Shapley Regression for Rare Disease Diagnosis Support: A Case Study on APDS

Safa Alsaidi[1,4,*], **Tomás Brogueira**[2,*], **Nizar Mahlaoui**[3], **Marc Vincent**[4], **Guilherme Pelegrina**[5], **Nicolas Garcelon**[4], **Adrien Coulet**[1] and **Miguel Couceiro**[2]
[1]Inria, Inserm, UPC, HeKA U1346, Paris, France
[2]Técnico, University of Lisbon, INESC-ID, Lisbon, Portugal
[3]Necker Enfants Malades University Hospital, AP-HP, Paris, France
[4] Data Science Platform, INSERM UMR1163, Imagine Institute, UPC, Paris, France
[5]Engineering School, Mackenzie Presbyterian University, São Paulo, Brazil
*The first two named authors had equal contribution.
{safa.alsaidi, adrien.coulet}@inria.fr, tomas.brogueira@tecnico.ulisboa.pt, nizar.mahlaoui@aphp.fr, {marc.vincent, nicolas.garcelon}@institutimagine.org, guilherme.pelegrina@mackenzie.br, miguel.couceiro@inesc-id.pt

Abstract

Activated PI3Kδ Syndrome (APDS) is a rare genetic immune disorder caused by variants in PIK3CD or PIK3R1, with highly heterogeneous symptoms that often delay diagnosis. Early recognition is hampered by overlapping clinical presentations and limited clinician awareness, motivating systematic, data-driven approaches to detect APDS-associated phenotypic patterns in routine electronic health records. Traditional linear scoring systems cannot capture complex symptom interactions, while deep learning models, though expressive, often lack interpretability. To bridge this gap, we propose *Shapley regression*, a novel game-theoretic model replacing the linear predictor with a k-additive cooperative game, explicitly modeling co-occurrence of symptoms while maintaining the transparency and convexity of logistic regression. We carry out an empirical study of our lightweight method on eight public biomedical datasets, showing that a 2-additive model with ℓ_2 regularization achieves an optimal trade-off between predictive power and noise robustness. We apply it to a real-world cohort of 222 patients, on which Shapley regression accurately distinguished APDS cases from matched controls, confirming and validating phenotypes known to be associated with APDS, and facilitating the exploration of pairwise interactions between symptoms, validated by clinical experts.

1 Introduction

Rare primary immunodeficiencies pose a substantial diagnostic challenge due to their low prevalence, heterogeneous presentation, and overlap with common immune disorders [Faviez *et al.*, 2020]. Activated PI3Kδ Syndrome (APDS) exemplifies these difficulties. Caused by pathogenic variants in the PIK3CD or PIK3R1 genes, APDS affects an estimated 1–2 individuals per million [Lougaris *et al.*, 2024; Thouenon *et al.*, 2021]. However, its true global prevalence is unknown and likely underestimated due to underdiagnosis and incomplete registry reporting [Maccari *et al.*, 2023; Mahlaoui *et al.*, 2017]. Moreover, APDS was only genetically characterized in 2013 [Angulo *et al.*, 2013]. Confirmed cases demonstrate that APDS patients can develop recurrent infections, lymphoproliferation, autoimmunity, bronchiectasis, and an increased risk of malignancy; however, the specific combination and severity of symptoms vary widely among individuals [Zhu *et al.*, 2024; Singh *et al.*, 2020]. Because many clinical manifestations overlap with other primary immunodeficiencies, early recognition is difficult and often delayed [Lougaris *et al.*, 2024; Tavakol *et al.*, 2020]. Moreover, limited awareness among healthcare providers frequently results in prolonged morbidity and missed opportunities for targeted interventions before a definitive genetic explanation is available [Vanselow *et al.*, 2023]. Given this, it is likely that a substantial number of APDS cases remain undiagnosed worldwide, further obscuring the true disease burden and delaying access to appropriate genetic testing and care.

From an ML perspective, diagnosing rare diseases like APDS presents a specific "small data, high complexity" dilemma. On the one hand, traditional linear models such as standard Logistic Regression offer stable learning on small datasets but are inherently limited by their purely additive structure (unless one includes interaction terms manually); they assume that each symptom contributes independently and cannot capture situations in which the joint presence of two symptoms conveys more information than the sum of their individual effects. On the other hand, high-capacity nonlinear models (like neural networks) can capture these interactions but are prone to severe overfitting on small cohorts and lack the transparency required for clinical adoption.

To address this challenge, we propose *Shapley regression*, a game-theoretic extension of logistic regression in which the affine form is replaced by a cooperative game. We provide the implementation on our github repository.[1]

[1]https://github.com/tomasbrogueira/shapley_regression

1. **Model Non-Linearity:** Our proposed approach can account for crucial interaction effects (synergy and redundancy) between features.
2. **Maintain Interpretability:** We parameterize the model using Shapley Interaction Indices. This means the model's weights are not abstract values, but direct game-theoretic measures of how much each feature (or feature pair) contributes to the diagnosis.
3. **Ensure Stability:** We utilize a k-additive formulation, limiting interactions to a specific order (e.g., $k = 2$ for pairwise interactions). This drastically reduces the parameter space, preventing the curse of dimensionality.
4. **Theoretical guarantees and explanations by design:** Our lightweight solution can account for the optimal trade-off between model performance and complexity. Moreover, its parameters correspond to importance criteria and their usefulness is attested on the medical domain, through an empirical study on the detection of APDS diganosis.

For extensive benchmarking, validation, and ablation studies, see [Alsaidi *et al.*, 2026], which provides additional insights into the model's behaviour and game-theoretic structure.

