Opening the ANN-blackbox with explainability components

Xavier Angerri[1][0009−0006−9021−4126], Dario Sacco[2][0009−0002−0989−6678], Xavier Garcia-del-Muro[1][0009−0001−4024−9906], Jordi Pascual-Fontanilles[1][0000−0002−7528−5819], and Karina Gibert[1][0000−0002−8542−3509]

[1] Knowledge Engineering and Machine Learning Group (KEMLg) at Intelligent Data Science and Artificial Intelligence Research Center at Universitat Politècnica de Catalunya—BarcelonaTech
xavier.angerri@upc.edu
[2] University of Naples Federico II

Abstract. Artificial Neural Networks (ANNs) are opaque models despite their predictive power. This causes a lack of trust in the ANN models in decision-making contexts where decisions must be justified and cannot rely on good but cryptic predictions. This paper introduces a task-agnostic framework designed to decode the internal logic of an ANN model to provide insights into the role played by explanatory variables so that a higher level of understanding of the model can raise trust in it. Our proposal is based on the analysis of the marginal impact of predictors through intensive computational procedures. To demonstrate the versatility of the proposal, the framework was tested across multiple datasets of different natures, featuring both qualitative and quantitative target variables. The results demonstrate the influence of predictors in the response variable, contributing to a significant improvement of the explanatory models. Current approaches often rely on rankings of the predictors but fail to analyze the form of their relationship with the response variable.

Keywords: Explainable Artificial Intelligence · Artificial Neural Networks· XAI· Functional Sensitivity · Decision Support

1 Introduction

Artificial Neural networks within the field of Artificial Intelligence (AI) are inspired by the human brain, where neurons convert input signals into electrical impulses sent to other neurons, forming sequential networks. The artificial neurons move the natural mechanisms of neurons to weights, activation functions, and message passing, transforming several inputs into numerical values. While the mechanisms of single artificial neurons are well understood and have clear mathematical expressions, the intricate connections of the ANN architecture establish complex functional compositions that preclude human comprehension.

This is a serious limitation when the ANN is used to support decision-making, as explaining the predictions becomes critical to trust in the ANN's results and to make the decision.

Despite Explainable AI advances in interpreting black-box models, most widely used methods primarily focus on feature importance rankings (e.g., Shapley values [1]), although some do capture detailed effects and interactions.

In this paper, we propose ANN-GFIV, a new generalized ANN task-agnostic methodology to enhance explainability by mapping how predictors impact the response variable using an experimental auxiliary dataset (EAD) derived from the original training data. The EAD isolates the marginal impacts of each predictor and proposes graphical representations to reveal the functional relationship with the response variable. Applicable to any ANN architecture, this approach supports both classification and regression tasks, revealing behavioral patterns and influence from thresholds. We also use resampling to ensure stochastic stability across multiple simulations. Regarding the structure of this article, Section 2 reviews the state of the art; Section 3 describes the proposed methodology; Section 4 presents its application across two real-world databases; and Section 5 offers the conclusions.

2 State of the art

2.1 Feature Importance

Feature importance methods explain model predictions by ranking the influence of input variables [2]. They include sensitivity-based approaches, which assess output changes through gradients or partial derivatives [3], and attribution-based approaches, which estimate each feature's contribution to a prediction. However, these methods may be unstable due to input perturbations, baseline choices, preprocessing, or model reparametrization, and may not always faithfully reflect the model's learned behaviour [4].

2.2 Surrogate Models

Surrogate models represent complex model behaviour through simpler and more interpretable representations, facilitating the understanding of how input features relate to outputs [5]. They can be global or local surrogate models.

Global techniques provide a complete description of the model. These techniques require understanding the entire model structure and are, in practice, difficult to implement for complex models [7]. An example is decision trees, as in [8], where they are used to explain convolutional neural networks by capturing the overall behaviour in a human-understandable form.

Local explanations focus on decisions made by specific instances [7]. LIME (Local Interpretable Model-agnostic Explanations) fits a linear model in the neighbourhood of an instance to identify the most influential features. For example, in [9], LIME explains artificial neural network predictions for credit scoring.

