Towards explainable artificial intelligence with potential games

Evangelos Spyrou[1], **Vassilios Kappatos**[2], **Afroditi Anagnostopoulou**[3], **Evangelos Bekiaris**[4]
[1, 2, 3, 4]Hellenic Institute of Transport, Centre for Research and Technology Hellas,
6th Km Charilaou-Thermi, Thessaloniki, 57001, Greece
[1]Department of Informatics and Telecommunications, University of Ioannina,
Kostakii Artas, Arta, 47150, Greece
[1]Corresponding author
E-mail: [1]*espyrou@certh.gr*, [2]*vkappatos@certh.gr*, [3]*a.anagnostopoulou@certh.gr*, [4]*abek@certh.gr*

Received 14 August 2024; accepted 6 November 2024; published online 9 March 2025
DOI https://doi.org/10.21595/mme.2024.24454

Abstract. Explainable Artificial Intelligence (XAI) emerged when researchers attempted to identify methods that would interpret the models that are used to perform classification and predictions, in order to avoid having a black box just informing about the result. Methods of XAI are crucial to determine details of the model feature contribution towards the result. One of these methods is attributed to cooperative game theory and especially Shapley values. With this method the features are considered as players and the marginal contribution of the features are employed. In this paper, we take onboard the Potential Game paradigm to show the interconnection between them and the Shapley values. We show that the Shapley values are interlinked with the potential function. Moreover, we setup a game with the marginal contribution of the players as their utility functions and we prove that the game is a potential game. Finally, we show that the price of stability of this game is 1. We utilise the Simulated Annealing (SA) method to find the optimal solution.

Keywords: explainable artificial intelligence, game theory, Shapley value, potential game, potential function.

1. Introduction

Situated at the vanguard of various domains, including wireless communication and telecommunication systems [1-3], smart cities [4] and Internet of Vehicles (IoV) [5] among others Machine Learning (ML) and Artificial Intelligence (AI) have emerged as pivotal research avenues to the of engineering problems. These innovative technologies are harnessed for predicting and classifying diverse attributes. However, the inherent opacity of AI and ML mechanisms often deprives engineers of exhaustive comprehension regarding the nuanced contributions or connections of model attributes with the resultant predictions.

Here, the concept of Explainable AI (XAI) takes center stage. A variety of models are available, offering insights into how the elements within a dataset contribute to predicting the desired outcome. XAI revolves around illuminating the impact these features wield over the anticipated outcome, providing engineers and domain experts with a powerful instrument to fully comprehend their findings. This constitutes just a facet of the broader and promising domain known as Human AI Teaming (HAIT), where the collaborative efforts of experts and machines drive decision-making, with explainable AI acting as a conduit to foster confidence in the model's results [6]. A plethora of research initiatives dedicated to XAI, as evidenced in the works of [7-9], collectively contribute to the ever-evolving array of XAI models. A wealth of such models can be explored in [10], further enhancing the toolbox available for generating XAI.

Amid the array of existing XAI methodologies, the SHapley Additive exPlanation (SHAP) approach takes a prominent position [11]. Rooted in game theory, this method presents a captivating proposition. SHAP, serving as a game-theoretic instrument, delves into the intricacies of Machine Learning predictions. Its endeavor involves quantifying the individual contribution of each feature to a decision, akin to casting features as players engaged in a coalition-forming game. Essentially SHAP uses the Shapley values [12] to provide an explainer to the model that is used.

In this paper, the Potential Game paradigm is encapsulated, utilizing it as a framework to illustrate the inherent interplay between this concept and Shapley values. This work results in the revelation of a profound linkage between Shapley values and the underlying potential function.

Moreover, a gaming scenario is devised where each player's utility function is derived from their unique marginal contributions. This meticulously constructed game is then rigorously proven to align with the definition of a potential game. As a culmination of the analysis, a notable finding is shown: the price of stability for this specific game is established at a value of 1. Notably, the authors build on the work of [13] and show that the potential function results in a potential game and the fact that the social optimum is accomplished. Here, the SA [14] is used to find the global optimum. SA has been utilised for a long time and applications in operational research can be found in [15]. Moreover, this method has been applied to several works in the literature as the read can see in [16, 17].

The potential games paradigm is utilised to optimise the potential function, which we show that has a relation to the Shapley values. In this way, we aim to bridge the Shapley values with the Potential Games. The contributions of this paper are the following:

• There exists a profound linkage between Shapley values and the potential function.

• A potential game scenario is devised where each player's utility function is derived from their marginal contributions.

• The price of stability for the game is established at a constant value of 1.

• Simulated Annealing (SA) is used to find the global optimum, referencing its long-standing use in operational research.

2. Motivation

Using Shapley values in a breast cancer dataset analysis provides essential insights into the model's decision-making process and is particularly valuable in healthcare contexts where interpretability is critical. There exists a plethora of research works aiming to detect and explain breast cancer using text and/or image data [29-32]. The work we are proposing uses the sklearn breast cancer dataset to work with its features using our novel approach for XAI. Shapley values allow us to quantify the contribution of each feature to the model's predictions. This enables a clear understanding of which factors drive the model's classification, for example, distinguishing between benign and malignant tumors or estimating survival rates. By identifying these important features, Shapley values make the model's reasoning more transparent, especially in complex black-box models like neural networks, which are often used for predictive analysis but lack inherent interpretability. This transparency is not only beneficial for researchers but also critical for clinicians and patients, as it fosters trust in the model's predictions and ensures that decisions are based on meaningful patterns rather than obscure mathematical outputs.