2 Related Work

Rare diseases affect hundreds of millions of individuals worldwide, yet they remain vastly underdiagnosed, often taking years to be correctly identified due to their low prevalence, clinical heterogeneity, and symptom overlap with common conditions [Phillips *et al.*, 2024; Alves *et al.*, 2016]. In recent years, machine learning (ML) and artificial intelligence (AI) have been increasingly explored to address these challenges [Topol and Verghese, 2019; Rajkomar *et al.*, 2019]. A 2020 scoping review identified over 200 studies applying ML methods to 74 different rare diseases, with a strong emphasis on diagnostic and prognostic tasks [Schaefer *et al.*, 2020]. However, the review also highlighted persistent methodological limitations, including small sample sizes, severe class imbalance, and a lack of external validation or systematic comparison with clinical expertise, underscoring the need for more robust and clinically grounded approaches. Despite the increasing interest in computational diagnosis of rare diseases generally, there is limited published work specifically addressing ML based classification of primary immunodeficiency disorders such as APDS. Current diagnostic practice primarily relies on clinical evaluation combined with genetic testing, while only a limited number of proof-of-concept studies in related primary immunodeficiency disorders have explored the use of ML models and feature-importance analyses to support differential diagnosis [Méndez Barrera *et al.*, 2023].

A fundamental obstacle in applying ML to rare diseases such as APDS is data scarcity and imbalance. Many rare conditions have fewer than 100 documented cases in electronic health records (EHRs), making it difficult to train models that generalize beyond the training cohort [Mitani and Haneuse, 2020]. To mitigate these challenges, prior work has proposed pipelines that integrate phenotype extraction, semantic similarity measures, and imbalance-aware learning strategies to amplify informative clinical signals while reducing noise from dominant control populations [Faviez *et al.*, 2024; Patrick *et al.*, 2022]. Equally critical is the issue of explainability and clinical credibility. High-performing black-box models often fail to provide transparent reasoning for their predictions, thus limiting clinical trust and adoption [Raposo, 2025]. As a result, there has been increasing interest in interpretable modeling approaches that produce human-readable explanations, align with domain knowledge, and can be evaluated in conjunction with expert clinical review [Ronicke *et al.*, 2019; Rider *et al.*, 2025]. Such explainable approaches are particularly valuable in diseases like APDS, where diagnostic patterns are subtle and require specialized immunological expertise to interpret.

Within this context, capacities and their associated Choquet integrals offer a principled framework for modeling complex feature interactions while maintaining interpretability. These aggregation functions have been extensively studied in multi-criteria decision making (MCDM) [Torra and Narukawa, 2007; Grabisch *et al.*, 2008], and their connections to cooperative game theory—particularly through the Shapley value and interaction indices—provide powerful tools for quantifying feature importance and interdependencies in a real-world setting. The integration of Choquet integrals as an aggregation function for a classification problem, in specific for logistic regression, was first proposed by [Tehrani *et al.*, 2011], who introduced "Choquistic regression". Their approach yielded a monotone, interaction-aware model that outperformed standard logistic regression while maintaining interpretability. More recently, [Pelegrina *et al.*, 2025] proposed a game-theoretic logistic regression framework that eliminated the need for monotonicity constraints while still capturing complex feature interactions. Various parameterizations and computational efficiency improvements have since been explored to address the practical challenges of learning these models [Tehrani, 2021; Pelegrina *et al.*, 2022].

3 Proposed Method

To tackle the "small data, high complexity" challenge inherent to rare disease diagnosis, we propose **Shapley Regression (SR)**. This framework extends the standard logistic regression (LR) model by incorporating non-linear interactions derived from Cooperative Game Theory.

LR assumes that features contribute additively to a diagnosis[2], however, biological systems are rarely additive. Instead, they often exhibit *synergistic* effects (*e.g.*, two features jointly provide much stronger evidence for a diagnosis than either feature alone) or *redundant* effects (*e.g.*, two features convey overlapping information, such that the presence of one reduces the additional diagnostic value of the other). To capture this while maintaining the interpretability of a linear model, we use cooperative games that can be parametrized by "Shapley Interaction Indices" [Grabisch, 2016]. In this section we recall the necessary background of our proposed model and

[2]I.e., the risk of Feature A + Feature B is Risk A + Risk B.

the subsequent theoretical analysis, and provide further details in Appendix A [Alsaidi *et al.*, 2026].

3.1 Shapley Regression Formulation

We consider a binary classification task with a feature vector $\mathbf{x} = (x_1, \ldots, x_n)$ normalized to $[0,1]^n$. We define the probability of the positive class (APDS) as:

$$P(y = 1 \mid \mathbf{x}) = \sigma\left(\beta + \sum_{A \subseteq \mathcal{F}, A \neq \emptyset} I(A) \cdot \phi_A(\mathbf{x})\right),$$

where $\mathcal{F} = \{1, \ldots, n\}$ is the set of clinical features, $\sigma(\cdot)$ is the sigmoid function, and β is the bias. Here, the learnable coefficients $I(A)$ are exactly the **Shapley Interaction Indices** associated with a subset of features A. The basis functions $\phi_A(\mathbf{x})$ are constructed to allow this direct game-theoretic interpretation. For any coalition of features A, the basis function is defined as [Pelegrina *et al.*, 2022]:

$$\phi_A(\mathbf{x}) = \sum_{C \subseteq A} \frac{(-1)^{|A|-|C|}}{|A| - |C| + 1} \min_{i \in C} x_i. \tag{1}$$

This formulation allows us to learn a non-linear decision boundary where the weights directly represent the marginal contribution of features (and feature combinations) to the positive class probability.

3.2 2-Additive Complexity and Interpretation

A fully general Choquet integral would require 2^n parameters, which is computationally intractable and prone to overfitting. To ensure stability on small medical cohorts, we enforce k**-additivity**, specifically setting $k = 2$. This choice provides the optimal performance-complexity as empirically attested. This restricts the model to learning only **main effects** (singletons, $|A| = 1$) and **pairwise interactions** ($|A| = 2$). Higher-order interactions are forced to zero. This reduces the parameter space from exponential to quadratic ($\mathcal{O}(n^2)$), specifically $n + \binom{n}{2}$ coefficients. The resulting model offers a transparent interpretation for clinicians:

Main Effects ($I(\{i\})$): The independent contribution of clinical feature i (analogous to standard LR coefficients).

