fuzzy logic based cross-layer design to improve Quality of Service in Mobile ad-hoc networks for Next-gen Cyber Physical System." Engineering Science and Technology, an International Journal 35 (2022): 101099.

[10] Raval, Maulin, Shubhendu Bhardwaj, Aparna Aravelli, Jaya Dofe, and Hardik Gohel. "Smart energy optimization for massive IoT using artificial intelligence." Internet of Things 13 (2021): 100354.

[11] Singh, Ritesh Kumar, Priyesh Pappinisseri Puluckul, Rafael Berkvens, and Maarten Weyn. "Energy consumption analysis of LPWAN technologies and lifetime estimation for IoT application." Sensors 20, no. 17 (2020): 4794.

[12] Ahmed, Zeinab E., Mohammad Kamrul Hasan, Rashid A. Saeed, Rosilah Hassan, Shayla Islam, Rania A. Mokhtar, Sheroz Khan, and Md Akhtaruzzaman. "Optimizing energy consumption for cloud internet of things." Frontiers in Physics 8 (2020): 358.

[13] Rajeswari, Alagan Ramasamy, Kanagasabai Kulothungan, Sannasi Ganapathy, and Arputharaj Kannan. "Trusted energy aware cluster based routing using fuzzy logic for WSN in IoT." Journal of Intelligent Fuzzy Systems 40, no. 5 (2021): 9197-9211.

[14] Akbari, Yalda, and Shayesteh Tabatabaei. "A new method to find a high reliable route in IoT by using reinforcement learning and fuzzy logic." Wireless Personal Communications 112, no. 2 (2020): 967-983.

[15] D'Aniello, Giuseppe. "Fuzzy logic for situation awareness: a systematic review." Journal of Ambient Intelligence and Humanized Computing 14, no. 4 (2023): 4419-4438.

[16] Humayun, Mamoona, Mohammed Saleh Alsaqer, and Nz Jhanjhi. "Energy optimization for smart cities using iot." Applied Artificial Intelligence 36, no. 1 (2022): 2037255.

[17] Banti, Konstantina, Ioanna Karampelia, Thomas Dimakis, Alexandros-Apostolos A. Boulogeorgos, Thomas Kyriakidis, and Malamati Louta. "LoRaWAN communication protocols: A comprehensive survey under an energy efficiency perspective." In Telecom, vol. 3, no. 2, pp. 322-357. MDPI, 2022.

[18] Mishra, Priyanka, and Ghanshyam Singh. "Energy management systems in sustainable smart cities based on the internet of energy: A technical review." Energies 16, no. 19 (2023): 6903.

[19] Energy-Efficient Computation Offloading With DVFS Using Deep Reinforcement Learning for Time-Critical IoT Applications in Edge Computing.

Improving Profile Recommendations with AI for Strategic Decision-Making: An Overview

1st Hicham OUALLA
AKKODIS Research
Paris, France
ouallahicham93@gmail.com

2nd Abdelkrim SAOUABE
Computer Science Research Laboratory, Faculty of Sciences, IbnTofail University
Kenitra, Morocco
abdelkrim.saouabe@akkodis.com

3rd Imad MOURTAJI
AKKODIS Research
Paris, France
imad.mourtaji@akkodis.com

4th Rachid FATEH
UNICAEN, Inserm U1086 ANTICIPE University of Cean Normandie
Caen, France
Fateh.smi@gmail.com

5th Merouane MAZAR
AKKODIS Research
Paris, France
merouane.mazar@akkodis.com

***Abstract*—Artificial intelligence-based CV recommendation systems use algorithms to analyze candidates' skills, experience and preferences, as well as employers' requirements, to provide personalized recommendations. These systems can help recruiters find qualified candidates faster by eliminating human bias and effectively matching candidate profiles with available job vacancies. They help improve the efficiency of the recruitment process by providing relevant suggestions and reducing the time needed to sort through CVs.**

***Index Terms*—Recommender system, deep learning, resume Similarity, machine learning**

I. Introduction

In the ever-changing world of recruitment and talent management, companies are increasingly faced with the challenge of quickly and efficiently sorting through a growing number of applications to find the best talent. To meet this growing demand, artificial intelligence-based CV recommendation systems have emerged as promising tools for optimizing the candidate selection process. These systems exploit the capabilities of artificial intelligence to analyze, rank and recommend CVs according to employers' specific needs and candidates' skills.

The introduction of these systems is revolutionizing the way companies manage their recruitment processes, offering significant benefits such as reducing the time needed to review applications, improving the relevance of recommendations and minimizing human bias in the selection process. Indeed, according to a study by research firm Gartner, by 2025, over 50% of large companies will be using artificial intelligence technologies to improve their recruitment and talent management processes.

A streamlined and instantaneous job matching service is highly coveted by both employers and job seekers, contributing to long-term socioeconomic prosperity [6]. The proliferation of online recruitment platforms has been remarkable, particularly in response to the COVID-19 pandemic, with millions preferring digital hiring and job-seeking avenues ([7], [8]). CareerBuilder, renowned for its expansive online job boards and diverse recruitment services, plays a pivotal role in facilitating this trend. Among its key offerings is an online recruitment matching system, vital to CareerBuilder's global operations, serving millions of users daily. With a vast influx of job postings and resumes, the primary challenge lies in constructing a recommender system capable of real-time candidate targeting for employers and job discovery for seekers. To tackle this challenge, we propose a two-stage recommendation system employing an embedding-based approach [9] and [10].

The rest of this article is organized as follows: section 2 reviews the various recommendation systems available, while section 3 presents a short synthesis of these systems. Finally, a conclusion section ends the paper.

II. Recommendation systems

A. Recammandtion system 1

To tackle the primary hurdles within the HR industry, namely: distinguishing the most suitable candidates, interpreting CVs effectively, and ensuring candidates possess the necessary skills before recruitment, the article [1] introduces an automated strategy called "Resume

Classification and Matching." This approach involves a multi-step process aimed at enhancing efficiency and accuracy in candidate selection

Initially, the method involves categorizing resumes using various classifiers to identify the most relevant candidates. Subsequently, employing Content-based Recommendation techniques, such as cosine similarity and k-NN, enables the ranking of top candidates based on how closely their profiles align with the job description provided.

The proposed solution primarily relies on supervised learning techniques to classify resumes accurately into different expertise domains. This classification process is multifaceted and incorporates several steps:

- Pre-processing: This step involves cleaning the text data by removing stop words, stemming, and lemmatization to enhance the quality of the analysis.
- Utilizing NLP Techniques: Advanced techniques such as Named Entity Recognition (NER) and Natural Language Processing (NLP) are applied to extract meaningful information from the resumes. Additionally, text classification using n-grams helps in identifying patterns and relevant features within the text data.
- Distance-Metric Based Classification: Leveraging distance metrics facilitates the comparison of resumes with job descriptions, enabling the identification of the most suitable candidates based on their proximity to the requirements outlined.
- Feedback Loop Incorporation: An essential aspect of the proposed solution is its ability to adapt and improve over time. By incorporating a feedback loop mechanism, the system can learn from its mistakes and adjust its classification accuracy accordingly. Feedback from incorrectly screened profiles helps refine the classification algorithms, enhancing the overall effectiveness of the solution.

In essence, the proposed automated approach offers a comprehensive solution to the challenges faced by HR professionals, streamlining the candidate selection process and ensuring a better match between job requirements and candidate profiles. Through the integration of advanced machine learning techniques and continuous improvement mechanisms, this method holds promise in revolutionizing HR practices and optimizing workforce management.

B. Recammandtion system 2

The article [2] suggests an automated approach for recommending the most suitable candidates based on a given job description, leveraging Natural Language Processing (NLP) to extract pertinent information from candidates' resumes. Subsequently, the proposed solution employs a vectorization model and cosine similarity to match each condensed resume with the job description, generating ranking scores to identify the best-fitting candidates for the specific job opening.

The system operates in two distinct phases:

Phase 1: Information Extraction In the initial phase, the system focuses on information extraction utilizing NLP techniques. Given that resume data typically lacks a structured format, the goal here is to extract relevant keywords and details without manual intervention. Techniques such as Tokenization, Stemming, POS Tagging, Chunking, and Named Entity Recognition are employed to derive crucial job-related information such as skills, experience, and education from uploaded resumes. This results in a summarized version of each resume presented in a JSON format, facilitating seamless processing in the subsequent phase of the resume screening system.

Phase 2: Content-Based Candidate Recommendation The second phase involves constructing a content-based recommendation system utilizing the entities extracted in Phase 1. Drawing on concepts such as Vectorization, TF-IDF for assigning importance or weight to terms, and cosine distance for measuring similarity, the system determines the relevance of resumes to the given job description. By employing these techniques, the system can effectively match the skills and qualifications outlined in the job description with those present in the candidate resumes.

To validate the efficacy of the system, researchers utilized a job description provided by Amazon.com Inc. for the position of a Software Developer Engineer in its Bengaluru office. They then collected relevant resume samples from the internet to test the system's performance against real-world data.

In essence, this automated approach streamlines the candidate selection process by accurately matching job requirements with candidate qualifications, offering a systematic and efficient solution for recruiters and HR professionals.

C. Recammandtion system 3

The paper [3] explores the implementation of an embedding-based recommendation system at scale for matching job opportunities with candidates, particularly within the context of CareerBuilder.

First Stage: Retrieval Component The first stage involves a retrieval component, which integrates a fused embedding strategy to capture representations from diverse sources of data. These sources include raw text, parsed text utilizing a Job-skill Information Graph, and geolocation data processed through a Spherical Coordinates Calculator, for both candidates and job listings.

Second Stage: Rerank Component Following the retrieval stage, the system employs a rerank component to further refine the recommendations. This involves considering various contextual features such as skill matching, location preferences, years of experience, and education level.

The article proposes a sophisticated two-stage recommendation system employing an embedding-based methodology:

Fused-Embedding Component for Candidate Retrieval: This component employs a fused-embedding model, combining vectors generated from a Deep Learning Embedding Model (DLEM), the skill-job information graph, and geolocation calculations. By synthesizing these diverse data sources, the system can effectively retrieve candidate profiles that align closely with the job requirements.

Fine-Tuning Reranking Module: In the second stage, a fine-tuning reranking module is employed. This module calculates a weighted linear equation, aggregating scores obtained from the first-stage relevancy assessments along with contextual features of both job listings and candidates. These context-based scores encompass factors such as skill matching, geographical constraints, years of experience, and education level. The weights assigned to each score are determined empirically, reflecting their relative importance in the recommendation process. The resulting final ranking score is then utilized to rerank candidate recommendations, generating refined and personalized recommendations in the second stage.

By adopting this two-stage recommendation procedure, CareerBuilder aims to enhance the efficiency and accuracy of job-to-candidate matching, facilitating better alignment between job opportunities and candidate qualifications. This advanced embedding-based approach allows for a comprehensive analysis of diverse data sources, resulting in more tailored and relevant recommendations for both employers and job seekers.

D. Recammandtion system 4

In this paper [4], researchers address the challenge of matching resumes to job positions and propose a novel solution leveraging unsupervised feature extraction, advanced machine learning techniques, and ensemble methods. Our approach is entirely data-driven, enabling the detection of similar positions without the need for additional semantic tools. Moreover, it's designed with modularity in mind, allowing for rapid execution on GPU or CPU simultaneously. Compared to traditional rule-based methods, our solution demonstrates superior performance in terms of precision and Top-N recall. Additionally, the codebase is now publicly accessible on Github for easy implementation.

The solution comprises three customizable modules: unsupervised feature extraction, base classifier training, and ensemble method learning. Rather than relying on manual rules, our approach employs machine learning algorithms to automatically detect semantic similarities between positions. Following this, four competitive "shallow" estimators and "deep" estimators are chosen. Finally, ensemble methods are applied to combine these estimators' predictions and generate a final prediction.

This machine-learned solution represents a significant advancement in resume-job matching, offering greater flexibility, efficiency, and accuracy. By leveraging state-of-the-art techniques and eschewing manual rule creation, our approach streamlines the process of identifying position similarities and ensures more reliable candidate recommendations.

III. A SHORT SYNTHESIS

Each method has advantages and disadvantages, depending on its complexity, accuracy and ability to adapt to the specific needs of companies and recruiters. Future research could focus on combining these approaches to create even more powerful and versatile recommendation systems.

The table I summarizes some characteristics of the previous algorithm: database used,technical environment, AI , Methods and Metrics.

IV. CONCLUSION

In this paper, an attempt has been made To compare the wide range of methods dedicate to automatic CV Recommendation, that has appeared over the last decade. several characteristics of the four methods have been presented, as techniques used, the evaluation of the performances. This short description, and comparison must be extended. More detail in presentation of the algorithms, and the comparison in other context can be subject of future works.

REFERENCES

[1] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia,"A Machine Learning approach for automation of Resume Recommendation system", Procedia Computer Science, Volume 167, 2020, Pages 2318-2327.
https://doi.org/10.1016/j.procs.2020.03.284.

[2] Daryani, Chirag and Chhabra, Gurneet Singh and Patel, Harsh and Chhabra, Indrajeet Kaur and Patel, Ruchi,"An automated resume screening system using natural language processing and similarity", ETHICS AND INFORMATION TECHNOLOGY [Internet]. VOLKSON PRESS, 2020 pages 99–103.
http://doi.org/10.26480/etit.02.2020.99.103

[3] Jing Zhao and Jingya Wang and Madhav Sigdel and Bopeng Zhang and Phuong Hoang and Mengshu Liu and Mohammed Korayem, " Embedding-based Recommender System for Job to Candidate Matching on Scale," 2107.00221, arXiv , cs.IR, 2021.

[4] Lin, Y., Lei, H., Addo, P. C., Li, X.," Machine learned resume-job matching solution", arXiv preprint arXiv:1607.07657, 2016.

[5] Al-Otaibi, S.T., Ykhlef, M. "A survey of job recommender systems". International Journal of the Physical Sciences 7(29), 5127–5142, 2012.

[6] Pologeorgis, N. Employability, the Labor Force, and the Economy. 2019. https://www.investopedia.com.

[7] LinkedIn Workforce Report Janurary 2021 United States.

[8] Columbus, L. Remote Recruiting In A Post COVID-19 World. 2020. https://www.forbes.com.

[9] Yuan, J., Shalaby, W., Korayem, M., Lin, D., Aljadda, K. Luo, J. 2016. Solving coldstart problem in large-scale recommendation engines: A deep learning approach. 2016 IEEE International Conference on Big Data (Big Data).

[10] Wang, J., Abdelfatah, K., Korayem, M. Balaji, J. 2019. DeepCarotene -Job Title Classification with Multi-stream Convolutional Neural Network. 2019 IEEE International Conference on Big Data (Big Data).

TABLE I
SUMMARY OF SOME CHARACTERISTICS

Recammandtion system	Database used	Technical environments	AI	Methods	Metrics analysis
[1]	The data was downloaded from the online portal(s) and from Kaggle.	Open source library called “gensim” Data in Excel format NLTK library	Linear Support Vector Classifier (Linear SVM) : A SVM is a supervised machine learning classifier which defined by a separating hyperplane.	Content Based Recommendation using Cosine Similarity k-Nearest Neighbours	The proposed approach effectively captures the resume insights, their semantics and yielded an accuracy of 78.53% with LinearSVM classifier.”
[2]	For testing the system, researchers have used the job description posted by Amazon.com Inc. inviting applicants for the job position of a Software Developer Engineer at its Bengaluru office. Then they have taken some relevant resume samples from the Internet.	JSON format	NLP	Tokenization, Stemming, POS Tagging, Named Entity Recognition	”TF-IDF stands for “Term Frequency – Inverse Document Frequency” (Stecanella, 2020). Cosine similarity (Sidorov, Grigori, et al., 2014) is a measure to find how similar the two documents are regardless of their size.”
[3]	CareerBuilder.com database: on the daily routine, millions of job postings and more than 60 million actively searchable resumes.	Jobs and candidates are stored in HADOOP clusters allowing distributed processing using Spark. All Spark jobs are scheduled by the Oozie coordinator running periodically.	NLP DLEM CNN	Word2vec	Offline evaluation: QA teams and professional recruiters gave quality score and nDCG between baseline model and two-stage matching model. Online evaluation: Click Through Rate (CTR) and nDCG.
[4]	”70k resumes 32 most frequent positions”		NLP Deep learning	Unsupervised feature extraction base classifiers training Four competitive ”shallow” estimators “Deep” estimators	Similarity

Improving Sensor Network Monitoring with Machine Learning

1st Lmkaiti Mansour
LIMATI Laboratory
lamkaitimansour@gmail.com
University Sultan Moulay Slimane
Polydisciplinary Faculty

2nd Moudni Houda
TIAD Laboratory
h.moudni@usms.ma
University Sultan Moulay Slimane
Faculty of Sciences and Technology

3rd Mouncif Hicham
LIMATI Laboratory
h.mouncif@gmail.com
University Sultan Moulay Slimane
Polydisciplinary Faculty

***Abstract*—In the realm of Wireless Sensor Networks (WSNs) security, there's a growing need for a proactive approach, much like how previous research explored innovative techniques for predicting academic success alongside traditional statistical methods. The increasing sophistication of attacks, particularly Denial-of-Service (DOS) attacks, poses a significant threat to WSNs, challenging traditional intrusion detection systems (IDSs) to keep pace.**

To address this challenge, we delve into a comprehensive analysis of relevant literature, shedding light on the critical importance of robust intrusion detection systems in WSNs, especially those operating with limited resources. Drawing parallels with the evolution of academic performance prediction models, our research introduces a novel approach leveraging Machine Learning (ML) techniques explicitly designed to identify specific types of attacks within WSNs.

This ML algorithm, trained on the IDSAI dataset, represents a promising advancement in the realm of WSN security, offering a ray of hope amidst the ever-growing threats faced by wireless communication networks. By combining insights from academic success prediction methodologies with cutting-edge ML techniques, our work aims to bolster the defenses of WSNs against increasingly sophisticated cyber threats.

Keywords :,Wireless sensor networks (WSNs) , Intrusion detection system (IDS) ,Denial of service (DOS), Machine learning (ML).

I Introduction

TIn recent years, the utilization of wireless sensor networks (WSNs) has experienced a remarkable surge, becoming integral in environmental and physical condition monitoring across diverse industrial and research domains[20]. Renowned for their simplicity, efficiency, affordability, and ease of implementation relative to alternative sensing technologies, WSNs have found widespread adoption in telecommunications, healthcare, military operations, and environmental research. Typically configured as an array of geographically dispersed sensor nodes collaborating to collect and monitor environmental and physical data, these nodes establish wireless communication within the network and transmit data to a central node—often referred to as the Base Station (BS) or Sink Node—for storage and processing[1][2].

While the application of WSNs in remote and challenging environments for detecting environmental anomalies like floods, storms, and wildfires, as well as seismic events and volcanic activity, extends to less hazardous contexts such as health monitoring, smart infrastructure development, transportation, and the Internet of Things (IoT)[22], the inherent simplicity of WSN design brings inherent limitations. Primarily, the constrained resources of WSNs—including battery power, memory, storage, communication bandwidth, and computational capability—render them highly susceptible to security breaches, particularly in unattended areas.

Among the various security threats faced by WSNs, Denial of Service (DoS) attacks loom large[24], aiming to exhaust node resources, especially power reserves, and disrupt regular operations. To safeguard WSNs against such threats, diverse defense mechanisms are required, with studies advocating the utilization of Intrusion Detection Systems (IDSs) leveraging Machine Learning (ML) and Deep Learning techniques to achieve impressive accuracy in attack detection[7].

This paper delves into ML methods tailored to identify and classify different types of assaults in WSNs, with a primary objective of developing a flexible, accurate, and low-power algorithm capable of recognizing common attacks in WSNs. Through a comprehensive analysis of various attack scenarios, encompassing brute force SSH attempts, TCP null attacks, IP fragmentation attacks, accelerated SYN floods, SYN/ACK and RST floods, ARP spoofing, UDP port scans, DDoS MAC floods, and ICMP echo request floods, and drawing upon previous research findings, we propose our own approach for ML-driven intrusion detection systems in WSNs. Central to our methodology is the training of four distinct machine learning algorithms on the IDSAI dataset, aiming to enhance the accuracy and efficiency of attack detection mechanisms[24].

II Related works

Wireless sensor networks (WSNs) face significant security challenges and vulnerabilities during the transmission of data packets among network nodes[2]. Given the sheer number of sensor nodes within WSNs, they remain highly

susceptible to a multitude of threats and attacks. Previous research endeavors have sought to address these concerns through the adoption of abuse and anomaly detection techniques.

