International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 391-404. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i2.1278

www.editada.org

An Explainable Artificial Immune System (XAIS) for Classification in Biomedical Classification Tasks

Paola Itzel Delena-García[1], *Yenny Villuendas-Rey*[2] ✉, *León S Mora-Guerrero*[1], *Antonio Alarcón-Paredes*[1]

[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, (CIC-IPN), Mexico City, Mexico
[2]Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo, (CIDETEC-IPN), Mexico City, Mexico
E-mails: pdelenag2023@cic.ipn.mx, yvilluendasr@ipn.mx✉, lmorag2024@cic.ipn.mx, aalarcon@cic.ipn.mx

Abstract. Biomedical datasets often contain noise, missing values, imbalance, and heterogeneous feature structures, making them difficult to model reliably and complicating the extraction of discriminative patterns required for effective classification. Although modern machine-learning models can achieve strong performance on such data, many of these approaches operate as opaque systems, offering little insight into how decisions are produced—an essential requirement in biomedical applications. This work introduces the Explainable Artificial Immune System (XAIS), an immune-inspired classification model that delivers prototype-based explanations derived from similarity-driven antibody responses and complemented by performance-aware indicators, providing users with direct evidential insight into each decision. XAIS was evaluated on eight publicly available biomedical datasets using stratified 5-fold cross-validation and compared against standard machine-learning classifiers. The results show that XAIS attains competitive predictive performance while offering structured, instance-level evidential explanations, underscoring its potential as a transparent and trustworthy foundation for biomedical decision-support systems.

Keywords: XAI, Biomedical data, AIS, classification, Heterogeneous data.

Article Info
Received January 4, 2026
Accepted Feb 4, 2026

1 Introduction

Biomedical data usually proves challenging for computational models, as it often exhibits incompleteness, high dimensionality, class imbalance, noise, high variance and nonlinear feature interactions. In addition, depending on the clinical scenario, it may be necessary to handle either scarce samples or very large volumes of data. These properties complicate the extraction of reliable patterns often requiring specialized computational approaches that can address data heterogeneity, manage uncertainty, and maintain model interpretability while ensuring clinical relevance and robustness. To address this issue many soft computing techniques have emerged, showing promising results and great performance in many healthcare applications (Han & Liu, 2021; Houssein et al., 2023).

Despite the remarkable performance achieved by state-of-the-art Machine Learning (ML) models, their widespread adoption in healthcare remains limited by the lack of inherent transparency. Most Artificial Intelligence (AI) approaches operate as black boxes, offering little to no insight into the mechanisms underlying their predictions. This lack of interpretability and transparency becomes particularly problematic in biomedical applications, where understanding the decision-making process of a model is not merely desirable but essential for clinical reliability, ethical accountability, and regulatory compliance. The demanding nature of biomedical data exposes the need for methods that are explicitly engineered to be robust under such conditions while providing transparent, interpretable reasoning to support reliable clinical decision-making. Consequently, the domain of explainable artificial intelligence (XAI) has gained increasing prominence, giving rise to methodological frameworks—such as SHAP and LIME—that attempt to render model behavior intelligible through local or feature-attribution explanations (Audemard et al., 2021; Zhou et al., 2022).

This demand for robust and interpretable analytical models has become even more pronounced over the past decade, as the rapid growth of heterogeneous biomedical data continues to expand both the scale and complexity of biomedical information. Together, these trends expose a persistent gap: the lack of computational frameworks specifically designed to accommodate biomedical data complexity while offering the interpretability required in clinical settings.

Among soft computing models Artificial Immune Systems (AIS) stand out, they are bio-inspired computational models grounded in the biological immune system's ability to defend the organism through the detection of pathogens and other foreign agents, develop cell-mediated adaptive responses, and their subsequent capacity to recognize and neutralize threats upon re-exposure (Aickelin, Dasgupta, et al., 2014). Numerous AIS variants have been developed, most are characterized by principles such as clonal selection, negative selection, and immune network theory. These models have been successfully applied to anomaly detection, feature selection, data mining, clustering, classification and optimization tasks (Aickelin et al., 2014; Sotiropoulos & Tsihrintzis, 2017).

However, as with other AI approaches, most AIS variants behave as black-box systems, whose internal decision pathways remain difficult to trace, providing limited insight into how prototypes evolve or how affinities are computed. This lack of transparency presents a significant barrier in sensitive areas such as healthcare and cybersecurity, where interpretability and traceability are essential due to the high risks associated with misclassification. Recent surveys in AIS research emphasize that both interpretability and explainability remain as open challenge in AIS, calling for models that either expose their decision-making logic, integrate explanation mechanisms to support their predictions or provide interpretable outputs (Myakala et al., 2025).

Moreover, the biomedical domain continues to be underexplored within AIS research, with only a few immune-inspired classifiers tailored to the specific challenges posed by clinical data.

Among the few AIS variants explored in biomedical classification, AISAC stands out as a supervised, prototype-based model that has demonstrated promising performance across several medical datasets. However, AISAC—like most immune-inspired classifiers—does not incorporate mechanisms for explainability, and its adaptive processes remain opaque, offering no explicit information linking prototypes and the evidence driving each classification outcome. Consequently, AISAC does not address the methodological gap identified above: the lack of immune-based models capable of producing clear, clinically interpretable explanations alongside robust predictive performance (González-Patiño et al., 2020).

To address this limitation, we introduce the *Explainable Artificial Immune System* (XAIS), an immune-based classifier that builds upon AISAC's foundational principles while integrating a model-intrinsic explicability. XAIS offers transparency by revealing the evidence that supports each prediction and by indicating the expected reliability of its decisions through prototype-based and performance-aware mechanisms. In addition, the model allows users to select key training criteria—such as the dissimilarity metric used to compute antibody–antigen affinity and the evaluation metric that guides optimization—thereby aligning the learning process with domain-specific requirements and enabling the model to adapt to the characteristics of each task.