Despite their interpretability advantages, some studies [10] have shown that surrogate models do not always reliably capture the underlying causal relationships between input variables and model outputs, which limits the robustness and reliability of the explanations.

2.3 Shapley-Based Methods

Shapley values originate from cooperative game theory to distribute the total gain among players by computing each player's contribution across all possible coalitions [1]. In machine learning, Shapley values computation requires evaluating all possible feature subsets, resulting in exponential computational complexity with respect to the number of input features.

SHAP [11] is a unified framework that connects Shapley values with additive feature attribution methods for machine learning models. SHAP methods can be broadly categorized into *model-agnostic approximation methods* and *model-type-specific approximation methods*. Several works have adapted SHAP to ANNs. For instance, Kernel SHAP reformulates the problem as a linear approximation based on Shapley values [12].

A key limitation of SHAP is the assumption that input features are independent, which can lead to biased or unrealistic feature attributions when strong dependencies exist between variables, as shown in [14].

2.4 Automatic Interpretation

Regarding quantitative outputs toward conceptual frameworks that enhance stakeholder understanding, previous work includes the Traffic Light Panel (TLP) (clustering interpretation based on traffic-light colour semantics methodology), The Annotated TLP (aTLP) (incorporates within-class variability through colour degradation based on the coefficient of variation [15]). The thermometer-based TLP uses thermometer to automatically build the TLP integrating expert-defined semantic ranges [16]

2.5 Explainability in Artificial Neural Networks

Deep learning models can effectively capture non-linear relationships and distributed representations, but their high internal complexity makes them inherently opaque. This opacity limits their interpretability and usability, particularly in decision-support contexts, where understanding the basis of predictions is essential for trust, accountability, and adoption [17]. Explainable ANN approaches seek to reduce this opacity through self-interpretable models and post-hoc explanations [18]; however, both strategies present limitations, either by constraining model complexity or by producing unstable and inconsistent explanations [19].

2.6 Partial Dependence Plots

Partial Dependence Plots (PDPs), introduced by Friedman [20], are post-hoc interpretability tools that estimate the marginal effect of one or two predictors on model predictions by averaging over the remaining features. They have been applied to improve ANN interpretability in contexts such as compositional modelling [21] and nonlinear relationship analysis [22]. However, PDPs assume feature independence, which may lead to misleading interpretations and unrealistic feature combinations when predictors are correlated; they can also be computationally expensive for large datasets or complex models. In contrast, ANN-GFIV isolates marginal effects via resampling and visualizes functional trajectories, revealing thresholds and non-linearities more distinctly than PDPs or SHAP dependence plots.

3 ANN-GFIV methodology

This study introduces ANN-GFIV (Graphical Functional Interpretation of Variables in ANN), a generalized methodology for the post-hoc interpretation of ANN models. The primary contribution of this research is a methodology designed to reveal the decision-making patterns of complex neural networks within the Explainable AI paradigm. The methodology is engineered as an independent, scalable workflow implemented, ensuring statistical robustness and cross-domain relevance. This workflow comprises four key steps: (1) select predictors, (2) construct EAD via resampling, (3) predict with ANN, and (4) visualize functional trajectories.

Technical accuracy alone is insufficient if model behavior remains opaque; thus, identifying and quantifying underlying drivers of a prediction is a fundamental requirement for translating high-dimensional model responses into actionable strategies.

Let $\chi = [X_1, \ldots, X_K]$ be a dataset of n observations and K predictive variables, and let $f : \mathbf{X} \rightarrow \hat{Y}$ be the function previously trained by an ANN that predicts a response variable Y based on χ. The objective of this strategy is to decode the internal logic of the ANN so that the role of the predictors in f becomes explicit. The steps proposed are:

1. Determine the subset T of predictors to be studied.
2. Building experimental domain values for predictors: To build EAD, a subset of valid values for the explanatory variables is defined, so that they cover the entire range of the variable and allow analysis of the functional relationships between X_k and Y. A representative subset of M values is defined as D_k. Specific processing is applied based on the statistical nature of the modalities of the predictors:
   - if X_k is qualitative: To optimize analysis relevance, D_k contains the M most prevalent modalities of X_k.