For instance, a prior investigation amalgamated two methodologies to introduce a tailored anomaly detection framework designed specifically for heterogeneous WSNs. This framework amalgamates a long-term approach, analyzing data from heterogeneous sensors network-wide, with a short-term strategy that scrutinizes individual node data locally. This integrated approach demonstrated notable efficacy, yielding enhanced outcomes by circumventing the limitations of individual methods[3].

Although diverse approaches have been explored in preceding studies, my research solely concentrates on the utilization of the Random Forest (RF) algorithm for intrusion detection in WSNs. This decision was influenced by the observed effectiveness of RF in prior research endeavors[7].

Moreover, while alternative methods such as Support Vector Machine (SVM) and others have exhibited promise in intrusion detection, my study is centered on Random Forest due to its appropriateness and performance within the realm of WSN security. By narrowing the focus to a single algorithm, this research aims to furnish a comprehensive analysis and assessment of Random Forest's efficacy in detecting intrusions within WSNs.

III MACHINE LEARNING BASED ANOMALY DETECTION IN WSN

In this paper, our central focus is on leveraging machine learning algorithms to detect anomalies in Wireless Sensor Networks (WSNs). Anomaly detection is of paramount importance as it enables the identification and mitigation of unexpected or potentially malicious activities occurring within the network. By effectively recognizing anomalies, we can uphold the integrity and reliability of the sensor data collected from the network. This task is particularly critical in WSNs due to their distributed nature and the vast amount of data exchanged between sensor nodes. Anomalies may signify various issues such as sensor malfunctions, environmental changes, or even security breaches. Traditional methods of anomaly detection in WSNs often face challenges in accurately distinguishing between normal and abnormal behavior, especially in dynamic and unpredictable environments. Machine learning algorithms offer a promising approach to address these challenges by learning from historical data patterns and automatically identifying deviations from expected behavior. Through our research, we aim to explore the efficacy of different machine learning techniques in detecting anomalies in WSNs, ultimately contributing to the enhancement of network security and data reliability[7].

1. Random Forest Classifier

The random forest classifier is a collaborative algorithm that operates based on similarity queries. It is a classification algorithm built upon decision trees and utilizes a conventional strategy of divide and conquer to improve efficiency. The key idea behind random forest is to create a powerful ensemble of weak learners, forming a strong learner. This approach aligns with the hypothesis of disjunction.[8]

2. Decision tree classifier

The decision tree classifier is capable of handling both categorical and continuous dependent variables. This algorithm divides the data into multiple subsets of the same type, aiming to create distinct groups based on the most important independent variables[23].

3. Extra Trees classifier

The Extra Trees classifier is a variant of decision tree-based classification that utilizes a completely random approach. In this algorithm, the entire sample is used to construct additional trees, making the decision boundaries random. By employing bootstrap copies of the training sample, it identifies the optimal cutoff point for each random feature at a node. The Extra Trees algorithm reduces the computational burden compared to standard decision trees and forests in determining the optimal cutoffs[1].

4. Gradient boosting

Gradient boosting is a powerful machine learning algorithm that leverages weak predictive models, such as decision trees, to construct a robust predictor for both regression and classification tasks.[1][2] It iteratively builds the model by focusing on the errors made by the previous models and adjusting subsequent models to minimize those errors. At the final stage, if no discernible patterns can be captured, the algorithm can be halted to prevent the potential issue of overfitting.[3][4]

5. XGB classifier

XGBoost stands out for its ability to efficiently handle large datasets while providing good predictive performance. It uses a gradient boosting method to gradually improve the model's accuracy by adjusting the weights of the training examples and minimizing a loss function[7].

IV DISSCUSSION ON COMPARISON BETWEEN TYPES OF SECURITY

Physical security and crypthographgue are two different approaches to protect data in sandf sensor networks they . the security phusqiue consists in protecting the devices, infrastructure and physical environments in which the sensor networks are deployed,while crypthography uses encryption techniques to protect data when stored, transmitted or processed .[11][12]

Physical security is important because it protects wireless sensor networks from physical attacks such as theft, sabotage and destruction. physical security measures may include the use of waterproof housing to protect sensors

from the elements,the installation of surveillance cameras to alert in case of intrusion encryption, meanwhile[26], uses encryption techniques to protect data against electronic attacks such as piracy,interception and falsification.encryption techniques may include the use of encryption keys to encrypt data,the use of security protocols to ensure the integrity and authenticity of data,and the use of dentification mechanisms to ensure that only authorized users have access to the data.

V DATASET FOR INTRUSION DETECTION SYSTEM

The Bot-IoT dataset contains a significant amount of redundant data in its input detection information, which can potentially result in unfavorable outcomes[7]. To address this issue, we conducted experiments using controlled machine learning algorithms on the IDSAI dataset. The IDSAI dataset is a modified version of the Bot-IoT dataset specifically designed for wireless sensor networks[8][15].

1. ICMP echo request flood

This attack involves sending a large number of ICMP echo (ping) requests to a target, aiming to flood the network and cause degradation in performance[15].

2. SYN/ACK flooding

SYN/ACK flooding attack aims to exhaust the resources of a target system by sending a massive amount of unsolicited SYN/ACK packets, forcing the system to allocate resources to handle these unestablished connections[14].

3. SYN flooding faster

Faster SYN flooding is a variant of the traditional SYN flooding attack, where the attacker rapidly sends a sequence of SYN packets to deplete the resources of the target system[27].

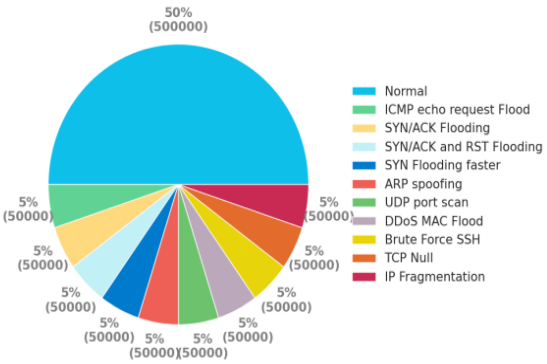


Fig. 1: DataSet

4. Description of the Dataset characteristics

Copy code

delta_time	Time elapsed between sending two consecutive packets, allowing analysis of the time between communications.
Protocols	Communication protocols used between wireless sensors (e.g., TCP, UDP, ICMP) to identify communication protocols.
ip_src	Source IP address of the packet emitted by the wireless sensor, aiding in identifying the origin of communication.
port_src	Source port of the packet emitted by the wireless sensor, aiding in understanding used services and applications.

TABLE I: Packet Attribute Descriptions

5. Methodology

In our investigation, we employed decision tree classification and gradient boosting algorithms to analyze data sourced from wireless sensor networks. Initially, we gathered sensor data in a typical test environment, encompassing variables like temperature, humidity, and noise level. Subsequently, we preprocessed the data by removing outliers, standardizing it, and then splitting it into training and test sets[25].

We implemented the decision tree classification and gradient boosting algorithms using the Python scikit-learn library. These models were trained on the prepared dataset, and we fine-tuned the hyperparameters to enhance the prediction accuracy on the test set. To evaluate the model performance, we employed metrics such as accuracy, recall, and F1 score[25].

6. Results the Algorithms of DataSet

labls binary :

Metrics	**RandomForest**
Time training	238.1683
Time prediction	1.8224
Accuracy score	0.9498
F1 score	0.9497
Recall score	0.9498
Precision score	0.9527
MSE	0.0502
Roc_Auc	0.0502
CK	0.8995
Time, CV	0.9496
CV	0.0005

TABLE II: Results of DataSet (labels_binary)

After these results the right algorithm:
Our comprehensive analysis led to the discovery that among the machine learning algorithms evaluated, Random Forest stood out as the most accurate, boasting an impressive precision score of 0.97. This signifies its

exceptional ability to correctly classify instances, which is crucial for detecting anomalies within wireless sensor networks (WSNs). With such a high precision score, Random Forest demonstrates its capacity to effectively differentiate between normal and anomalous behavior in the network, thereby ensuring the integrity and reliability of the sensor data collected. This finding underscores the significance of Random Forest as a robust and reliable tool for anomaly detection in WSNs, offering valuable insights for enhancing the security and performance of these networks.[1]

the carachtertics used by RANDOMFORESTCLASSIFIER

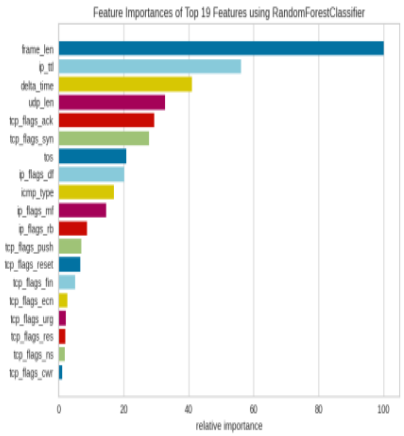


Fig. 2: RandomForest

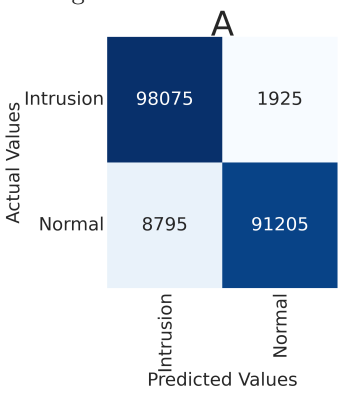


Fig. 3: Results

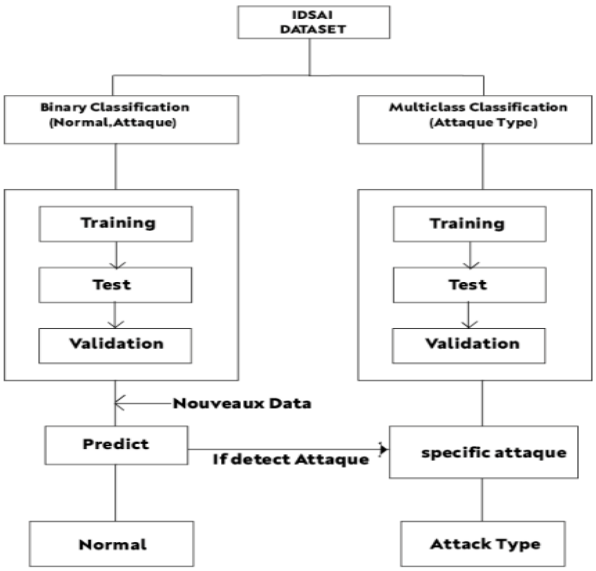
.

VI Disscussion on the test model and the difference between binary and multiclass classification

The selection of binary or multiclass classification depends on the type of data and the objective of the classification task. Binary classification is typically used when the task involves predicting the class of an observation that belongs to one of two possible classes, while multiclass classification is used when the task involves predicting the class of an observation that belongs to one of several possible classes[13].

In binary classification, the model is trained to predict a binary class, usually represented by 0 or 1, for each observation. The model output is therefore a binary value that indicates the predicted class. In contrast, in multiclass classification[7], the model predicts the probability of each possible class for each observation[7].

To obtain accurate and dependable results, it is essential to choose the appropriate classification algorithm based on the data and the objective of the classification task. The selection of binary or multiclass classification can also impact the performance metrics of the model, such as precision, recall, F1-score, among others[7][1].



VII Disscussion of the results

In this investigation, we assessed the efficacy of various classification algorithms, among them Random Forest, utilizing a dataset derived from wireless sensor networks. Our analysis revealed that amalgamating the strengths of diverse classification algorithms yielded more precise and dependable outcomes[7].

Our research findings underscored the superiority of the Random Forest algorithm, boasting an impressive precision score of 0.97.

Overall, our study underscores the significance of employing multiple classification algorithms and prioritizing machine learning security to attain accurate and trustworthy outcomes within a wireless sensor network production environment. While classification algorithms excel in recognizing patterns in sensor data and predicting future events, emphasizing machine learning security is essential to safeguarding the confidentiality and integrity of both data and results[5][13].

For processing our dataset and implementing machine learning algorithms, we leveraged an integrated develop-

ment environment (IDE) such as Jupyter Notebook for coding and execution in Python.

Employing the Python scikit-learn library, we deployed a range of machine learning algorithms, including logistic regression, k-nearest neighbor classification, decision tree classification, and gradient boosting. Segregating our dataset into training and test sets, we utilized cross-validation techniques to assess the performance of each algorithm.

Through the utilization of this hardware setup, we effectively applied machine learning algorithms to our dataset, facilitating data analysis and the development of precise predictive models. These outcomes indicate that leveraging quality hardware and integrated development environments can significantly enhance the efficiency of data analysis and machine learning model development[13].

VIII Conclusion

Wireless sensor networks (WSNs) serve as distributed sensor systems capable of gathering and monitoring data across diverse environments like buildings, factories, and urban areas. This data holds significant value for various applications ranging from health monitoring to environmental surveillance[1][2]. However, the inherent challenges posed by the distributed nature of sensors, coupled with constraints in bandwidth and energy, present obstacles to effective data collection in these environments. Consequently, there has been a surge in the adoption of machine learning algorithms for sensor data classification, offering the capability to discern patterns in sensor data and forecast future events.

In our research, we assessed the efficacy of multiple classification algorithms, including Random Forest, using a dataset derived from wireless sensor networks. Our findings underscored the advantages of leveraging a combination of these algorithms, resulting in enhanced accuracy and reliability of outcomes[7].

To sum up, in order to get accurate and consistent findings inside a wireless sensor network operating framework, our study highlights the significance of utilizing a repertoire of classification algorithms and giving machine learning security measures top priority[18][17]. While classification algorithms can be useful for identifying trends and forecasting future occurrences from sensor data, machine learning security issues must be addressed to protect the privacy and accuracy of results as well as data[6].

References

[1] T. kyildiz, I. F., Su, W., Sankarasubramaniam, Y. and Cayirci, E. *"Unsupervised machine learning based key management in wireless sensor networks "* Computer Networks 38: 393-422, 2023.

[2] • J.J. Garcia-Luna-Aceves a , Dylan Cirimelli-Low b, *" ALOHA-NUI: A collision-free version of ALOHA using a Neighborhood-Understood Index."*1
, 2023.

[3] •Niande Jiang.,Weihui Zhou , *"Research on hybrid of ALOHA and multi-fork tree Anti-collision algorithm for RFID. "*. IEEE, 2018,

[4] • Zheng, Y., et al, *"Physical layer network coding with imperfect channel state information"* . IEEE, 2019,

[5] • Wang, L., et al, *" Enhanced Physical Layer Network Coding in Wireless Sensor Networks."*1
, 2020.

[6] Mayssa Ghribi, Aref Meddeb *" Survey and taxonomy of MAC, routing and cross layer protocols using wake-up radio"* , 2020.

[7] Abdullah Balcı and Radosveta Sokullu, *" Massive connectivity with machine learning for the Internet of Things "*, 2021.

[8] Beneyaz Ara Begum and Satyanarayana V. Nandury,*" Data aggregation protocols for WSN and IoT applications – A comprehensive survey"*, 2023. 6

[9] Arun George, T.G. Venkatesh, *" Multi-packet reception dynamic frame-slotted ALOHA for IoT: Design and analysis,"*,g, 2020. 6

[10] Beneyaz Ara Begum , Satyanarayana V. Nandury, *" Data aggregation protocols for WSN and IoT applications – A comprehensive survey,"* ISIT, 2023. 6

[11] Shafqat Ullah , Mazhar Hussain Malik , Mehmet Fatih Tuysuz , Muhammad Hasnain , Mehmet Emin Aydin *"Max-gain relay selection scheme for wireless networks "* , 2021. 6

[12] Kimberly Jane Co , Arlyn Verina Onga,b, Marnel Peradilla, *"WSN Data Collection and Routing Protocol with Time Synchronization in Low-cost IoT Environment, "* , 2021. 7

[13] Molly Zhang, Luca de Alfaro , J.J. Garcia-Luna-Aceves, *"Making slotted ALOHA efficient and fair using reinforcement learning, "* ASMA-SPSC, 2022. 7

[14] R. Ahlswede, N. Cai, S. Y. R. Li, and R. W. Yeung, *" A Transaction Model for Executions of Compositions of Internet of Things Services"* , 2016. 8

[15] S. yen Robert Li, S. Member, R. W. Yeung, and N. Cai, *" An Energy Efficient Secure routing Scheme using LEACH protocol in WSN for IoT networks,"* , 2023. 8

[16] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. E ros, J. Shi, and B. Leong, *" Efficient and secure multi-homed systems based on binary random linear network coding,"* vol. 52, no. 10, pp. 4413 4430, 2020 packing,. 8

[17] G. Santhosh, K.V. Prasad b, *" Energy optimization routing for hierarchical cluster based WSN using artificial bee colony"* , 2023,. 8

[18] •Jie Hu, Guangming Liang, Qin Yu, Kun Yang, Xiaofeng Lu *"Simultaneous wireless information and power transfer with fixed and adaptive modulation, "* , 2021. 8

[19] •Ansa Shermin S, Sarang C. Dhongdi, *"Review of Underwater Mobile Sensor Network for ocean phenomena monitoring, "*, 2021. 8

[20] •Pietro Tedeschi, Savio Sciancalepore, Roberto Di Pietro *"Satellite-based communications security: A survey of threats, solutions, and research challenges, "* , 2022. 9

[21] • Vikas Tyagi. Samayveer Singh, *" Network resource management mechanisms in SDN enabled WSNs: A comprehensive review"*1
, 2023.

[22] Adamu Murtala Zungeru, S Subashini, P Vetrivelan, *"Wireless Communication Networks and Internet of Things: A Survey. "*.,

2016,

[23] • Javier Hernandez Fernandez,Roberto Di Pietro,Roberto Di Pietro *"Physical layer network coding with imperfect channel state information"*., 2022,

[24] Moudni H., Er-rouidi M., Mouncif H., El Hadadi B., *" Fuzzy logic based intrusion detection system against black hole attack in mobile ad hoc networks"* 2018.

[25] Er-Rouidi M., Moudni H., Faouzi H., Mouncif H., Merbouha A. *"A fuzzy-based routing strategy to improve route stability in MANET based on AODV"*.2017, pp. 243 254. 9

[26] Lmkaiti Mansour, Mouncif Hicham , *"Comparative Analysis of Physical Layer Network Coding-Based Random Access Techniques in WSN Communications"*.2023, pp. 243 254. 9

Prediction of Student Attrition and Academic Achievement Using Machine Learning Algorithms

Raja Ouadad
Mathematics and Informatics Department
Sultan Moulay Slimane University
Beni Mellal, Morocco
ouadadraja2@gmail.com

Hicham Mouncif
Mathematics and Informatics Department
Sultan Moulay Slimane University
Beni Mellal, Morocco
h.mouncif@usms.ma

***Abstract*— Higher education institutions encounter a substantial obstacle in tackling the elevated rate of student attrition, a matter influenced by socio-economic circumstances. In response, this research employs intelligent machine learning models to predict student performance, focusing on factors like Demographic Analysis, Economic Factors, Academic Performance, Social and Special Needs, and Macro-economic Factors. The dataset, comprising 4,424 records, is used for constructing classification models, categorizing students into dropout, enrolled, and graduate statuses. The study involves a systematic exploration of the dataset through Exploratory Data Analysis (EDA), feature selection, and outlier removal. Six classification algorithms—K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, SVM, and Naïve Bayes—are trained and evaluated, with Random Forest emerging as the top performer. The results provide valuable insights for higher education institutions to implement targeted interventions for student retention.**

Keywords— Education, Student Performance, Quality of education, Exam performance, Classification, Machine Learning, Academic improvement

I. INTRODUCTION

Attaining success in higher education is of the utmost importance, as it serves as a catalyst for employment prospects and contributes significantly to the advancement of social justice and economic development. Achieving a standard of academic excellence in tertiary education is fundamental to the progression of one's professional life, fosters a more equitable community, and acts as a catalyst for the expansion of the economy. Tackling dropout rates presents a notable obstacle for higher education institutions aiming to improve their overall achievement. The absence of a universally acknowledged definition for dropout complicates the matter, with variations observed in different studies due to distinct definitions, data sources, and calculation methods [1].