XAIS was validated on several publicly available biomedical datasets and benchmarked against standard machine-learning classifiers using stratified k-fold cross-validation. The heterogeneity of these datasets allows for a comprehensive assessment of the algorithm's robustness and enables a fair comparison under diverse class distributions and data complexities. By structuring its decisions through similarity-driven prototype selection and a transparent adaptive process learned during training, XAIS provides an accountable and traceable alternative to traditional classifiers, addressing modern requirements for explainability in biomedical decision-support systems.

This article is organized as follows. Section 2 reviews related work on Artificial Immune Systems, with emphasis on AIS approaches for classification and their applications in biomedical contexts. Section 3 details the methodology, including the datasets, baseline models, and the internal dynamics of XAIS—prototype construction, adaptation mechanisms, and the mechanisms through which the model provides explainability. Section 4 presents the experimental setup and evaluation procedures. Section 5 discusses the results, and Section 6 presents the conclusions and outlines potential directions for future research.

2 Related Works

2.1 Artificial Immune Systems

Artificial Immune Systems (AIS) are inspired by the biological immune system, particularly by its ability to defend the organism through the detection of pathogens and other foreign agents, the development of cell-mediated adaptive responses, and the subsequent capacity to recognize and neutralize the same or similar threats upon re-exposure (Aickelin, Dasgupta, et al., 2014).

2.2 Artificial Immune Systems for Classification

AIS have been widely explored as bio-inspired approaches for optimization and pattern recognition tasks; however, their application to supervised classification has received comparatively less attention, as many AIS variants were not originally conceived with classification as a primary objective. This tendency is reflected in several immune-inspired proposals that emphasize feature selection rather than instance-level prediction (Dudek, 2012; Wang & Li, 2020) underscoring the versatility of AIS in representation-learning tasks while simultaneously highlighting the scarcity of models tailored specifically for classification.

Early efforts demonstrated that immune-inspired algorithms could serve as viable alternatives to classical learning paradigms, particularly in scenarios where robustness and diversity are essential. One of the first AIS specifically designed for classification was Immunos-81, an abstraction of T-cell and B-cell interactions that models how antibodies respond to antigens. In its original evaluation on the *Cleveland Heart-Disease Dataset*, Immunos-81 achieved 83.2% accuracy using 10-fold cross-validation (Carter, 2000), indicating that immune-inspired abstractions could capture discriminative patterns in clinical features. Subsequent work formalized the clonal selection principle in models such as CLONALG (De Castro & Zuben, 2002), which iteratively refine candidate solutions through cloning and hypermutation, showing competitive performance on supervised learning tasks.

A major milestone in AIS classification came with the introduction of the Artificial Immune Recognition System (AIRS) (Watkins & Boggess, 2002). AIRS incorporates artificial recognition balls (ARBs), resource-limitation dynamics and memory-cell formation, enabling the classifier to generalize effectively from limited data. In two simulated binary-classification datasets, AIRS achieved 86% and 94% accuracies. Later variants, such as AIRS2, refined cloning and resource allocation mechanisms to reduce computational cost while preserving accuracy. Further advances extended AIS beyond antibody–antigen interactions. Do et al., (2009) proposed AIS-AC, an associative classifier based on clonal selection that integrates rule discovery and classification within a unified immune framework. Their experiments on datasets such as *Adult*, *Letter*, *Nursery*, *Digit* and *Reuters-R8* reported accuracies ranging from 76.9% to 98.1%, demonstrating that immune-inspired systems can serve as alternatives to traditional associative classifiers in high-dimensional or sparse domains.

Altogether, these contributions establish AIS as a family of models capable of competitive classification performance relative to conventional machine-learning methods, while offering desirable properties such as adaptivity, population diversity, resilience to noise, and compatibility with high-dimensional data. These characteristics have motivated growing interest in immune-based strategies as foundations for new classification frameworks, including applications in biomedical contexts.

2.3 Artificial Immune Systems for Classification in Biomedical Data

Biomedical datasets typically exhibit high dimensionality, noise, missing values, and class imbalance; therefore, computational models must be robust yet flexible enough to accommodate these challenges (Houssein et al., 2023). In this context, AIS-based classifiers tailored to biomedical data emerge as promising alternatives, as their population-based dynamics, clonal expansion, and memory selection mechanisms provide great performance while mitigating overfitting by preserving diversity. Early efforts in this direction, such as the work of Chikh et al. (2012) using AIRS2 and AIRS2-fuzzy-kNN for diabetes classification, reported accuracies of 82.69% and 84%, showing that immune-inspired processes can achieve competitive results even with limited or noisy clinical attributes.

More advanced immune models have further strengthened this evidence. AISAC, for instance, builds representative prototypes through an adaptive response involving macrophage-like aggregation, B-cell activation, prototype adjustment and clonal refinement. Its associative classification mechanism stores antibody-like prototypes that are iteratively optimized using validation-driven fitness, moving them closer to instances of their class and away from others. According to its original evaluation, AISAC achieved competitive or superior accuracy compared to well-established machine-learning methods and other immune-based classifiers, including AIRS, Immunos and CLONALG, across ten cancer-related datasets. These characteristics make AISAC one of the few immune-based classifiers extensively validated on diverse medical datasets, including breast cancer detection tasks, consolidating the potential of immune-inspired learning in biomedical applications (González-Patiño et al., 2020).

Recent work has also explored hybrid strategies that combine AIS components with classical machine-learning classifiers to address imbalance and noise. Slimani, (2023), for example, integrated a Negative Selection Algorithm (NSA) with Naïve Bayes, SVM and Logistic Regression for diabetes classification, achieving accuracies between 75% and 84%. These approaches highlight the growing interest in immune dynamics to enhance predictive reliability in complex settings.

However, despite their performance, most AIS-based classifiers lack built-in explainability. Their decision processes remain opaque, limiting their suitability for high-stakes settings where traceability is required. Unfortunately, only a few immune-inspired classification models have incorporated explicit explanatory mechanisms, and these efforts remain isolated within the broader AIS literature. As a result, explainable AIS methods, particularly those designed for biomedical classification tasks, are still markedly underexplored, leaving an open research gap for models that can provide both competitive performance and clinically meaningful explanations. Motivated by this gap, we introduce XAIS, an immune-inspired classifier that builds upon AISAC while incorporating a redesigned architecture that enables intrinsic explainability and maintains performance.