- if X_k is quantitative: D_k contains M values between min and max calculated as the mean value of the corresponding quantile. If $M = 10$, D_k contains the mean of the values between deciles of X_k and so on.

3. Construction of an Experimental Auxiliary Dataset ($\mathbf{X}_E$) for testing: $\mathbf{X}_E$ is defined to systematically analyze the marginal impact of every predictor X_k on Y:
   (a) Building a baseline experimental testing dataset ($\mathbf{X}_{E,0}$): A random subset of n_ν instances is extracted from the original training dataset $\mathbf{X}$. The n_ν is selected much smaller than n to gain efficiency, but sufficiently high to provide robustness [23]
   (b) Creating the EAD: Instead of building a complete factorial experimental design matrix, which would be too big and long to build, we propose to substitute the values of the T selected X_k variable to be studied by a random trial from the corresponding D_k in the $X_{E,0}$. Let us call χ_E the resulting EAD. In fact, χ_E is a sample of the complete factorial design matrix for the T variables.
4. Use the target ANN model to predict the response $\hat{Y}$ for all instances in χ_E. At this point, it is irrelevant if the combination of X values shown in a row of χ_E makes sense or not, since we are only trying to understand the functional behaviour of predictors.
5. Compute $Y|D_k$: The average of predictions conditional on each specific modality or decile of each variable is calculated. Formally, small new matrices are created for each of the T selected predictors:

$$\chi_k = [D_k, E(Y|D_k = d)], \forall d \in D_k \tag{1}$$

6. Represent the marginal influence of X_k on Y: The concluding phase involves the graphical representation of the χ_k matrices, each of which contains information on the evolution of Y upon changes on X_k, revealing the direction and magnitude with which X_k influences the forecast.
7. Additional robustness layer: To ensure scientific robustness and track the prediction's influence, resampling techniques are used to replicate the process from step 3 to step 5 several times (R times). So step 5 represents the aggregated mean of all curves provided by all independent simulation cycles in a colored curve.

This methodological approach achieves a dual objective:

- Predictive Logic Interpretation: The graphical representation of every χ_k shows how Y changes with respect to the range of values of X_k, so that we can appreciate if the marginal relationship is linear, quadratic, inverse, sinusoidal, or others. Also, the slope of the representation indicates whether X_k has a strong or weak influence on Y since horizontal shapes represent situations where the changes in X_k do not impact Y, and this is associated with independence.
- Stability Verification: The convergence of the independent simulation curves toward the mean ensures that the extracted explanation is structural and not dependent on random fluctuations or sampling noise within the EAD.

This strategy facilitates a rigorous comparative analysis of influence trajectories, aiding in the identification of critical drivers that govern the final forecast.

4 Case Study, Model Architecture, Estimation Process, and ANN-GFIV

ANN-GFIV was evaluated on two case studies involving classification and regression tasks. The first predicted 30-day hospital readmission risk, while the second estimated life expectancy using socioeconomic and health-related indicators.

To generate the EAD defined in section 3 the selected variables T are the most influential variables ($T \ll K$) within the model f. This selection utilizes SHAP values to isolate factors with the highest marginal impact on the prediction. Also, the fixed parameters are $n_\nu = 2000$ and $M = 10$.

4.1 Modelling the probability of readmission after discharge in the adult population

We tested our strategy in hospital readmission. The goal shifts from mere binary classification to quantifying the dynamics by which variables influence post-discharge vulnerability. In adult populations where frailty is a primary determinant [24], risk stems from complex, non-linear interactions, requiring a transition toward identifying individual risk trajectories rather than isolated outcomes [25].