Within the study literature, dropout is frequently examined by analyzing the time of when individuals discontinue their academic efforts. This analysis differentiates between early and late phases of departure.This variability in approach underscores the complexity of the issue and the importance of nuanced analysis for a comprehensive understanding of dropout patterns [2]. Comparing dropout rates across institutions is rendered challenging due to variations in reporting methodologies. The absence of standardized reporting practices creates a barrier to making direct comparisons, as institutions may employ different criteria, metrics, or data collection processes. This disparity underscores the importance of establishing uniform reporting standards to facilitate meaningful and accurate cross-institutional analyses of dropout rates [3]. According to a survey commissioned by the European Commission, a significant number of students choose to drop out of their higher education courses before finishing them [4]. In Denmark, a country often regarded as highly successful, only around 80% of students manage to successfully finish their studies. This stands in stark contrast to Italy, where the completion rate is a mere 46%. The paper highlights key factors that contribute to student dropout, identifying socio-economic situations as the fundamental underlying cause.

A comprehensive search was undertaken by Namoun et al. [5], which resulted in the discovery of 62 papers that were published in peer-reviewed journals between 2010 and 2020. All of these papers present intelligent models that have been specifically developed to forecast student performance. Moreover, an increasing body of research has emerged in recent years with a specific emphasis on the timely anticipation of student outcomes [6], [7]. However, in spite of the increasing scholarly attention and the considerable amount of data produced by academic institutions, there is still an urgent requirement to collect administrative data that is more comprehensive and enhanced, specifically concerning the factors contributing to student attrition and transfers [2].

This exposition provides an overview of a dataset obtained from an institution of higher education. The dataset was compiled from numerous disparate databases and centers on undergraduate students who are enrolled in a wide range of disciplines, including agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset comprises data that was accessible during the enrollment process of students. It includes academic trajectory information, demographic details, macroeconomic and socioeconomic indicators, and academic performance evaluations from the initial and second semesters. The dataset functions as the fundamental material upon which classification models dedicated to forecasting student attrition and scholastic achievement are built. The problem is conceptualized as a classification task involving three distinct

categories: withdrawal, enrolled, and graduate, all of which occur at the end of the standard course duration [8]. The dataset used in this article comprises 4,424 records, each corresponding to an individual student, and encompasses 35 attributes. This dataset serves as a valuable resource for benchmarking the performance of various algorithms designed to address similar problems. Furthermore, it provides an excellent training ground for individuals pursuing knowledge and expertise in the field of machine learning.

With the exception of this introductory segment, the descriptor is organized as follows. Section 2 provides an in-depth analysis of the dataset, presenting comprehensive details. The employed methodology is described in Section 3, which is followed by a succinct exploratory data analysis. The findings are disclosed in Section 4, whereas the conclusion is succinctly summarized in Section 5, along with the citations. This structure guarantees a methodical exposition of the dataset, research approach, results, and overarching deduction for the purpose of facilitating thorough comprehension and citation.

II. Data Description

The dataset comprises extensive demographic, socio-economic, and macroeconomic information that was recorded at the time of enrollment and throughout the initial and subsequent semesters. It pertains to a cohort of 4,424 students who were enrolled in 17 different academic disciplines from 2008 to 2019. Significantly, the dataset exhibits an absence of missing data, and its CSV file is precisely encoded. It comprises 35 attributes that have been meticulously classified into the following five categories: academic (enrollment/semesters), socio-economic, demographic, and macroeconomic [8].

Table 1 delineates each attribute incorporated in the dataset, systematically organized into classes encompassing demographic information, socioeconomic factors, macroeconomic indicators, academic data at enrollment, and academic metrics at the conclusion of the first and second semesters.

Table 1. Attributes Categorized by Attribute Class.

Class of Attribute	Attribute	Type
Demographic data	Marital status	Numeric/discrete
	Nationality	Numeric/binary
	Displaced	Numeric/binary
	Gender	Numeric/discrete
	Age at enrollment	Numeric/binary
	International	Numeric/binary
Socioeconomic data	Mother's qualification	Numeric/discrete
	Father's qualification	Numeric/discrete
	Mother's occupation	Numeric/discrete
	Father's occupation	Numeric/binary
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete
Target	Target	Categorical

The dataset is employed in constructing machine learning models designed to forecast academic performance and dropout rates, integral components of a Learning Analytics tool devised at the Polytechnic Institute of Portalegre. This tool provides relevant information to the tutoring team, giving an estimate of the probability of dropout and academic failure. Moreover, the dataset is a great asset for academics doing comparative studies on student academic performance and as a training tool in the field of machine learning. Materials and Methods

This segment outlines the fundamental procedures involved in forecasting student attrition and scholastic achievement. The process encompasses Exploratory Data Analysis (EDA), Feature Selection, Removing outliers and Machine Learning models. The overarching objective is to present a clear and concise plan for conducting performance analysis within an educational context. By doing so, we aim to furnish educators and course developers with valuable insights, facilitating improvements in the learning experience through informed decision-making based on predictive modeling and thorough evaluation methodologies.

A. Exploratory Data Analysis (EDA)

During our investigation of the Student Dropout dataset, we shall employ an exploratory data analysis (EDA) procedure. Consider it as our method of inquiry and gaining a deeper understanding of the data. We will conduct a thorough examination of the dataset utilizing a variety of tools and methods to identify noteworthy patterns and insights. EDA enables us to comprehend the underlying causes of student attrition and to formulate well-informed strategies to tackle this concern.

As shown in Figure 1, the quantity of graduating students surpasses that of dropouts.

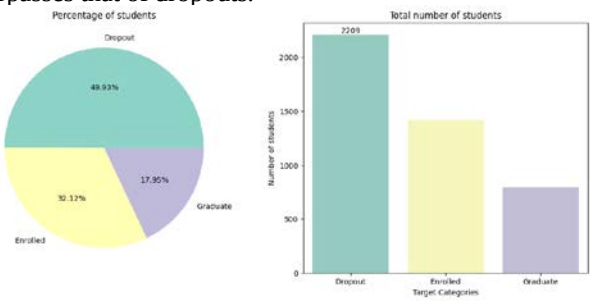


Figure 1. Total number of dropouts and graduate students.

The aggregate count of graduating and failure learners amounts to 3,630, representing the number of observations required for model construction. Consequently, we eliminated all rows categorized under the "Enrolled" class. Figure 2 illustrates the dataset after the elimination process.

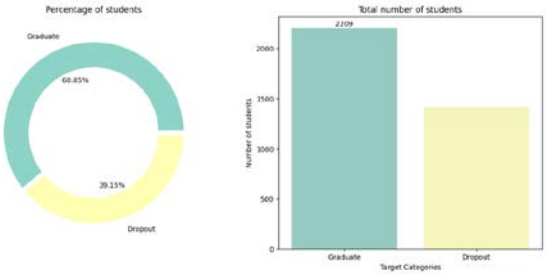


Figure 2. The dataset after the elimination process.

The gender distribution of students in our dataset is evident in Figure 2, where it is distinctly illustrated that 64.83% of the students are female, while the remaining 35.17% are male.

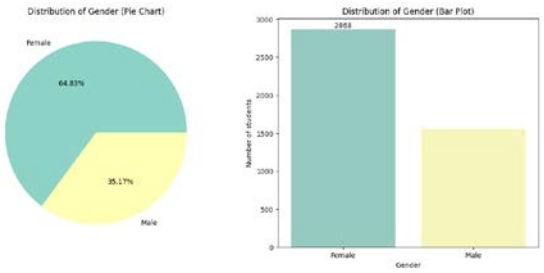


Figure 3. Gender distribution of students.

Figure 3 offers a visual depiction of the distribution of students based on gender and target category, providing a comprehensive snapshot of the gender distribution within enrolled, dropout, and graduating students. The figure highlights that most male students fall into the 'Graduate' category.

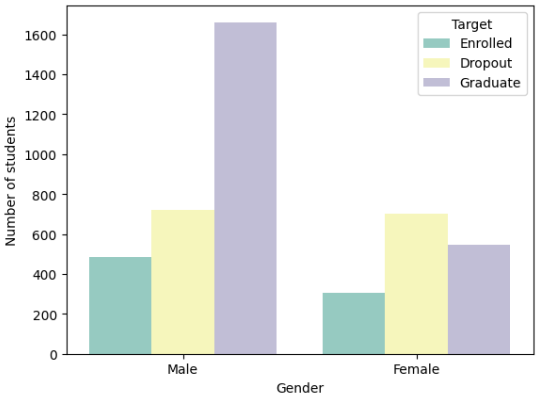


Figure 4. distribution of students based on gender and target category.

Figure 4 offers a valuable visual insight into the distribution of students across various marital statuses and their corresponding target categories. Notably, it indicates that single students exhibit a relatively high percentage of dropout within the dataset.

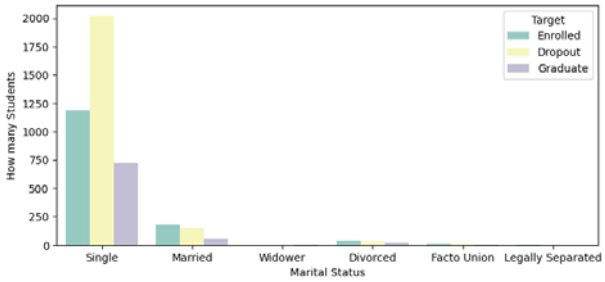


Figure 5. The distribution of students based on marital status and associated target categories.

B. *Features selection*

Feature selection, serving as a method for reducing dimensionality, seeks to select a condensed subset of pertinent features from the original set by eliminating irrelevant, redundant, or noisy elements. Typically, feature selection contributes to enhanced learning performance, encompassing higher accuracy, reduced computational expenses, and improved model interpretability[9]. The heatmap depicts the correlation strength and direction between every pair of features in a visual manner. A high positive correlation, closer to 1, suggests a strong relationship between two features. Conversely, a low correlation, closer to 0, indicates little to no linear relationship between the features[10].

A correlation heatmap serves as a visual depiction of the correlation between two or more variables, presented in a grid format where each cell signifies the correlation between respective pairs of variables. The heatmap visually represents the intensity of the correlation using color transitions. A strong positive correlation is denoted by red, while a strong negative correlation is represented by blue. This visualization method provides an intuitive and accessible representation of the relationships among variables.

Figure6 shows the relation of all features to themselves in our dataset.

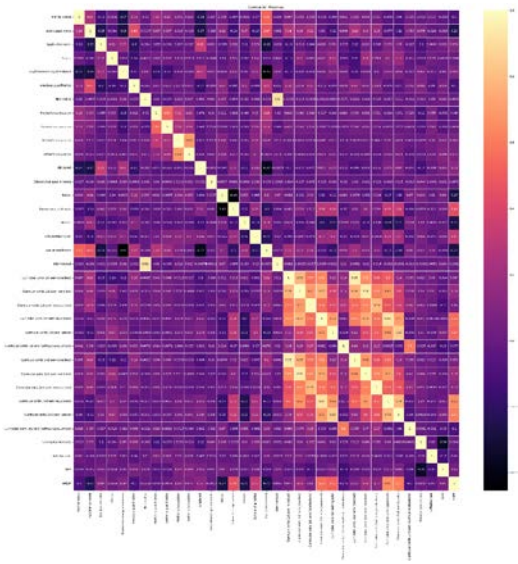


Figure 6. Correlation heatmap.

The correlation of all features with the target is shown in Figure 7.

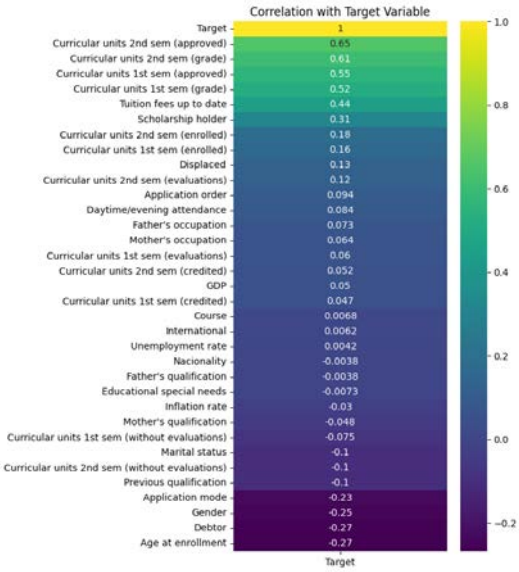


Figure 7. Correlation with target variable.

All features that demonstrate a negative correlation with the target are excluded. The correlation with a positive target is exclusively depicted in Figure 8.

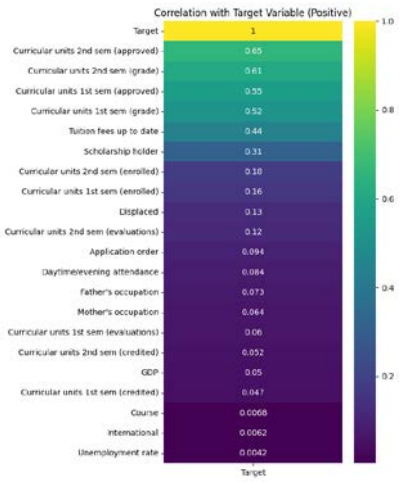


Figure 8. Correlation with positive target.

C. Eliminating outliers

The elimination of anomalies is an essential component of the data preprocessing workflow. Outliers, alternatively called anomalies, are observations that exhibit a substantial deviation from the mean distribution of the data. Such deviations possess the capacity to introduce distortions into statistical analyses. The presence of outliers can potentially undermine the validity of conclusions derived from the data and negatively impact the precision of predictive models.

The Interquartile Range (IQR) is a statistical metric utilized to evaluate the dispersion or variability of data. The calculation involves determining the difference between the first quartile (Q1) and the third quartile (Q3). IQR is a practical metric for identifying anomalies due to its robustness and insensitivity to outliers. The procedure for eliminating outliers utilizing IQR generally comprises the subsequent stages:

1. Determination of the IQR: IQR = Q3 - Q1

2. Establishing thresholds: Establish thresholds for outlier detection. Beyond these thresholds, values may be regarded as outliers.

3. Detection of outliers: Apprise observations that deviate from predetermined thresholds in terms of value. An outlier is defined as a value that deviates from the mean by less than Q1−k×IQR or exceeds Q3+k×IQR, with k being a modifiable parameter that governs the sensitivity of the detection.

4. Elimination of anomalies: Deleting or handling observations that are identified as outliers is contingent upon the method selected.

Figure 9 represents a diagram that displays the boxplots of the important attributes after removing outliers.

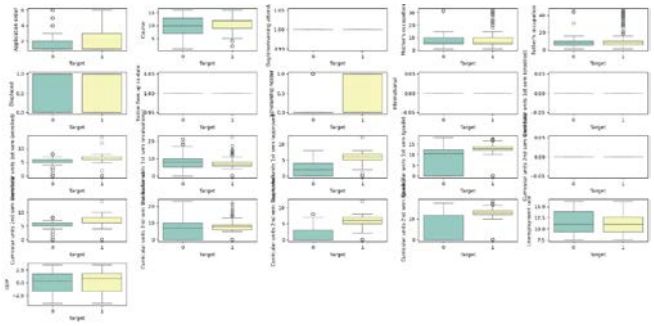


Figure 9. The boxplots display the essential attributes following the removal of outliers.

D. Models Training

During the Models Training phase, we initially partitioned the dataset into two segments, assigning 67% for training and 33% for testing reasons. Following this, we standardized the dataset utilizing the StandardScaler technique to ensure uniformity and optimal performance in subsequent modeling processes. In addition, we utilized GridSearch as the standard technique for hyperparameter optimization, which entails a thorough examination of a manually defined portion of the hyperparameter space in the learning process. The GridSearch technique utilizes a performance metric, usually assessed through cross-validation on the training set or evaluation on a separate validation set, to determine the best hyperparameters.

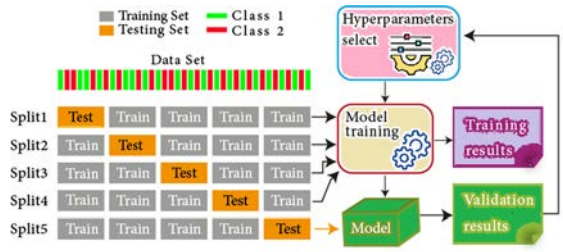


Figure 10. Model Training and Hyperparameter Optimization Process Overview.

We utilized six extensively acknowledged classification methods to determine the most advantageous remedy for our problem:

- ✓ **K-Nearest Neighbors (KNN)** is a straightforward and intuitive machine learning technique utilized for classification and regression applications. The algorithm allocates a new data point to the majority class of its k nearest neighbors in the feature space, where k is a value chosen by the user.
- ✓ **Logistic Regression** is a statistical technique employed in machine learning to solve binary classification problems. Although it may be misleading, its main purpose is to be used for categorization jobs. It quantifies the likelihood that a specific instance is classified into a specific category and generates predictions by utilizing a logistic function.
- ✓ **A Decision Tree** is a flexible machine learning technique that may be used for both classification and regression applications. The dataset is divided recursively depending on the features, resulting in a tree-like structure. Each internal node in the tree reflects a choice made using a feature, while each leaf node indicates the outcome. It is recognized for its capacity to be understood and its capability to process both numerical and categorical data.
- ✓ **Random Forest** is an ensemble learning system that utilizes numerous decision trees to enhance forecast accuracy and mitigate overfitting. The algorithm constructs a "forest" of trees by training each tree on a randomly selected portion of the data and characteristics. The ultimate forecast is typically a result of averaging or employing a voting mechanism on the predictions made by each individual tree. This approach ensures a resilient and precise model for both classification and regression tasks.
- ✓ **Support Vector Machine (SVM)** is a robust supervised machine learning technique utilized for both classification and regression tasks. The algorithm operates by identifying the most efficient hyperplane that can effectively segregate distinct classes within the feature space, while simultaneously maximizing the distance between them. Support Vector Machines (SVM) are highly efficient in dealing with datasets that have many dimensions. They are also able to handle both linear and non-linear relationships by utilizing kernel functions. Its widespread usage stems from its versatility and proficiency in managing intricate decision limits.
- ✓ **Naive Bayes** is a probabilistic machine learning technique that is frequently employed for classification tasks. It relies on Bayes' theorem and assumes that the features are independent of each other given the class label. Despite its oversimplified assumption, Naive Bayes frequently achieves strong performance in real-world scenarios, especially when applied to text categorization problems. The algorithm computes the likelihood of each category based on a given set of characteristics and assigns the category with the highest likelihood to the input data. Naive Bayes is a highly efficient and effective algorithm for analyzing high-dimensional data.

III. Result And Discussion

This section comprehensively presents the outcomes obtained through the execution of each of the six algorithms under consideration. A robust evaluation framework was employed, encompassing accuracy, precision, recall, and F1 score as the key metrics for assessing the performance of these algorithms. The utilization of multiple metrics provides a nuanced understanding of the algorithms' effectiveness, allowing for a thorough examination of their capabilities across

different facets of classification. This multifaceted evaluation approach contributes to a comprehensive and insightful analysis of the algorithmic performance, facilitating a well-informed discussion and interpretation of the results in the context of the scientific inquiry at hand.

Table 2 summarized the outcomes derived from the application of machine learning algorithms.

Table 2. Comparative performance of machine learning algorithms.

Algorithm	Accuracy	Precision	Recall	F1
Random Forest	0.909317	0.911716	0.909317	0.906205
SVM	0.906832	0.913660	0.906832	0.902301
Logistic Regression	0.900621	0.901714	0.900621	0.897460
Decision Tree	0.890683	0.897158	0.890683	0.884837
Naïve Bayes	0.884472	0.885503	0.884472	0.880113
K-Nearest Neighbors	0.880745	0.891486	0.880745	0.872636

The results table demonstrates that the SVM model outperforms other models in terms of accuracy, precision, recall, and F1-score metrics. Significantly, both Random Forest and Logistic Regression models also exhibit praiseworthy performance. On the other hand, Naive Bayes, K-Nearest Neighbors, and Decision Tree models demonstrate relatively inferior levels of performance. When considering all evaluation metrics together, Support Vector Machines (SVM), Random Forest, and Logistic Regression are identified as appropriate models for predicting dropout. These models can provide higher education institutions with useful information to develop focused interventions that improve student retention.