3 Methodology

3.1 Datasets

The experimental evaluation of XAIS was conducted using eight publicly available biomedical datasets that differ in dimensionality, class distribution, and clinical context. These datasets were selected to examine the behavior of the proposed method under heterogeneous conditions, ranging from moderate to high dimensionality and from balanced to markedly imbalanced class scenarios. The selection covers clinical and diagnostic scenarios related to oncology, dermatology, metabolic disorders and immunology.

To ensure consistency across experiments, inclusion criteria for datasets were defined as follows: (i) provide a well-defined class label suitable for supervised classification, (ii) contain numerical-only feature representations (either integer, boolean or floating-point values), (iii) can have missing values in its features and (iv) be commonly used as benchmarks in biomedical machine-learning research.

Three of the selected datasets contain missing values within their feature space, reflecting common patterns of incompleteness in real-world scenarios. In addition, two datasets are considered high-dimensional due to their features being ≥ 1000 and their relationship with their instance cardinality. A comparative summary of their main dataset characteristics is presented in **Table 1**. Additionally, a brief description of each dataset is provided below.

Breast Cancer Wisconsin (Diagnostic)
Breast Cancer Wisconsin Dataset comprises quantitative features extracted from digitized images of fine-needle aspirates (FNA) of breast masses. The attributes describe the morphological properties of cell nuclei. The dataset is widely used as a benchmark for binary classification in medical diagnosis and it is available in UCI Repository (Street et al., 1993).

Dermatology
Dermatology dataset was created to address the differential diagnosis of erythemato-squamous diseases, a group of skin conditions that share many symptoms. Target in the dataset are the six classes that represents the diseases considered in differential diagnosis: psoriasis, seborrheic dermatitis, pityriasis rosea, chronic dermatitis, lichen planus and pityriasis rubra pilaris. Skin biopsies are usually required to distinguish between them. Due to its multi-class structure, class imbalance and missing attribute values, it represents a challenging classification task. Dermatology dataset is available in UCI Repository (Ilter & Guvenir, 1998).

Pima Indians Diabetes
Pima Indians Diabetes is a dataset that belongs to National Institute of Diabetes and Digestive and Kidney Diseases. Comprises clinical information and measurements collected from female patients of Pima Indian heritage above the age of 21. Since it was created to diagnose diabetes mellitus, this dataset has been extensively used as a binary classification benchmark, challenging due to the incompleteness. It is available in Kaggle and previously available in UCI (Smith et al., 1988).

Bone Marrow Mononuclear Cells with AML.
The dataset consists of gene expression profiles from bone marrow mononuclear cells obtained from 3 patients, one patient diagnosed with acute myeloid leukemia (AML) and two healthy controls. The samples were generated across three independent experiments available online in 10x Genomics under the names: AML027 pre-transplant BMMCs, Frozen BMMCs (Healthy Control 1), and Frozen BMMCs (Healthy Control 2) (10x Genomics, 2016c, 2016b, 2016a).

Smoking Effects on B Lymphocytes
This dataset consists of gene-expression data from peripheral circulating B cells from smoking and non-smoking healthy US white females. Since smoking-induced diseases are directly associated with B cells, the dataset provides useful information about the impact of smoking in women (Pan et al., 2010).

Diabetes
The Diabetes dataset contains medical information and laboratory analyses from Iraqi patients who are classified in 3 categories, non-diabetic, diabetic, and prediabetic. The data was originally acquired by the Laboratory of Medical City Hospital and is available on Mendeley Data. (Ahlam Rashid, 2020).

Breast Cancer Digital Repository (BCDR)
The BCDR-F01 subset from the Breast Cancer Digital Repository was used in this study. BCDR is provided by the Faculty of Medicine of the University of Porto and consists of 362 quantitative descriptors extracted from manually segmented lesions in digitised film mammograms. These descriptors include morphological measures, boundary-based statistics, and moment-based features derived from the regions of interest. The subset contains 200 biopsy-confirmed lesions from 190 Portuguese women aged 28 to 82, of which 175 are labelled malignant and 187 benign. Although BCDR is a publicly available resource, full access requires a data usage agreement (Moura & Guevara López, 2013).

Table 1. *Overview of the main characteristics of the biomedical datasets.*

Dataset	Instances	Features	Missing Values	Classes	Imbalance Ratio
Diabetes	1000	11	No	3	15.92
Smoking effect on B lymphocytes	79	3000	No	2	1.025
Dermatology	366	34	Yes, 8 values, 0.1%	6	5.6
Pima Indians Diabetes	768	8	Yes, 652 values, 10.6%	2	1.86
Breast Cancer Wisconsin (Diagnostic)	683	9	No	2	1.85
Bone marrow mononuclear cells with AML	1000	1000	No	2	1.11
Lung Cancer	32	56	Yes, 5 values, 0.3%	3	1.44
BCDR	362	38	Yes, 43 values, 0.3%	2	1.06

3.2 Classification Models

To contextualize the performance of XAIS, several well-known machine-learning classifiers were implemented as baseline models. Each algorithm represents a different inductive bias and learning strategy, providing a diverse comparative framework.

Naïve Bayes
Naïve Bayes is a probabilistic classifier based on Bayes' theorem under the assumption of conditional independence between features. Despite its simplicity, it often performs competitively in high-dimensional spaces and serves as a strong baseline for tabular biomedical data (Domingos & Pazzani, 1997).

Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a feed-forward neural network composed of fully connected layers trained via backpropagation. Its ability to model nonlinear decision boundaries makes it suitable for supervised classification tasks, although it typically requires careful tuning and sufficient training data (Rumelhart et al., 1986).

Decision Tree

Decision trees partition the feature space through recursively defined axis-aligned splits selected to maximize class purity. Their interpretability and ability to capture nonlinear relationships make them a widely used baseline in classification studies, particularly in heterogeneous datasets (Breiman et al., 2017).

k-Nearest Neighbors (kNN)

The kNN classifier assigns a label to a query instance based on the labels of its k-closest neighbors under a predefined distance metric. As a non-parametric, instance-based method, kNN is sensitive to local structure and provides a complementary perspective on the dataset (Cover & Hart, 1967).

