Explaining evolutionary feature selection via local optima networks

Jason Adair
jason.adair@stir.ac.uk
University of Stirling
Computing Science and Mathematics
Stirling, UK

Sarah L. Thomson
s.thomson4@napier.ac.uk
Edinburgh Napier University
Computing, Engineering & The Built Environment
Edinburgh, UK

Alexander E.I. Brownlee
alexander.brownlee@stir.ac.uk
University of Stirling
Computing Science and Mathematics
Stirling, UK

ABSTRACT

We analyse fitness landscapes of evolutionary feature selection to obtain information about feature importance in supervised machine learning. Local optima networks (LONs) are a compact representation of a landscape, and can potentially be adapted for use in explainable artificial intelligence (XAI). This work examines their applicability for discerning feature importance in supervised machine learning datasets. We visualise aspects of feature selection LONs for a breast cancer prediction dataset as case study, and this process reveals information about the composition of feature sets for the underlying ML models. The estimations of feature importance obtained from LONs are compared with the coefficients extracted from logistic regression models (interpretable AI), and also against feature importances obtained through an established XAI technique: SHAP (explainable AI). We find that the features present in the LON are not strongly correlated with the model coefficients and SHAP values derived from a model trained prior to feature selection, nor are they strongly correlated within similar groups of local optima after feature selection, calling into question the effects of constraining the feature space for wrapper-based techniques based on such ranking metrics.

CCS CONCEPTS

• **Mathematics of computing** → **Graph algorithms**; *Combinatorial algorithms*; • **Theory of computation** → **Evolutionary algorithms**.

KEYWORDS

Fitness Landscapes, Explainable AI, Local Optima Networks (LONs)

ACM Reference Format:
Jason Adair, Sarah L. Thomson, and Alexander E.I. Brownlee. 2024. Explaining evolutionary feature selection via local optima networks. In *Genetic and Evolutionary Computation Conference (GECCO '24 Companion), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3638530.3664183

GECCO '24 Companion, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0495-6/24/07.
https://doi.org/10.1145/3638530.3664183

1 INTRODUCTION

Local Optima Networks (LONs) [33] are a compact representation of fitness landscapes, and are well-established as a way of understanding the relationship between a metaheuristic algorithm and a problem. LONs allow us to make insights such as understanding the relative performance of algorithms that use different sets of operators. LONs also reveal the common paths taken by algorithms as they search the space on the way to high-quality solutions. It has been suggested [2] that the information represented by a LON can also be exploited for eXplainable AI (XAI): explaining the quality of the solutions that have been identified by the search. The features present in the local optima, and the order in which they were included or excluded by the search, reveal what the algorithm has learned about the problem at hand. Thus analysis of the LON itself provides one way to explain solutions for the target problem. In this paper, we propose two approaches towards mining explanatory information from LONs: statistical analysis of LON characteristics and clustering of the local optima. Many applications necessitate the use of multiple runs, which enables the construction of a LON. Thus in such a situation much of the explanatory information is generated as a byproduct of the search process itself, rather than requiring an additional search or probing of the model after it is fitted.

We test the proposed approaches by applying them to feature selection in machine learning. Metaheuristics have often been applied to feature selection [7]. Typically a 'wrapper' approach is employed, whereby the metaheuristic selects a subset of features for the ML model and the fitness of solutions is the cross-validation accuracy (or similar) of the resulting model using those features. This provides an ideal context in which to explore the potential of LONs for explainability: explanations of feature importance derived from a LON can be tested against feature importance measures already in use among the XAI community. The key novelty of our approach is that the explanations are derived from the search; there is nothing preventing the approach being applied for parametric optimisation problems other than feature selection. In the present paper we focus on mining feature importance (of the trained ML model, rather than fitness landscape features) from LONs assuming that the features are independent; consideration of linkage between features (e.g., [1, 31]) is also important and a future consideration.