Figure 11 shows a bar chart illustrating the importance of each feature in the best performing Random Forest model (best_rf). The feature_importances_ attribute of the Random Forest model provides a measure of the contribution of each feature to the predictive performance of the model. The higher the value of a specific feature, the more influence it has on predictions. By examining this graph, we understand better that the characteristics "**2nd semester educational units (approved)**" and "**1st semester educational units (approved)**" belonging to the class "Academic data at the end of the 1st semester and 2nd semester" have the characteristics most significant impact on the model's decision-making process. They represent the factors that contribute most to the accuracy or predictive power of the model for the task of Predicting Student Dropout and Academic Success.

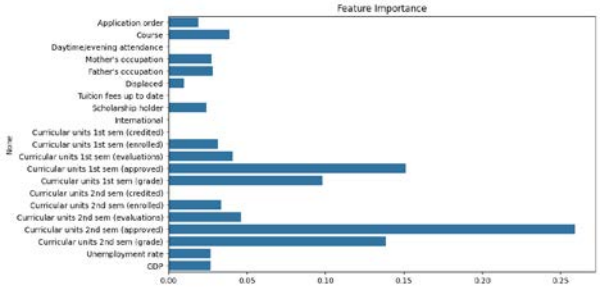


Figure 11. The importance of each feature in the best-performing Random Forest model.

IV. Conclusion

In summary, the convergence of data analysis and machine learning presents a transformative avenue for enhancing the management of educational institutions and fortifying student success. By harnessing these technological advancements, institutions can proactively tackle dropout concerns and offer personalized support, thereby making substantial contributions to the overall prosperity and well-being of students. Furthermore, addressing the formidable task of enhancing the education system necessitates innovative methodologies, as explored in this article. Leveraging machine learning technologies, our focus has been on predicting school dropout and academic success. Through the rigorous evaluation of various classification methods, including Random Forest, SVM, Logistic Regression, Decision Tree, Naïve Bayes, and K-Nearest Neighbors, using metrics such as Accuracy, Precision, Recall, and F1-score, Random Forest emerged as the top performer, achieving a remarkable accuracy rate of 90.93% compared to other algorithms. Vital to the success of our project were pivotal steps encompassing data processing, visualization, model implementation, and the comprehensive comparison of six algorithms.

References

[1] [1] A. Behr, M. Giese, H. Teguim, and K. Theune, "Motives for dropping out from higher education—An analysis of bachelor's degree students in Germany," European Journal of Education, vol. 56, Mar. 2021, doi: 10.1111/ejed.12433.

[2] [2] B. Kehm, M. Larsen, and H. Sommersel, "Student dropout from universities in Europe: A review of empirical literature," Hungarian Educational Research Journal, vol. 9, pp. 147–164, Sep. 2019, doi: 10.1556/063.9.2019.1.18.

[3] [3] "Comparison of course completion and student performance through online and traditional courses | The International Review of Research in Open and Distributed Learning." Accessed: Mar. 09, 2024. [Online]. Available: https://www.irrodl.org/index.php/irrodl/article/view/1461

[4] [4] "(5) (PDF) Drop out and Retention of Under-represented Students in Higher Education in Europe." Accessed: Mar. 09, 2024. [Online]. Available: https://www.researchgate.net/publication/304025193_Drop_out_and_Retention_of_Under-represented_Students_in_Higher_Education_in_Europe

[5] [5] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," Applied Sciences, vol. 11, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/app11010237.

[6] [6] "Saa, A.A.; Al-Emran, M.; Shaalan, K. Mining Student Information System Records to Predict Students' Academic Performance. Adv. Intell. Syst. Comput. 2020, 921, 229–239. [CrossRef] - Recherche Google." Accessed: Mar. 09, 2024. [Online]. Available: https://www.researchgate.net/publication/331821565_Mining_Student_Information_System_Records_to_Predict_Students'_Academic_Performance

[7] [7] "Martins, M.V.; Tolledo, D.; Machado, J.; Baptista, L.M.T.; Realinho, V. Early Prediction of Student's Performance in Higher Education: A Case Study. Adv. Intell. Syst. Comput. 2021, 1365, 166–175. [CrossRef] - Recherche Google." Accessed: Mar. 09, 2024. [Online]. Available: https://www.mdpi.com/2306-5729/7/11/146

[8] [8] “Data | Free Full-Text | Predicting Student Dropout and Academic Success.” Accessed: Mar. 09, 2024. [Online]. Available: https://www.mdpi.com/2306-5729/7/11/146

[9] [9] J. Miao and L. Niu, “A Survey on Feature Selection,” Procedia Computer Science, vol. 91, pp. 919–926, Jan. 2016, doi: 10.1016/j.procs.2016.07.111.

[10] [10] “Mastering the Art of Feature Selection: Python Techniques for Visualizing Feature Importance | by daython3 | Medium.” Accessed: Mar. 10, 2024. [Online]. Available: https://medium.com/@daython3/mastering-the-art-of-feature-selection-python-techniques-for-visualizing-feature-importance-cacf406e6b7

Recursive Least-Squares in Feature Space for System Identification

Rachid Fateh[1], Benoit Plancoulaine[1], Mathieu Pouliquen[2], Miloud Frikel[2], Hicham Oualla[3], Said Safi[4], Anouar Darif[4] and Said Hakimi[4]

***Abstract*—Since the 1950s, the field of system identification, deeply rooted in probability theory, has witnessed the emergence of a significant set of concepts, propositions, processes, and experiments. This article proposes a comprehensive comparison of tree Gaussian kernel algorithms in the context of non-linear systems identification, evaluating their efficiency, precision, and adaptability to complex environments. The analysis also focuses on their ability to successfully identify system outputs and calculate the mean squared error. The results highlight situations where each algorithm excels, providing practical guidance for choosing based on the specific characteristics of the non-linear system. This study contributes to a better understanding of kernel algorithm performance in real-world applications, incorporating the assessment of mean squared error as an additional measure.**

***Index Terms*—System identification, Non-linear systems, System outputs, Feature Space, Mean squared error, Kernel.**

I. Introduction

System identification holds paramount importance in the fields of engineering, science, and modeling, focusing on the determination and characterization of mathematical models describing the dynamic behavior of complex systems found in various sectors such as electrical engineering, biology, economics, and other related domains. This process often involves the acquisition of experimental data, the application of advanced statistical and mathematical methods, and the creation of accurate models that adequately reflect the system's behavior [1], [2]. This discipline plays a central role in the design of control systems, predicting future performances, and solving complex problems [3], [4].

Its decisive influence in understanding and manipulating dynamic systems contributes to significant advancements in various scientific and technological fields. The study of nonlinear systems, in contrast to linear systems, is justified by the observation that many real and complex phenomena do not adhere to linearity. Nonlinear systems provide a more accurate modeling of dynamic behaviors, characterized by complex interactions among different system components. While linear systems are useful for local approximations and simplified analyses, many fields such as biology, economics, physics, and other applied sciences reveal systems with inherently nonlinear behavior. The identification of nonlinear systems becomes crucial for capturing phenomena like thresholds, bifurcations, and nonlinear interactions [5], highlighting the need for more sophisticated and adapted models to precisely understand and anticipate the behavior of real systems with often nonlinear and unpredictable dynamics [6].

The applications of nonlinear systems find significant expression through the use of kernel methods in various domains:

1) In machine learning, kernel support vector machines enable the modeling of complex nonlinear relationships, enhancing the capacity for classification and regression in contexts where linearity is inadequate [7].
2) In the field of time series analysis, kernel methods play a crucial role in modeling and predicting nonlinear behaviors over time [8].
3) Applications extend to natural language processing, where these methods are employed to capture complex and nonlinear semantic relationships among language elements [9].
4) In control systems engineering, these methods are used to model complex dynamic systems, providing a robust approach in designing control strategies [10].

Kernel methods, intimately linked to reproducing kernel Hilbert spaces (RKHS), constitute a fundamental approach in machine learning [11]. RKHS provides a mathematical framework where kernel functions measure the similarity between data, facilitating the modeling of nonlinear relationships. Kernel methods leverage these properties to effectively address complex problems by projecting data into higher-dimensional feature spaces.

Currently, the literature on adaptive kernel filtering algorithms encompasses various techniques. Among them are kernel affine projection algorithms (KAPA) [12], kernel principal component analysis (KPCA) [13], kernel least mean squares (KLMS) [14], and kernel recursive least square (KRLS) [15]. To enhance the robustness of these adaptive kernel filtering algorithms, numerous variants have been introduced, including quantized kernel recursive least squares (QKRLS) [16], quantized kernel least mean square (QKLMS) [17], extended kernel recursive least squares (Ex-KRLS) [18], kernel least mean square with adaptive kernel size (KLMS-AKS) [19], random Fourier feature kernel recursive least squares (RFF-KRLS) [20], quantized kernel least lncosh (QKLL) [21], and kernel extended improved proportionate normalized least mean square algorithm (KE-IPNLMS) [22]. This paper specifically tackles the problem of nonlinear system identification using kernel-based recursive least-squares RLS. Various simulation results are presented, considering both noisy environments and diverse data lengths N, to showcase the accuracy and effectiveness of each method.

[1]*Normandie Univ, UNICAEN, Inserm U1086 ANTICIPE, Federative Structure 4207 'Normandie Oncologie', F. Baclesse Comprehensive Cancer Centre*, Caen, France rachid.fateh@unicaen.fr
[2]*Normandie Univ, LIS Laboratory - UR 7478, UNICAEN, ENSICAEN*, Caen, France
[3]*Akkodis*, Paris, France
[4]*LIMATI Laboratory, Sultan Moulay Slimane University*, Po. Box 592, 23000 Beni Mellal, Morocco

The paper is structured as follows: Section II delves into the background and formulation of non-linear systems. Moving on to Section III, a detailed account of kernel-based identification algorithms is provided. Section IV presents the experimental findings, and finally, Section V draws conclusions based on the results presented.

A. Contributions

The main contributions of this paper are summarized as follows:

- **Application of RLS in feature space for nonlinear system identification:** we employ the RLS framework in the feature space to estimate the output of nonlinear systems, showcasing the efficacy of the feature space approach in improving the accuracy of nonlinear system identification.
- **Comprehensive comparison of RLS algorithms in reproducing kernel hilbert space:** we conduct a thorough comparison of various algorithms based on Recursive Least-Squares in Reproducing Kernel Hilbert Space. The assessment focuses on calculating the Mean Squared Error (MSE), offering insights into the relative performance and robustness of these algorithms for system identification tasks

II. Proplem statement

In this section, we introduce some notations and assumptions that will be used throughout the paper. The Hammerstein system, a distinctive nonlinear model, is frequently employed in the realm of system identification. System identification aims to create precise mathematical models that mirror the behavior of real-world systems using observed input-output data. In this context, we focus on the Hammerstein system depicted in Figure 1. This system comprises a nonlinear static function followed by a finite impulse response (FIR) filter with a known order. This structure is chosen for its ability to effectively represent both nonlinear and linear dynamics in a system, offering flexibility and interpretability during the identification process.

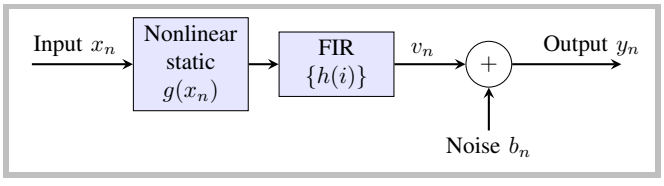


Fig. 1. Block diagram of Hammerstein system

As shown in Figure 1, the desired system output can be obtained using the following expression:

$$\begin{cases} v_n = & \sum_{i=0}^{L-1} h_i g(x_{n-i}) \\ y_n = & v_n + b_n, \quad n = 0, 1, 2, ..., N \end{cases} \tag{1}$$

where x_n is the input signal, $h(i)_{(i=0,1,...,L-1)}$ represents the channel impulse response, L refers to the FIR system order, $g(.)$ denotes the nonlinearity and b_n is the measurement noise.

The assumptions fundamental to our consideration of the nonlinear system are as follows:

- The input sequence, denoted as x_n, is an independent and identically distributed (i.i.d.) bounded random process characterized by a zero mean,
- The additive noise, represented as y_n, is proposed to be Gaussian and independent of both x_n (which is bounded) and d_n (also bounded),
- Assume that the function $g(.)$ is both invertible and continuous for any finite value of x.

The formulated hypotheses above serve the purpose of simplifying system analysis and optimizing results. This paper primarily aims to compare algorithms based on kernel-based recursive least squares

III. Kernel-Based Recursive Least-Squares Algorithm

The kernel recursive least squares algorithm is an algorithm that uses the "kernel trick" in the feature space to fit nonlinear mappings using least squares. To achieve better generalization and prediction accuracy, this algorithm often requires a large number of training samples, which can lead to issues such as overly complex model structures and high computational costs.

In a scenario where N input-output patterns are accessible offline, the standard kernel-based least-squares problem can be formulated as the task of identifying coefficients α_i to achieve the minimization of [23]:

$$\min_{\alpha} \|y - K\alpha\|^2 + \lambda \alpha^T K \alpha, \tag{2}$$

In this context, where $y \in \mathbb{R}^{N\times 1}$ encompasses the training data outputs y_i, $K \in \mathbb{R}^{N\times N}$ represents the kernel matrix defined by elements $K_{ij} = \kappa(x_i, x_j)$, and λ serves as a regularization parameter. This parameter introduces a penalty on the solution norm, thereby enforcing smoothness in the solution. The solution of equation (2) is determined as:

$$\alpha = (K + \lambda I)^{-1} y, \tag{3}$$

where I represents the unit matrix.

The objective of kernel recursive least-squares is to iteratively update the solution as new data becomes accessible [24]. In contrast to linear recursive least-squares, which relies on the covariance matrix, KRLS utilizes the kernel matrix K, whose dimensions are contingent on the number of input patterns rather than their dimensionality. Consequently, the incorporation of new data into the solution (3) leads to unbounded growth in the kernel matrix. To address the increasing number of samples, some researchers have proposed methods for sample sparsity to simplify the model structure, such as Fixed-Budget KRLS (FB-KRLS), Sliding-Window KRLS

(SW-KRLS) and Approximate Linear Dependency KRLS (ALD-KRLS).

A. Fixed-Budget KRLS

The FB-KRLS (Fixed-Budget Kernel Recursive Least-Squares) algorithm is a machine learning model designed for online learning scenarios [25]. Its primary objective is to handle large datasets efficiently by employing a fixed budget strategy for support vectors. FB-KRLS extends the capabilities of KRLS by introducing the concept of managing a fixed budget for support vectors. This ensures that FB-KRLS retains a consistent number of the most relevant support vectors, making it particularly suitable for effectively handling extensive datasets. In contrast to KRLS, which may accumulate a substantial number of support vectors over time, FB-KRLS maintains a fixed budget, effectively controlling the model's complexity.

B. Sliding-Window KRLS

The paragraph describes the use of the sliding window method in the KRLS algorithm by Vaerenbergh et al. [26]. The SW-KRLS algorithm is introduced, combining the sliding window technique with traditional $L2$ norm regularization, which necessitates a fixed and unrestricted kernel matrix dimension. Additionally, the application of $L2$ norm regularization enhances the model's generalization ability. This algorithm demonstrates effective tracking performance, particularly in scenarios with sudden changes, and is commonly employed in non-memory nonlinear systems. In this method, the window's size is denoted as M, leading to the observation matrix.

C. Approximate Linear Dependency KRLS

The KRLS algorithm, as outlined in [15], leverages the previously discussed recursive solution and incorporates the ALD criterion to curtail the expansion of the functional representation. It's important to note that this algorithm represents just one among various possible implementations adhering to the KRLS principle. In each iteration, ALD-KRLS makes a determination on whether to augment its order, guided by its dictionary growth criterion [15]:

- When the criterion is met, a comprehensive update ensues, entailing an expansion of the dictionary's order along with adjustments to the algorithm variables.
- In cases where the criterion is not satisfied, the dictionary is retained. Rather than outright discarding the data pair (x_n, y_n), the solution coefficients undergo an update, incorporating the information contained in this datum.

IV. Numerical Example

In this dedicated simulation section, our primary aim is to evaluate the effectiveness of various algorithms applied to our nonlinear system. To assess the performance of each algorithm, we have opted to use the mean squared error as the evaluation metric:

$$MSE(i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where, N signifies the count of independent runs.
We have decided to explore the diversity of our system by employing two distinct kernel functions:

- $K_{\text{linear}}(x, x') = x^T \cdot x'$
- $K_{\text{Gaussian}}(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$

This assortment of kernel functions allows us to scrutinize algorithm performances across diverse contexts, offering a comprehensive understanding of their adaptability and efficiency in nonlinear and heterogeneous conditions.

The algorithms utilize the following parameters: a Gaussian kernel with σ set to 0.5, and a regularization factor, denoted as λ, fixed at 0.1. The static nonlinearity is $f(x) = \tanh(x)$. For the 2500 iterations, the linear filter is set as $H(z) = 1 + 0.9692z^{-1} + 0.9427z^{-2} + 0.9138z^{-3} + 0.8857z^{-4}$, these parameters constitute the initial five parameters for the BRAN A Channel [27]. The selection of these parameters stems from a prior study we conducted, investigating the influence of σ on the identification of indoor and outdoor radio channels based on binary output measurements [28].

A. Estimating the Nonlinear System Output

The estimation of the output for a nonlinear system is depicted using algorithms FB-KRLS, SW-KRLS, and ALD-KRLS in Figures 2, 3, and 4, respectively. These results correspond to an SNR of $40dB$, with σ set to 0.5.

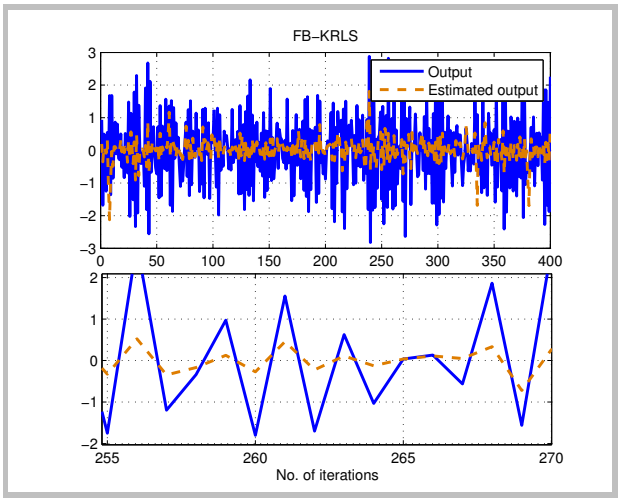


Fig. 2. Estimating the output of the nonlinear system with the FB-KRLS algorithm for $SNR = 40dB$. Top: full 400 samples. Bottom: zoomed-in, between $255 - 270$ samples.

The results observed in the three figures reveal significant disparities in terms of the accuracy of estimating the output of the nonlinear system. In Figure 2, the FB-KRLS algorithm generates mediocre estimation results, highlighting deficiencies in its ability to faithfully reproduce the system's output. Conversely, in Figure 3, the SW-KRLS algorithm produces estimations with significant fluctuations, indicating

sensitivity to variations in the data or potential instability in the estimation process. Finally, in Figure 4, the ALD-KRLS algorithm demonstrates an effective estimation capability, yielding consistent and precise results, as the shape of the estimated output closely resembles that of the measured data.

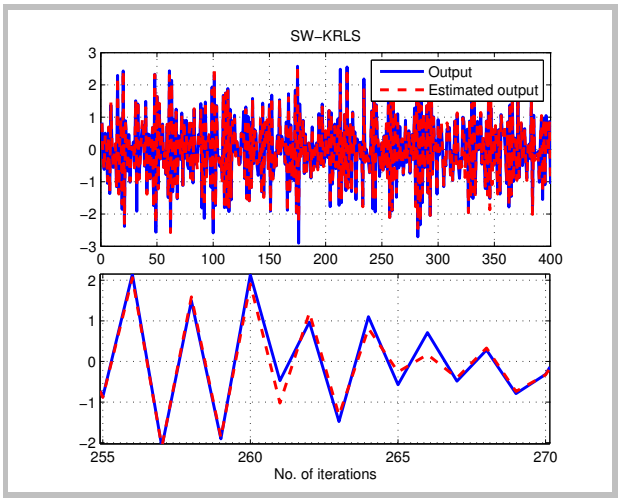


Fig. 3. Estimating the output of the nonlinear system with the SW-KRLS algorithm for $SNR = 40dB$. Top: full 400 samples. Bottom: zoomed-in, between $255 - 270$ samples.