Beyond general interpretability, Shapley values offer a personalized understanding of risk by highlighting feature importance at the individual level, explaining why a model predicts a higher or lower risk of cancer for a specific patient. These patient-specific insights are invaluable in personalized medicine, allowing clinicians to make informed decisions about further testing or targeted treatment options based on an individual's unique risk profile. Additionally, Shapley values can reveal potential biases, ensuring that factors like demographic variables do not unduly influence predictions. This is especially important in sensitive areas like breast cancer, where models must be equitable and generalize well across diverse patient populations. Shapley values also aid in model validation, confirming whether the model depends on clinically relevant information or merely statistical noise, thereby enhancing its robustness. In sum, Shapley values contribute to the accuracy, fairness, and reliability of predictive models in breast cancer diagnostics, offering a well-rounded approach to understanding complex medical data.

The potential games paradigm is utilized to optimize the potential function, which we show has a direct relationship to Shapley values. By exploring this relationship, we aim to bridge the cooperative game-theoretic concept of Shapley values with the non-cooperative framework of potential games. The Shapley value, known for fairly distributing gains among players in a cooperative setting, is leveraged here as an insight into payoff allocation among players in potential games, effectively allowing each player's contribution to be valued relative to the total system performance. Through this connection, we extend the interpretation of potential functions to account for individual player contributions in a way that parallels the fairness achieved by Shapley values in cooperative games. This approach not only enhances the interpretability of potential games but also aligns the collective optimization goals with fair distribution principles, making it possible to derive equilibria that are both optimal and balanced in terms of player contributions.

3. XAI conceptual background

As evidenced by the literature documented in [18] and its related sources as well as from [19], the domain of Explainable AI (XAI) has garnered substantial attention from both researchers and businesses in recent times. This surge of interest can be attributed to the increasing integration of AI/ML across various industries. However, this integration has also raised a significant concern: the lack of comprehensive insights into the underlying rationale of the decision-making process that underpins the predictions and recommendations yielded by these systems.

It is evident that a multitude of factors, spanning from societal to ethical pressures, accentuate the necessity for decisions to be both comprehensible and defensible. This necessity arises from the absence of a standardized or precisely defined notion of XAI within technical contexts. Instead of being a strictly formal technical concept, XAI encompasses a spectrum of movements, initiatives, and undertakings aimed at addressing transparency and trust issues in AI. At its essence, XAI revolves around the unraveling of the mysterious inner workings of AI. Consequently, the principles of Responsible AI play a pivotal role in the construction of models that exhibit transparency. This attribute must be achieved without compromising the accuracy of AI and ML systems. Striking a delicate equilibrium between accuracy and interpretability often becomes an essential consideration.

Ultimately, XAI forms a cornerstone of the emerging wave of AI methodologies, commonly denoted as the third wave of AI. The primary objective of this wave is to cultivate precise algorithms endowed with an innate capacity for self-explanation. The culmination of these endeavors seeks to bridge the divide and eventually converge toward achieving human-level intelligence.

4. Shapley value

As read in [20], the employment of cooperative game theory, namely the Shapley value [12] can offer a remedy for assessing the contributions of individual features in the context of a machine learning model's single predictions.

What's particularly intriguing is understanding the influence of each feature on the prediction of a data point. When employing a linear model, it becomes relatively simple to gauge the individual effects. The reduction of a linear model for a single data instance is represented as:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j, \tag{1}$$

where x represents the specific instance for which aims to calculate the contributions. Every x_m is a value of a feature with $m = 1, \ldots, j$. Lastly, the β_m represents the weight that corresponds to feature j.

The $j$ths feature contribution ξ_j regarding $\hat{f}(x)$, which is the prediction, is:

$$\xi_j(\hat{f}) = \beta_j x_j - E(b_j X_j) = \beta_j x_j - \beta_j E(X_j), \tag{2}$$

where $E(\beta_j X_j)$ is the average estimated value of the feature j. In essence, the contribution represents the difference between the feature and the average effect. If all the feature contributions are summed up for a single instance, the result will be as follows:

$$\begin{gathered}\sum_{j=1}^{p} \xi(\hat{f}) = \sum_{j=1}^{p} \left(\beta_j x_j - E\left(\beta_j X_j\right)\right), \\ \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) - \left(\beta_0 + \sum_{j=1}^{p} E\left(\beta_j X_j\right)\right) = \hat{f}(x) - E(\hat{f}(X)),\end{gathered} \tag{3}$$

which is the predicted value for a data point minus the average predicted value.

The Shapley value is calculated using a value function, denoted as val, applied to players in the set S. For each feature value, its contribution to the payoff is determined, weighted, and then summed across all possible combinations of feature values to provide the calculation as follows:

$$\xi(val) = \sum_{S\subseteq\{1,\ldots,p\}\{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left(val(S\cup\{j\}) - val(S)\right), \tag{4}$$

where, S represents a subset of the features used in the model, while x is the vector containing the feature values for the specific instance that needs to be explained, and p denotes the feature number. Additionally, $val_x(S)$ represents the prediction for the feature values within the set S taking into account only the features that are part of the set S, and marginalizing over those features that are not included in S [21]:

$$val_x(S) = \int \hat{f}\left(x_1, \ldots, x_p\right) dP_{x \notin S} - E_X\left(\hat{f}(X)\right). \tag{5}$$

The key idea is that numerous integrations are performed for each feature that is not included in the set S.

The Shapley value corresponds to the method that satisfies four properties, namely Efficiency, Symmetry, Dummy and Additivity. These properties combined can characterise the payoff as being fair. Below, a brief explanation of the properties is provided:

- Efficiency: The features' contributions must be summed to the difference of the x prediction and the mean.
- Symmetry: If the contribution of two feature values are equal to all possible coalitions, the contributions of these two feature values should be the same.
- Dummy: A feature that does not alter the value for the prediction, should have a Shapley value of 0.
- Additivity: For a game with combined payoffs the respective Shapley values can be added.