The contributions of this work are as follows:

(1) the proposal of two approaches for mining LONs for explanations: summary statistics and clustering

(2) experimental comparison of these approaches with established XAI techniques: logistic regression coefficients and SHAP
(3) the selection of suitable visualisation techniques to make the explanations accessible

2 BACKGROUND

2.1 Fitness Landscapes

A fitness landscape [26] is composed of three parts: $(S, N, f) : S$ is the full set of possible solutions; $N : S \longrightarrow 2^S$ is the neighbourhood function, which assigns a set of adjacent solutions $N(s)$ to every $s \in S$; and f is a fitness function $f : S \longrightarrow \mathbb{R}$ that provides a mapping from solution to associated fitness. That fitness can be conceptualised as the solution *height* within the landscape metaphor.

2.2 Local Optima Networks

Local Optima Networks (LONs) [18] are a means to study the global structure of a fitness landscape. We will describe their constituent components before introducing the LON as a whole.

Neighbourhood. The *neighbourhood* of a solution, s_i, are the solutions which are adjacent to s_i according to a neighbourhood function: $N(s)$. In this work, the notion of adjacency is defined as single bit-flip in the binary solution.

LON nodes. A local optimum has superior or equal fitness to its neighbours according to a fitness function *f*. In this work, we do not exhaustively search the neighbourhood as this would be computationally infeasible. Instead, we consider that a solution lo_i is a local optimum if it has superior or equal fitness to its *sampled* neighbourhood SN. Formally: $\forall n \in SN(lo_i) : f(lo_i) >= f(n)$ (assuming maximisation, as is the case for this study) where $SN(lo_i)$ is the sampled neighbourhood, n is a particular neighbour. The nodes in a LON, LO, are the local optima as just defined.

LON edges. There is an edge from local optimum lo_i to local optimum lo_j, if lo_j can be obtained after applying a random perturbation to lo_i followed by local search, and $f(lo_j) \geq f(lo_i)$. In LON terminology, these are called *escape* edges [33]. The edges are determined to be *monotonic* because they record only non-deteriorating, directed connections between local optima. Edges are weighted with the frequency of transition: the number of times during searches that lo_j was reached by applying perturbation then local search to lo_i. The set of edges is denoted by E.

Local optima network (LON). A local optima network, LON = (LO, E), consists of nodes $lo_i \in LO$ which are the local optima, and edges $e_{ij} \in E$ between pairs of nodes lo_i and lo_j with weight w_{ij} if $w_{ij} > 0$.

2.3 Related Work

Fitness Landscapes and XAI. Fitness landscapes are used as a vehicle for understanding or *explaining* metaheuristic algorithms; it follows that they are an intuitive bridge between evolutionary computation and explainable artificial intelligence [4, 30]. Indeed, authors have proposed using XAI to gain insight into algorithm performance prediction models by using SHAP (an XAI method for feature importance) [32]. Local optima networks have been used in the past to analyse neural architecture search spaces [19, 22]; we argue that this endeavour was at least XAI-adjacent, given that the analysis led to information regarding construction of a good model. In the studies, the landscapes were found to be straightforward and well-suited to iterated local search. Machine learning pipelines have also been subject to LON construction [28, 29]. The most closely-related work to the present study, however, captures LONs for evolutionary feature selection [16]; they found that there were local optima plateaus, indicating the presence of irrelevant/non-informative features in some models. Our study takes inspiration from that work, but differs in the following ways: we explore and propose ways of mining the LON data for the purpose of XAI and we consider a larger search space than those considered in the aforementioned study. Although not strictly fitness landscape analysis, the work of Tinós, Przewozniczek *et al.* [23, 24] is strongly related: in one recent study [31], the authors use a linkage-based genetic algorithm which learns a variable interaction graph during optimisation; one of the contexts they applied this was to evolutionary feature selection, which allowed them to discover feature interactions in machine learning datasets.

Shapley Additive Explanations (usually referred to as SHAP) [12] are a prevalent XAI method [3, 8, 36] which estimate the contribution of features to a prediction. SHAP trains models for different sets of features. The *marginal contribution* of a feature — for a particular observation — is obtained by subtracting the prediction of a model which *excludes* that feature from the prediction of the same model which *includes* the feature. Marginal contributions of the feature across all models which contain it are added together — resulting in a SHAP value for the feature-observation pair. While SHAP values constitute local explanations, these can be aggregated for a set of observations to provide a global model explanation. SHAP has been used for XAI relating to neural networks previously [10, 37, 38]; in this work we compute SHAP values and compare them with insights gained from the LON analysis.