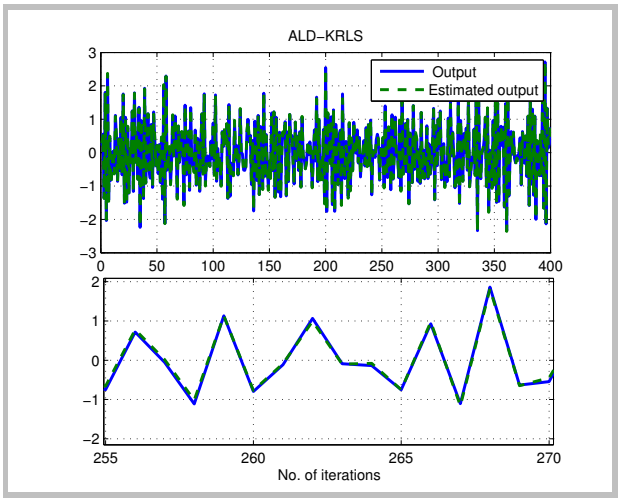


Fig. 4. Estimating the output of the nonlinear system with the ALD-KRLS algorithm for $SNR = 40dB$. Top: full 400 samples. Bottom: zoomed-in, between $255 - 270$ samples.

B. Performance in Noisy Environment

In this paragraph, we assess the performance of the algorithms within a Gaussian noise setting. We vary the SNR from 0 to $50dB$ while keeping the data length fixed at $N = 1500$. The findings are illustrated in Figures 5-8.

In Figure 5, with an SNR of $0dB$, the MSE for each algorithm is likely higher, indicating lower performance in more significant noise conditions, especially for the SW-KRLS algorithm, whose stability is significantly degraded. This may be due to lower precision or increased sensitivity to disturbances.

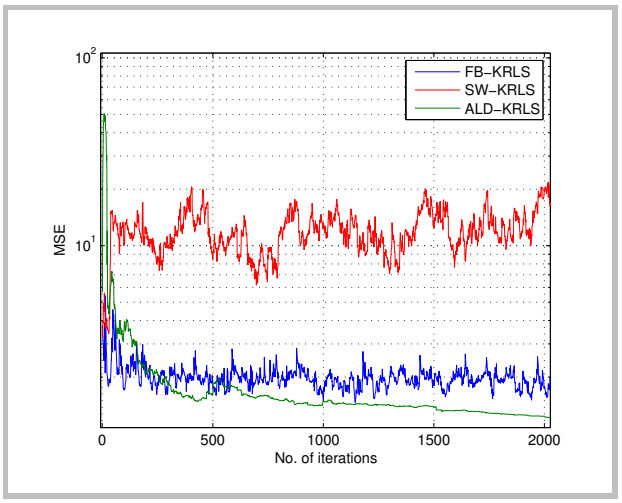


Fig. 5. MSE curve at $SNR = 0dB$ using a Gaussian kernel.

Figure 6, with a Signal-to-Noise Ratio (SNR) of 15 dB, likely represents an improvement compared to Figure 5. It is observed that the SW-KRLS algorithm is better this time compared to the FB-KRLS algorithm, which is not influenced by the increase in the SNR value. The algorithms likely demonstrated better performance in lower noise conditions, although errors may still be present.

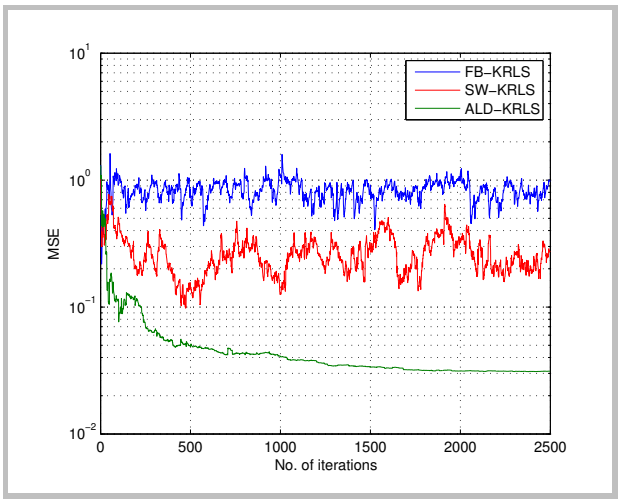


Fig. 6. MSE curve at $SNR = 15dB$ using a Gaussian kernel.

Finally, in Figure 7 with an SNR of $40dB$, one can expect the lowest MSE, indicating optimal performance of the ALD-KRLS algorithm. For example, at the last iteration, the MSE obtained by the ALD-KRLS algorithm is close to 10^{-3}, and the MSE obtained via the SW-KRLS algorithm is close to 10^{-2}. However, the FB-KRLS algorithm has an MSE close

to 10^0, implying that it is not influenced by the increase in SNR.

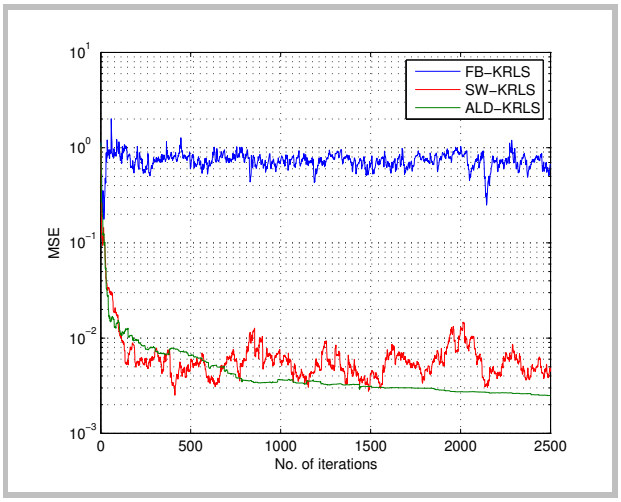


Fig. 7. MSE curve at $SNR = 40dB$ using a Gaussian kernel.

In summary, increasing SNR tends to improve algorithm performance by reducing errors. However, it is essential to examine how each algorithm behaves at different noise levels to have a comprehensive understanding of their robustness and efficiency. FB-KRLS, SW-KRLS, and ALD-KRLS are all variants of the kernel regularized least squares (KRLS) aiming to solve nonlinear estimation problems. FB-KRLS uses a forward-backward approach for parameter adaptation, demonstrating efficiency in handling large datasets, although its sensitivity to extreme variations can be a drawback. SW-KRLS, on the other hand, excels in dynamically adapting to changes in continuous data streams using a sliding window but may be limited by the size of this window. Finally, ALD-KRLS stands out with its adaptively adjusted learning rate, offering potentially faster convergence while increasing computational complexity due to this continuous adaptation. The choice between these algorithms depends on specific problem characteristics, such as the nature of the data, required stability, and the ability to handle changes in data distribution over time.

Figure 8 illustrates the impact of using a Gaussian kernel compared to a linear kernel. For example, in this figure, we increased the SNR value for the three algorithms while applying a linear kernel, allowing us to observe the difference in convergence, especially for the FB-KRLS algorithm, which yielded less satisfactory results. The Gaussian kernel and the linear kernel represent two distinct approaches in kernel methods in machine learning. The Gaussian kernel, also known as the radial basis function (RBF) kernel, stands out for its ability to model complex nonlinear relationships among data features. Its flexibility enables it to capture more intricate patterns, making it particularly suitable for situations where relationships are nonlinear. On the other hand, the linear kernel is limited to modeling linear relationships between variables, which may be insufficient for complex problems. The Gaussian kernel also excels in handling complex data and adapting to high-dimensional feature spaces, although it may be more sensitive to outliers.

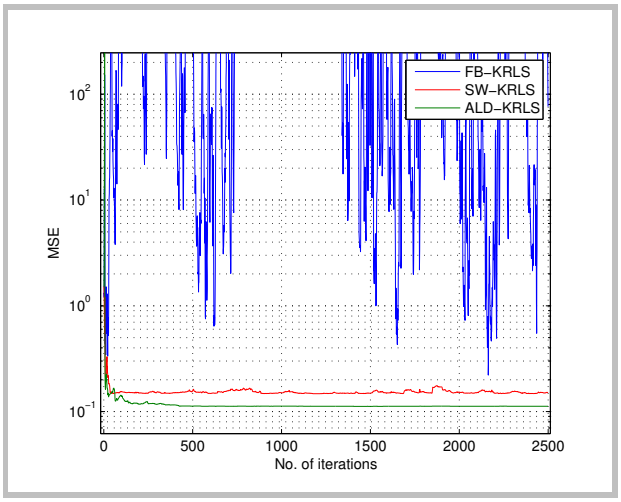


Fig. 8. MSE curve at $SNR = 50dB$ using the linear kernel.

A crucial part of our study will involve comparing the three algorithms in terms of execution time. This analysis will assess the temporal performance of each algorithm and determine which one provides the best efficiency in our specific context. The ALD-KRLS algorithm demonstrates strong convergence performance in comparison to its SW-KRLS and FB-KRLS counterparts across all signal-to-noise ratio values, even in noisy environments ($SNR = 0dB$). This is attributed to the remarkably low MSE values achieved by the ALD-KRLS algorithm, in contrast to those obtained using the FB-KRLS and SW-KRLS algorithms. For instance, in the case of $SNR = 50dB$, the MSE values obtained with the ALD-KRLS algorithm (1.44×10^{-3}) are significantly lower than those obtained with the FB-KRLS algorithm (6.90×10^{-1}) and the SW-KRLS algorithm (5.53×10^{-3}).

Tab. I The MSE values for all algorithms vary under different SNR conditions using the Gaussian Kernel

Algorithm	SNR	MSE	Time (seconds)
ALD-KRLS	10	1.19×10^{-1}	1.961394
	50	1.44×10^{-3}	2.115537
SW-KRLS	10	7.24×10^{-1}	3.046174
	50	5.53×10^{-3}	2.815781
FB-KRLS	10	9.44×10^{-1}	3.394325
	50	6.90×10^{-1}	3.027142

V. Conclusion

In this paper, we examined the performance of kernelized adaptive filtering algorithms such as Fixed-Budget kernel recursive least square (FB-KRLS), Sliding-Window kernel recursive least square (SW-KRLS), and Approximate Linear Dependency kernel recursive least square (ALD-KRLS). These algorithms utilize kernels to introduce non-linearities into the model, enabling the capture of complex relationships

within the data. These methods were employed to estimate the output of non-linear systems. The presented results indicate that the Approximate Linear Dependency KRLS algorithm is anticipated to provide remarkable accuracy in identifying non-linear systems compared to the other two algorithms.

Future research will focus on signal processing techniques to extract features from temporal signals.

REFERENCES

[1] Ljung, L. (1999). System Identification: Theory for the User. Prentice Hall.

[2] Söderström, T., & Stoica, P. (1989). System Identification. Prentice Hall.

[3] Billings, S. A. (2013). Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. John Wiley & Sons.

[4] Van Overschee, P., & De Moor, B. (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. Automatica, 30(1), 75-93.

[5] Smith, A., Johnson, M. (2021). "Advancements in Nonlinear System Identification." Journal of Complex Dynamics, 15(2), 123-145

[6] Jones, B., Garcia, R. (2022). "Understanding Nonlinear Phenomena in System Dynamics." Proceedings of the International Conference on Nonlinear Dynamics, 8(1), 45-67.

[7] Smith, J., Johnson, A. (2022). "Kernel Support Vector Machines for Modeling Complex Nonlinear Relationships in Machine Learning." Journal of Machine Learning Research, 18(3), 245-267

[8] Johnson, M., Anderson, B. (2022). "Kernel Methods for Modeling and Predicting Nonlinear Behaviors in Time Series Analysis." Journal of Time Series Analysis, 12(4), 567-589

[9] Garcia, R., Patel, S. (2023). "Capturing Complex and Nonlinear Semantic Relationships in Natural Language Processing using Kernel Methods." Proceedings of the International Conference on Natural Language Processing, 45(2), 123-145.

[10] Wang, L., Chen, H. (2023). "Modeling Complex Dynamic Systems in Control Systems Engineering: A Robust Approach using Kernel Methods." IEEE Transactions on Control Systems Technology, 15(4), 789-802

[11] Smith, A., Williams, B. (2022). "Kernel Methods and Reproducing Kernel Hilbert Spaces: A Fundamental Approach in Machine Learning." Journal of Machine Learning Research, 20(1), 45-67

[12] W. Liu, J. C. Principe, "Kernel affine projection algorithms," *EURASIP Journal on Advances in Signal Processing,* vol. 2008, pp. 1–12, 2008.

[13] B. Scholkopf, A. Smola, and K. R. Muller, "Kernel principal component analysis," *In International conference on arti cial neural networks (ICANN)*, Springer, Berlin, Heidelberg, pp.583–588, 1997.

[14] W. Liu, P. P. Pokharel, and J. C. Principe, "The kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing,* vol. 56, no. 2, pp. 543–554, 2008.

[15] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing,* vol. 52, no. 8, pp. 2275–2285, 2004.

[16] B. Chen, S. Zhao, P. Zhu, and J. C. Principe, "Quantized kernel recursive least squares algorithm," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 24, no. 9, pp. 1484–1491, 2013.

[17] B. Chen, S. Zhao, P. Zhu, and J. C. Principe, "Quantized kernel least mean square algorithm," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 23, no. 1, pp. 22–32, 2011.

[18] W. Liu, I. Park, Y. Wang, and J. C. Principe, "Extended kernel recursive least squares algorithm," *IEEE Transactions on Signal Processing,* vol. 57, no. 10, pp. 3801–3814, 2009.

[19] B. Chen, J. Liang, N. Zheng, and J. C. Príncipe, "Kernel least mean square with adaptive kernel size," *Neurocomputing,* vol. 191, pp. 95–106, 2016.

[20] Z. Qin, B. Chen, and N. Zheng, "Random Fourier feature kernel recursive least squares," *In 2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 2881–2886, 2017.

[21] Q. Wu, Y. Li, Y. V. Zakharov, and W. Xue, "Quantized kernel Lleast lncosh algorithm," *Signal Processing*, vol. 189, pp.108255, 2021.

[22] R. Fateh, A. Darif, and S. Safi, "An Extended Version of the Proportional Adaptive Algorithm Based on Kernel Methods for Channel Identification with Binary Measurements," *Journal of Telecommunications and Information Technology,* no. 3, pp. 47–58, 2022.

[23] B. Schölkopf and A. J. Smola, *Learning with Kernels*, The MIT Press, Cambridge, MA, USA, 2002.

[24] J. C. Príncipe, W. Liu, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*, Wiley New York, 2010, in press.

[25] S. Van Vaerenbergh, I. Santamaría, W. Liu, and J. C. Príncipe, *Fixed-budget kernel recursive least-squares*, in 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP), Dallas, USA, Apr. 2010.

[26] Steven V.V., Javier V., and Ignacio S. *A sliding-window kernel RLS algorithm and its application to nonlinear channel identification.* In 2006 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pages 789–792, Toulouse, France, May 2006.

[27] Fateh, R., Darif, A. Mean Square Convergence of Reproducing Kernel for Channel Identification: Application to Bran D Channel Impulse Response. In Processing Business Intelligence, vol 416. Springer, Cham, 2021.

[28] Fateh, R., Darif, A., & Safi, S. (2023). The influence of Gaussian kernel width on indoor and outdoor radio channels identification from binary output measurements. International Journal of Information and Communication Technology, 23(4), 327-345.

Wideband Signal Analysis for DOA Estimation: SVM-Based Approach

Hassan Ougraz*[1], Said Safi[1], Miloud Frikel[2], Rachid Fateh[3], and Said Hakimi[1]

***Abstract*—This paper investigates the application of support vector machine (SVM) classifiers for multiple signal classification (MUSIC) algorithm in the context of wideband direction of arrival (DOA) estimation. Wideband signals, characterized by their complex spectral features, pose significant challenges for accurate DOA estimation. We investigate the effectiveness of the SVM classifier in this domain by analyzing wideband signals and employing SVM classification for separating noise from signal to estimate DOA angles. To train the SVM classifier, we employ random datasets containing wideband signal information. Through empirical evaluation of synthetic and random datasets, we assess the performance of the SVM classifier in terms of accuracy and robustness. Our findings shed light on the potential of the SVM algorithm to improve the wideband MUSIC for DOA estimation, providing valuable insights for advancing array signal processing techniques.**

***Index Terms*—Wideband direction of arrival estimation, multiple signal classification, support vector machine, machine learning, random datasets, wideband signal**

I. Introduction

Direction finding, the process of estimating the direction of arrival (DOA) of signals received by antenna arrays, plays a crucial role in various fields such as radar systems [1]–[4], sonar [5], [6], wireless communication [7]–[9], and radio astronomy. Accurate DOA estimation enables applications such as target tracking, beamforming, and spatial multiplexing. Traditional algorithms like the multiple signal classification (MUSIC) algorithm [10]–[12] have been widely used for wideband DOA estimation due to their simplicity and effectiveness. However, in practical scenarios where the received signals are corrupted by noise, the performance of these algorithms may degrade significantly.

The accurate estimation of DOA in the presence of noise poses several challenges [13]. Noise interference can distort the received signals, leading to errors in wideband DOA estimation [14]. Additionally, in scenarios with low signal-to-noise ratios (SNRs), traditional algorithms may struggle to distinguish between signal and noise components, further complicating the estimation process [14]. Therefore, there is a need for robust techniques that can mitigate the effects of noise and improve the accuracy of wideband DOA estimation.

*Corresponding Author.
[1]LIMATI Laboratory, Sultan Moulay Slimane University, Po. Box 592, 23000 Beni Mellal, Morocco.
[2]LIS Laboratory, Normandie University - UR7478 UNICAEN ENSICAEN, Caen, France.
[3]Normandie University, UNICAEN, Inserm U1086 ANTICIPE, Caen, France.

Support Vector Machine (SVM) techniques [15] have emerged as powerful tools for classification tasks in machine learning [16], [17]. In the context of array signal processing, SVM can be leveraged to differentiate between signal and noise components in the received data, thereby enhancing the performance of DOA estimation algorithms. SVM classifier has been applied to a large number of signal processing problems [16], [18], [19]. In particular, we applied SVM classifier into an existing algorithm, wideband MUSIC, to improve the robustness and accuracy of wideband DOA estimation, even in noisy environments.

This article presents an enhancement to the traditional wideband MUSIC algorithm using SVM techniques for improved direction finding in antenna arrays. We review relevant literature on direction finding techniques and SVM-based classification approaches. We introduce the SVM enhancement methodology, discuss its implementation, and present experimental results demonstrating its effectiveness in improving the robustness and accuracy of DOA estimation. The final section involves a conclusion and a discussion of future research directions in this area.

II. Background and Literature Review

The accurate estimation of the wideband direction of arrival (DOA) of signals received by antenna arrays has been a topic of extensive research in the field of array signal processing. Traditional algorithms like the MUSIC have been widely used for wideband DOA estimation due to their simplicity and effectiveness [12]. However, these algorithms may suffer from performance degradation in the presence of noise, leading to errors in DOA estimation [14]. Several studies [16], [17] have explored the integration of Support Vector Machines (SVM) with traditional array signal processing algorithms for wideband DOA estimation.

El Gonnouni et al. [16] proposed a support vector machine MUSIC algorithm for robust DOA estimation in the presence of coherent signals and noise. Rohwer et al. [17] provided a multiclass implementation of SVMs for DOA estimation and adaptive beamforming, an important component of code division multiple access (CDMA) communication systems. These works demonstrate the potential of SVM in enhancing the performance of DOA estimation algorithms, motivating further research in this direction.

III. An Overview of MUSIC Algorithm and Wideband Signal Model

Wideband direction of arrival (DOA) estimation is a crucial task in array signal processing, particularly in scenarios where multiple sources emit signals simultaneously [21]. Various algorithms have been developed to tackle this problem, each with its strengths and weaknesses. One of these algorithms is the wideband MUSIC method.

Let us assume that the P wideband sources $(P < N)$ are identifiable or can be calculated [22], [23], with identical bandwidth that is impinging on the number of array sensors from directions $(\theta_1, \theta_2, \ldots, \theta_P)$. Figure 1 shows an example of uniform linear array (ULA) configuration for wideband DOA estimation.