Support Vector Machine (SVM)

SVM constructs a maximum-margin hyperplane to separate classes, optionally employing kernel functions to project data into a higher-dimensional space where linear separability is improved. Its stability and strong performance in high-dimensional biomedical features make it an essential baseline (Cortes & Vapnik, 1995).

Random Forest

Random Forest is an ensemble of decision trees trained on bootstrapped subsets of data and features. By aggregating predictions through majority voting, it reduces variance and often achieves robust performance on tabular biomedical datasets, especially in the presence of noise or nonlinearity (Breiman, 2001).

Artificial Immune System for Associative Classification (AISAC)

AISAC is an immune-inspired associative classifier that generates prototype-like antibodies through macrophage aggregation, B-cell activation, clonal refinement, and validation-driven adaptation. Its compact prototype representation and iterative refinement mechanism allow it to achieve competitive performance on biomedical classification tasks, making it an appropriate immune-inspired baseline for comparison (González-Patiño et al., 2020).

3.3 Explainable Artificial Immune System (XAIS)

The proposed "eXplainable Artificial Immune System" (XAIS) is an immune-inspired classification model that extends the conceptual structure of AISAC by redesigning its response dynamics to incorporate intrinsic interpretability. XAIS follows an eager-learning paradigm: rather than retaining the training set for later classification, it constructs internal data representations—embodied as antibodies—that replace the original instances during prediction.

XAIS formalizes two complementary computational mechanisms that mirror biological immunity: the *acquired immune response,* responsible for prototype-antibodies creation and optimization to represent data as a whole—training stage, and the *innate immune response,* which handles the classification of unseen instances through affinity evaluation —predictions stage. **Figure 1** shows a flowchart describing XAIS framework with the innate and acquired responses and its phases.

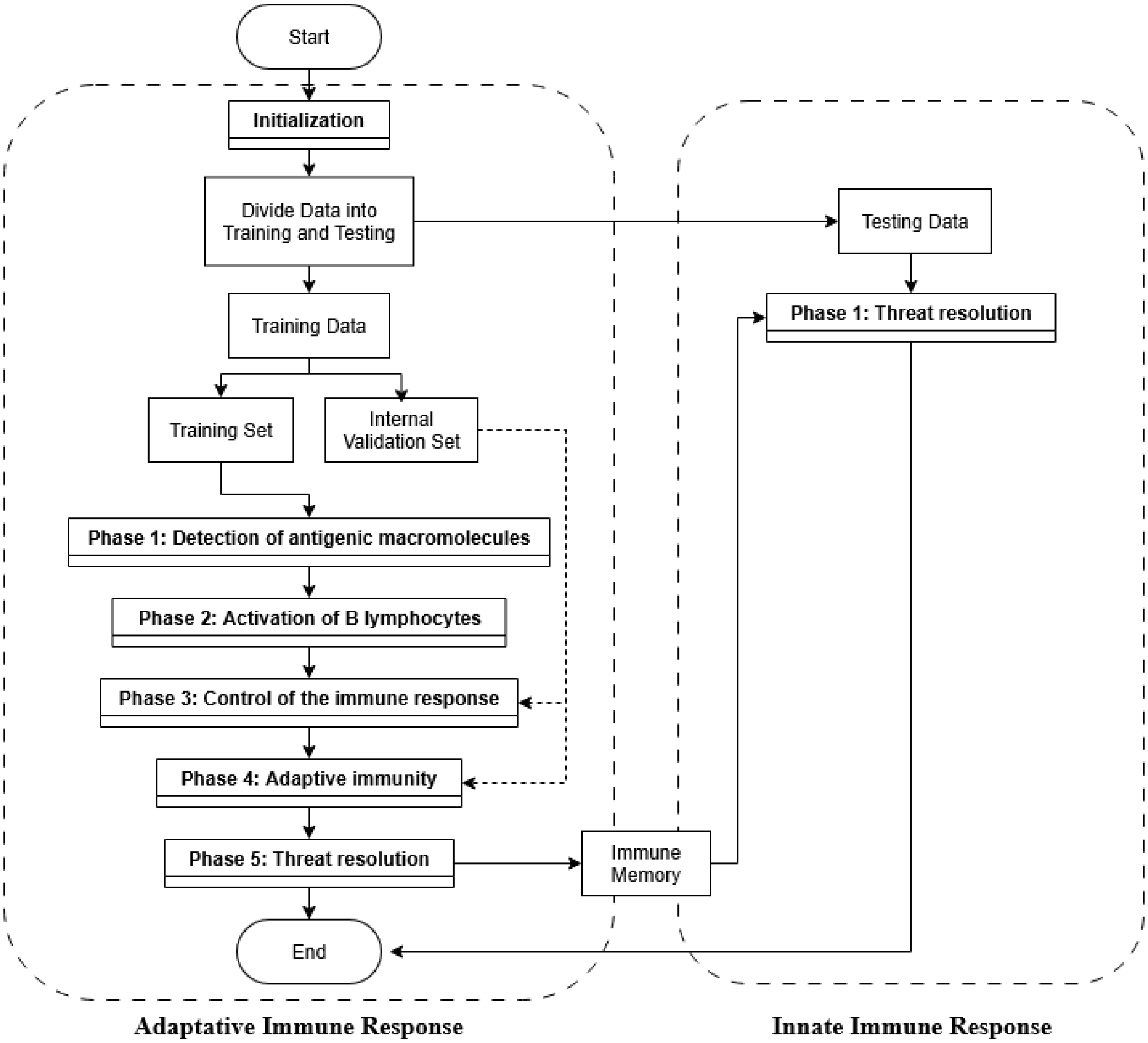
Start
Initialization
Divide Data into Training and Testing
Training Data
Training Set
Internal Validation Set
Phase 1: Detection of antigenic macromolecules
Phase 2: Activation of B lymphocytes
Phase 3: Control of the immune response
Phase 4: Adaptive immunity
Phase 5: Threat resolution
Immune Memory
End
Testing Data
Phase 1: Threat resolution
Adaptative Immune Response
Innate Immune Response

***Figure 1**. Flowchart of the XAIS framework, detailing the phases of the adaptive and innate immune responses.*

3.3.1 Acquired Immune Response

The acquired immune response corresponds to the training stage of the classifier and comprises five distinct phases: (i) the detection of antigenic macromolecules, (ii) the activation of B-lymphocytes, (iii) the regulation of the immune response, (iv) the development of adaptive immunity via clonal expansion and mutation, and (v) threat resolution through the consolidation of prototypes in the immune memory.