3 METHODOLOGY

3.1 Dataset

We consider tabular classification only and select a well-known real-world dataset which did not require pre-processing: Wisconsin Diagnostic Breast Cancer. This is a binary classification task; the features characterise different aspects of a breast mass, with the classes being *benign* and *malignant*. We import the dataset through Scikit-learn [21]. There are 569 observations, 30 independent features, and no missing values.

3.2 Learning Algorithm

Logistic regression (LR) [35] formulates a logistic equation to capture a dependent variable based upon supplied data. Instead of the straight line which is fit during linear regression, in logistic regression an *S*-shaped logistic function is mapped. Although the model produces a numeric value during prediction — the probability of belonging to a particular class — logistic regression is typically used for classification tasks.

LR is chosen as the machine learning model which serves as the foundation for the optimisation problem. There are a few reasons

for this: it is non-stochastic, which will lower the amount of noise in the fitness landscape; the number of hyperparameters to consider is comparatively low; and, principally, it is "inherently interpretable" [14] through the extraction of feature coefficients. The use of a machine learning method with this quality allows contrast of the insights from the coefficients with those gained from the XAI and LON explanations.

3.3 LON Construction

Iterated Local Search (ILS) is a metaheuristic which is well-suited to constructing sampled LONs owing to its two-level search strategy. Indeed, LONs have been constructed from the traces of ILS runs in the literature [17, 20]. ILS combines random perturbations applied to local optima with hill climbing. To construct a LON sample, r independent ILS runs from random starting solutions are executed. For each run, local optima are logged. Additionally, transitions between local optima (using perturbation and then hill climbing) are noted. Nodes and edges from the r runs are amalgamated into a single LON for the associated problem.

3.4 Explainable Artificial Intelligence

Broadly, explainable AI methods can be delineated into 'global' (the model as a whole) or 'local' (a particular observation) [6]. The LON approach to explainability proposed in this paper is global in nature: each local optimum encodes the features comprising a machine learning model, without a focus on any specific prediction. Additionally, we study the LON as a whole object — this provides explanations about *groups* of models, which means that LONs are perhaps better characterised as an "aggregate" global XAI method.

Shapley Additive Explanations (usually referred to as SHAP or SHAP values) [12] are a highly prominent XAI method [3, 8, 9, 25, 36] which estimate the contribution of features to a prediction; SHAP has also been used as a means for feature selection [13]. SHAP trains models for different sets of features. The *marginal contribution* of a feature — for a particular observation — is obtained by subtracting the prediction of a model which *excludes* that feature from the prediction of the same model which *includes* the feature. Marginal contributions of the feature across all models which contain it are added together — resulting in a SHAP value for the feature-observation pair. While SHAP values constitute local explanations, these can be aggregated for a set of observations to provide a global model explanation. We compute SHAP values in this work and compare them with insights gained from the LON analysis.

4 EXPERIMENTAL SETUP

The Python library Scikit-learn [21] is used for modelling.

4.1 Data Splitting and Preprocessing

In this work, we do not consider an independent test set — only training and validation sets. The reason for this is that feature selection, which is the optimisation problem under study, is typically carried out on the training set during model selection (without knowledge of a test set). We conduct no preprocessing on the machine learning datasets in this study and use k-fold cross-validation where $k = 5$. For a given dataset, the same data split is used for every model built, and stratified folds are employed.

4.2 LON Construction

Iterated Local Search. The ILS which serves as the foundation for LON extraction is designed as follows: the perturbation consists of $\frac{N}{10}$ random bit flips, where N is the number of features (a large perturbation magnitude is chosen to ensure diversification, in particular because a low number of runs are conducted due to computational expense). A first-improvement pivot rule is used, and the local search uses single bit-flip mutation. New local optima are accepted if their fitness is better than the current local optimum. If both have equivalent fitness then the solution with less features is chosen (if they have the same number then the new one is rejected to prevent stagnation on plateaus). This mechanism was implemented to try and steer the search away from solutions containing irrelevant features. Thirty independent ILS runs from random starting solutions are conducted. For random samples like this, it has been argued that 30-50 are sufficient [15]. This aligns with the Central Limit Theorem, which stipulates that from around 30 observations, sample means begin to resemble a normal distribution [11]. Individual LON sampling runs terminate after 100 iterations with no improvement in local optimum quality; this termination condition has been used in a previous study on LONs [19].