Pairwise Interactions ($I(\{i,j\})$):

Synergy: $I > 0$. The presence of *both* features increases the probability of APDS more than the sum of their individual parts (complementarity).

Redundancy: $I < 0$. The features provide overlapping information; observing the second adds less diagnostic value if the first is already present.

Independence: $I \approx 0$. The features act independently.

3.3 Model Training

To train the model, we construct a design matrix $\Phi \in \mathbb{R}^{N \times p}$ where each column corresponds to a basis function $\phi_A(\mathbf{x})$ for $|A| \leq 2$. The model is then trained as a standard logistic regression on this expanded feature space by minimizing the regularized negative log-likelihood:

$$\mathcal{L}(\boldsymbol{\theta}, \beta) = -\sum_{i=1}^{N} \Big[y^{(i)} \ln \hat{p}^{(i)} + \big(1 - y^{(i)}\big) \ln\big(1 - \hat{p}^{(i)}\big) \Big] + \lambda \|\mathbf{I}\|_q,$$

where $\mathbf{I}$ is the vector of Shapley indices. We employ ℓ_2 regularization ($q = 2$, Ridge), which encourages smoothness and, as shown in the following section, guarantees the stability of the explanations.

3.4 Theoretical Generalisation and Stability Analysis

To interpret the empirical performance of the k-additive Choquet model, we analyse the generalisation gap—the difference between expected and empirical risk—using Structural Risk Minimization (SRM). We examine three distinct complexity regimes. Our goal is to explain why regularised models (Regimes 2 and 3) succeed in high-order settings where unregularised models (Regime 1) fail.

Notation. Let $\mathcal{H}_k$ denote the hypothesis class of k-additive Choquet integrals defined with respect to a cooperative game v. The complexity of this class depends on the dimension $D_k = \sum_{j=1}^{k} \binom{n}{j}$. We seek to bound the generalisation gap $R(h) - \hat{R}_N(h)$, where R and $\hat{R}$ denote respectively the true and empirical risk (error rate).

Validation Protocol for Theoretical Bounds

To empirically validate the generalisation bounds derived in the following sections, we conducted controlled experiments using **random synthetic data** ($\mathbf{X} \sim U[0,1]^n, y \sim$ Bernoulli(0.5)). Using pure noise ensures that our measurements capture the intrinsic **capacity** and **stability** of the hypothesis class $\mathcal{H}_k$ itself, independent of any signal in the data. This "worst-case" testing confirms that the bounds hold for *any* data distribution. Unless otherwise stated, validation plots use $N = 100$ samples and $n = 10$ features.

Regime 1: Unregularised (Baseline VC Bound)

In the absence of constraints, the model's capacity is determined purely by the count of its parameters. The generalisation gap is bounded using the Vapnik-Chervonenkis (VC) dimension $d_{\text{VC}} = D_k + 1$: $R(h) - \hat{R}_N(h) \leq \mathcal{O}\left(\sqrt{\frac{D_k}{N}}\right)$.

Limitation: This bound scales linearly with $\sqrt{D_k}$. Since D_k grows polynomially with n ($\mathcal{O}(n^k)$), this bound becomes loose rapidly. It predicts that without regularisation, increasing k requires an exponentially larger dataset to prevent overfitting.

Regime 2: ℓ_1 Regularisation (Rademacher Bound)

When applying ℓ_1 regularisation, we restrict the hypothesis class to functions with bounded norm $\|\boldsymbol{\theta}\|_1 \leq B$. The VC bound is overly pessimistic here because it ignores this constraint. A tighter bound is provided by the *Rademacher Complexity*, which measures the ability of the constrained class to fit random noise: $R(h) - \hat{R}_N(h) \leq \mathcal{O}\left(B\sqrt{\frac{\ln(D_k)}{N}}\right)$.

Comparison: The crucial improvement over the VC bound is that the dependence on dimension D_k improves from **polynomial** ($\sqrt{D_k}$) to **logarithmic** ($\sqrt{\ln D_k}$). This implies that ℓ_1-regularised models are significantly less sensitive to the "curse of dimensionality." Mathematically, this justifies why we can explore high-order interactions (large k) using Lasso:

the cost of adding extra dimensions is negligible, provided the underlying true model is sparse.

Regime 3: ℓ_2 Regularisation (Algorithmic Stability)
The ℓ_2 penalty (Ridge) induces strict convexity in the loss function. This allows us to move away from counting parameters (VC dimension) and instead analyze the *Algorithmic Stability* of the optimiser. Stability bounds guarantee that if the algorithm is robust to small changes in the training set, it will generalize well.
The Bound: For a loss function with Lipschitz constant L and regularisation strength λ, the generalisation gap is bounded by the stability coefficient β:

$$\mathbb{E}[R(h) - \hat{R}_N(h)] \leq \beta \leq \frac{2L^2}{\lambda N},$$

where the term L represents the Lipschitz constant of the loss function with respect to the parameters. In our context, L scales with the maximum Euclidean norm of the feature vectors, $L \approx \sup_x \|\Phi_k(\mathbf{x})\|_2$.
Comparison: While the numerator L^2 does grow with k (as feature vectors get longer), this bound offers a crucial control mechanism that VC theory lacks: the regularisation parameter λ in the denominator.In the **Unregularised (VC)** regime, the complexity is fixed by the architecture ($\sqrt{D_k}$). Overfitting is inevitable as k grows.In the ℓ_2 **Stability** regime, we can actively counteract the growth of the feature space by increasing λ.

The Effective Dimension Bound (Refined Capacity)
The combinatorial VC bound (D_k) overestimates capacity because the Shapley basis functions are structurally correlated. We propose a tighter bound based on the **Effective Dimension** (d_{eff}).