The biologically inspired response initiates with the detection of antigenic macromolecules by macrophages. Each macrophage specializes in phagocytosing a specific number of antigenic determinants of the antigenic molecule. These macrophages then present the antigenic determinants to the T-helper lymphocytes, marking the beginning of the immune response (Phase 1). In phase 2, these lymphocytes will generate an immune response, activating a certain number of B-lymphocytes. The activated B lymphocytes will produce and release specific antibodies to the antigens presented by the macrophage. Then, the immune response will be controlled (Phase 3). If the immune response is satisfactory, the antibodies generated are preserved. If not, a process of readjustment of the antibodies generated is carried out, so that they can combine with the antigenic determinants presented. The development of adaptive or acquired immunity (Phase 4) is a meticulous process, where each antibody undergoes reconstitution, driven by clonal and mutation strategies to improve its immune response. Finally, in Phase 5, antigenic macromolecules are completely eradicated, and the adjusted antibodies are stored in the immune memory.

XAIS, as a computational model, require a training set $U = \{u_1, \dots, u_n\}$, that represents the *antigenic macromolecules*, where each instance is represented by $u_i = [u_{i1}, \dots, u_{im}] \in R^m$. This set of data constitutes the *antigenic determinants* (epitopes) of the *antigens* to be detected. Each instance belongs to a single class $l(u_i) \in L$, where L represents the set of all classes in U, denoted by $L = \{l_1, \dots, l_k\}$. Each of these classes is considered as *antigenicity-carrying macroprotein*.

At initialization phase, the training data input U will be divided into two subsets by stratified hold-out method. EU corresponds to the training data denoted as $EU = \{e_1, \dots, e_t\}$ where $e_i = [e_{i1}, \dots, e_{im}] \in R^m$, while PU corresponds to an internal validation dataset that is denoted as $PU = \{p_1, \dots, p_t\}$ where $p_i = [p_{i1}, \dots, p_{im}] \in R^m$. This process is depicted in **Figure 1**, prior to phase one.

Additionally, a set of antibodies $A = \{a_1, \dots, a_f\}$ will be created and optimized through generations, each antibody is denoted as $a_i = [a_{i1}, \dots, a_f] \in R^m$. During the training antibodies will be adjusted to recognize antigenic determinants of the antigens. The pseudocode of the adaptive immune response in the proposed XAIS model is shown below.

```
Adaptive Immune Response
Inputs:     Training set $U$
            Number of iterations $G$
            Number of macrophages $f$
            Training percent $\alpha$
            Update rate $ur$
            Learning rate $lr$
            Clones $cl$
            Number of adjustments a$dj$
            Performance metric $M$
            Dissimilarity function $diss$
            Mutation rate $mr$
            Mutation Range (min, max)
Outputs:    Immune memory $IM$
            $PU$ set
            Confusion matrix of PU, $CM$
Initialization:
▪ Divide the training set U into a training set EU and an internal validation set PU,
  according to the value $\alpha$ defined by the user.
▪ Store $min$, $max$ and $mean$ of each feature in $EU$.
▪ If there are missing values in $PU$, impute missing values with $mean$.
Phase 1: Detection of antigenic macromolecules
1. Determine the number of instances needed to represent all the classes (macrophages will
   phagocytose the antigenic determinants of each antigenic macromolecule), as $f_{count} = \left\lfloor \frac{f}{L} \right\rfloor$.
2. For each class $L_i$ (macrophage $mac_i$):
   2.1. Phagocyte the corresponding antigenic determinants by randomly assigning $f_{count}$
        corresponding instances.
   2.2. Present phagocytosed antigenic determinants to T-Helpers lymphocytes by forwarding
        the list of assigned instances to the merging procedure.
Phase 2: Activation of B lymphocytes
3. For each list of assigned instances (T Helper lymphocyte) $linf_i$:
   3.1. Activate the corresponding B-lymphocyte $B_i$ by applying the merging procedure.
        3.1.1. The B lymphocyte will release an antibody $\bar{a}_i$ corresponding to the antigenic
               determinants presented by the macrophage $mac_i$, by computing the $mean$ of
               corresponding instances to create a representative prototype.
   3.2. Add the merged prototype $\bar{a}_i$ to the set of antibodies $A$.
4. $it = 0$;
5. While $it < G$
Phase 3. Control of the immune response
   5.1. Evaluate the immune response (fitness) by quantifying the classification
        performance of the antibody set $A$ on $PU$ set under performance metric $M$.
   5.2. Control the immune response by adjusting the prototypes set $A$ to detect the
        antigenic determinants instances in $PU$ using $Adapt(A, PU, diss)$.
   5.3. If the new set of antibodies produced by the adjustments $A' = \{\bar{a'}_i, \dots, \bar{a'}_f\}$ has a better
        performance according to $M$ than $A$, then $A \leftarrow A'$.
Phase 4. Adaptative immunity
   5.4. For each antibody $\bar{a}_i$:
        5.4.1. Create $cl$ clones.
        5.4.2. For each $cl$ clone:
            5.4.2.1.   Choose which components (attributes) of the clone will mutate based
                  on $mr$.
            5.4.2.2.   Mutate the components of the clones, considering that they will have
                  a random mutation increasing or decreasing a value chosen randomly between
                  the chosen mutation range and the limits for each attribute.
        5.4.3. Obtain a new antibody $\overline{ac}_i$ by averaging the parent and all the clones, per
              parent, and add it to $Ac$.
    5.5. If the new set of antibodies produced by cloning $Ac = \{\overline{ac}_i, \ldots, \overline{ac}_f\}$ has a better immune
         response (fitness) than $A$, then $A \leftarrow Ac$.
    5.6. $it = it + 1$
Phase 5. Threat resolution
6.  Store the final set of antibodies $A$ in the immune memory, $IM \leftarrow A$.
7.  Store $PU$ set.
8.  Store the confusion matrix of $PU$ as $CM$.
```