All coalitions that may emerge need to be evaluated with the presence and the absence of the j-th feature, in order to compute the exact Shapley value. Notably the exact solution to the given problem creates an issue since the number of all possible coalitions are exponential with the number of features. In [22], the authors propose an approximation with Mote-Carlo sampling:

$$\hat{\xi}_j = \frac{1}{M} \sum_{m=1} M(\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)), \tag{6}$$

where $\hat{f}(x_{+j}^m)$ corresponds to the x prediction. However, a random selection of feature values is substituted with values from a random data point, with the exception of the feature's value j. The vector $\hat{f}(x_{-j}^m)$ is similar to the $\hat{f}(x_{+j}^m)$ but the value x_j^m is taken from the sampled data point.

First, choose a particular instance labeled as x and designate a specific characteristic as j. Set the number of iterations as M. In each iteration, a random instance z is sampled from the dataset, and a new random sequence of characteristics is generated. This process results in the creation of two new instances by combining the attributes of instance x and the selected instance z.

The instance denoted as x_{+j} corresponds to the original instance of interest. In this instance, all attributes following the characteristic j are substituted with their corresponding values from the selected instance z. On the other hand, the instance x_{-j} is similar to x_{+j}, with the additional alteration of replacing the value of characteristic j with its corresponding value from the instance z.

Afterward, the change in predictions is calculated, made by the undisclosed model. This change in prediction from the black-box model is determined as:

$$\xi_j^m = \hat{f}\left(x_{+j}^m\right) - \hat{f}\left(x_{-j}^m\right). \tag{7}$$

All the differences like above are average and give:

$$\xi_j(x) = \frac{1}{M}\sum_{m=1}^{M} \xi_j^m. \tag{8}$$

The procedure has to be repeated for each of the features to get all Shapley values.

5. Game theory and potential games

As read in [23], game theory delves into the study of mathematical models depicting both conflict and cooperation, as illustrated by interactions among nodes in our example. In our context, the term 'game' encompasses any type of social interaction involving two or more nodes.

A player is considered rational when it strives to achieve its preferences by selecting appropriate strategies. Decision theory posits that the preferences of a decision maker (node) must adhere to certain fundamental rationality principles. Consequently, the node's behavior can be captured through a utility function, which quantitatively describes its preferences. The primary objective of each decision maker is to maximize their utility function. The notation for non-cooperative games is provided below.

A strategic game $G = \langle N, A, u \rangle$ comprises the following components:

• N is the set of nodes participating in the game $N = \{1,2,\dots n\}$ where n is the number of nodes.

• The set of strategies $A: \sigma \in A = x_{i=1}^{n} A_i$ is the space of all strategy vectors, where each component σ_i of the vector σ belongs to the set A_i. The strategy profile is denoted as $\sigma = (\sigma_i, \sigma_{-i})$, where σ_i is node i's strategy and σ_{-i} are the strategies of the remaining nodes.

In the same way, $A_{-i} = x_{j \neq i} A_j$ denotes the strategy profiles of the nodes excluding i.

• Each node of the game $i \in N$ has a utility function $u_i: A \to \mathbb{R}$ expressing its preferences depending on the strategy profiles. $u = (u_i, \dots u_n): A \to \mathbb{R}^n$ is the vector of the nodes' utility functions which will be maximized to find the Nash equilibrium.

The Nash equilibrium is the most important equilibrium strategic form games. It is defined as the point where no node will increase its utility u by unilaterally changing its action σ.

A strategy profile $\sigma^* = (\sigma_i^*, \sigma_{-i}^*)$ is a Nash equilibrium if $\forall i \in N$ and $\forall \sigma_i \in A_i$:

$$u_i(\sigma^*) \geq u_i(\sigma_i, \sigma_{-i}^*). \tag{9}$$

Games may one or have a unique or multiple Nash equilibria, or even none.

In 2008, Daskalakis et al. established that determining a Nash equilibrium falls under the category of PPAD-complete problems [24]. PPAD, which stands for Polynomial Parity Arguments on Directed graphs, represents a class of total search problems. While it is proven that solutions exist for these problems, pinpointing a particular solution is notably challenging and often impractical. It's worth noting that PPAD-complete problems suggest a level of intractability, though they imply a lesser degree of difficulty compared to NP-complete problems, even though finding a solution remains a formidable task.

This advancement shifted the perspective of researchers seeking to apply game theory to their specific challenges, leading to the rise of 'Potential Games' [25]. These games gained prominence due to their significant characteristics, namely the assurance of the existence of pure equilibria and the (almost certain) convergence of best response dynamics.

A game $\langle N, A, u\rangle$ is an exact potential game if there exists an exact potential function $V: A \rightarrow \mathbb{R}$ subject to $\forall i \epsilon G, \forall \sigma_{-i} \epsilon A_{-i}$, $\forall \sigma_i, \sigma_i' \epsilon A_i$, such as, $V(\sigma_i, \sigma_{-i}) - V(\sigma_i', \sigma_{-i}) = u_i(\sigma_i, \sigma_{-i}) - u_i(\sigma_i', \sigma_{-i})$.

Thus, if the current game strategy is σ, and node i switches from strategy σ_i to strategy σ_i', then the resulting savings i exactly matches the change in the value of the potential function $V(A)$.

Another category of potential games is known as ordinal potential games. In these games, having an ordinal potential function is crucial because it aligns in a sign direction with the individual utility function when a player deviates unilaterally. It's important to highlight that, similar to the precise potential game, $V(A)$ exhibits a similar behavior to the precise potential function. However, the distinction lies in the fact that this function tracks the 'direction' of node improvement rather than the actual degree of improvement. In other words, when a node makes a beneficial move, the potential function decreases.