Fitness. Recall that the solution representation is binary, and denotes whether features are included in the model or not. The fitness function is the mean five-fold cross-validation accuracy of the model configuration which uses the solution's feature set, rounded to six decimal places. The same cross-validation splits are used for all fitness evaluations associated with a particular dataset - this is to minimise noise in the fitness landscape.

All models use Logistic Regression (LR) with default values as imported from the Scikit-learn library. This is with the exception of the maximum number of iterations - this required an increase to 4000 as the default value of 100 often resulted in non-convergence. All other hyperparameters were left as the default values: *solver*=L-BFGS; *C*=1.0; *penalty*=l2; *multiclass*=auto; *tol*=1e-4.

Threshold parameter. When presenting the results, a parameter is sometimes used to threshold the fitness of considered local optima. Where this parameter is in place, it means that the group of local optima included have fitness greater than or equal to a specified threshold: β.

4.3 SHAP and Logistic Regression Coefficients

The modelling setup for computing SHAP values and extracting logistic regression coefficients is exactly the same as the models evaluated during LON construction: five-fold cross-validation (with the same data splits which were used in LON sampling); Logistic Regression with the parameters specified in Section 4.2.

SHAP. SHAP analysis is conducted using the Python package shap [12] and using its PermutationExplainer. This explainer [27] produces values obtained from sampling of a permutation variant of the SHAP equations, and it was chosen for this work because it does not need parameter tuning. We compute SHAP

values for every node (i.e., feature configuration) found in the LON sample. Output from running SHAP includes SHAP values for each observation-and-feature pair; the observations are from the validation set. The five-fold cross-validation yields five observation-feature SHAP matrices, and we take the mean of these to obtain a single SHAP value matrix for a model. In this matrix, there is a row for each observation and a column for each feature. Although the values in this matrix are local model explanations (they are for a single observation), we obtain global explanations for a feature by computing the mean of the absolute values for the SHAP entries which comprise a column. The resultant values capture the magnitude of a feature's importance.

Logistic Regression Coefficients. Logistic regression coefficients are extracted from every model built for SHAP analysis just described. For each coefficient, the absolute mean value across the five folds of cross-validation is computed — resulting in a value which represents the magnitude of importance for the feature.

4.4 Comparing Ranks of Features

The feature importances, as estimated/ranked by (a) logistic regression coefficients, (b) SHAP, and (c) feature presence in the LON, are compared. For this, we are interested in whether the rankings are aligned between the different approaches. By *ranking*, we mean the ordered list of features where the most-important feature (according to the method) is first and the least-important is last. Rankings can be compared using Kendall's Tau [5], which is defined between -1 and 1. A value of 1 would indicate a perfect match. The measure, τ, is computed as:

$$\tau = \frac{|CP| - |DP|}{|TP|} \tag{1}$$

where $|CP|$ is the number of *concordant* pairs (a pair is concordant if the ranks match), $|DP|$ is the discordant pairs, and $|TP|$ is the total number of pairs. We use the SCIPY [34] implementation of Kendall's τ.