Adjustment of the immune response ($Adapt$)

Inputs:	Antibody set A Test antigens set PU Dissimilarity function $diss$
Output:	Adjusted antibody set A'

```
1.  $it = 0; A' = \emptyset$
2.  While $it < I$
    2.1. For each antigenic determinant $ag \in PU$
        2.1.1. Determine the corresponding antibody $ca$, which is the closest to the
              antigenic determinant $ag$, according to $diss(ca, ag)$
        2.1.2. For each component $j$ of the antibody $ca$
            2.1.2.1.   $ca'_{ij} = \begin{cases} ca_{ij} + \left(lr * \left(ag_j - ca_{ij}\right)\right), & \text{if } l(ca) = l(ag) \\ ca_{ij} - \left(lr * \left(ag_j - ca_{ij}\right)\right), & \text{if } l(ca) = l(ag) \end{cases}$
        2.1.3. Add the modified antibody $ca'$ to the set $A'$
    2.2. $lr = lr * ur$
    2.3. Permute $PU$ randomly.
    2.4. $it = it + 1$
3.  Return $A'$
```

3.3.2 Innate Immune Response

The innate immune response consists of a single phase dedicated exclusively to threat resolution: the antibodies stored in immune memory are matched against the newly presented instance, and the closest prototype determines its class. Functionally, this process is equivalent to performing a prediction step, where the stored antibodies act as the model's decision structures and the antigen is classified as soon as it is presented. The pseudocode of the innate immune response in the proposed XAIS model is shown below.

Innate Immune Response

Inputs:	Unknown antigenic determinant d Number of desired antigens per decision n Boolean value to select if the user wants to see the set used for calibrating the algorithm: $showPU$ Boolean value to select if the user wants to see the resulting immune memory after calibrating the algorithm: $showIM$
Ouputs:	FPR: False Positive Rate FNR: False Negative Rate Top-n and Bottom-n antigenic macromolecules for each decision and their corresponding distance to diss IM PU

```
Phase 1. Threat resolution
1.  For each antibody in the immune memory $a \in IM$
    1.1. Calculate and store the affinity of said antibody with the unknown instance
         (antigenic determinant), as $aff_a(d) = diss(a, d)$, where $diss$ is the function defined when
         training the algorithm.
2.  For each antigenic macromolecule:
    2.1. Return the top-n and bottom-n antibodies according to their affinity to $d$, as well
         as their affinity value.
```

```
3. Compute and Return the False Negative Rate and the False Positive Rate of the confusion
   matrix $CM$.
4. If show$PU$ then Return $PU$
5. If show$IM$ then Return $IM$
```

3.3.3 Explainability

The explainability mechanisms of XAIS operate through two complementary channels that expose both the internal behavior of the model and the evidential basis supporting each prediction. The explainability of XAIS is presented in two ways:

(a) ***Performance-aware transparency***

XAIS preserves the internal validation set (PU) used during training together with its associated confusion matrix, computed under the user-selected performance metric. This information enables early detection of systematic biases, class-specific weaknesses, or undesirable shifts occurring during antibody adaptation. By making the calibration behavior of the model explicitly accessible, XAIS provides a performance-aware transparency layer that is uncommon in traditional AIS classifiers.

(b) ***Prototype-based evidential explanations***

For every unseen instance, XAIS returns the n most similar antibodies (*top-n*) to the query and the least similar ones (*bottom-n*), based on the dissimilarity function chosen by the user. These prototype-based explanations are complemented with the false-negative and false-positive rates derived from the stored confusion matrix, allowing the user to gauge the reliability of each decision path. Through this combination of similarity-driven evidence and performance indicators, XAIS frames each prediction within a transparent and contrastive decision space.

These mechanisms position XAIS not merely as an explainable model, but as one aligned with the emerging paradigm of evaluative artificial intelligence. According to (Miller, 2023), evaluative AI systems should provide:

(i) Options: XAIS presents, for each class, the top-n candidate antibodies supporting that decision;

(ii) Judgement support: the model supplies class-specific error profiles (FNR and FPR) that contextualize the plausibility of each alternative.

(iii) Trade-off support: XAIS offers both evidence for and against a hypothesis by returning top-n and bottom-n similar instances in the antibodies set, enabling users to weigh competing explanations independently of their predicted likelihood.

Through this evaluative structure, XAIS delivers explanations that are both evidentially grounded and performance-aware, addressing key transparency requirements in biomedical decision-support systems.

4 Results

All classifiers were evaluated under a unified experimental protocol to ensure comparability across datasets. Each dataset was assessed using *stratified 5-fold cross-validation*, preserving class proportions in every split and reducing variability due to sampling. No preprocessing, normalization, or previous imputation was applied; this decision was intentional, as it allows the evaluation to reflect each model's intrinsic robustness when confronted with incomplete, noisy, or heterogeneous biomedical attributes.

For every fold, models were trained on four partitions and evaluated on the remaining one, and performance metrics were averaged across five folds. These aggregated results provide a stable estimate of each classifier's behavior. All numerical experiments were conducted on a standard computing environment using Python and scikit-learn, and the same random seed was applied to ensure reproducibility.

Model hyperparameters were tuned manually through an iterative exploration of configurations that balanced stability and performance across datasets. For XAIS, the *F1-score* was adopted as the performance measure M guiding the evaluation of antibody fitness during the adaptive immune response. The same metric was also employed in the prediction stage of each fold to determine the class assigned to test instances, ensuring consistency between prototype optimization and decision-making criteria.

The comparative accuracy values for all classifiers across the selected datasets are summarized in **Table 2Table 1**, while **Table 3** presents the corresponding F1-scores, offering a complementary perspective on performance in the presence of class imbalance. Together, these results form the basis for the analysis discussed in the following section.