More formally:

A game $\langle N, A, u\rangle$ is an ordinal potential game if there exists an ordinal potential function:

$$V: A \rightarrow \mathbb{R}, \tag{10}$$

subject to $\forall i \epsilon G$, $\forall \sigma_{-i} \epsilon A_{-i}$, $\forall \sigma_i, \sigma_i' \epsilon A_i$ such as, $V(\sigma_i, \sigma_{-i}) - V(\sigma_i', \sigma_{-i}) > 0 \Leftrightarrow u_i(\sigma_i, \sigma_{-i}) - u_i(\sigma_i', \sigma_{-i}) > 0$.

6. Game formulation

The assumption is that there exist N players for the game formulation we are establishing, which denote the features of the AI model. There are M coalitions. Let $DP_m = \operatorname{argmax}_{m \in N} g_m$ be the selected value of the player when it is chosen as the best possible one for joining a coalition and $E_i(g_m)$ the empirical distribution sample of a feature when it is removed from a coalition. Also, another assumption is that a value from the empirical distribution of the feature's values is inserted if a player does not join the coalition and the value of the marginal contribution if she does. Denote $g_m = \{g_m^1, g_m^2, \ldots, g_m^N\}$ as the coalition selection strategy of player m. Let the set of vectors $s = \{g_1, g_2, \ldots, g_{|N|}\}$ the joined strategy profile of all the players. Denote as $DP(s)$ or $DP(g_m)$ denote the player selection profile or strategy. Moreover, denote as g_{-m} the strategies played by other players except m. Also, denote as $s = \{g_m, g_{-m}\}$ to indicate the player value selection profile when player m is of interest. DP_m is the marginal contribution of player m when it plays joining a specific coalition. Let Γ give the set of all the games. Then by denoting a function $P: \Gamma \rightarrow \mathbb{R}$ that links a real number $P(N, v)$ to every game (N, v), the marginal contribution of a player i in a game (N, v) is given by:

$$DP_i(N, v) = P(N, v) - P(N\{\backslash i\}, v), \tag{11}$$

where $i \in N$.

The utility function of player m is given by:

$$u_i(g_m) = \begin{cases} DP_i(g_m), & i \in g_m, \\ \dfrac{E_i(g_m)}{\epsilon}, & \text{player value is removed from coalition,} \end{cases} \tag{12}$$

where ϵ is a constant to regulate the value of the empirical distribution function.

Thereafter, the potential function is formulated as the summation of all marginal contributions of the players that form the coalition as:

$$V(g) = \sum_{i}^{N} u_i, \tag{13}$$

for all games (N, v) where $v(N)$ is the characteristic function of the grand coalition.

The potential function when all the players are in the coalition provides us with the fact that its marginals are always efficient; thus they add up to the valuation of the grand coalition. Thereafter, some evidence is given about the use of the potential function and its uniqueness. In Theorem A of [13], the reader can see that for every game (N, v) the payoff vector of marginal contributions coincides with the Shapley value and the potential is uniquely given by Eq. (13) applied to the game and its sub-games.

The claim that the game of coalition selection is a potential game will be shown. In order to do that initially the exact potential game will be checked. If a player/feature i is removed from the coalition, then the coalition that remains has marginal contribution with a number coming from sampling the empirical distribution over a set of samples coming from the dataset; Note that if the departing player was not contributing much to the coalition's goals, the remaining players might experience positive effects. If the leaving player was a significant contributor, the remaining players might face challenges. This could result in a loss of resources. Here the assumption is that the player who is removed the coalition loses value or when added it gains value.

In the essence of the game there exist the following cases:

Case 1: Player i is removed from the coalition:

$$\begin{aligned} &V(g_i, (g_{-i}) - V(g_{i'}, g_{-i}) > 0 \\ &\quad \Leftrightarrow u(g_i, g_{-i}) - u(g_{i'}, g_{-i}) + V_{m \neq i}(g_m, g_{-m}) - V_{m \neq i}({g'}_m, g_{-m}) > 0 \\ &\quad \Leftrightarrow u(g_i, g_{-i}) - u(g_{i'}, g_{-i}) > 0. \end{aligned} \tag{14}$$

The utility of the players m when payer i unilaterally deviates from (N, v), to (N, v') the term $V(g_m, g_{-m}) - V({g'}_m, g_{-m}) > 0$ since the utilities of the other features lose value from the reduction of the coalition; hence the game is an ordinal potential game with a potential Eq. (13).

Case 2: Player i is added to the coalition (Here addition is considered the empirical distribution value after the player removal):

$$\begin{aligned} &V(g_i, (g_{-i}) - V(g_{i'}, g_{-i}) < 0 \\ &\quad \Leftrightarrow u(g_i, g_{-i}) - u(g_{i'}, g_{-i}) + V_{m \neq i}(g_m, g_{-m}) - V_{m \neq i}({g'}_m, g_{-m}) < 0 \\ &\quad \Leftrightarrow u(g_i, g_{-i}) - u(g_{i'}, g_{-i}) < 0. \end{aligned} \tag{15}$$

In this case, the utility of the players m when payer i unilaterally deviates from (N, v), to (N, v') the term $V(g_m, g_{-m}) - V({g'}_m, g_{-m}) < 0$ since the utilities of the other features gain value from the addition of the replacement from the empirical distribution or the player's value in the coalition; hence the game is an ordinal potential game with a potential Eq. (13).

Next the price of stability (PoS) is 1 is shown, whereby the Nash equilibrium coincides with the potential function maximisers.

Theorem 1. For any coalition selection potential game Γ, the best Nash equilibrium is the global optimum solution, which means that the $PoS = 1$ when the players form the grand coalition

Proof. Denote the optimum solution of the marginal contribution of a coalition as g_i^* and that is not a Nash equilibrium. Assuming that the player i unilaterally deviates. Then:

$$u_i(g_{i'}) + \sum_i^N u_i(g_i^*) > u_i(g_i^*) + \sum_i^N u_i(g_i^*), \tag{16}$$

which comes as a contradiction that g_i^* is the global optimum solution. In other words the optimum solution must be a Nash equilibrium. For this potential game there are multiple equilibria. The fact that the marginal contribution of a feature is substituted by the expected empirical cumulative distribution value to the coalition this gives us a sub optimal Nash equilibrium, since the value of player i is not as large as the value of the substitution.