5 RESULTS

5.1 Feature proportions

Figure 1 shows a beeswarm plot indicating the distribution for fitness of the sampled local optima. Notice that several separate feature sets (solutions) have the equivalent fitness, and that there are distinct 'levels' to the fitness distribution. We also note that there is a group of low-quality local optima and a group of higher-quality local optima. In Figure 2 we consider how the composition of feature sets (local optima) in the sampled LON change with respect to the validation accuracy (fitness). The y-axis is the percentage of LON nodes which contain a feature; the x-axis is β (recalling its definition in Section 4.2, this means that at 0.96 the local optima included have a fitness greater than, or equal to, 0.96); and each line represents a feature, as indicated in the legend. The x axis begins at 95%, excluding the lower fitness optima group, as seen in Figure 1 and Table 1, as there are no local optima discovered in the fitness range of 89.3% and 95%. Notice from Figure 2 that the proportion of nodes containing the features remains relatively constant from the lowest value for β up to $\sim$ 95.6%. This implies that most local optima fitness levels contain similar feature sets. For higher values of β, the proportions begin to shift, sometimes dramatically. This is exemplified by Feature 7: *'Mean concavity'* is present in almost 90% of all local optima until the fitness threshold of 96.1% where it drops to less than 20% - a trend shared by a number of other features. This can be contrasted with Feature 4 *'mean area'* which was only present in 42% of the 428 identified local optima, but present in all of the highest fitness local optima. Considering the placement of the lines on the y-axis, it becomes clear that some features are in very few sampled local optima regardless of fitness — such as feature 13 (which is *'perimeter error'*) and feature 14 (*'area error'*). Other features are present in most local optima — such as feature 1 (*'mean radius'*) and feature 3 (*'mean perimeter'*). There is a dramatic reconfiguring of feature proportions among local optima when β is between 96.1% and 96.4%. We argue that this shows that changes of a significant magnitude to good feature sets may be needed in order to obtain feature sets which are of an even higher quality. This may convey that a very particular set of features are needed together in order to obtain a validation accuracy of higher than approximately 96.1%.

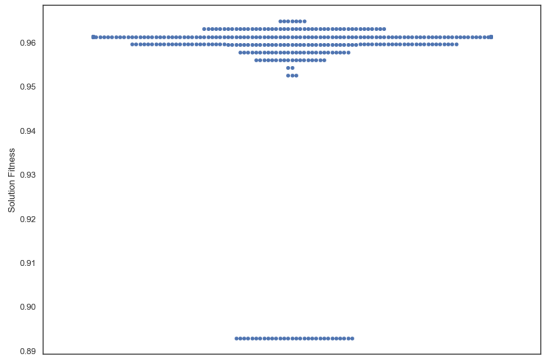


Figure 1: Fitness of discovered local optima

cluster	node count	mean fitness
low-fitness optima	30	0.892874
high fitness optima	428	0.960728

Table 1: Description for clusters of nodes from the breast cancer LON

5.2 Clustering

We would like to explore the potential of clustering on the LON for explainability. To this end, hierarchical clustering is applied to the nodes of the LON.

The clustering considers the distance between solutions in the binary space, thereby associating solutions which have similar feature composition. Clusters from this analysis will be used in subsequent comparisons of feature presence in LON nodes with SHAP values and logistic regression (LR) coefficients. A dendrogram

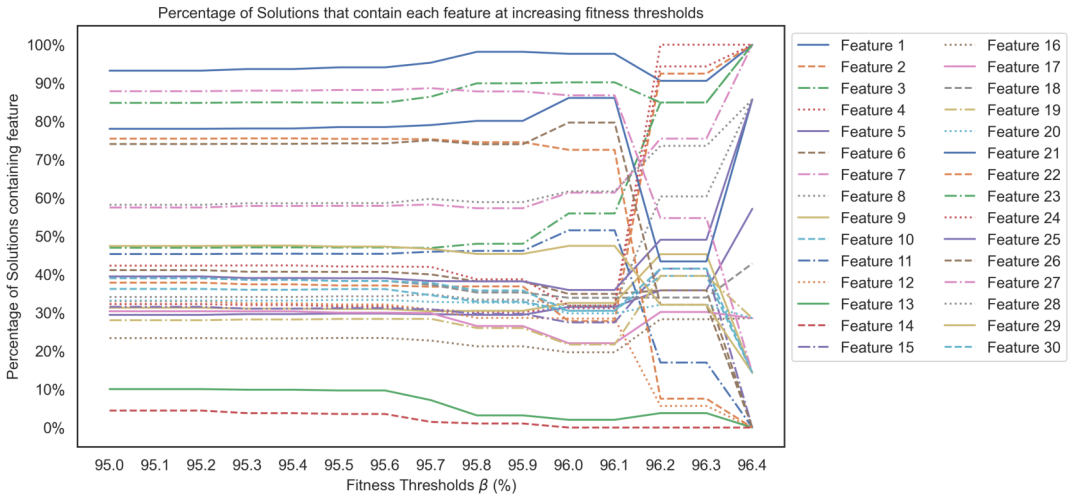


Figure 2: Change in proportion of nodes (local optima of the breast cancer dataset) which contain a feature with increasing β