Table 2. *Average accuracies of classification models, AISAC, and XAIS across eight biomedical datasets. Best results are highlighted in bold.*

Dataset	Naive Bayes	MLP	Decision Tree	kNN	SVM	Random Forest	AISAC	XAIS
Diabetes	88.61	90.59	97.34	78.48	78.61	**96.59**	82.30	86.50
Smoking Effect on B Lymphocytes	83.54	**89.87**	73.41	78.48	50.63	79.74	67.00	69.83
Dermatology	**98.08**	96.72	92.62	88.79	30.60	97.81	75.40	93.72
Pima Indians Diabetes	7643	73.30	71.74	70.44	68.09	**77.47**	67.18	72.14
Breast Cancer Wisconsin (Diagnostic)	**97.51**	96.61	94.14	97.21	95.90	97.07	96.93	96.93
Bone Marrow Mononuclear Cells With AML	84.90	93.90	95.70	92.70	52.70	**96.90**	86.50	94.00
Lung Cancer	50.00	46.90	50.00	46.32	40.60	46.90	62.38	**71.88**
BCDR	74.58	79.00	72.37	56.35	57.18	80.93	70.27	**82.97**

Table 3. *Average F1-scores of classification models, AISAC, and XAIS across eight biomedical datasets. Best results are highlighted in bold.*

Dataset	Naive Bayes	MLP	Decision Tree	kNN	SVM	Random Forest	AISAC	XAIS
Diabetes	88.61	90.71	**97.38**	78.55	81.77	96.61	59.47	69.98
Smoking Effect On B Lymphocytes	83.54	**89.87**	73.40	78.41	34.03	79.74	66.57	69.38
Dermatology	**98.09**	96.72	92.53	88.91	14.34	97.81	73.06	93.67
Pima Indians Diabetes	76.63	73.19	70.65	70.23	67.01	**77.12**	57.47	69.23
Breast Cancer Wisconsin (Diagnostic)	**97.52**	96.64	94.10	97.21	95.93	97.07	96.64	96.65
Bone Marrow Mononuclear Cells With AML	84.90	93.90	95.69	92.70	36.37	**96.89**	86.48	93.99
Lung Cancer	48.30	46.81	51.00	46.32	23.40	46.40	60.48	**72.01**
BCDR	74.57	78.99	72.35	56.36	47.41	80.94	70.18	**82.95**

5 Discussion

A distinguishing advantage of XAIS relative to standard ML models is its intrinsic transparency. Because each prediction is supported by the top-n and bottom-n antibodies most similar to the input instance, users can directly examine which prototype patterns support—or contradict—the final decision. This form of evidence-based explanation aligns the method with current expectations for trustworthy AI, particularly in healthcare, where interpretability is a regulatory and ethical requirement rather than a secondary feature.

Another notable property of XAIS is its natural compatibility with multiclass and imbalanced datasets. By explicitly representing every class through its antibody set, the model preserves minority-class information even when samples are scarce. This mitigates the tendency of traditional classifiers to bias predictions toward majority classes, an issue that is especially problematic in biomedical domains where minority classes often correspond to clinically critical conditions.

XAIS's adaptability arises primarily from two user-controlled components: (i) the performance metric M that guides fitness evaluation and (ii) the dissimilarity function used to compute antibody–antigen affinity. Selecting M defines the optimization objective and directly influences prototype evolution. Metrics such as F1-score or balanced accuracy promote uniform class performance, while sensitivity or specificity can bias the learning process toward reducing false negatives or false positives—an important consideration for diagnostic tasks. In parallel, the choice of dissimilarity function allows the algorithm to adapt to heterogeneous feature spaces by modifying how similarity relations are computed. Together, these two design freedoms provide a mechanism for aligning the learning dynamics of XAIS with the structure and requirements of diverse biomedical problems.

XAIS also exposes a broader range of hyperparameters than AISAC, including mutation range, mutation rate, learning rate, and update rate. These parameters offer additional control over the exploration–exploitation balance during prototype formation. However, this flexibility increases the dimensionality of the hyperparameter search space and may complicate model tuning compared with baseline classifiers such as naïve Bayes, which require little or no parameter adjustment. Moreover, because prototype construction and mutation depend on iterative refinement, the solution space is large and cannot be explored exhaustively, potentially affecting convergence speed in high-dimensional scenarios. Future work should therefore examine systematic or automated strategies for selecting metrics, tuning hyperparameters, and adapting mutation dynamics.

The explainability mechanisms embedded in XAIS extend beyond simple traceability. By preserving the internal validation set and its confusion matrix under the selected performance measure, users can assess whether certain classes were systematically more difficult to model during training. When this information is combined with the top-n and bottom-n prototype lists and their affinity values, XAIS provides a richer evaluative framework: the model reveals which prototypes support a given decision, which contradict it, and how similar cases behaved historically. These characteristics position XAIS within the notion of evaluative AI proposed by Miller (2023), offering explicit options, judgment support, and observable trade-offs for each decision.

Overall, the findings position XAIS as an explainable alternative to black-box classifiers while maintaining competitive performance with established machine-learning models and offering methodological advances over prior AIS-based approaches. Its capacity to construct meaningful class representations, accommodate complex datasets, and make its decision process inspectable makes XAIS a strong candidate for biomedical classification tasks.

6 Conclusions

This work introduced XAIS, an explainable immune-inspired classifier designed to produce compact, discriminative prototype representations from heterogeneous biomedical data. The experimental results show that the model is capable of constructing stable internal memories for each decision class even in the presence of noise, imbalance, and limited sample availability. In contrast with conventional approaches, XAIS provides model-intrinsic explainability by revealing prototype-based evidence that supports each prediction, and performance-aware strategies to interpret the model's ability to perform that task, thereby enabling users to inspect the rationale behind individual decisions.

XAIS extends previous AIS-based classifiers by permitting the selection of both the fitness metric and the dissimilarity function, two elements that directly shape how prototypes evolve and how affinity is computed. This flexibility allows the algorithm to adapt to diverse biomedical scenarios, particularly those where class imbalance or overlapping feature distributions challenge the reliability of traditional models. Across the evaluated datasets, XAIS achieved competitive predictive performance while maintaining transparent decision pathways consistent with current requirements for trustworthy and explainable clinical AI.