Definition Let $\mathbf{\Sigma} = \mathbb{E}[\Phi(\mathbf{x})\Phi(\mathbf{x})^\top]$ be the covariance matrix of the feature mapping $\Phi(\mathbf{x})$. The effective dimension, or stable rank, is defined as:

$$d_{\text{eff}} = \frac{(\text{Tr}(\mathbf{\Sigma}))^2}{\text{Tr}(\mathbf{\Sigma}^2)} = \frac{\left(\sum_{i=1}^{D_k} \lambda_i\right)^2}{\sum_{i=1}^{D_k} \lambda_i^2},$$

where λ_i are the eigenvalues of $\mathbf{\Sigma}$.

Result For kernel-based methods (including Ridge regression on the Shapley basis), the refined bound replaces D_k with d_{eff}:

$$R(h) \leq \hat{R}_S(h) + \mathcal{O}\left(\sqrt{\frac{d_{\text{eff}}}{N}}\right).$$

Stability of Game-Theoretic Explanations
While the previous sections established bounds on the predictive error, we now assess the robustness of our model's *explanations*. An advantage of our architecture is that it is parameterised directly in the Shapley basis (Eq. 1). Indeed, unlike models that require post-hoc attribution methods, which often introduce approximation noise, the coefficients of our model are the Shapley Interaction Indices.

We now formalise the connection between the regularisation used during training and the reliability of these game-theoretic explanations.

Proposition 1 (Interaction Index Stability)**.** *Let $\hat{\boldsymbol{\theta}}$ be the estimated parameter tuple, and $\boldsymbol{\theta}^*$ be the true tuple of the underlying game. In the Shapley-basis formulation, where the parameter θ_A, $A \subseteq [n]$, is the Shapley Interaction Index $I(A)$, the ℓ_2 regularisation provides a direct bound on the explanation variance. More precisely, if the parameter estimation error is bounded by ε in the Euclidean norm, then the error in the interaction index for any coalition A is bounded by ε.*

Proof. We leverage the fundamental property of the Shapley basis parameterisation [Grabisch, 1997]: there is an identity mapping between the regression coefficient θ_A and the Shapley Interaction Index $I(A)$ for any coalition A. Therefore, an error in the explanation is exactly an error in the corresponding coefficient: $|\hat{I}(A) - I^*(A)| = |\hat{\theta}_A - \theta^*_A|$.

We now relate this individual error to the total model error. Let $\mathbf{e} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ be the vector of estimated errors. The magnitude of any single component $|e_i|$ is strictly bounded by the Euclidean norm (ℓ_2-norm) of the vector, i.e., $|e_i| \leq \|\mathbf{e}\|_2$. Since ℓ_2 regularisation explicitly constraints the global parameter norm $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \epsilon$, it follows immediately that:

$$|\hat{I}(A) - I^*(A)| \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \epsilon.$$

This shows that constraining the global model complexity via Ridge regression imposes a hard ceiling on the instability of every individual interaction explanation. □

This result establishes a theoretical isomorphism between the model's algorithmic stability and its interpretability. While standard stability bounds (Regime 3) only guarantee robust *predictions*, Proposition 1 guarantees robust *explanations*. Because ℓ_2 regularisation explicitly minimizes the norm $\|\hat{\boldsymbol{\theta}}\|_2$, it imposes a hard ceiling on the variance of every Shapley interaction index (see Appendix E [Alsaidi *et al.*, 2026]). As empirically demonstrated in Figure 1, tightening the regularisation (lower C) linearly reduces the sensitivity β, which directly translates to trustworthy, noise-invariant game-theoretic explanations.

4 Experimental Design: APDS Cohort

In this section, we detail the data sourcing, phenotype extraction, and classification framework for the detection of APDS.

4.1 Data Source and Cohort Selection

We used French EHRs from the Necker hospital data warehouse (Dr. Warehouse) [Garcelon *et al.*, 2018] and identified 222 patients (29 APDS and 193 Controls) from the Immunology department, with at least one hospital admission between 2018 and 2023. For APDS patients, the diagnosis date was approximated using the first mention of relevant terms in clinical notes, including: "APDS", "APDS1", "APDS2", "PIK3CD", "PIK3R1", "PI3K", and "PI3KCD". To avoid information leakage, all notes from the diagnosis date onward were excluded. In addition, we also excluded mentions of the phenotypic term "primary immunodeficiency," as this typically indicates that a diagnosis had already been established.

Controls were selected from the same department and matched to APDS cases based on sex, age, follow-up duration

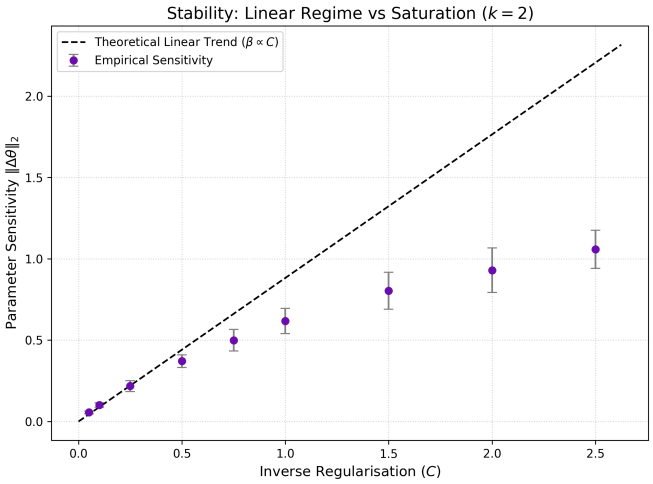

Figure 1: **Empirical Validation of Algorithmic Stability.** We measure sensitivity (Euclidean shift $\|\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}_{S'}\|_2$ after flipping a single label) across varying regularisation strengths C. The dashed line represents the **theoretical linear bound** ($\beta \propto C$) derived in Regime 3. Crucially, the empirical points follow this linear trend in the high-regularisation regime (small C) but naturally deviate and **saturate** around $C \approx 1.5$ as the logistic gradients vanish. This confirms that the model adheres to the stability bound while being strictly *more* robust than predicted in low-regularisation settings.