presenting the groups can be found in Figure 3. Notice from the Figure that the sampled LON nodes can be organised into five large groups (threshold determined using the elbow method). This shows that there are five distinct "types" of feature set which have high validation accuracy (observe the five colours present). Simple statistics for the clusters are provided in Table 2: the number of local optima (*node count*), pseudo global optima (the number of solutions with the highest fitness which was sampled; *global count*), and the mean fitness.

Considering Table 2, we can see that *Cluster 1* contains all of the pseudo-global optima (see third column). The closest clusters to *Cluster 1* are *Clusters 2* and *3* (as seen in Figure 3); despite being the nearest groups of solutions, their mean fitnesses are lower than that of the more distant *Clusters 4* and *5*. In terms of *explaining* these groups of models, we argue that, according to the LON sample constructed, there are five groups of feature sets which contain similar feature combinations. In terms of the solution space, the pseudo-global optima have a large number of surrounding sub-optimal solutions that perform worse than that of the substantially different sets of features (*Clusters 4* and *5*). The implications of this are better stated in terms of a hypothetical problem: suppose we are tasked with designing a battery of tests for predicting heart disease. Heart disease is a complex multi-factor illness and selecting to record only the most relevant contributing factors reduces the practical and financial burdens; therefore feature selection offers a viable solution. This experiment potentially highlights pseudo-global optima that contain counter-dependencies between specific tests — if refining medical tests to their minimal contributing factors for machine learning, we should use caution when substituting tests for similar measurements.

Cluster	Node Count	Global Count	Mean Fitness (Range)
1 (orange)	74	7	0.9623 (0.956-0.9649)
2 (green)	44	0	0.9594 (0.9526-0.9631)
3 (red)	101	0	0.9594 (0.9526-0.9631)
4 (purple)	69	0	0.9613 (0.9596-0.96137)
5 (brown)	140	0	0.9610 (0.9596-0.96137)

Table 2: Description for clusters of nodes from the breast cancer LON s

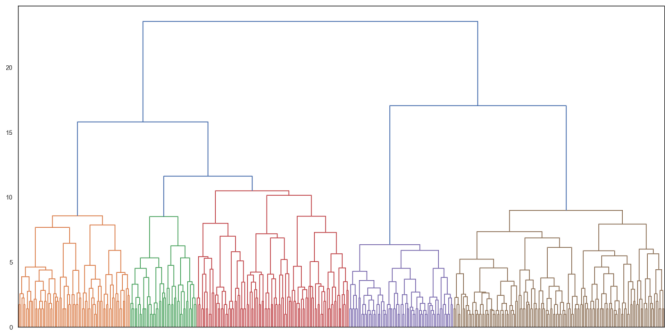
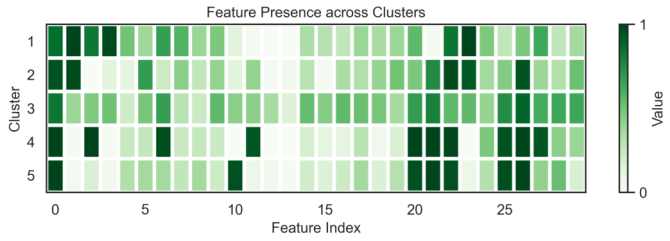