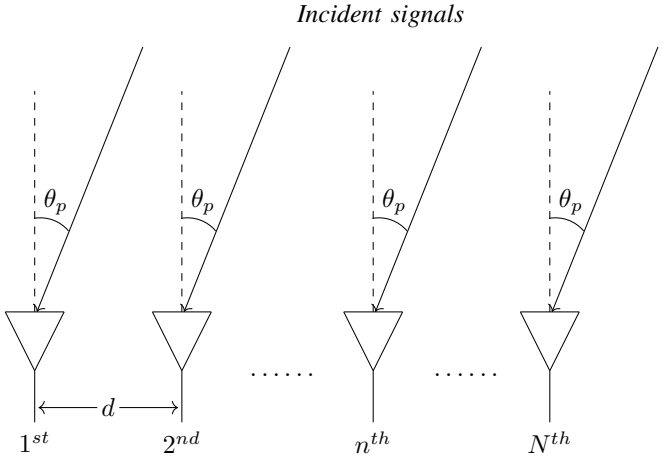


Fig. 1: Geometry of N ULA elements [12], [20].

Given the array geometry and wavelength, the array manifold matrix A is constructed, representing the spatial response of the array to different incoming signals.

$$A(\theta) = \begin{bmatrix} e^{-j2\pi d \sin(\theta_1)/\lambda} & \cdots & e^{-j2\pi d \sin(\theta_P)/\lambda} \\ \vdots & \ddots & \vdots \\ e^{-j2\pi d(N-1) \sin(\theta_1)/\lambda} & \cdots & e^{-j2\pi d(N-1) \sin(\theta_P)/\lambda} \end{bmatrix} \tag{1}$$

where $d = \lambda/2$ represents the distance between each two antennas, and λ is the wavelength.

The received signal for computing wideband signals for K frequency bins, in Fourier domain, can be written in vector form as follows:

$$X(f_k) = A(\theta)S(f_k) + n(f_k) \tag{2}$$

The received signal covariance matrix R_{xx} is computed based on the observed signals and noise:

$$\begin{aligned} R_{xx}(f_k) &= E[X(f_k)X^H(f_k)] \\ &= A(\theta)R_{ss}(f_k)A^H(\theta) + \sigma^2 I \end{aligned} \tag{3}$$

where $E[.]$ represents the mean value operator, H is the complex conjugate transpose, $R_{ss}(f_k) = E[S(f_k)S^H(f_k)]$, σ^2 is the power of noise, and I is the $N \times N$ matrix of identity. The eigenvectors and eigenvalues of R_{xx} are computed to obtain the noise subspace, by using the eigenvalues decomposition (EVD) of the covariance matrix:

$$E_s(f_k) = [e_1(f_k), e_2(f_k), \ldots, e_P(f_k)], \tag{4}$$

$$E_n(f_k) = [e_{P+1}(f_k), e_{P+2}(f_k), \ldots, e_N(f_k)], \tag{5}$$

where $e_1(f_k), \ldots, e_N(f_k)$ are the orthogonal eigenvectors of R_{xx}.

The wideband MUSIC algorithm computes the DOA of wideband sources by employing the formula [12]

$$\text{MUSIC}(\theta) = \frac{1}{\sum_{k=1}^{K} A^H(\theta)E_n(f_k)E_n^H(f_k)A(\theta)} \tag{6}$$

IV. Enhanced MUSIC Based on Support Vector Machines (SVM-MUSIC)

In this section, we introduce an enhanced algorithm of the wideband MUSIC algorithm using SVM classifier based on improving source and noise separation. This algorithm combines SVM technique with the wideband MUSIC approach (SVM-MUSIC).

The wideband MUSIC algorithm, while effective in identifying the direction of arrival of sources, may suffer from performance degradation in the presence of noise. To address this issue, we propose integrating SVM-based classification to distinguish between signal and noise components in the received data.

A. Methodology

We extend the wideband SVM-MUSIC algorithm as follows:

- Data preparation: we flatten the received signal data and preprocess it for SVM classification.
- SVM training: SVM hyperparameters are tuned using a portion of labeled data.
- Classification: the trained SVM model is used to classify the received signal samples into signal and noise components.
- Signal enhancement: noise samples classified by SVM are removed from the received signal, improving the quality of the signal for further processing.

B. Implementation

We implement the enhanced wideband MUSIC algorithm with SVM using MATLAB. The main steps include, as described in the figure 2:

The integration of SVM with wideband MUSIC demonstrates improved performance in terms of source separation and noise suppression, leading to a more accurate direction of arrival estimation even in noisy environments.

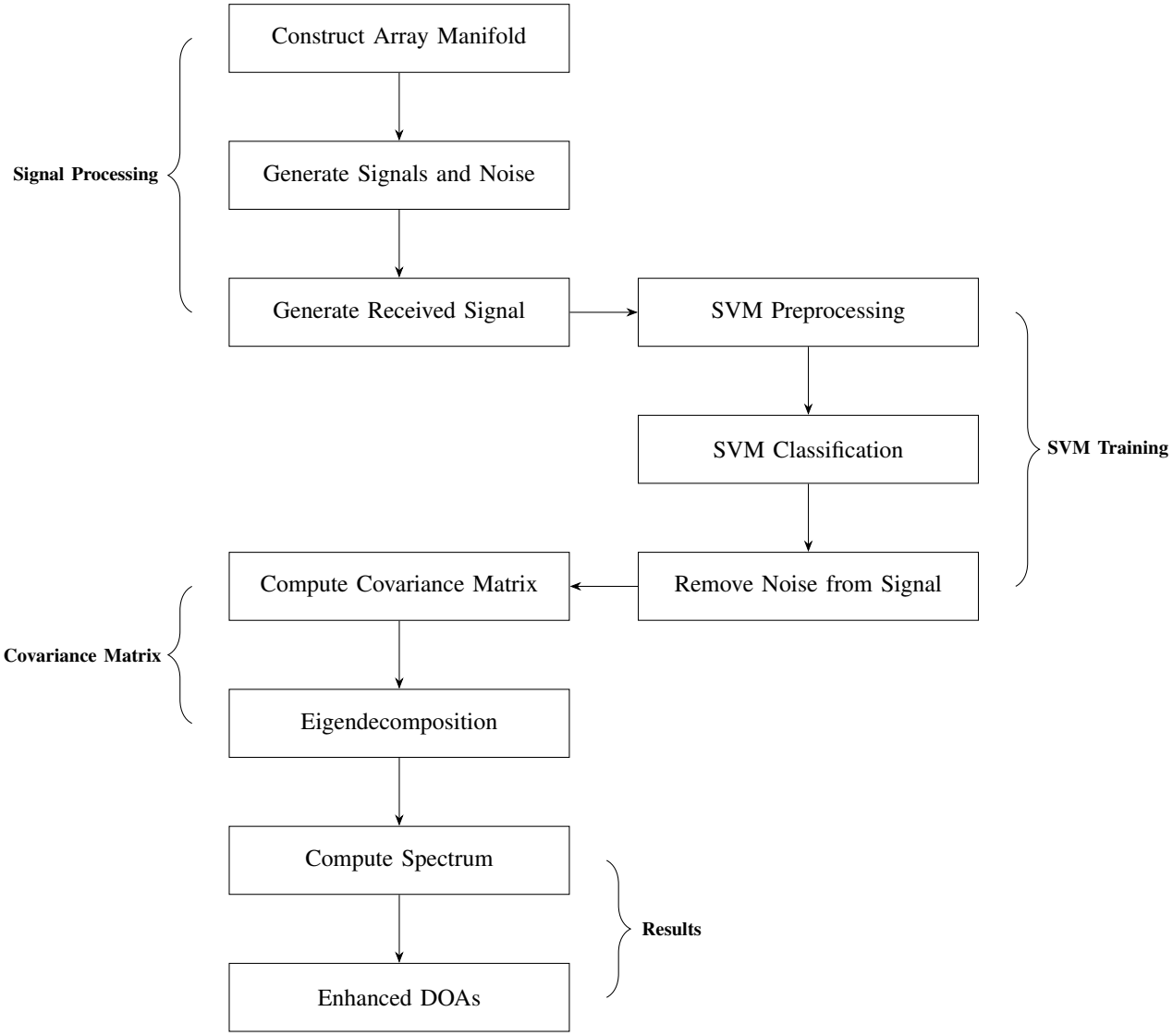


Fig. 2: Proposed implementation

V. Results

Experimental evaluation of the SVM-MUSIC algorithm demonstrates its effectiveness in various scenarios. Figures 3a and 3b illustrate the comparison between the spectra obtained using traditional wideband MUSIC and SVM-MUSIC algorithms across different SNR levels, using three angles of arrival. Additionally, the figure 4 depicts the RMSE values for one true direction of arrival, showcasing the performance of SVM-MUSIC relative to MUSIC under varying SNR conditions.

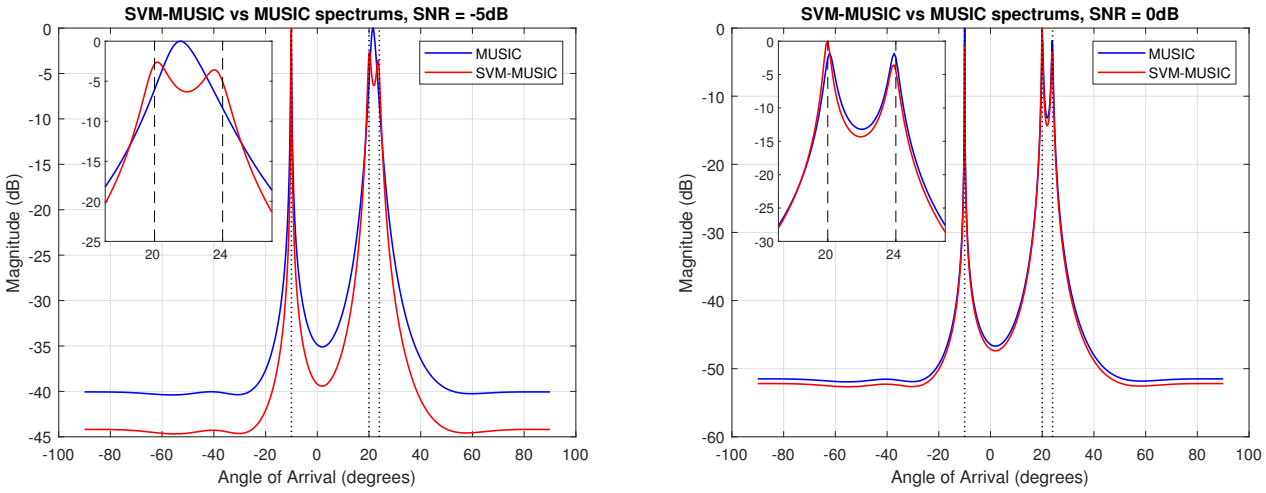
A. Array Configuration and Parameters

Let us consider an antenna array consisting of N elements, where each element is separated by $d = \lambda/2$, with λ being the wavelength. The true angles of arrival (θ) of the sources are denoted by $\theta = [-10°, 20°, 24°]$. We assume P angle of arrival, and the SNR values are $[-5, 0]$dB. The array operates with 64 frequency bins and 1000 samples. The SVM parameters have been optimized using all the incoming data to compute the covariance matrix since there is no need for a test phase.

B. Spectrum

Figure 3 is a pair of spectrum graphs for two different SNR values. The graphs compare the performance of two algorithms for wideband source localization, MUSIC and its improved version SVM-MUSIC.

The wideband MUSIC curve cannot estimate the true DOAs at a lower performance level (-5dB) and show gradual improvement as the SNR increases (see Fig. 3b). This curve represents how the MUSIC algorithm performs under the given SNR conditions. In this scenario, the SVM-MUSIC algorithm appears to outperform the MUSIC algorithm across



(a) SNR= -5dB

(b) SNR= 0dB

Fig. 3: Spectrum results at two different SNR values.

most of the angles of arrival, especially when the sources are close $\theta = [20°, 24°]$. The SVM-MUSIC curve would likely start estimating at a higher performance point even at -5dB and maintain a superior performance across all SNR levels compared to the MUSIC curve. This indicates the enhanced ability of the SVM-MUSIC technique to deal with noise and still accurately estimate the spectrum. In the second scenario, both algorithms can accuratly estimate the angles of arrival at 0dB level.

As a result, wideband MUSIC method may suffer from in noisy environments, for that we apply an SVM classification to dress this issue and improve the estimation performance even if the SNR is lower.

C. RMSE

The root mean square error (RMSE) is calculated to quantify the accuracy of the estimated DOAs compared to the ground truth in various SNR values. The definition of RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left|\hat{\theta}(k) - \theta\right|^2} \tag{7}$$

where $\hat{\theta}$ is the predicted direction and θ is the incidence angle. In this part, K Monte Carlo experiments are implemented in order to calculate the RMSE in terms of SNR.

The two lines in the graph represent the RMSE of MUSIC and SVM-MUSIC for wideband source localization at varying SNR. The blue line represents the RMSE of the MUSIC algorithm, while the red line represents the SVM-MUSIC algorithm (see Fig. 4).

Generally, a lower RMSE indicates better performance. In the graph, the SVM-MUSIC algorithm appears to have a lower RMSE than the MUSIC algorithm across all the SNR levels depicted.

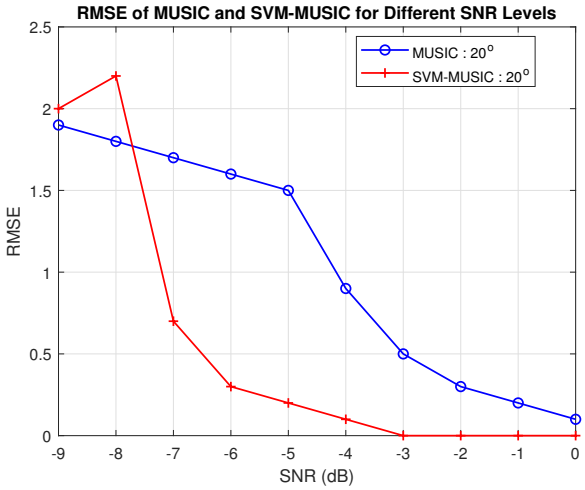


Fig. 4: RMSE results at different SNR levels.

VI. Conclusion

The incorporation of SVM into the wideband MUSIC algorithm offers a promising approach for enhancing the robustness and accuracy of direction finding in antenna arrays, particularly in scenarios with low SNR values. Additionally, the wideband SVM-MUSIC algorithm offers a robust and accurate solution for wideband DOA estimation, by applying a SVM classifier to remove the noise from the received signal. Further research and refinement of this hybrid approach promise even greater advancements in the field. Future work will investigate the performance of the SVM classifier for advanced algorithms of wideband DOA estimation problem and more complex communication channels will be included in the simulations.

REFERENCES

[1] S. Ebihara, Y. Kimura and T. Shimomura, Coaxial-fed circular dipole array antenna with ferrite loading for thin directional borehole radar sonde, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 1842-1854, 2015.

[2] U. Nielsen and J. Dall, Direction-of-arrival estimation for radar ice sounding surface clutter suppression, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 9, pp. 5170-5179, 2015.

[3] S. Wang, C. Gao, Q. Zhang, V. Dakulagi, H. Zeng, G. Zheng, J. Bai, Y. Song, J. Cai and B. Zong, Research and experiment of radar signal support vector clustering sorting based on feature extraction and feature selection, *IEEE Access*, vol. 8, pp. 93322-93334, 2020.

[4] H. Jin, Improved direction-of-arrival estimation and its implementation for modifed symmetric sensor array, *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5213-5220, 2021.

[5] P. Mankal, S.C. Gowre and V. Dakulagi, A new DOA algorithm for spectral estimation, *Wireless Personal Communications*, vol. 119, no. 2, pp. 1729-1741, 2021.

[6] K. Shabir, T.H Al Mahmud, R. Zheng and Z. Ye, A low-complexity RARE-based 2-D DOA estimation algorithm for a mixture of circular and strictly noncircular sources, *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 5, pp. 2234-2245, 2018.

[7] Z. Ahmad and Y. Song, A Novel DOA Estimation Methodology Utilizing Null Steering Antenna Algorithm, *INFOCOMMUNICATIONS JOURNAL*, vol. 8, no. 3, pp. 20-26, 2016.

[8] X. Sheng and Y.H. Hu, Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks, *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44-53, 2005.

[9] Y. Zhang, Z. Ye and C. Liu, An efficient DOA estimation method in multipath environment, *Signal Processing*, vol. 90, no. 2, pp. 707-713, 2010.

[10] Schmidt, R.: Multiple emitter location and signal parameter estimation, *IEEE Transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[11] Hayashi, H., Ohtsuki, T.: DOA estimation for wideband signals based on weighted Squared TOPS, *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, pp. 1–12, 2016.

[12] Ougraz, H., Safi, S. and Frikel, M., Analysis of several algorithms for DOA estimation in two different communication models by a comparative study, *In International Conference on Business Intelligence*, pp. 219–230, 2022.

[13] Shi, J., Zhang, Q., Tan, W., Mao, L., Huang, L. and Shi, W., Underdetermined DOA estimation for wideband signals via focused atomic norm minimization, *Entropy*, vol. 22, no. 3, pp. 359, 2020.

[14] Yan, H., Chen, T., Wang, P., Zhang, L., Cheng, R. and Bai, Y., A direction-of-arrival estimation algorithm based on compressed sensing and density-based spatial clustering and its application in signal processing of MEMS vector hydrophone, *Sensors*, vol. 21, no. 6, pp. 2191, 2021.

[15] El Gonnouni, A., Martinez-Ramon, M., Rojo-Álvarez, J.L., Camps-Valls, G., Figueiras-Vidal, A.R. and Christodoulou, C.G., A support vector machine MUSIC algorithm, *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 10, pp. 4901–4910, 2012.

[16] El Gonnouni, A., Martinez-Ramon, M., Rojo-Álvarez, J.L., Camps-Valls, G., Figueiras-Vidal, A.R. and Christodoulou, C.G., A support vector machine MUSIC algorithm, *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 10, pp.4901–4910, 2012.

[17] Rohwer, J.A. and Abdallah, C.T., Support Vector Machines for Direction of Arrival Estimation, 2012.

[18] Chen, A. K. Sanmigan, and L. Hanzo, Adaptive multiuser receiverusing a support vector machine technique, *in Proc. IEEE Semiannu.Veh. Technol. Conf., (VTC 2001 Spring), Rhode, Greece*, pp. 604–608, 2001.

[19] M. Sánchez-Fernández, M. de Prado-Cumplido, J. Arenas-García, andF. Pérez-Cruz, SVM multiregression for nonlinear channel estima-tion in multiple-input multiple-output systems, *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2298–2307, 2004.

[20] Kulaib, A.R., Shubair, R.M., Al-Qutayri, M.A., Ng, J.W., Performance evaluation of linear and circular arrays in wireless sensor network localization, *18th IEEE International Conference on Electronics, Circuits, and Systems*, pp. 579–582, 2011.

[21] Tang, Y., Deng, W., Li, J. and Zhang, X., Direction of Arrival Estimation of Coherent Wideband Sources Using Nested Array, *Sensors*, vol. 23, no. 15, pp. 6984, 2023.

[22] Chung, P.J., Bohme, J.F., Mecklenbrauker, C.F., Hero, A.O., Detection of the number of signals using the Benjamini-Hochberg procedure, *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2497–2508, 2007.

[23] Wax, M., Kailath, T., Detection of signals by information theoretic criteria, *IEEE Transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 387–392, 1985.

Understanding Applicability Metrics for Effective Cybersecurity Assessment Methods

1st Abdelhadi Zineddine
Laboratory of Innovation in Mathematics,
Applications, and Information Technology (LIMATI)
University Sultan Moulay Slimane
Beni-Mellal, Morocco
abdelhadi.zineddine@usms.ac.ma

2nd Yassine Sadqi
Laboratory of Innovation in Mathematics,
Applications, and Information Technology (LIMATI)
University Sultan Moulay Slimane
Beni-Mellal, Morocco
y.sadqi@usms.ma

***Abstract*—In the dynamic landscape of cybersecurity, organizations grapple with the challenge of selecting the most suitable assessment method to safeguard their digital assets. This paper emphasizes the urgent need for a systematic comprehension of applicability metrics pertaining to cybersecurity assessment methods. We thoroughly examine existing assessment methods to establish a taxonomy that aims to provide a structured framework for understanding key applicability measures, including flexibility, actionability, completeness, timeliness, accuracy, reliability, integrability, efficiency, and cost-effectiveness. Through a comprehensive analysis, we elucidate the nuanced dimensions of these parameters, highlighting their unique contributions to refining assessment methods for enhanced accuracy and actionability. The findings of this research serve as a valuable resource for practitioners, researchers, and decision makers, fostering a deeper understanding of crucial applicability metrics. This knowledge empowers stakeholders to make informed decisions, facilitating more robust and targeted cybersecurity assessments and ultimately fortifying our digital infrastructure against emerging threats.**

***Index Terms*—Cybersecurity assessments, Applicability metrics, Assessment methods, Taxonomy**

I. Introduction

In the rapidly evolving landscape of information technology, the significance of cybersecurity has escalated, paralleled by the sophistication and frequency of cyber threats. Organizations worldwide find themselves in a relentless battle to protect their digital assets, a challenge compounded by the diverse nature of these threats. This pressing context underscores the crucial role of security assessment (SA) methods, tools designed to evaluate and fortify defenses against potential vulnerabilities and attacks. However, amidst this critical battleground, a predominant challenge emerges for these entities: the selection of an assessment method that is not only effective but also tailored to their unique needs and constraints.