($\pm$10%), time between first and last visit, number of clinical events, and care pathways using a structured query, with a 2:1 control-to-case ratio.

4.2 Preprocessing

We applied interquartile range (IQR) filtering to the number of clinical documents on the full dataset. Based on the IQR-derived lower and upper bounds, we retained only the first 20 notes in chronological order for each patient, and excluded patients with fewer than three notes to ensure sufficient clinical history for classification.

4.3 Automated Phenotyping

As structured phenotype data was unavailable, clinical notes were automatically phenotyped following [Vincent *et al.*, 2022]. This approach combines a dictionary-based search for phenotypic terms with context-aware detection using deep learning. The dictionary relies on a subset of terms from the UMLS 2023AB vocabulary. The deep learning modules further enhance the extraction by performing disambiguation, subject attribution, and negation detection. This process yielded a feature space of 2,563 unique phenotypes. Consequently, each patient i is represented by a binary vector $\mathbf{x}_i \in \{0,1\}^{2563}$, where $x_{i,j} = 1$ indicates the presence of phenotype j in the patient's history. No dimensionality reduction was applied to preserve interpretability.

4.4 Classification Framework

We formulated APDS detection as a binary classification task ($y = 1$ for APDS, $y = 0$ for controls).

Baselines. We compared the proposed 2-additive SR with: **Logistic Regression (LR)** [Hosmer Jr *et al.*, 2013], **Random Forest (RF)** [Breiman, 2001], **XGBoost** [Chen and Guestrin, 2016], and a **feed-forward neural network (NN)** [Goodfellow *et al.*, 2016].

Protocol. All models were trained using **nested 5-fold cross-validation**, with 3-fold inner-loop used for hyperparameter tuning and outer loop assessed performance. Due to the class imbalance (29 *vs.* 193), we optimized for F1-score and report balanced accuracy, ROC-AUC, and PR-AUC.

5 Results and Discussion

This section presents the empirical evaluation of SR, covering predictive performance, interpretability, and interaction.

5.1 Predictive Performance

Table 1 reports the classification performance across the nested cross-validation. The 2-additive SR model achieved the highest performance across several metrics, specifically reaching a Balanced Accuracy of 0.941, outperforming the linear baseline (0.907) and the non-linear XGBoost (0.884). Crucially for rare disease screening, the SR model demonstrated the highest Sensitivity (over 0.933), meaning it successfully identified the highest proportion of actual APDS patients. While RF achieved higher specificity (fewer false positives), it did so at the cost of missing significant true cases (Sensitivity of 0.613). XGBoost demonstrated intermediate behavior, with higher variance across folds and lower balanced accuracy compared to the SR model.

Despite comparable ROC-AUC values across models, differences in PR-AUC and F1-score highlight the importance of metrics that explicitly account for class imbalance in this setting. These results indicate that models achieving similar global discrimination may nonetheless differ substantially in their ability to detect rare cases. By optimizing for F1-score on the imbalanced cohort, our method strikes a clinically safer balance, prioritizing the detection of rare cases while maintaining high specificity (0.948).

Model	Balanced Acc.	Sensitivity (Recall)	Specificity	Precision	F1-Score	ROC-AUC	PR-AUC
LR	$\underline{0.907 \pm 0.063}$	$\underline{0.860 \pm 0.127}$	$\underline{0.953 \pm 0.050}$	0.789 ± 0.212	$\underline{0.798 \pm 0.123}$	$\underline{0.985 \pm 0.013}$	$\underline{0.941 \pm 0.044}$
RF	0.804 ± 0.052	0.613 ± 0.113	$\mathbf{0.995 \pm 0.010}$	$\mathbf{0.960 \pm 0.080}$	0.737 ± 0.060	$\underline{0.962 \pm 0.043}$	$\underline{0.893 \pm 0.071}$
XGBoost	0.884 ± 0.081	0.793 ± 0.163	0.974 ± 0.001	0.810 ± 0.033	$\underline{0.795 \pm 0.104}$	0.914 ± 0.061	0.853 ± 0.086
NN	$\underline{0.916 \pm 0.057}$	$\underline{0.893 \pm 0.088}$	$\underline{0.938 \pm 0.053}$	0.723 ± 0.198	$\underline{0.784 \pm 0.138}$	$\mathbf{0.988 \pm 0.012}$	$\mathbf{0.952 \pm 0.035}$
SR (k=2)	$\mathbf{0.941 \pm 0.064}$	$\mathbf{0.933 \pm 0.133}$	$\underline{0.948 \pm 0.046}$	0.760 ± 0.168	$\mathbf{0.821 \pm 0.121}$	$\underline{0.982 \pm 0.019}$	$\underline{0.936 \pm 0.062}$

Table 1: Detailed clinical performance metrics across models using ℓ_2 regularization. Results are reported as mean $\pm$ standard deviation over 5-nested cross-validation folds. Best-performing results are highlighted in bold. Underlined values indicate competing methods with comparable performance when standard deviation intervals overlap.

Finally, we address practical clinical deployment constraints through empirical evaluation on several benchmark datasets with varying characteristics. Results on the Pima Indians Diabetes dataset[3] are provided in Appendix C [Alsaidi *et al.*, 2026]. The proposed model proved computationally lightweight for both training and inference on a standard CPU-based system, supporting its feasibility for broad clinical deployment without specialized hardware. CPU configuration details for the APDS experiments are provided in Appendix G [Alsaidi *et al.*, 2026].