Data Preprocessing, Feature Engineering and Leakage Prevention This study utilizes Hospital Discharge Records from AORN Sant'Anna e San Sebastiano (Caserta), spanning July 2023 to December 2024 (N=12,844). To enhance model stability and reduce statistical noise, *ICD-9-CM* codes were aggregated into macro-areas and *Major Diagnostic Categories* (MDCs) [25]. Beyond standard clinical variables, the dataset was enriched with meteorological data to account for environmental stressors affecting the adult population [24]. Clinical complexity was addressed by integrating the *Charlson Comorbidity Index* [26] and the *Elixhauser Comorbidity Index* [27]. The numerical features were normalized to the Z-score to optimize the convergence of the artificial neural network. Categorical variables were managed via dummy encoding with a 1% collapsing threshold for rare modalities, effectively reducing feature space dimensionality and mitigating noise from low-frequency classes.

Designing the Artificial Neural Network Architecture The model uses a Multilayer Perceptron (MLP) architecture.The network comprises two dense blocks (128 and 64 neurons) using ReLU activation for efficient feature abstraction . To prevent overfitting, we implemented a multi-level regularization strategy: L2 Weight Decay constrains parameter magnitude, while Dropout layers (30% and 25%) reduce neuronal co-adaptation.

Training Strategy and Threshold Calibration To address class imbalance, cost-sensitive learning with smoothed class weights was applied:

$$W_{pos} = \frac{\sqrt{Neg/Pos} + 1}{2} \tag{2}$$

The model was optimized using Adam with Early Stopping and parameter recovery to improve generalization and reduce overfitting. The decision threshold was calibrated by maximizing the $F_{0.5}$ score, prioritizing clinical precision while enforcing Recall $\geq 40\%$ [31].

Model Performance and Operational Metrics The deep neural network showed strong discriminative performance, achieving an AUC-ROC of 0.79 and, at the optimal $F_{0.5}$-based threshold of 0.23, a Recall of 44.9%.

Experimental results: visualization of dynamic trajectories The logic of model f is illustrated in Figure 1, showing how changes in D_K affect the predicted readmission probability. The colored curves represent aggregated mean predictions, while the horizontal dashed line indicates the global baseline risk, enabling the comparison of feature states that increase or decrease vulnerability relative to the global mean.

The model exhibits a clear monotonic increase in vulnerability: the *CCI Score* shifts the risk from a baseline of 0.12 (D_1) to a peak of 0.26 (D_6), representing an exponential positive function. The *Van Walraven Score* rises from 0.11 (D_1) to 0.25 (D_6), showing an exponential function. *Length of Stay* shows a stable risk profile for short-to-medium durations but reveals a sharp threshold effect at D_{10} (147 days), where the probability escalates to 0.21. The model shows an exponential function. *Age* exhibits a downward trajectory, with the highest risk observed in younger adults (0.18 at age 25.5; D_1) and the lowest in the geriatric population (0.11 at age 93; D_{10}), showing a kind of negative logistic function. A similar downward trend is observed for the *Number of Secondary Diagnoses*, decreasing from 0.16 (D_1) to 0.11 (D_8), having a negative hyperbolic function.

4.2 Modelling the prediction of Life Expectancy from Socioeconomic and Health Indicators

We also tested our approach in the context of predicting life expectancy, aiming to capture the influence of several factors on population longevity.

Data Preprocessing, Feature Engineering, and Bias Control This study uses the World Health Organization's Global Health Observatory data repository, obtained from the Kaggle repository [35]. It contains health monitoring variables The data spans from 2000 to 2015 for 193 countries (N = 2,938).

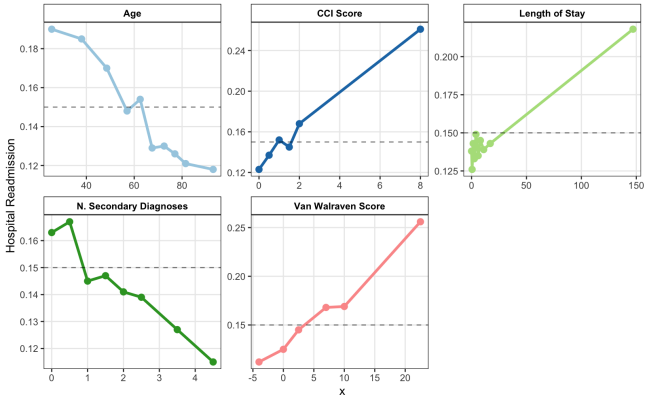


Fig. 1. Marginal influence of X_k on Y from case 1

Artificial Neural Network Architecture Design The model uses a MLP with 3 dense ReLU layers of 64 neurons. Its linear output layer enables continuous prediction, while the architecture balances modelling capacity and simplicity to reduce overfitting risk.