7. Approximating the global optimum with simulated annealing

In a potential game, the emergence of multiple Nash equilibria is contingent upon certain conditions. As each player (feature) adjusts their strategy (g_i) in an iterative manner, in response to the strategies of others $(g_{-i}$, where $i \in N)$, the utility function $u_i(g_i, g_{-i})$ is optimized. This iterative process may lead the N features to converge to a stable state, recognized as a Nash equilibrium. However, it's essential to note that the potential function $V(S)$ associated with this equilibrium represents a local optimum rather than a global one. This distinction arises due to the discrete nature of the strategy space $\{S_i\}$, $i \in N$, where S_i denotes the set of strategies for feature i.

To address the challenge of finding an approximation to the global optimum of the potential function $V(S)$, the utilization of simulated annealing (SA) is proposed as an effective game updating scheme, as discussed in [26] and [27].

The SA algorithm orchestrates the evolution of a discrete-time in-homogeneous Markov chain denoted as $x(n)$. In this context, $x(n) = \{g_1, \dots, g_N\}$ encapsulates the strategy combination of the N features at discrete time n. Each feature i faces the decision to either persist with the remaining at the coalition or transition to being replaced by the sample of the empirical distribution in the coalition. To introduce an element of randomness or heat, it is posited that the ith feature can randomly shift its current strategy, with an equal probability represented by $q_{s_i} = 1/2$. The sequential updating of strategies by each feature follows a set of rules that are given below [28]:

- If $u(g'_i, g_{-i}) \geq u_i(g_i, g_{-i})$ then $x(n+1) = \{g'_i, g_{-i})$.
- If $u(g'_i, g_{-i}) < u_i(g_i, g_{-i})$ then $x(n+1) = \{g'_i, g_{-i})$ with probability:

$$p = \exp\left\{\frac{V(g'_i, g_{-i}) - V(g_i, g_{-i})}{T(n)}\right\}, \tag{17}$$

which is equals to the difference in the utility function over $T(n)$ as well. Otherwise, $x(n+1) = x(n) = \{g_i, g_{-i}\}$.

Then the transition probability is provided by the equation:

$$P[x(n+1)] = \frac{1}{2}\exp\frac{\min\left(0, V(g'_i, g_{-i}) - V(g_i, g_{-i})\right)}{T(n)}, \tag{18}$$

where, $T(n)$ denotes the temperature at time n, gradually decreasing during the updating process according to a cooling schedule $\{T(n)\}$. A higher T introduces more randomness in selecting the next strategy, while a diminishing T favors a superior choice with a higher V:

$$T(n) = \frac{T_0}{1 + \alpha \times n}, \quad 0.8 \leq \alpha \leq 0.9. \tag{19}$$

8. Concluding remarks

The primary remark is that the potential function approach introduces a novel and compelling axiomatic characterization of the Shapley value, offering significant advantages from two distinct angles. Firstly, it stands out by requiring just a single axiom, which can be perceived as a combination of the principles of "efficiency" and "marginal contributions". Notably, this approach avoids assuming additivity, symmetry, or other conditions, setting it apart from traditional methods.

Secondly, the potential approach focuses exclusively on the specific game at hand. By applying Eq. (13) to the game and its sub games, it uniquely determines both the potential and the corresponding Shapley values, considering only one characteristic function. This particular feature proves exceptionally valuable in real-world applications, where a single specific problem is typically the subject of investigation. Moreover, the reader can see that with the potential game approach the summation of the marginals when is maximised coincide with the Nash equilibrium.

In contrast, the conventional axiomatizations of the Shapley value require applying various axioms (such as additivity, symmetry, etc.) across a broad domain, encompassing all games or all simple games. This broader scope is necessary to uniquely determine the value for any individual game within traditional methods. The potential approach's efficiency and focus on the specific game make it a compelling alternative in practical scenarios.

Moreover, re-write Eq. (13) as:

$$V(N, v) = \frac{1}{|N|}\left[v(N) + \sum_{i \in N} V(N \backslash \{i\}, v)\right]. \tag{20}$$

The Eq. (20) presents a clear and efficient recursive approach to compute the potential, Shapley values and associated values for all sub games within the given game. This is thoughtfully tailored for the specific task of calculating Shapley values and has demonstrated its effectiveness. Remarkably, applying Eq. (20) merely once to each of the $2^{|N|} - 1$ nonempty coalitions proves sufficient applied to (S, v) for $S \subset N$. A value of the recursion can be set to have the value of the empirical distribution value then the sub games coincide with the features being removed from the dataset.

9. Results

In this paper the load breast cancer dataset is utilised, which is publicly available at the sklearn datasets. The dataset contains 30 features and we compute the marginal contribution of each feature, in order to calculate the Shapley value for each feature. Note that the dataset utilised is an example to satisfy the objective of this paper which is the connection of the Shapley values computation with the class of the potential games. Moreover, during the execution of SA, the Shapley value is dynamically substituted by a better value if it surpasses the previous one, preventing the algorithm from getting stuck in local minima.

We know that the potential when given by Eq. (13) it can compute the Shapley values efficiently. This is the case when all features get the value from the marginal contributions. The Shapley value, during the SA execution, is substituted by the best value if it is greater than the previous one; thus, avoiding getting stuck to a local minimum.

In Fig. 1 we see that the coalition of a set of 30 features, which we examined, features exhibit a value close to 0.4827. From the continuous calculation of the individual Shapley values we observe that the potential function increases and the value of the SA is the maximum is 0.5943.