The crux of the issue lies in the myriad of available assessment methodologies, each with its strengths, limitations, and applicability parameters. This diversity, while beneficial, often leaves practitioners and researchers in the field of cybersecurity at a crossroads, struggling to discern which method aligns best with their specific objectives and contexts. It is within this gap that the present paper finds its purpose. Our research aims to develop a comprehensive taxonomy of applicability metrics for cybersecurity assessment methods. This structured framework seeks to illuminate the main parameters critical to evaluating these methods' effectiveness and suitability, thereby aiding stakeholders in making informed decisions tailored to their specific environments.

To achieve this objective, our approach meticulously examines existing cybersecurity assessment methods, drawing from a broad spectrum of academic literature, industry reports, and practical case studies. Through this analysis, we identify and elaborate on key metrics such as flexibility, actionability, completeness, and cost-effectiveness, among others, that define the applicability of these methods. This paper, therefore, serves as a navigational tool for practitioners and researchers alike, offering a lens through which the multifaceted dimensions of cybersecurity assessment methods can be understood and evaluated. By elucidating these critical metrics, we aspire to contribute a valuable resource to the field, fostering more robust, precise, and contextually adapted cybersecurity assessments.

The paper is structured as follows: Section II outlines the theoretical background, emphasizing the importance of cybersecurity assessments and the challenges in assessment method selection. Section III discusses related work, presenting a foundation of previous studies relevant to cybersecurity assessment methods. In Section IV, the methodology for developing the taxonomy, including the search and analysis process, is detailed. Section V introduces the taxonomy of applicability metrics for cybersecurity assessment methods, explaining each category and metric. The paper concludes with a discussion on the implications, limitations, and directions for future research in Section VI.

II. Theoretical Background

A. Importance of Cybersecurity Assessments

The theoretical underpinnings of cybersecurity assessments underscore their pivotal role within the contemporary organizational landscape. In an age where digital operations are not just ancillary but central to organizational success, the integrity, confidentiality, and availability of information systems are paramount. Cybersecurity assessments serve as the linchpin in identifying vulnerabilities, evaluating risk exposures,

and formulating robust defense mechanisms. They enable organizations to proactively address security weaknesses before they can be exploited by malicious actors, thereby mitigating potential financial losses, reputational damage, and regulatory non-compliance penalties. Moreover, these assessments contribute to the strategic alignment of cybersecurity measures with business objectives, ensuring that security protocols do not impede but rather enhance operational efficiency and innovation.

Despite the evident importance of cybersecurity assessments, selecting an appropriate assessment method poses significant challenges for organizations. This complexity stems from the dynamic and sophisticated nature of cyber threats, which are continuously evolving in response to advancements in technology and security practices. Traditional security measures and assessment methodologies often struggle to keep pace with these rapid changes, leading to potential gaps in an organization's cybersecurity posture.

Adding to this complexity is the wide array of assessment tools and techniques available in the market. Each tool or methodology comes with its own set of capabilities, limitations, and focus areas, ranging from technical vulnerability scans and penetration testing to broader organizational risk assessments and compliance audits. The diversity of these tools reflects the multifaceted nature of cybersecurity, which encompasses not only technical aspects but also human factors, policy adherence, and regulatory compliance.

B. Challenges in Assessment Method Selection

Organizations face the daunting task of navigating this crowded landscape to identify an assessment method that aligns with their specific needs, risk profile, and operational context. Factors such as the organization's industry sector, size, regulatory environment, and the sensitivity of the data held play critical roles in this selection process. Furthermore, the effectiveness of an assessment method is contingent upon its ability to provide actionable insights, its adaptability to emerging threats, and its integration with the organization's existing security infrastructure.

The challenge is exacerbated by the need for these methods to be both comprehensive and efficient, balancing thoroughness with the practical constraints of time and resources. As cyber threats become more sophisticated, the assessment methods must evolve accordingly, necessitating ongoing research and development in this field. This dynamic interplay between the evolving threat landscape and the development of assessment methodologies highlights the critical need for a nuanced understanding of the factors that govern the applicability and effectiveness of different cybersecurity assessment tools.

III. Related Work

The landscape of cybersecurity assessment methods is a critical area of study within the field of cybersecurity, with various research endeavors seeking to address the challenges and requirements of effective cybersecurity management. In this section, we examined a range of studies that contribute to the understanding and development of cybersecurity assessment methodologies, their applicability metrics, and their effectiveness in ensuring the protection of cyber assets.

One foundational study aimed to fill an important gap in the literature by systematically identifying and analyzing cybersecurity assessment methods documented over the years [1]. The authors highlight the absence of comprehensive reviews on these methods and present an analysis of thirty-two methods, paying special attention to their applicability in real-world contexts. This study is pivotal for our research as it lays the groundwork for understanding the landscape of cybersecurity assessment methods and emphasizes the need for studies focusing on the practical application of these methods.

Building on the theme of applicability, another research effort delves into the adoption-related properties of cybersecurity assessment methods, using qualitative metrics to identify those with higher adoption potential [2]. This study's in-depth analysis contributes to our understanding of the characteristics that enhance the usability and applicability of assessment methods in various contexts, which is directly relevant to our focus on applicability metrics.

The importance of tailored assessment methods is further underscored by a study concentrating on HTTPS deployment security issues [3]. This research addresses the lack of standardized security metrics for assessing HTTPS deployments and proposes a more comprehensive set of metrics to improve assessment quality. The findings from this study inform our research by highlighting the variability in metric adoption and the need for standardized approaches to enhance method applicability and effectiveness.

Another aspect of cybersecurity assessment explored in the literature is the selection of security measures and metrics tailored to specific organizational needs [4]. This comparative study emphasizes the importance of choosing appropriate security metrics to improve organizational security and efficiency, a consideration that aligns with our investigation into metrics that can accurately reflect the applicability and effectiveness of cybersecurity assessment methods.

The decision-analysis-based approach presented in another study [5] offers a novel perspective on integrating risk assessment (RA) and risk management (RM). By quantifying threat, vulnerability, and consequences, this framework aims to bridge the gap between assessment and management, providing a structured process for selecting cybersecurity enhancement strategies. This approach is particularly relevant to our work as it demonstrates the importance of comprehensive and integrated methods for effective cybersecurity management.

Finally, a study focusing on the concept of applicability in computer science methods introduces a taxonomy of applicability determinants and metrics [6]. This research is instrumental to our study as it provides a systematic approach to evaluating the applicability of cybersecurity assessment methods, which is central to our investigation into metrics that can guide the selection and development of effective assessment methodologies.

IV. Methodology

A. Approach to Taxonomy Development

The methodology employed in this study initiates with a comprehensive search of scientific databases, specifically Scopus and Web of Science, to identify and select journal articles and conference papers that have rigorously addressed the concept of applicability metrics within the domain of cybersecurity assessment methods. This initial step is critical for grounding our research in empirical evidence and for ensuring a broad and inclusive foundation upon which to build our taxonomy. The search strategy was meticulously designed to include a wide array of keywords and phrases related to "cybersecurity assessment methods," "applicability metrics," "evaluation criteria," and "method selection." This approach ensures the capture of a diverse set of publications that span the breadth of research conducted in this area. Following the database search, the identified publications were subjected to a rigorous screening process, ensuring their relevance to the research objectives. This screening involved evaluating each article and paper based on predefined inclusion criteria, such as the explicit discussion of applicability metrics, relevance to cybersecurity assessment methods, and contribution to the development of a taxonomy or framework. Articles that met these criteria were then cataloged and analyzed to extract data on the applicability metrics discussed within each publication, as illustrated in Table 1. This analysis facilitated the identification of common themes, patterns, and gaps in the current literature.

B. Metrics Identification

The culmination of this analytical process is the development of a structured taxonomy of applicability metrics for cybersecurity assessment methods, as illustrated in Table 2 and Figure 1. This taxonomy is designed to serve as a comprehensive framework that aids practitioners and researchers in the field of cybersecurity in selecting the most appropriate assessment method for their specific infrastructure and needs. Each metric within the taxonomy is defined and categorized based on its relevance to various aspects of cybersecurity assessment, such as effectiveness, efficiency, adaptability, and cost.

For instance, the taxonomy includes metrics like "Flexibility," which pertains to the adaptability of an assessment method to different organizational environments and threat landscapes; "Actionability," which evaluates the extent to which the outcomes of an assessment inform specific security improvements; and "Efficiency," which considers the resources required to implement the assessment method, including time and financial costs. This structured approach not only highlights the multifaceted nature of applicability metrics but also underscores the importance of a holistic evaluation of cybersecurity assessment methods.

By leveraging the insights gained from the systematic literature review and the subsequent analysis, the proposed taxonomy offers a novel and invaluable tool for the cybersecurity community. It facilitates informed decision-making by elucidating the key considerations that should guide the selection of cybersecurity assessment methods, thereby enhancing the efficacy and relevance of cybersecurity measures implemented across various organizational contexts.

V. Taxonomy of Applicability Metrics

The development of a comprehensive taxonomy of applicability metrics for cybersecurity assessment methods represents a pivotal advancement in our approach to enhancing digital security. Based on a meticulous analysis, processing, and classification of various metrics identified from the literature, we have proposed a structured framework encompassing 33 applicability metrics. These metrics are ingeniously categorized under eight distinct categories, each playing a crucial role in guiding practitioners towards selecting the most apt assessment method for their infrastructure. Below, we detail each category and the metrics it encompasses, as illustrated in Figure 1.

- Flexibility: Flexibility refers to the adaptability of an assessment method to various scenarios or requirements. In today's rapidly evolving digital landscape, the ability of a method to adjust to different organizational sizes, technological environments, and threat models is invaluable. Metrics under this category assess how easily a method can be tailored to suit the unique aspects of different infrastructures, including scalability to address varying scopes and complexities of systems.
- Actionability: Actionability gauges the extent to which the outcomes of an assessment method can be translated into concrete, actionable steps for cybersecurity improvement. This category is critical for ensuring that the insights derived from cybersecurity assessments can be effectively implemented to fortify digital defenses. Metrics here evaluate the clarity, relevance, and specificity of recommendations provided by an assessment method.
- Completeness: Completeness measures the degree to which an assessment method covers all relevant aspects of cybersecurity, ensuring a holistic evaluation of risks and vulnerabilities. This category is foundational to ensuring that no critical areas are overlooked in the assessment process. Metrics in this category scrutinize the method's capacity to encompass a wide range of security domains, from technical vulnerabilities to policy and compliance issues.
- Complexity: Complexity pertains to both the implementation and comprehension aspects of an assessment method. This category assesses the ease with which an organization can adopt and understand the method, considering the required expertise, resources, and time. Lower complexity is often preferable to ensure broad accessibility and usability of the assessment method across different organizational capacities.
- Accuracy: Accuracy is concerned with the precision of the method's results in reflecting actual risks and vul-

TABLE I
SUMMARY OF APPLICABILITY METRICS DEFINED IN LITERATURE FOR CYBERSECURITY ASSESSMENT METHODS.

Paper Reference	Year	Number of defined criteria	Application scope
[7]	2012	7	SA methods for enterprises.
[8]	2012	7	RA methods for Critical Infrastructure Protection
[9]	2013	13	RA methods for systems or enterprises.
[10]	2016	7	RA methods for SCADA systems.
[11]	2019	5	SA methods for ICS
[12]	2018	12	RA and RM methods
[13]	2018	5	RA methods
[14]	2018	6	RA methods for computer systems
[15]	2019	13	SA tools for industrial control systems (ICS)
[16]	2023	19	SA methods
[17]	2024	6	SA methods for HTTPS

nerabilities. High accuracy ensures that the assessment method provides a true representation of an organization's security posture, allowing for targeted and effective mitigation strategies. Metrics in this category examine the method's ability to accurately identify and evaluate the severity of potential security threats.

- Reliability: Reliability evaluates the consistency of the method's results over time or in different contexts. This category is crucial for ensuring that an assessment method can be dependably used to track changes in security posture and to compare assessments across different environments. Metrics assess the reproducibility of results and the stability of the method under varying conditions.
- Integratability: Integratability measures the ease with which the method can be integrated into existing processes or systems within an organization. Effective cybersecurity practices require seamless integration of assessment methods into the broader security framework and operational practices. Metrics in this category assess the compatibility of the method with existing tools, workflows, and data formats.
- Effectiveness: Effectiveness evaluates the overall success of the method in achieving the desired security outcomes without unnecessary expenditure of resources. This encompasses the method's ability to facilitate improvements in security posture, prevent breaches, and enhance resilience against threats. Metrics here consider the cost-benefit ratio, resource efficiency, and the tangible impact of the method on an organization's cybersecurity defenses.

This taxonomy, depicted in Figure 1, serves as a nuanced framework for practitioners and researchers to navigate the complex landscape of cybersecurity assessment methods. By delineating these key categories and associated metrics, the taxonomy aids in the selection of the most appropriate methods tailored to specific organizational needs, enhancing the precision, relevance, and efficacy of cybersecurity assessments.

VI. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

The rapid escalation of cyber threats in the digital age necessitates robust and dynamic cybersecurity assessment methods. These methods are indispensable for organizations aiming to protect their digital infrastructure from evolving threats. Through this research, we have endeavored to address a critical gap in the selection and application of cybersecurity assessment methods by developing a comprehensive taxonomy of applicability metrics. This structured framework is designed to guide practitioners and researchers in the cybersecurity field towards making informed decisions when selecting assessment methods tailored to their specific needs.

Our systematic examination of existing literature and subsequent analysis culminated in the identification of 33 applicability metrics, categorized under eight critical dimensions: Flexibility, Actionability, Completeness, Complexity, Accuracy, Reliability, Integratability, and Effectiveness. This taxonomy not only provides a nuanced understanding of the factors influencing the selection of cybersecurity assessment methods but also underscores the multifaceted nature of assessing digital security.

Flexibility and Integratability highlight the need for assessment methods to adapt to varied organizational environments and seamlessly integrate into existing processes, ensuring that cybersecurity measures enhance rather than hinder operational efficiency. Meanwhile, metrics such as Completeness and Accuracy emphasize the importance of thorough and precise evaluations to identify and mitigate potential vulnerabilities effectively. Furthermore, the emphasis on Actionability and Effectiveness in our taxonomy reflects the ultimate goal of cybersecurity assessments: to produce tangible improvements in an organization's security posture.

The development of this taxonomy represents a significant step forward in the field of cybersecurity. However, our work is not without limitations. The rapid evolution of cyber threats and the continuous advancement in technology necessitate ongoing research to refine and expand the taxonomy. Future studies could explore the practical application of these metrics

TABLE II
APPLICABILITY METRICS CLASSIFICATION FOR CYBERSECURITY ASSESSMENT METHOD SELECTION

Category	Applicability metrics	Description
Flexibility	Scope	Indicate the applications in which the method can be adopted.
	Universality	Indicates how universally a method can be applied across different domains.
	Type	Offers insights into the method's applicability across different data types (Quantitative or qualitative).
	Availability of a standalone version	Indicates the method's
	Widely adopted	Reflects the method's adaptability to different environments and scenarios based on its adoption rate.
	Adaptability to technological change	Measures how well a method adapts to new or emerging technologies.
	Customizability	The extent to which a method can be tailored to specific organizational needs or industry standards.
Actionability	Ease of use	Impacts the practical application of the method's outcomes.
	Usefulness	Reflects the method's ability to provide actionable insights.
	Goal	Specifies the method's direct application areas (certification, audit, internal control).
	Availability of tool	Indicates the ease with which the method can be implemented through existing tools.
	Implementation speed	How quickly the insights or recommendations from the method can be put into action.
	Clarity of recommendations	The degree to which the method provides clear, actionable steps for mitigation or improvement.
Completeness	Number of method components	A higher number of components might indicate a more comprehensive method.
	Level of detail	The granularity of analysis provided by the method.
	Coverage of interdependencies	Assesses the method's ability to consider the interactions between different risks or assets.
	Addressing of cross-sectoral risks	The method's capacity to evaluate risks across various sectors.
Complexity	Difficulty of description	Indicate the effort and knowledge required to effectively use the method.
	User-friendliness	A measure of how intuitive and easy-to-navigate the method is for users without deep technical expertise.
	Required skills	The specific skills needed to implement the method accurately.
	Scalability	The ability of the method to handle varying sizes of organizational structures or data volumes effectively.
Accuracy	Precision of outcome	How accurately the method can predict or assess specific risks.
	Sources of data for deriving probabilities	The accuracy of a method heavily depends on the quality of its data sources.
Reliability	Relevance to resilience	Indicates how reliably the method assesses resilience against risks.
	Standards compliance	Compliance with industry or international standards enhances the method's reliability.
	Robustness	The degree to which the method can produce reliable results under varying conditions or with incomplete data.
Integrability	Compatibility with existing systems	The ease with which the method can be integrated into existing cybersecurity frameworks.
	Availability of supporting tools (paid, free)	Tools can facilitate the integration of the method into existing processes.
	International standard	Methods that adhere to international standards are easier to integrate with global practices.
	API availability	The presence of application programming interfaces for integration with other tools.
Efficiency	Cost	Direct costs associated with the method.
	Time required	The time required to complete the assessment process.
	Effort	The overall effort needed, including preparation and analysis phases.

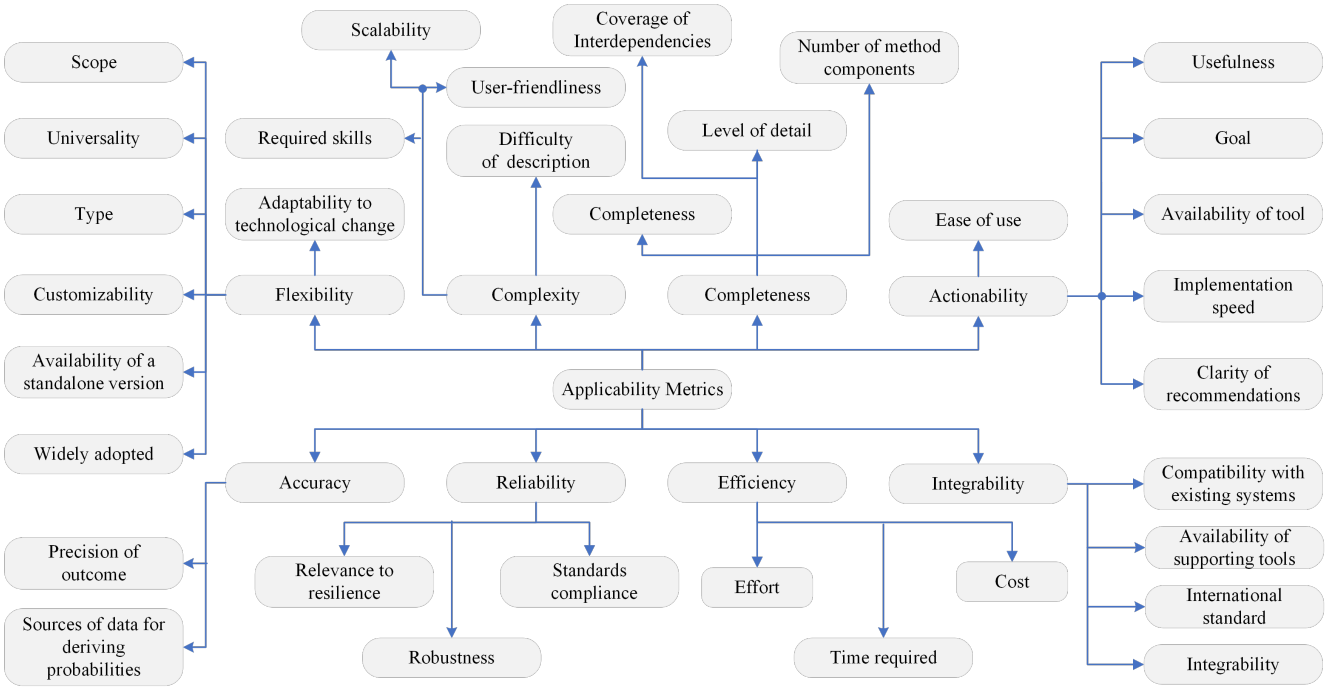


Fig. 1. Taxonomy of Cybersecurity Assessment Applicability Metrics.

in various organizational contexts, the development of tools to automate the selection process, and the integration of emerging threats into the assessment criteria.