Training Strategy and Model Performance The model was trained using the Adam optimizer, with Mean Squared Error as the loss function and Mean Absolute Error monitored for interpretation in life expectancy years. An internal validation split of 20% was used to control convergence and prevent overfitting, achieving strong predictive performance with $R^2 = 0.916$.

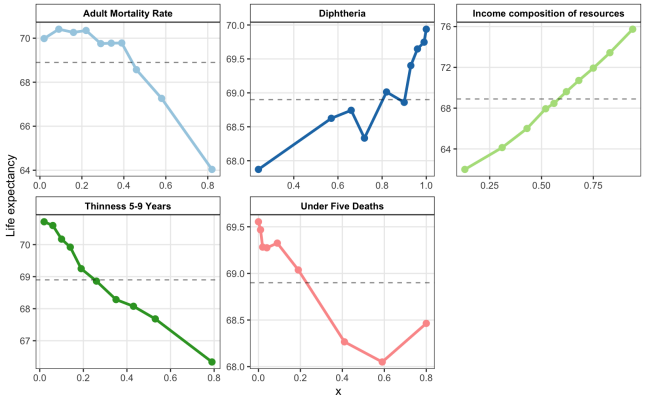


Fig. 2. Marginal influence of X_k on Y of case 2

Experimental Results: Visualization of Dynamic Trajectories The model logic is visualized in Figure 2, illustrating how variations in input features influence life expectancy. The dotted horizontal line represents the average life expectancy and serves as a reference to identify feature values that increase or decrease the target variable.

Adult Mortality Rate begins at the D_1 (0.02) with a life expectancy of 69.98 years. Until D_4 (0.22), it increases to 70.35, then until D_7 (0.39) decreases to 69.78. From there, it is shown a steep downward slope, falling to 64.03 at the D_{10} (0.82). It indicates that higher adult mortality rates are associated with lower life expectancy. The relation follows a concave decreasing exponential function. *Under-five deaths* represents the ratio of deaths among children under five years of age. From D_1 (0) to D_8 (0.59), a nearly linear trend is observed. Finally, an increase is observed at D_9 (0.8), reaching a value of 68.46. *Thinness 5–9 years* shows the proportion of undernourished children and displays a monotonically decreasing linear function. At D_1 (0.02), life expectancy is 70.71, declining to 66.33 at D_10 (0.79). Overall, as the prevalence of thinness increases, life expectancy decreases. *Income composition of resources*, reflects the socioeconomic development of a country, shows a linear relationship. Life expectancy starts at 61.99 and increases to a final value of 75.74. High values of this feature mean better access to resources and are associated with substantially higher life expectancy. *Diphtheria* corresponds to the coverage of childhood vaccination against diphtheria, tetanus, and pertussis. From D_1 (0.24) to D_3 (0.66), the function exhibits a linear shape. However, from D_3 to D_7 (0.93), it shows a zigzag pattern with slight fluctuations before stabilizing again. Beyond this point, the relationship becomes lineal, reaching a life expectancy of 69.938 at D_{10} (1).

5 Conclusions

While the proposed ANN-GFIV is designed to be model-agnostic and can be generalized to understand the marginal functional relationships of any supervised predictive method, this study deliberately validates it on Artificial Neural Networks. ANNs are highly opaque models; their intricate connections establish complex functional compositions that preclude human comprehension, despite their undeniable predictive power. For intrinsically interpretable methods, such a reconstructive approach would be largely redundant. However, for extreme black-box architectures, it becomes strictly necessary.