Here we do not show the entire set of the coalitions but the unilateral deviation of a feature from the grand coalition which is the potential function.

In Fig. 2 we can see the individual Shapley values. The careful observation here is that the Shapley value that shows an increased value is obtained when the empirical distribution value substitutes the original value. This requires further investigation and we leave it for future work. The values that are close to 0.1 are the values sampled to obtain the original value since it is larger than the empirical distribution value.

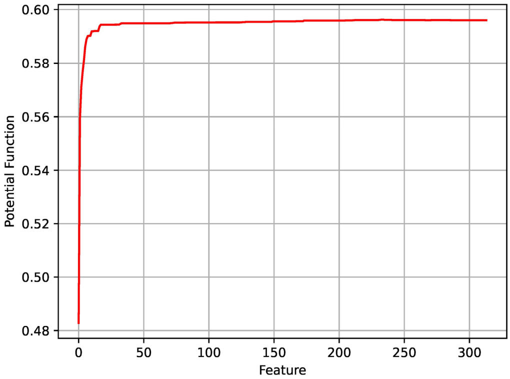


Fig. 1. Potential function of the execution of the SA

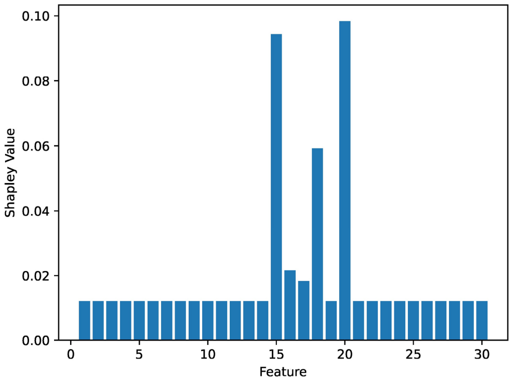


Fig. 2. Shapley values of each feature

In this study, we compute the marginal contribution of each feature and, subsequently, the Shapley value for each. Our goal is to demonstrate how Shapley value computation can connect with potential games, specifically illustrating how features' contributions in a breast cancer dataset can be interpreted in the framework of cooperative game theory. Here, the Shapley value computation is likened to a potential function in a potential game. Each feature's Shapley value acts as an indicator of how much that feature adds to the potential of the dataset (in terms of predictive accuracy). By connecting Shapley values to potential games, we can view the optimization process as a step toward maximizing this potential function, ensuring that each feature's role is properly accounted for and optimized within the overall model. Using simulated annealing (SA) for optimization, we dynamically update the Shapley value of each feature to a higher value if a better marginal contribution is identified, which helps prevent the algorithm from becoming trapped in local minima. This dynamic adjustment is key in ensuring the SA process progresses toward an optimal solution in terms of feature contribution, as represented by the potential function.

10. Conclusions

In this article, we deeply explore the Potential Game paradigm, employing it as a structural framework to showcase the inherent interaction between this concept and Shapley values. Through our efforts, a significant connection is unveiled, revealing the profound interrelationship between Shapley values and the fundamental potential function. Additionally, we formulate a gaming scenario wherein the utility functions of individual players stem from their distinct marginal contributions. This intricately crafted game is subsequently subjected to rigorous analysis, confirming its alignment with the essential characteristics of a potential game. As the culmination of our study, we present a discovery: the price of stability for this specific game is solidly established at a value of 1.

Acknowledgements

This research was done as part of the HAIKU project. This project has received funding from the European Union's Horizon Europe research and innovation programme HORIZON-CL5-2021-D6-01-13 under Grant Agreement no 101075332 but this document does not necessarily reflect the views of the European Commission.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Evangelos Spyrou: conceptualization, methodology, formal analysis and investigation, writing - original draft preparation, resources. Vassilios Kappatos: writing-review and editing, funding acquisition, supervision. Afroditi Anagnostopoulou: writing-review and editing. Evangelos Bekiaris: project administration, resources, supervision, writing-review and editing

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia, "Machine learning for wireless communications in the internet of things: a comprehensive survey," *Ad Hoc Networks*, Vol. 93, p. 101913, Oct. 2019, https://doi.org/10.1016/j.adhoc.2019.101913

[2] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: a survey," *IEEE Communications Surveys and Tutorials*, Vol. 21, No. 3, pp. 2224–2287, Jan. 2019, https://doi.org/10.1109/comst.2019.2904897

[3] C.-X. Wang, M. D. Renzo, S. Stanczak, S. Wang, and E. G. Larsson, "Artificial intelligence enabled wireless networking for 5g and beyond: recent advances and future challenges," *IEEE Wireless Communications*, Vol. 27, No. 1, pp. 16–23, Feb. 2020, https://doi.org/10.1109/mwc.001.1900292

[4] A. Heidari, N. J. Navimipour, and M. Unal, "Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review," *Sustainable Cities and Society*, Vol. 85, p. 104089, Oct. 2022, https://doi.org/10.1016/j.scs.2022.104089

[5] X. Xu, H. Li, W. Xu, Z. Liu, L. Yao, and F. Dai, "Artificial intelligence for edge service optimization in Internet of Vehicles: A survey," *Tsinghua Science and Technology*, Vol. 27, No. 2, pp. 270–287, Apr. 2022, https://doi.org/10.26599/tst.2020.9010025

[6] R. Zhang, N. J. Mcneese, G. Freeman, and G. Musick, “An ideal human,” *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, No. CSCW3, pp. 1–25, Jan. 2021, https://doi.org/10.1145/3432945

[7] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, “Applications of explainable AI for 6G: technical aspects, use cases, and research challenges,” *arXiv:2112.04698*, Jan. 2021, https://doi.org/10.48550/arxiv.2112.04698

[8] A. Renda et al., “Federated learning of explainable AI models in 6G systems: towards secure and automated vehicle networking,” *Information*, Vol. 13, No. 8, p. 395, Aug. 2022, https://doi.org/10.3390/info13080395

[9] T. Zebin, S. Rezvy, and Y. Luo, “An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks,” *IEEE Transactions on Information Forensics and Security*, Vol. 17, pp. 2339–2349, Jan. 2022, https://doi.org/10.1109/tifs.2022.3183390

[10] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, “Explainable artificial intelligence: an analytical review,” *WIREs Data Mining and Knowledge Discovery*, Vol. 11, No. 5, Jul. 2021, https://doi.org/10.1002/widm.1424

[11] Scott M. Lundberg and Su-In Lee, “A unified approach to interpreting model predictions,” in *31st International Conference on Neural Information Processing Systems*, 2017.

