I. INTRODUCTION

Phishing is a cyberattack technique originating in the 1990s and remains one of the most common and damaging security threats. Its strategies and procedures continuously evolve, with common methods including email fraud, spear- phishing, and whaling. These attacks deceive victims by impersonating trusted individuals or organizations to manipulate them into disclosing confidential credentials, such as banking details, credit card numbers, and passwords [1]. A typical phishing attack uses fraudulent emails to redirect users to counterfeit websites that mimic legitimate entities, with the goal of stealing sensitive personal and financial data [2]. This method represents a critical form of social engineering.

To combat this threat, machine learning methodologies are widely employed for the autonomous identification of phishing websites. Among these, Random Forest (RF) is recognized as a highly efficient algorithm due to its strong feature selection capabilities, effectiveness in modeling non- linear relationships, and robust generalization. RF can independently extract important patterns from large, diverse datasets, making it well suited for real-world applications [3]. However, despite its advantages, the performance of RF can be suboptimal without proper parameter tuning. Consequently, optimization techniques like Particle Swarm Optimization (PSO) are introduced to enhance its predictive accuracy. PSO is favored for its simplicity, flexibility, and effectiveness in both continuous and discrete optimization problems. Its ease of implementation and independence from gradient information make it an attractive method for improving machine learning models [4].

Although the number of machine learning studies for phishing detection continues to grow, research that comprehensively examines the synergy between RF and PSO remains limited. To the best of our knowledge, there is a research gap regarding how the performance of the RF-PSO combination is affected by a fundamental factor in experimental design: the data split ratio. A common practice in many studies is to adopt a standard ratio such as 80:20 without justification or further testing. This condition highlights the need for a more in-depth analysis to determine the most optimal model configuration.

To address the challenges of phishing detection, this paper proposes a phishing detection strategy based on the RF algorithm, with its performance enhanced by the PSO method. This integrated approach allows for automatic hyperparameter optimization to improve accuracy [5] and has proven effective in various classification tasks [7]. Furthermore, PSO has demonstrated significant efficacy in related applications, such as feature selection for phishing detection [6]. Therefore, the combination of RF and PSO presents a highly promising approach for building a more robust detection system, a justification this study aims to validate [8].

This paper is structured as follows: Section 2 reviews the relevant literature on phishing and machine learning-based detection. Section 3 details the methodology, including data collection, feature selection, and the implementation of the RF and PSO algorithms. Section 4 presents experimental results and discussion. Finally, Section 5 provides the conclusions and suggests directions for future work.