Research on artificial neural network explainability has often been limited to quantifying the relative importance of individual predictors, neglecting the description of the functional dynamics linking input to output. This study attempts to bridge this gap by tracing the trajectories of the predicted response values relative to each value of the predictor variables. The ANN-GFIV methodology is not limited to measuring a static impact; rather, it isolates the marginal contributions of the regressors X_k to visually reconstruct the behavior learned by the network. Compared to the use of a traditional test set, the employment of an intensive experimental matrix allows for questioning the system within controlled

and systematic scenarios, acting as a stress-test protocol that reveals insights into marginal feature effects through independent simulation cycles. The resulting bivariate plots transform the system's interpretive inaccessibility into explicit risk profiles. Furthermore, compared to traditional feature-response methods like Partial Dependence Plots (PDP), ANN-GFIV significantly mitigates both computational and visual bottlenecks. By employing an intensive experimental matrix, our framework ensures stability: the simultaneous visualization of multiple independent simulation trajectories provides explicit evidence of the model's robustness. The framework has been validated on datasets of diverse natures, demonstrating its effectiveness in complex and real-world problems, such as classification within the healthcare domain. The analysis enabled the identification of different functional behaviors—ranging from exponential and linear trends to complex sinusoidal functions, such as the drastic change in the predictive trend observed across the duration of the hospital stay. The convergence of multiple independent simulation trajectories confirms the stability of these findings, allowing for the precise isolation of high-vulnerability profiles across both quantitative and categorical predictors.

While ANN-GFIV provides intuitive visual explanations, it has some constraints. First, it requires pre-selecting predictors, which may omit important variables. Second, like PDPs, it may generate unrealistic feature combinations during resampling, which could affect interpretation in datasets with strong dependencies. Users should validate findings against domain knowledge. Also, our empirical validation is currently limited to MLP architectures. Extending to deeper networks (e.g., CNNs) is left for future work. Currently, ANN-GFIV builds an automatic system that directly identifies an estimation made by a model, which can represent Y as a composition of analytical models that provide a good approximation of the original Y, following what is done in surrogate Models, but by combining different graphical shapes. So a future development, in our case, is that we can have a functional, mathematical expression that is different for each variable X_k. Currently, the discretization of quantitative variables relies on fixed quantiles D_k. Future developments aim to introduce adaptive strategies driven by the local scale, dynamically intensifying the density of evaluation points in regions characterized by high volatility or sudden transitions. This approach allows for higher-resolution capture of critical non-linearities and inflection points while simultaneously reducing the computational burden in stationary zones of the response function. This contribution is intended as a preliminary step in the field of XAI, providing a functional window into black boxes and transforming complex models into interpretable evidence for decision support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Rozemberczki, B., Watson, L., Bayer, P., Yang, H.T., Kiss, O., Nilsson, S., Sarkar, R.: The Shapley value in machine learning. In: The 31st International Joint Con-