5.2 Shapley Regression Interaction Analysis

To analyze relationships among clinical phenotypes, we leveraged the game-theoretic structure of the SR model to decompose predictions into main effects and pairwise interactions. Firstly, we extracted Shapley-based coefficients from each fold of the nested cross-validation to identify phenotypes contributing to positive APDS predictions. The top 20 phenotypes, ranked by mean coefficients across folds, were subsequently reviewed by a clinical expert to distinguish known APDS manifestations (Appendix G [Alsaidi *et al.*, 2026]). Secondly, we examined pairwise interactions learned by the SR model. For each trained model, the interaction coefficients were arranged into a symmetric interaction matrix, where entry (i, j) represents the learned interaction strength between phenotypes i and j. These interaction matrices were then aggregated across cross-validation folds by computing the mean interaction strength for each phenotype pair, yielding a consensus interaction matrix that reflects stable interaction patterns across training splits.

We focused the analysis on the most influential phenotypes, we computed a global interaction strength score for each phenotype by summing the absolute values of its interactions with all other phenotypes. The top 30 phenotypes with the highest total interaction strength were selected, and the interaction matrix was restricted to this subset. To ensure robustness, we further assessed the stability of each interaction by calculating the fraction of folds in which the corresponding interaction coefficient was non-zero. Interactions observed in fewer than 70% of folds were considered unstable and excluded from visualization. For the selected set of phenotypes, we examined both individual main effects, illustrated as a bar plot (Figure 2), and pairwise phenotype interactions, represented as an interaction matrix (Figure 4 [Alsaidi *et al.*, 2026]). The filtered interaction matrix was visualized as a heatmap, with color intensity indicating the mean interaction strength and sign across folds. Both the main effects plot and the interaction heatmap were jointly reviewed with a clinical expert to evaluate the plausibility of the interaction patterns identified by the SR model. Among the 30 phenotypes analyzed, 17 were identified as core clinical manifestations associated with APDS.

As shown in Figure 2, *tetanus* and *rhonchi* emerged as positive predictors of APDS. This finding aligns with previous analyses discussed in Appendix G [Alsaidi *et al.*, 2026], where *tetanus* was attributed to vaccination records, and *rhonchi* reflected respiratory signs associated with other symptoms. Interestingly, two phenotypes recognized by the clinical expert as APDS-associated (*pleural effusion* and *allergy*) appeared to help classifying individuals as not having APDS. This counterintuitive pattern likely reflects overlapping phenotypes present in both APDS and control cohorts.

To further explore these relationships, the filtered interaction matrix was visualized as a heatmap (Figure 4 [Alsaidi *et al.*, 2026]), illustrating the pairwise interactions among the selected phenotypes. In this heatmap, five core APDS phenotypes—*immunodeficiency*, *splenomegaly*, *lymphopenia*, *lymphadenopathy*, and *superinfection*—displayed predominantly negative interactions with many other phenotypes. Within the SR framework, such negative interactions reflect redundancy rather than antagonism: these features are already highly informative individually, and their co-occurrence with additional phenotypes contributes limited incremental information to the model. This behavior is consistent with clinical manifestations commonly reported for APDS patients in Orphanet [Pavan *et al.*, 2017], indicating that the model appropriately captures their central role in disease characterization. Other APDS-associated phenotypes, including *otorrhea*, *cervical lymphadenopathy*, *bronchitis*, *CD4 T-cell lymphopenia*, *hypogammaglobulinemia*, *bronchiectasis*, *combined immunodeficiency*, *hyperplasia*, *pneumonia*, *allergy*, and *infection*, exhibited a more heterogeneous pattern of positive, negative, and neutral interactions. Notably, several of these phenotypes showed negative interactions with core APDS features, again suggesting partial redundancy with already informative immunological and lymphoproliferative manifestations. This variability indicates that their contribution to APDS characterization might be context-dependent and modulated by the presence of other clinical findings.

In contrast, although *pleural effusion* and *allergy* appeared negatively associated with APDS in isolation, they demonstrated strong positive interactions with phenotypes not traditionally linked to APDS. For example, *pleural effusion*

[3]https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

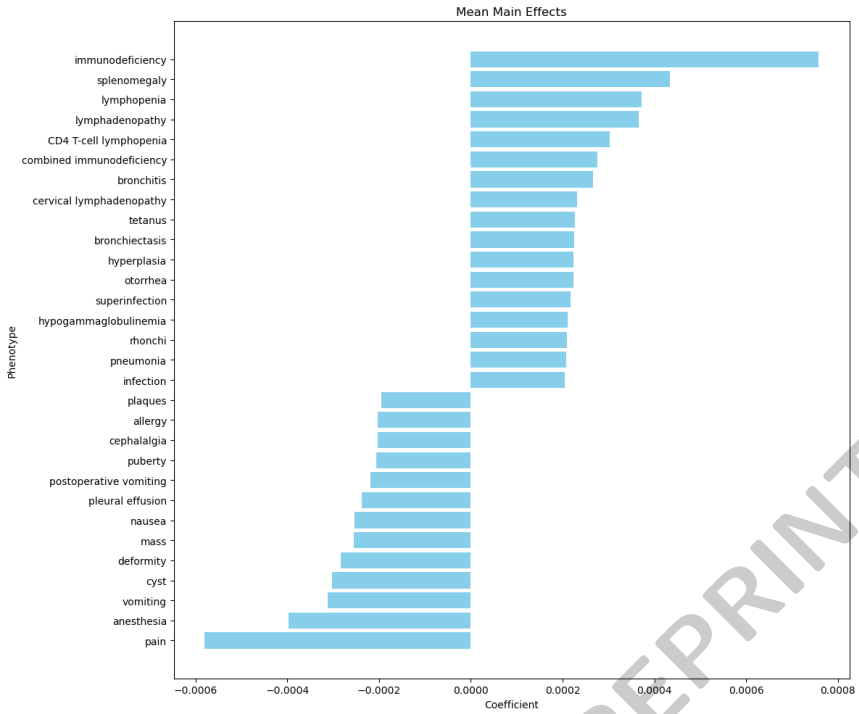

Figure 2: Marginal contribution (main effects) among phenotypes in APDS dataset.

showed positive associations with *pain*, *anesthesia*, and *deformity*, among others. The interaction matrix also highlights the role of phenotypes not classically associated with APDS, including *pain*, *anesthesia*, *cyst*, *deformity*, and *cephalalgia*, which tended to interact positively with multiple other phenotypes, both APDS-associated and non-associated. For instance, positive interactions were observed among non-APDS phenotypes themselves, such as *nausea* and *cephalalgia*. These patterns suggest that such phenotypes may reflect downstream clinical consequences, comorbidities, or procedural contexts rather than primary drivers of the disease. Interestingly, *pain* exhibited widespread positive interactions with both APDS and non-APDS phenotypes, despite not emerging as a strong individual predictor of APDS. This observation is consistent with clinical expectations, as pain is a non-specific symptom whose diagnostic significance depends largely on its context and association with other clinical manifestations rather than its isolated occurrence.