[12] L. S. Shapley, *Notes on the N-Person Game – II: The Value of an N-Person Game*. Santa Monica, CA: RAND Corporation, 1951.

[13] S. Hart and A. Mas-Colell, “The potential of the Shapley value,” in *The Shapley Value*, Cambridge University Press, 1988, pp. 127–138, https://doi.org/10.1017/cbo9780511528446.010

[14] R. W. Eglese, “Simulated annealing: A tool for operational research,” *European Journal of Operational Research*, Vol. 46, No. 3, pp. 271–281, Jun. 1990, https://doi.org/10.1016/0377-2217(90)90001-r

[15] C. Koulamas, S. Antony, and R. Jaen, “A survey of simulated annealing applications to operations research problems,” *Omega*, Vol. 22, No. 1, pp. 41–56, Jan. 1994, https://doi.org/10.1016/0305-0483(94)90006-x

[16] H. Oliveira and A. Petraglia, “Establishing Nash equilibria of strategic games: a multistart Fuzzy Adaptive Simulated Annealing approach,” *Applied Soft Computing*, Vol. 19, pp. 188–197, Jun. 2014, https://doi.org/10.1016/j.asoc.2014.02.013

[17] M. Saadatpour, A. Afshar, H. Khoshkam, and S. Prakash, “Equilibrium strategy based waste load allocation using simulated annealing optimization algorithm,” *Environmental Monitoring and Assessment*, Vol. 192, No. 9, Sep. 2020, https://doi.org/10.1007/s10661-020-08567-w

[18] A. Adadi and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI),” *IEEE Access*, Vol. 6, pp. 52138–52160, Jan. 2018, https://doi.org/10.1109/access.2018.2870052

[19] E. D. Spyrou and V. Kappatos, “XAI using SHAP for outdoor-to-indoor 5G mid-band network,” in *IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 862–866, Apr. 2023, https://doi.org/10.1109/csnt57126.2023.10134625

[20] C. Molnar, “Interpretable machine learning,” 2020.

[21] C. Molnar, “Interpretable machine learning: A guide for making black box models explainable,” 2022.

[22] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and Information Systems*, Vol. 41, No. 3, pp. 647–665, Aug. 2013, https://doi.org/10.1007/s10115-013-0679-x

[23] E. D. Spyrou, “Link Quality Optimisation and Energy Efficiency in Wireless Sensor Networks using Game Theory,” Ph.D. thesis, Aristotle University of Thessaloniki, 2019.

[24] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, “The complexity of computing a Nash equilibrium,” *SIAM Journal on Computing*, Vol. 39, No. 1, pp. 195–259, Jan. 2009, https://doi.org/10.1137/070699652

[25] D. Monderer and L. S. Shapley, “Potential games,” *Games and Economic Behavior*, Vol. 14, pp. 124–143, 1996.

[26] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, Vol. 220, No. 4598, pp. 671–680, May 1983, https://doi.org/10.1126/science.220.4598.671

[27] V. Černý, “Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm,” *Journal of Optimization Theory and Applications*, Vol. 45, No. 1, pp. 41–51, Jan. 1985, https://doi.org/10.1007/bf00940812

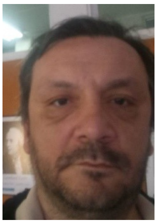
[28] Yanqing Liu, Liang Dong, and R. J. Marks, "Common control channel assignment in cognitive radio networks using potential game theory," in *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 315–320, Apr. 2013, https://doi.org/10.1109/wcnc.2013.6554583

[29] S. Rajpal et al., "XAI-CNVMarker: Explainable AI-based copy number variant biomarker discovery for breast cancer subtypes," *Biomedical Signal Processing and Control*, Vol. 84, p. 104979, Jul. 2023, https://doi.org/10.1016/j.bspc.2023.104979

[30] P. S. Oztekin et al., "Comparison of explainable artificial intelligence model and radiologist review performances to detect breast cancer in 752 patients," *Journal of Ultrasound in Medicine*, Vol. 43, No. 11, pp. 2051–2068, Jul. 2024, https://doi.org/10.1002/jum.16535

[31] F. Silva-Aravena, H. Núñez Delafuente, J. H. Gutiérrez-Bahamondes, and J. Morales, "A hybrid algorithm of ML and XAI to prevent breast cancer: a strategy to support decision making," *Cancers*, Vol. 15, No. 9, p. 2443, Apr. 2023, https://doi.org/10.3390/cancers15092443

[32] R. M. Munshi, L. Cascone, N. Alturki, O. Saidani, A. Alshardan, and M. Umer, "A novel approach for breast cancer detection using optimized ensemble learning framework and XAI," *Image and Vision Computing*, Vol. 142, p. 104910, Feb. 2024, https://doi.org/10.1016/j.imavis.2024.104910



Dr **Evangelos Spyrou** is currently a Research Associate at the Hellenic Institute of Transport (HIT) of the Center for Research and Technology Hellas, and a Postdoctoral Researcher at the University of Ioannina in Hellas. He is working mainly on wireless communications, as well as optimization models, applied to the transport domain. His research interests include wireless networks and applying game theoretic models to them, stochastic network optimization and practical algorithms of wireless parameters from the physical to the network layer. He has a BSc in Computing and an MSc in Embedded Computer Systems Engineering from the Northumbria University of Newcastle, UK and a PhD in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece. He has experience in wireless communications working in several projects as well as software development for safety critical systems. He has served as a reviewer of the Internet of Things and Transactions of Internet Transportation Systems IEEE Journals. He is a current reviewer of the various IEEE and MDPI journals.