Despite these strengths, the method enlarges the hyperparameter space and relies on iterative refinement procedures whose behavior may become sensitive in high-dimensional settings. These characteristics highlight the need for systematic or automated strategies for metric selection, hyperparameter tuning, and adaptive mutation control. Future work should also examine how different combinations of performance and dissimilarity metrics influence prototype formation in specific biomedical contexts, especially in problems where several classes exhibit substantial overlap. Moreover, extending these analyses to broader families of immune-inspired architectures may further clarify the advantages of the evaluative mechanisms introduced here.

Overall, the findings demonstrate that XAIS constitutes a competitive and *explainable* alternative to conventional ML models and to existing AIS-based classifiers, offering a principled methodological foundation for interpretable decision-support systems in biomedical applications.

References

10x Genomics. (2016, July 24). *AML027 pre-transplant BMMCs / Human (Dataset by Cell Ranger 1.1.0)*. https://www.10xgenomics.com/datasets/aml-027-pre-transplant-bmm-cs-1-standard-1-1-0

10x Genomics. (2016, July 24). *Frozen BMMCs (Healthy Control 1) / Human (Dataset by Cell Ranger 1.1.0)*. https://www.10xgenomics.com/datasets/frozen-bmm-cs-healthy-control-1-1-standard-1-1-0

10x Genomics. (2016, July 24). *Frozen BMMCs (Healthy Control 2) / Human (Dataset by Cell Ranger 1.1.0)*. https://www.10xgenomics.com/datasets/frozen-bmm-cs-healthy-control-2-1-standard-1-1-0

Aickelin, U., Dasgupta, D., & Gu, F. (2014). Artificial immune systems. In E. K. Burke & G. Kendall (Eds.), *Search methodologies* (pp. 187–211). Springer. https://doi.org/10.1007/978-1-4614-6940-7_7

Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., & Marquis, P. (2021). *On the explanatory power of decision trees*. http://www.cril.univ-artois.fr/expekctation/

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge.

Carter, J. H. (2000). The immune system as a model for pattern recognition and classification. *Journal of the American Medical Informatics Association, 7*(1), 28–41. https://doi.org/10.1136/jamia.2000.0070028

Chikh, M. A., Saidi, M., & Settouti, N. (2012). Diagnosis of diabetes diseases using an artificial immune recognition system (AIRS2) with fuzzy k-nearest neighbor. *Journal of Medical Systems, 36*(5), 2721–2729. https://doi.org/10.1007/s10916-011-9748-4

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

de Castro, L. N., & von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation, 6*(3), 239–251.

Do, T. D., Hui, S. C., Fong, A. C. M., & Fong, B. (2009). Associative classification with artificial immune system. *IEEE Transactions on Evolutionary Computation, 13*(2), 217–228. https://doi.org/10.1109/TEVC.2008.923394

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*(2–3), 103–130. https://doi.org/10.1023/A:1007413511361

Dudek, G. (2012). An artificial immune system for classification with local feature selection. *IEEE Transactions on Evolutionary Computation, 16*(6), 847–860. https://doi.org/10.1109/TEVC.2011.2173580

González-Patiño, D., Villuendas-Rey, Y., Argüelles-Cruz, A. J., Camacho-Nieto, O., & Yáñez-Márquez, C. (2020). AISAC: An artificial immune system for associative classification applied to breast cancer detection. *Applied Sciences, 10*(2), 515. https://doi.org/10.3390/app10020515

Han, H., & Liu, X. (2021). The challenges of explainable AI in biomedical data science. *BMC Bioinformatics, 22*(Suppl 12), 443. https://doi.org/10.1186/s12859-021-04368-1

Houssein, E. H., Hosney, M. E., Emam, M. M., Younis, E. M. G., Ali, A. A., & Mohamed, W. M. (2023). Soft computing techniques for biomedical data analysis: Open issues and challenges. *Artificial Intelligence Review, 56*(Suppl 2), 2599–2649. https://doi.org/10.1007/s10462-023-10585-2

Ilter, N., & Güvenir, H. A. (1998). *Dermatology dataset*. UCI Machine Learning Repository. https://doi.org/10.24432/C5FK5P

Miller, T. (2023). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *ACM Proceedings*. https://doi.org/10.1145/3593013.3594001

Moura, D. C., & Guevara López, M. A. (2013). An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *International Journal of Computer Assisted Radiology and Surgery, 8*(4), 561–574. https://doi.org/10.1007/s11548-013-0838-2

Myakala, P. K., Bura, C., & Jonnalagadda, A. K. (2025). Artificial immune systems: A bio-inspired paradigm for computational intelligence. *Journal of Artificial Intelligence and Big Data, 5*(1), 1–13. https://doi.org/10.31586/jaibd.2025.1233

Pan, F., Yang, T.-L., Chen, X.-D., et al. (2010). Impact of female cigarette smoking on circulating B cells in vivo. *Immunogenetics, 62*(4), 237–251. https://doi.org/10.1007/s00251-010-0431-6

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications in Medical Care* (pp. 261–265).

Sotiropoulos, D. N., & Tsihrintzis, G. A. (2017). Artificial immune systems. In *Intelligent Systems Reference Library* (Vol. 118, pp. 159–235). Springer. https://doi.org/10.1007/978-3-319-47194-5_7

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). *Breast Cancer Wisconsin (Diagnostic) dataset*. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

Wang, Y., & Li, T. (2020). Local feature selection based on artificial immune system for classification. *Applied Soft Computing, 87*, 105989. https://doi.org/10.1016/j.asoc.2019.105989

Watkins, A. B., & Boggess, L. C. (2002). A resource limited artificial immune classifier. In *Proceedings of the Congress on Evolutionary Computation* (pp. 926–931). https://doi.org/10.1109/CEC.2002.1007049

Zhou, Q., Li, R., Xu, L., et al. (2023). Towards explainable meta-learning for DDoS detection. *SN Computer Science*. https://doi.org/10.1007/s42979-023-02383-y