Overall, the interaction analysis distinguishes strong standalone APDS predictors from context-dependent phenotypes whose importance arises through interactions. This highlights the utility of the game-theoretic framework in disentangling redundant, complementary, and context-dependent clinical signals, providing an interpretable view of symptom co-occurrence patterns in APDS.

6 Conclusion

We presented a framework for interpretable rare disease diagnostic support, evaluated on APDS. By employing a 2-additive Shapley model with robust regularization, we effectively addressed the high dimensionality and data scarcity inherent to this domain. Our results show that SR is competitive with state-of-the-art models such as XGBoost in predictive accuracy. In our use case, it also improves sensitivity, albeit with a modest decrease in specificity, while providing a key advantage: explanatory global Shapley Interaction Indices that offer insights into phenotype interactions.

This framework confirms known APDS-associated phenotypes reported in Orphanet [Pavan *et al.*, 2017] and by clinicians (Appendix G [Alsaidi *et al.*, 2026]). More importantly, it enables systematic analysis of how phenotypes jointly contribute through interactions rather than in isolation. This interaction-focused analysis provides a complementary perspective on symptom co-occurrence patterns not captured by traditional models and has the potential to improve our understanding of disease mechanisms and clinical manifestations. Although cohort size limits predictive conclusions, expanding the APDS patient pool would enable more robust validation, improved performance, and greater confidence in clinical relevance. These findings suggest that game-theoretic models such as SR serve as useful exploratory tools for understanding complex phenotypic interactions in rare diseases.

As future work, we plan to further investigate theoretical bounds (Sec. 3.4), considering dataset characteristics and capacity distributions. Also, we saw that SR outperforms the other models considered in terms of time complexity, when the number of features is reasonable (see Appendix C [Alsaidi *et al.*, 2026]). However, sampling techniques should be developed for high number of features. Finally, negative interactions, which are of particular interest for medical applications like APDS detection, will be explored in collaboration with our clinical team.

Ethical Statement

We propose a lightweight approach that achieves performance competitive while remaining computationally efficient. Notably, the proposed method can be easily implemented on standard personal computers using only CPU resources.

This research was approved by the Necker Hospital Data Privacy Officer under the registration number #2026 04171640. The study relied exclusively on collected clinical data that were fully de-identified prior to analysis, and all computational findings were reviewed in conjunction with expert clinical interpretation. A language model (ChatGPT) was used to assist in editing text for clarity and readability.

Acknowledgments

This work was partially supported by the French National Research Agency (ANR)[4] project "Analogies: from theory to tools and applications" (AT2TA), ANR-22-CE23-0023, and partially funded under the "Investissements d'Avenir" program, ANR-10-IAHU-01. This work was also supported by Portuguese national funds through FCT under project UIDB/50021/2020 (DOI: 10.54499/UIDB/50021/2020), and by the São Paulo Research Foundation (FAPESP) under the grants 2025/00700-0 and 2025/25154-8. Additionally, the work was partially supported by the project Center for Responsible AI, reference C628696807-00454142.

References

[Alsaidi *et al.*, 2026] Safa Alsaidi, Tomás Brogueira, Nizar Mahlaoui, Marc Vincent, Guilherme Pelegrina, Nicolas Garcelon, Adrien Coulet, and Miguel Couceiro. Shapley regression for rare disease diagnosis support: a case study on apds, 2026.

[Alves *et al.*, 2016] Rui Alves, Marc Piñol, Jordi Vilaplana, Ivan Teixidó, et al. Computer-assisted initial diagnosis of rare diseases. *PeerJ*, 4:e2211, July 2016.

[Angulo *et al.*, 2013] Ivan Angulo, Oscar Vadas, Fabien Garçon, et al. Phosphoinositide 3-Kinase δ Gene Mutation Predisposes to Respiratory Infection and Airway Damage. *Science*, 342(6160):866–871, November 2013.

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

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[Faviez *et al.*, 2020] Carole Faviez, Xiaoyi Chen, Nicolas Garcelon, et al. Diagnosis support systems for rare diseases: a scoping review. *Orphanet Journal of Rare Diseases*, 15(1):94, December 2020.

[Faviez *et al.*, 2024] Carole Faviez, Marc Vincent, Nicolas Garcelon, et al. Performance and clinical utility of a new supervised machine-learning pipeline in detecting rare ciliopathy patients based on deep phenotyping from electronic health records and semantic similarity. *Orphanet Journal of Rare Diseases*, 19(1):55, February 2024.

[Garcelon *et al.*, 2018] Nicolas Garcelon, Antoine Neuraz, Rémi Salomon, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. *Journal of Biomedical Informatics*, 80:52–63, April 2018.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[Grabisch *et al.*, 2008] Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, et al. *Aggregation Functions*. Cambridge University Press, 2008.

[Grabisch, 1997] Michel Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997.

[Grabisch, 2016] Michel Grabisch. *Set Functions, Games and Capacities in Decision Making*. Springer, 2016.

[Hosmer Jr *et al.*, 2013] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.

[Lougaris *et al.*, 2024] Vassilios Lougaris, Federico Le Piane, Caterina Cancrini, et al. Activated phosphoinositde 3-kinase (PI3Kδ) syndrome: an Italian point of view on diagnosis and new advances in treatment. *Italian Journal of Pediatrics*, 50(1):103, May 2024.