ference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, pp. 5572–5579. International Joint Conferences on Artificial Intelligence Organization (2022)
2. Mandler, H., Weigand, B.: A review and benchmark of feature importance methods for neural networks. ACM Computing Surveys **56**(12), 1–30 (2024)
3. Sundararajan, M., Taly, A., Yan, Q.: Gradients of counterfactuals. arXiv preprint arXiv:1611.02639 (2016). https://arxiv.org/abs/1611.02639
4. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., et al.: The (un)reliability of saliency methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 267–280. Springer International Publishing, Cham (2019)
5. Williams, B., Cremaschi, S.: Novel tool for selecting surrogate modeling techniques for surface approximation. In: Computer Aided Chemical Engineering, vol. 50, pp. 451–456. Elsevier (2021)
6. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI methods – a brief overview. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, pp. 13–38. Springer International Publishing, Cham (2020)
7. Mersha, M., Lam, K., Wood, J., Alshami, A.K., Kalita, J.: Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing **599**, 128111 (2024)
8. Sigut, J., Fumero, F., Arnay, R., Estévez, J., Díaz-Alemán, T.: Interpretable surrogate models to approximate the predictions of convolutional neural networks in glaucoma diagnosis. Machine Learning: Science and Technology **4**(4), 045024 (2023)
9. Munkhdalai, L., Wang, L., Park, H.W., Ryu, K.H.: Advanced neural network approach, its explanation with LIME for credit scoring application. In: Asian Conference on Intelligent Information and Database Systems, pp. 407–419. Springer International Publishing, Cham (2019)
10. Park, C.H., Park, C.S.: Limitations and issues of conventional artificial neural network-based surrogate models for building energy retrofit. Journal of Building Performance Simulation **17**(3), 361–370 (2024)
11. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017)
12. Ogami, C., Tsuji, Y., Seki, H., Kawano, H., To, H., Matsumoto, Y., Hosono, H.: An artificial neural network pharmacokinetic model and its interpretation using Shapley additive explanations. CPT: Pharmacometrics & Systems Pharmacology **10**(7), 760–768 (2021)
13. Bhattarai, P., Thakuri, D.S., Nie, Y., Chand, G.B.: Explainable AI-based DeepSHAP for mapping the multivariate relationships between regional neuroimaging biomarkers and cognition. European Journal of Radiology **174**, 111403 (2024)
14. Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence **298**, 103502 (2021)
15. Gibert, K., Conti, D.: aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. AI Communications **28**(1), 113–126 (2015)
16. Angerri, X., Gibert, K.: Variable selection for meaningful clustering of multitopic territorial data. Mathematics **11**(13), 2863 (2023)
17. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence **5**(5), 726–742 (2021)

18. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89. IEEE (2018)
19. Wei, L., Liu, H., Guo, R., Wang, W., Tian, X.: A hybrid-driven soft sensor model with symbolic representation for enhanced self-interpretability. Measurement Science and Technology **36**(3), 035103 (2025)
20. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189–1232 (2001)
21. Alharthi, A., Alfares, A., Sha'aban, Y.A., Adegbite, D.A.: Improving the interpretability of ANN-based predictions of lattice constants in aliovalently doped perovskites using partial dependence plots. Crystals **15**(6), 538 (2025)
22. Roy, T., Das, P., Jagirdar, R., Shhabat, M., Abdullah, M.S., Kashem, A., Rahman, R.: Prediction of mechanical properties of eco-friendly concrete using machine learning algorithms and partial dependence plot analysis. Smart Construction and Sustainable Cities **3**(1), 2 (2025)
23. Efron, B.: Jackknife-after-bootstrap standard errors and influence functions. Journal of the Royal Statistical Society: Series B (Methodological) **54**(1), 83–111 (1992)
24. Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., et al.: Frailty in older adults: Evidence for a phenotype. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences **56**(3), M146–M157 (2001)
25. Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., Kripalani, S.: Risk prediction models for hospital readmission: A systematic review. JAMA **306**(15), 1688–1699 (2011)
26. Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R.: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of Chronic Diseases **40**(5), 373–383 (1987)
27. Elixhauser, A., Steiner, C., Harris, D.R., Coffey, R.M.: Comorbidity measures for use with administrative data. Medical Care **36**(1), 8–27 (1998)
28. Van Walraven, C., Austin, P.C., Jennings, A., Quan, H., Forster, A.J.: A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical Care **47**(6), 626–633 (2009)
29. Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. Journal of the American Medical Informatics Association **25**(10), 1419–1428 (2018)
30. Ashfaq, A., Sant'Anna, A., Lingman, M., Nowaczyk, S.: Readmission prediction using deep learning on electronic health records. Journal of Biomedical Informatics **97**, 103256 (2019)
31. Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., Ranganath, R.: A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings **2020**, 191 (2020)
32. Futoma, J., Morris, J., Lucas, J.: A comparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics **56**, 229–238 (2015)
33. Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al.: Health system-scale language models are all-purpose prediction engines. Nature **619**(7969), 357–362 (2023)
34. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
35. Sabry, R.: Life expectancy regression with ANN. Kaggle (2024)