Dr. **Vassilios (Vassilis) Kappatos** is the Head of Constructural and Infrastructure Research in Maritime and Air Transport Laboratory at Hellenic Institute of Transport (HIT), Center for Research and Technology Hellas (CERTH), Greece. He is a member of the board of CERTH, he is Visiting Fellow at School of Aerospace, Transport and Manufacturing, Cranfield University, UK and member of oN DuTy program supported by the Collaborative Research and Training Experience (CREATE) of the National Sciences and Engineering Research Council of Canada (NSERC) and industrial partners. He is currently Fellow of The Welding Institute (FWeld) and Member of the - Engineering Council of United Kingdom (CEng) / Technical Chamber of Greece (TEE-TCG)/ Institute of Electrical and Electronics Engineers (MIEEE)/ International Society for Condition Monitoring (ISCM) / Hellenic Society for Non Destructive Testing (H.S.N.T)/ Hellenic Institute of Acoustics (HELINA)/ Danish Center for Applied Mathematics and Mechanics (DCAMM)/ Society for Machinery Failure Prevention Technology (MFPT)/ Greek-German Chamber of Commerce and Industry. He obtained his PhD in the area of Non Destructive Evaluation (NDE) and also holds the Diploma of Mechanical and Aeronautical Engineering (MEng) since 2002. His research areas are NDE, Structural Health Monitoring, Condition Monitoring, Structural Integrity, Pattern Recognition and Signal Processing. His research has been supported by the European Commission and other organizations. As a research leader, he has delivered and coordinated 77 European (European Defence Fund, Horizon Europe, Erasmus, H2020, FP7), 16 UK (TSB, Innovate UK), 5 Danish and 11 Greek National funded research projects with a total budget of more than 150 million €. He has extensive expertise on proposal writing for European funded projects and in developing partnerships within the industry. Throughout the years, he has reviewed and submitted many proposals, providing technical input, writing proposals and sections, forming partnerships, securing funds of around 20 million €. He has over 145 publications in International Journals, Conference Proceedings, Book Chapters and Articles. He is a reviewer of several international journals and conference proceedings. Throughout his career, he has lectured at University level on a broad range of modules in the area of Applied Engineering in Denmark, the United Kingdom and Greece. He was Associate Professor (permanent position) at Mechanical Engineering Section, Department of Technology and Innovation, University of Southern Denmark (SDU), Denmark. He has managed more than 100 Research Fellows, Associates and Assistants, Technicians, Ph.D. Students and Project Technical Assistants (Placement Students) over his career.

Dr. **Afroditi Anagnostopoulou** is a Principal Researcher at the Hellenic Institute of Transport / Centre for Research and Technology Hellas, experienced in project management and transport economics overseeing the “Economic and Social Impacts of Transport Systems” laboratory. She is also Deputy of Unit on Transport Economics and Environment, Maritime and Air Transport in the HIT. Her research interest lies in the analysis and evaluation of the economic and social impacts from the operation of transport systems and services. She participates in several research projects and she has also been teaching in the MSc Logistics and Supply Chain Management at the EPSILON COLLEGE supervised by Northampton University since 2022. In addition, she has worked as course instructor in the MSc in Shipping at the University of Piraeus from 2018 to 2022. She holds a BSc in Management Science and Technology from the Athens University of Economics and Business and an MSc in Computer Science from the University College London. She has also received a PhD in Operations Research from the Department of Management Science and Technology of the Athens University of Economics and Business. She is a member of the Hellenic Operational Research Society, the Hellenic Institute of Transport Engineers, the Association of Greek Researchers and the Economic Chamber of Greece. In addition, she was a member of committee for developing the Gender Equality Plan of Centre for Research and Technology Hellas (2022-2023). Dr. Afroditi Anagnostopoulou is also a research fellow in the Operations Research & Decision Systems Centre of the Athens University of Economics and Business. She is listed in “Who's Who in the World” (2015, 2016, 2017, 2018, 2019) and her work has been published in international academic journals (including the first-tiered Transportation Science journal) and presented in more than 30 national and international conferences (including the TRISTAN VII, 1st & 3rd VEROLOG, EWGT2014, EWGT2023, TRA2014, TRA2018, TRA2020, TRA2022, CoDIT2019 & CoDIT2020).

Dr. **Evangelos Bekiaris**, Ph.D. on Mechanical Engineering NTUA, has been elected as Researcher of the Hellenic Institute of Transport (HIT) of the Centre for Research and Technology Hellas (CERTH) in 2001; and elected as HIT Director General as of 2016 and CERTH Vice-President as of 2021. Since 1992 he has participated in 105 research projects, in 42 of which at the role of Coordinator. He is member of the National Council for Research and Innovation (N.C.R.I.) since 2019, National Representative in the European Commission Climate Neutral and Smart Cities Mission Board of Horizon Europe (HE) since 2020 and the National representative of Greece in the H2020 Transport Committee since 2014. He is also the President of the Hellenic Institute of Electric Vehicles (HELIEV) and the European Transport Research Alliance (ETRA) since 2021 and the President of the European Conference of Transport Research Institutes (ECTRI) and the European Rail Research Network of Excellence (EURNEX) since 2019. In the past he’s also been the President of the European Associations FERSI (on Road Safety) and HUMANIST (on Human Factors in Transport).