[Maccari *et al.*, 2023] Maria Elena Maccari, Martin Wolkewitz, Charlotte Schwab, et al. Activated phosphoinositide 3-kinase δ syndrome: Update from the ESID Registry and comparison with other autoimmune-lymphoproliferative inborn errors of immunity. *Journal of Allergy and Clinical Immunology*, 152(4):984–996.e10, October 2023.

[Mahlaoui *et al.*, 2017] Nizar Mahlaoui, Jean-Philippe Jais, Pauline Brosselin, et al. Prevalence of primary immunodeficiencies in France is underestimated. *Journal of Allergy and Clinical Immunology*, 140(6):1731–1733, December 2017.

[Mitani and Haneuse, 2020] Aya A. Mitani and Sebastien Haneuse. Small Data Challenges of Studying Rare Diseases. *JAMA Network Open*, 3(3):e201965, March 2020.

[Méndez Barrera *et al.*, 2023] Jose Alfredo Méndez Barrera, Samuel Rocha Guzmán, Elisa Hierro Cascajares, et al. Who's your data? Primary immune deficiency differential diagnosis prediction via machine learning and data mining of the USIDNET registry. *Clinical Immunology*, 255:109759, October 2023.

[Patrick *et al.*, 2022] Matthew T. Patrick, Redina Bardhi, Wei Zhou, et al. Enhanced rare disease mapping for phenome-wide genetic association in the UK Biobank. *Genome Medicine*, 14(1):85, August 2022.

[Pavan *et al.*, 2017] Sonia Pavan, Kathrin Rommel, Mateo Marquina, et al. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLOS ONE*, 12(1):e0170365, January 2017.

[4] Agence Nationale de la Recherche.

[Pelegrina *et al.*, 2022] Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, and Michel Grabisch. A k-additive choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artificial Intelligence*, 325:104014, 2022.

[Pelegrina *et al.*, 2025] Guilherme Dean Pelegrina, Miguel Couceiro, Cristiano Torezzan, et al. Game theoretic extensions of logistic regression. *Theory Decis.*, 99(1-2):71–95, 2025.

[Phillips *et al.*, 2024] Christine Phillips, Anne Parkinson, Tergel Namsrai, et al. Time to diagnosis for a rare disease: managing medical uncertainty. A qualitative study. *Orphanet Journal of Rare Diseases*, 19(1):297, August 2024.

[Rajkomar *et al.*, 2019] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, April 2019.

[Raposo, 2025] Vera Lúcia Raposo. The fifty shades of black: about black box AI and explainability in healthcare. *Medical Law Review*, 33(1):fwaf005, January 2025.

[Rider *et al.*, 2025] Nicholas L. Rider, Yingya Li, Aaron T. Chin, et al. Evaluating large language model performance to support the diagnosis and management of patients with primary immune disorders. *Journal of Allergy and Clinical Immunology*, 156(1):81–87, July 2025.

[Ronicke *et al.*, 2019] Simon Ronicke, Martin C. Hirsch, Ewelina Türk, et al. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. *Orphanet Journal of Rare Diseases*, 14(1):69, December 2019.

[Schaefer *et al.*, 2020] Julia Schaefer, Moritz Lehne, Josef Schepers, et al. The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases*, 15(1):145, December 2020.

[Singh *et al.*, 2020] Ankita Singh, Vibhu Joshi, Ankur Kumar Jindal, et al. An updated review on activated PI3 kinase delta syndrome (APDS). *Genes & Diseases*, 7(1):67–74, March 2020.

[Tavakol *et al.*, 2020] Marzieh Tavakol, Mahnaz Jamee, Gholamreza Azizi, et al. Diagnostic Approach to the Patients with Suspected Primary Immunodeficiency. *Endocrine, Metabolic & Immune Disorders - Drug Targets*, 20(2):157–171, February 2020.

[Tehrani *et al.*, 2011] Afshin Fallah Tehrani, Weiwei Cheng, and Eyke Hüllermeier. Choquistic regression: Generalizing logistic regression using the Choquet integral. In *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-11)*, pages 868–875. Atlantis Press, 2011.

[Tehrani, 2021] Afshin Fallah Tehrani. Heuristics-based learning approach for choquistic regression models. *Pattern Recognition Letters*, 149:137–142, 2021.

[Thouenon *et al.*, 2021] Romane Thouenon, Nidia Moreno-Corona, Lucie Poggi, et al. Activated PI3Kinase Delta Syndrome—A Multifaceted Disease. *Frontiers in Pediatrics*, 9:652405, June 2021.

[Topol and Verghese, 2019] Eric J. Topol and Abraham Verghese. *Deep medicine: how artificial intelligence can make healthcare human again*. Basic Books, New York, NY, first edition edition, 2019.

[Torra and Narukawa, 2007] Vicenç Torra and Yasuo Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators*. Cognitive Technologies. Springer-Verlag Berlin Heidelberg, 2007.

[Vanselow *et al.*, 2023] Sven Vanselow, Volker Wahn, and Catharina Schuetz. Activated PI3Kδ syndrome – reviewing challenges in diagnosis and treatment. *Frontiers in Immunology*, 14:1208567, July 2023.

[Vincent *et al.*, 2022] Marc Vincent, Maxime Douillet, Ivan Lerner, Antoine Neuraz, et al. Using Deep Learning to Improve Phenotyping from Clinical Reports. In Paula Otero, Philip Scott, Susan Z. Martin, and Elaine Huesing, editors, *Studies in Health Technology and Informatics*. IOS Press, June 2022.

[Zhu *et al.*, 2024] Ke Zhu, Qifan Li, Lingli Han, et al. Activated phosphoinositide 3-kinase delta syndrome: Pathogenesis, clinical manifestations, and treatment. *Pediatric Discovery*, 2(3):e2504, September 2024.