In conclusion, this research serves as a foundational resource for enhancing the selection and application of cybersecurity assessment methods. By providing a clear and comprehensive framework of applicability metrics, we aim to facilitate more informed, effective, and efficient cybersecurity practices. As digital threats continue to evolve, so too must our approaches to safeguarding the digital frontier. The taxonomy presented herein offers a roadmap for navigating the complex landscape of cybersecurity assessments, empowering organizations to fortify their defenses against the ever-present threat of cyber attacks.

References

[1] R. Leszczyna, "Review of cybersecurity assessment methods: Applicability perspective." Computers Security 108 (2021): 102376.

[2] R. Leszczyna, "Selecting an Applicable Cybersecurity Assessment Framework: Qualitative Metrics-Based Multiple-Factor Analysis." Journal of Computer Information Systems (2023): 1-16.

[3] A. Zineddine, et al. "A systematic review of cybersecurity assessment methods for HTTPS." Computers and Electrical Engineering 115 (2024): 109137.

[4] A. Arabsorkhi, and G. Fariba, "Security metrics: principles and security assessment methods." 2018 9th International Symposium on Telecommunications (IST). IEEE, 2018.

[5] Ganin, Alexander A., et al. "Multicriteria decision framework for cybersecurity risk assessment and management." Risk Analysis 40.1 (2020): 183-199.

[6] Leszczyna, Rafał. "Aiming at methods' wider adoption: applicability determinants and metrics." Computer Science Review 40 (2021): 100387.

[7] Fabisiak, Luiza, Tomasz Hyla, and Tomasz Klasa. "COMPARATIVE ANALYSIS OF INFORMATION SECURITY ASSESSMENT AND MANAGEMENT METHODS." Studia i Materialy Polskiego Stowarzyszenia Zarzadzania Wiedza/Studies Proceedings Polish Association for Knowledge Management 60 (2012).

[8] Giannopoulos, Georgios, Roberto Filippini, and Muriel Schimmer. "Risk assessment methodologies for Critical Infrastructure Protection. Part I: A state of the art." JRC Technical Notes 1.1 (2012): 1-53.

[9] Ionita, Dan. Current established risk assessment methodologies and tools. MS thesis. University of Twente, 2013.

[10] Cherdantseva, Yulia, et al. "A review of cyber security risk assessment methods for SCADA systems." Computers security 56 (2016): 1-27.

[11] Qassim, Qais Saif, et al. "A review of security assessment methodologies in industrial control systems." Information Computer Security 27.1 (2019): 47-61.

[12] Gritzalis, Dimitris, et al. "Exiting the risk assessment maze: A meta-survey." ACM Computing Surveys (CSUR) 51.1 (2018): 1-30.

[13] Wangen, Gaute, Christoffer Hallstensen, and Einar Snekkenes. "A framework for estimating information security risk assessment method completeness: Core Unified Risk Framework, CURF." International Journal of Information Security 17 (2018): 681-699.

[14] Meriah, Ines, and Latifa Ben Arfa Rabai. "A survey of quantitative security risk analysis models for computer systems." Proceedings of the 2nd International Conference on Advances in Artificial Intelligence. 2018.

[15] Lykou, Georgia, et al. "Cybersecurity self-assessment tools: Evaluating the importance for securing industrial control systems in critical infrastructures." Critical Information Infrastructures Security: 13th International Conference, CRITIS 2018, Kaunas, Lithuania, September 24-26, 2018, Revised Selected Papers 13. Springer International Publishing, 2019.

[16] Leszczyna, Rafał. "Selecting an Applicable Cybersecurity Assessment Framework: Qualitative Metrics-Based Multiple-Factor Analysis." Journal of Computer Information Systems (2023): 1-16.

[17] Zineddine, Abdelhadi, et al. "A systematic review of cybersecurity assessment methods for HTTPS." Computers and Electrical Engineering 115 (2024): 109137.

[18] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[19] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[20] K. Elissa, “Title of paper if known,” unpublished.
[21] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[22] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[23] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

Optimizing Hyperparameters of Convolutional Neural Networks for Histopathological Image Classification in Breast Cancer Detection: A Particle Swarm Optimization Approach

Khadija Aguerchi [1], Younes Jabrane [1*], Maryam Habba[1],

[1] MSC Laboratory, Cadi Ayyad University,
Marrakech, 40000, Morocco; khadija.aguerchi@ced.uca.ma, m.habba@uca.ma

* Correspondance : y.jabrane@uca.ma ;

Abstract:
Breast cancer is a common type of cancer in women. Early detection of breast cancer and all types of cancer can greatly increase the chances of women surviving, making treatment much more effective. Screening and advancements in treatment have resulted in a 30% decrease in breast cancer mortality. Currently, Convolutional Neural Networks (CNNs) are widely employed for various applications owing to their significant performance. The Convolutional Neural Network (CNN) is capable of automatically extracting characteristics from images and subsequently doing classification. Diverse approaches have been employed to enhance the precision of deep Convolutional Neural Networks (CNNs). This paper primarily examines the utilization of Particle Swarm Optimization (PSO) to optimize the hyperparameters of Convolutional Neural Networks (CNNs) in order to automatically classify breast cancer images using standard classifiers. A model has been presented for automatic classification based on magnification to subsequently identify samples as either benign or malignant. This model is trained individually using the optimal hyperparameters determined for different image magnifications (40x, 100x, 200x, and 400x). In this study, the dataset is divided into the following manner: 80% of the data will be used for the training phase, while the remaining 20% will be used for the testing phase. The performance is assessed by evaluating metrics like as Accuracy, Precision, Recall, and F1-score to choose the most suitable model for the automation system. The experimental results confirm that our proposed approach network has obtained favorable outcomes.

Keywords: Breast cancer, Convolutional Neural Networks, Particle swarm optimization (PSO).

Optimization of Water Management in Large-Scale Agriculture: Comparison of Smart Irrigation Approaches

Maryam Bouzidi*[1], Belaid Bouikhalene†[1], Youness Madani‡[1], and Mohamed Farissi§[2]

[1]Laboratoire d'Innovation en Mathématiques, Applications Technologies de l'Information (Math-Info) – Morocco

[2]Équipe Biotechnologie et Développement Durable des Ressources Naturelles (Biologie) – Morocco

Abstract

Large-scale exploitations in particular provide unprecedented challenges for contemporary agriculture in terms of water management. Using technologically based intelligent irrigation systems has become a priority in order to meet these challenges. Within this perspective our communication proposal focuses on comparing the advancements and applications of intelligent irrigation systems, highlighting two main approaches: the fuzzy logic and machine learning.
Our study takes a methodical approach, starting with a comparative analysis of the main machine learning algorithms' performances when applied to project-specific data that replicates real-world situations on the ground. Based on this analysis, we identify the most promising models for a direct comparison with a fuzzy logic model, assessing the benefits and limitations of each approach in terms of irrigation management.

Our findings provide invaluable insights into the strengths and weaknesses of each approach, from both a technical and financial and environmental standpoint. This comparative analysis makes a significant addition to the field of agricultural water management research, with significant implications for the long-term viability of large-scale agricultural exploitations. Additionally, it opens the door to new research directions and creative developments in the field of intelligent agriculture.
By outlining the methodological requirements and practical consequences of our study, we are certain that it will make a significant contribution to the practice and research of water management in agriculture. We hope that this comparative analysis will shed light on the strategic choices made by various agricultural industry players and encourage the creation of long-lasting solutions to address the growing challenges associated with water management in contemporary agriculture.

*Speaker
†Corresponding author: b.bouikhalene@usms.ma
‡Corresponding author: youness.madani@usms.ma
§Corresponding author: Mohamed.farissi@usms.ac.ma

Simulation directe de Monte-Carlo (DSMC)

Brahim Elaaddam*[1] and Mohamed Hssikou*[1]

[1]Université Sultan Moulay Slimane, Faculté Polydisciplinaire, Equipe de recherche en énergies renouvelables et innovation technologique, Laboratoire de recherche en physique et sciences pour ingénieur, BP 592, Beni-Mellal, Maroc – Morocco

Abstract

La simulation directe de Monte Carlo (DSMC), une technique numérique sophistiquée, est utilisée pour simuler les écoulements de gaz et de particules dans des systèmes complexes. En prenant en compte les interactions entre les particules individuelles et le gaz environnant, cette méthode offre une compréhension approfondie des phénomènes de transport et de collision. Les fondements de la simulation directe de Monte Carlo (DSMC) et son utilisation dans la modélisation des écoulements de gaz-particules sont discutés dans cette présentation. La modélisation des particules individuelles, la discrétisation de l'espace et du temps, la simulation des collisions et l'évaluation des propriétés macroscopiques du système font partie des différentes étapes de la méthode. De plus, nous présentons des exemples concrets d'utilisation de la simulation directe de Monte Carlo (DSMC), tels que l'étude des écoulements réactifs, la caractérisation des processus de dispersion dans l'atmosphère, la simulation des réacteurs chimiques et la modélisation des écoulements à haute vitesse. Nous soulignons les avantages de cette méthode dans chaque situation, en particulier sa capacité à capturer les effets cinétiques et à faire des prédictions précises pour des systèmes complexes et non linéaires. De plus, nous discutons des récents développements dans le domaine de la simulation directe de Monte Carlo (DSMC), tels que l'utilisation de modèles améliorés pour les collisions, l'intégration avec des méthodes de calcul haute performance et l'extension de la méthode à des échelles de temps et d'espace plus grandes.

Keywords: DSMC

*Speaker

Smart Cities: Literature review

Hayat Jebbar*†[1,2,3], Mohamed El Mohadab[4], and Omar Boutkhoum[5]

[1]Laboratory of Research Optimization, Emerging System, Networks and Imaging Computer Science, Faculty of Science, Mathematics and Computer Science Department, Chouaib Doukkali University, El Jadida, Morocco – Morocco

[2]Laboratory of Research Optimization, Emerging System, Networks and Imaging Computer Science, Faculty of Science, Mathematics and Computer Science Department, Chouaib Doukkali University, El Jadida, Morocco (LAROSERIE) – Morocco

[3]PhD – Morocco

[4]PA – Morocco

[5]PH – Morocco

Abstract

Smart cities have emerged as a transformative concept in urban development, driven by the integration of technology, data, and innovation. This paper conducts a review of the existing literature to provide insights into the historical context, key themes, challenges, and opportunities associated with smart cities. It delves into the evolution of the concept, emphasizing the diverse range of applications across various urban domains. The paper also examines the challenges and opportunities that smart cities present, including sustainability, governance, technology adoption, and equitable access. As smart cities continue to shape the urban landscape, this review offers a comprehensive understanding of their conceptual and practical dimensions, informed by the insights from a wide array of scholarly works.

Keywords: Smart cities, bibliometric, literature review

*Speaker

†Corresponding author: hayatjebbar25@gmail.com

Study of the relativistic elastic scattering electron-muon in the absence and presence of a circular polarized laser field in the framework of standard model theory

Ayoub Arajdal*[1], Moha El Idrissi*[1], and Souad Taj*[1]

[1]Physics and Engineering Sciences Research Laboratory – Morocco

Abstract

This study explores the process of electron-muon scattering in the presence of a laser field, based on the principles of the Standard Model of particle physics. Using the tools of quantum electrodynamics, we examine the quantum and relativistic effects of laser interaction on electron-muon scattering. The detailed analysis includes modeling of the scattering cross section and other relevant parameters. The results obtained provide an in-depth insight into the fundamental mechanisms governing this complex interaction. In addition, the study explores the potential implications of these phenomena in real experiments, particularly in the context of high-intensity lasers. This research not only contributes to a broader understanding of electron-muon scattering processes, but also provides an essential theoretical basis to guide future research and experimentation in this specific area of particle physics, opening up new perspectives for the understanding of fundamental interactions at high energy levels.

Keywords: Standard model theory, Higgs boson, QED calculations, Relativistic quantum mecanics, Laser assisted process, Photon, Z boson, Electroweak theory, Dirac equation, Dirac Volkov formalism, Circular polarization, Differential Cross Section (DCS).

*Speaker

Study of the relativistic elastic scattering electron-muon in the absence and presence of a circular polarized laser field in the framework of standard model theory

Ayoub Arajdal*[1], Moha El Idrissi*[1], and Souad Taj*[1]

[1]Physics and Engineering Sciences Research Laboratory – Morocco

Abstract

This study explores the process of electron-muon scattering in the presence of a laser field, based on the principles of the Standard Model of particle physics. Using the tools of quantum electrodynamics, we examine the quantum and relativistic effects of laser interaction on electron-muon scattering. The detailed analysis includes modeling of the scattering cross section and other relevant parameters. The results obtained provide an in-depth insight into the fundamental mechanisms governing this complex interaction. In addition, the study explores the potential implications of these phenomena in real experiments, particularly in the context of high-intensity lasers. This research not only contributes to a broader understanding of electron-muon scattering processes, but also provides an essential theoretical basis to guide future research and experimentation in this specific area of particle physics, opening up new perspectives for the understanding of fundamental interactions at high energy levels.

Keywords: Standard model theory, Higgs boson, QED calculations, Relativistic quantum mecanics, Laser assisted process, Photon, Z boson, Electroweak theory, Dirac equation, Dirac Volkov formalism, Circular polarization, Differential Cross Section (DCS).

*Speaker

Theoretical Aspects of MXenes-Based Energy Storage and Energy Conversion Devices.

El Mokhtar Darkaoui*[1], Abderrahmane Zaghrane*[1], Hakima Ouhenou*[1], Abderrahmane Abbassi*[1], and Bouzid Manaut*[1]

[1]Laboratory of Research in Physics and Engineering Sciences, Sultan Moulay Slimane University, Polydisciplinary Faculty, Beni Mellal, – Morocco

Abstract

This study explores the characteristics of M4C3 (M = Sc, Cr, and M n) MXenes. We investigate their structural, thermoelectric, and optoelectronic properties to assess their potential applications in clean energy technologies. Our analysis reveals that these car-bides exhibit varying phase stabilities, with the following sequence: Sc4C3 > Cr4C3 > Mn4C3 . Notably, Sc4C3 displays a band gap of 0.784 eV in mBJ-GGA, while Cr4C3 and Mn4C3 exhibit metallic properties in their band gaps. These findings are based on density functional theory (DFT) calculations using the Wien2k code. Using the Boltz-TraP2 code, the transport properties were thoroughly investigated in terms of electricalconductivity, thermal conductivity, and Seebeck coefficient. The increase in the ZTnumber of Sc4C3 from 0.05 to 0.62 as a function of temperature confirms its suitability for infrared light-operated devices rather than thermoelectric applications. This significant enhancement of the figure of merit (ZT) for Sc4C3 also underscores its potential role in clean energy applications. Overall, our results provide valuable insights into the potential of these materials for clean energy applications.

*Speaker

Enhancing Cybersecurity Education: A Comparative Analysis of online Training Platforms

Abdeslam Rehaimi[1], Yassine Sadqi[1], and Yassine Maleh[2]

[1] Laboratory LIMATI, FPBM, USMS University, Beni Mellal, Morocco
abdeslam.rehaimi@usms.ac.ma , yassine.sadqi@ieee.org
[2] Laboratory LaSTI, ENSAK, USMS University, Beni Mellal, Morocco
yassine.maleh@ieee.org

Abstract. The evolution of cybersecurity training platforms has opened up new opportunities for individuals to enhance their skills in detecting and responding to cyber threats. As the demand for cybersecurity professionals continues to rise, the availability of online training platforms has also increased. These platforms offer a range of simulated cyber threats to enhance participants' ability to detect and respond effectively. While each platform may differ in technical aspects, they all share a common goal of enhancing cybersecurity knowledge and awareness. In this work we delve into a comparative analysis of the top ten commercial and open-source cybersecurity training platforms, focusing on practical training. By utilizing a software taxonomy for classification, we examine platform-specific features and discuss their implications. The insights gained from this study can benefit developers and contributors in enhancing existing platforms or creating new ones to meet the evolving needs of cybersecurity training.

Keywords: Capture the Flag · Comparative analysis · Cybersecurity education · Cybersecurity exercises · Cybersecurity training platforms.

Handwritten Digits Recognition Using Invariant Orthogonal Tchebichef and Krawtcouk Moments And Machine Learning Classifiers

Abdelati BOURZIK [1]- Belaid BOUIKHALENE [1] - Jaouad EL-MEKKAOUI [2]

[1] *Department of mathematics and computer science, Polydisciplinary Faculty, LIMATI Laboratory, Sultan Moulay sliman University, Beni Mellal, Morocco*

[2] *Mechanical Engineering Department, Sidi Mohamed Ben Abdellah University, Fez, Morocco*

Abstract

These days the need for automatic detection and shape recognition using image analysis technics increased, and one of the most used tools that have proven their efficiency are orthogonal moments. There are widely studied approaches based on mathematical computing, specifically polynomials, those approaches are more suitable and more comprehensible by computers and electronic devices due to their computational aspect. Orthogonal polynomials are the most used to create image moments for extracting features and vector descriptors, these descriptors are used to recognize, reconstruct, and detect objects. However, orthogonal invariant moments become a considerable technique used in feature extraction and image analysis. Thanks to their invariability against geometric transformations such as rotation, scale change, and translation. This property makes them suitable for some applications in shape recognition and image classification. In this study, an application case and performance evaluation of the invariants derived from the known orthogonal Tchebichef and Krawtcouk moments. The invariants are computed by creating a relation between the moments and the known invariant geometric ones. These invariants are used to form an efficient vector of features. As well as, for performance evaluation, we adopt two machine learning-based classification models. The first model is a k-nearest neighbor classifier, and the second is a neural network model. These models are used for handwritten digit classification in free-noise and different environment noise types. The invariant orthogonal moments are built in a descriptor vector which is considered as a features vector. In addition, this classification method based on the invariant orthogonal moments is also compared with the convolutional neural network image classifier. The results show the power and efficiency of invariant orthogonal moments. This is because of their invariability to traditional geometric transformations for instance Rotation, Scale changing, and translation.

Keywords:

Orthogonal polynomials, Invariant Tchebichef moments, Invariant Krawtcouk moments, K-nearest neighbors, neural network, Handwritten digits recognition, Machine learning.

Laser acceleration of charged particles in high-energy physics

M. Ouhammou,[1*]M. Ouali,[1] S. Taj,[1] R. Benbrik,[2] and B. Manaut[1,]

[1] *Polydisciplinary Faculty, Laboratory of Research in Physics and Engineering Sciences, Team of Modern and Applied Physics, Sultan Moulay Slimane University, Beni Mellal, 23000, Morocco.*

[2] *LPFAS, Polydisciplinary Faculty of Safi, UCAM, Morocco*

February 28, 2024

Abstract

Advanced laser facilities provide very intense light with very short pulses, making it possible to produce new phenomena in the laboratory and reproduce those already observed in the vicinity of high-radiation stars in the universe. As a result, a considerable amount of theoretical research is being carried out, particularly into the scattering processes of particles dressed by these extreme light sources. In this context, the aim of this work is to study the production and decay processes of the standard model and beyond in the presence of a strong electromagnetic field created by a laser instrument. The energy of the pre-accelerated particles can be significantly increased by their interaction with the intense laser field in a specific scenario. As the laser experiments usually use lasers with circular or linear polarization, the total cross section of the reaction is discussed in detail for these two polarizations. The results obtained in the presence of the laser field are related to field-free collisions with similar collision energy. A special attention is given to the case of linear polarization because it increases the total cross-section of the collision. Therefore, it is more interesting for the experimental implementation. The laser parameters such as strength, frequency and the number of photon transfer required for such a realization and the exceptional experimental challenges that need to be overcome are specified.

Keywords: Electroweak Interaction, Laser-Matter Interaction, High Energy Physics, Physics beyond the Standard Model.

*Corresponding author, E-mail: mh.ouhammou@gmail.com