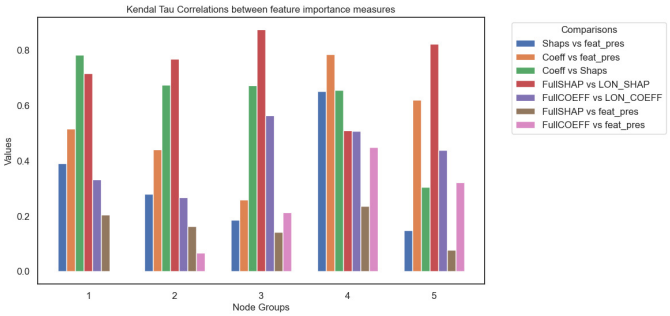
Figure 3: Dendogram showing the local optima can be split into five clusters according to their locality (hamming distances)

5.3 Comparison with established techniques

In this Section we compare explanations from LON analysis with SHAP values and LR coefficients, which are better-established explainability and interpretability techniques (respectively). Experimental details for the SHAP value and LR coefficient calculation process were provided in Section 4.3. We consider three scenarios: *(a)* the metrics applied to the high performing optima discussed in Section 5.1, *(b)* the metrics applied to the cluster identified to contain the pseudo-global optima (*Cluster 1)*, and *(c)* the metrics applied to a LR model trained using the full dataset (no feature



(a) Occurrence of features in clusters of LON in high fitness optima



(b) Correlations between rankings of features according to each technique in clusters of high fitness optima

Figure 4: Describing the distribution of features in the identified clusters and their correlations with existing metrics

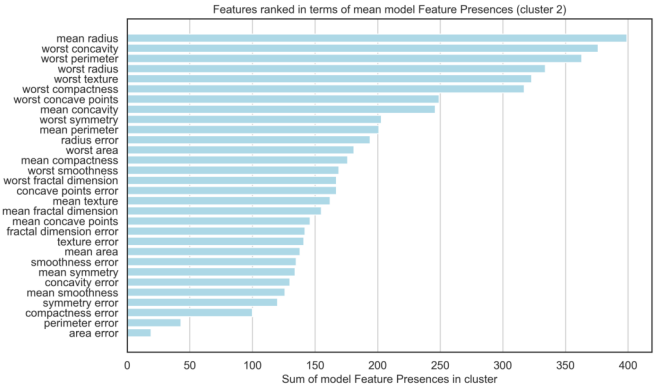
selection). First, we begin by considering *(a)* the metrics when applied to the high performing optima. Specifically, we compare the presence of features that are most commonly selected in high performing solutions, their mean coefficients and their SHAP values. As seen in Figure 5, the distribution of commonly selected features (5a) was much wider than what we may have expected through consideration of the SHAP values (5c) and the model coefficients (5b) alone. This indicates that regions of the search landscape that are of higher fitness are not particularly well explained by the *interpretability* or *explainability* aspects of the models generated from solutions found within them. This is further exemplified by the correlations between the feature rankings. Figure 5d demonstrates the rankings of features according to their SHAP values in comparison to those most commonly found within the LON, while Figure 5e demonstrates the rankings of the features according to mean model coefficients of solutions within the LON. The relationship between feature presence and model coefficients is moderately correlated (Kendall Tau correlation of 0.5661) while the relationship between the explanability technique, SHAP, was only weakly correlated with a Kendall Tau correlation of 0.3038. This indicates that there may factors that determine the LON that are not captured by common model *interprability* and *explainability* techniques.

In order to more accurately describe regions of the solution space, let us consider the clusters derived from section 5.2. As seen in Figure 4a, the mean feature presence in each cluster differs but correspond to what we might expect to observe from the dendrogram (Figure 3) — *Clusters 2* and *3* are most similar, followed by Cluster 1, but each remaining quite distinct. As in the previous analysis, we compare the model coefficients and SHAP values with the feature presence in the LON, but in this case we will compare it to more homogeneous clusters of nodes. Specifically, we present *(b)* the metrics applied to the cluster identified to contain the pseudo-global optima (*Cluster 1)*, and provide the results for *Clusters 2, 3, 4, and 5* as supplementary material. Even in this more restrained region of the space, we can still see that the feature presence (Figure 6a) is not adequately reflected in the rankings according to the SHAP values (6b) or model coefficients (6c). In fact, when we consider the correlations between the rankings as described by the metrics, we see almost identical relationship strengths as in the larger subset (correlations with feature presence of 0.3898 and 0.5161 between SHAP and model coefficients, respectively).

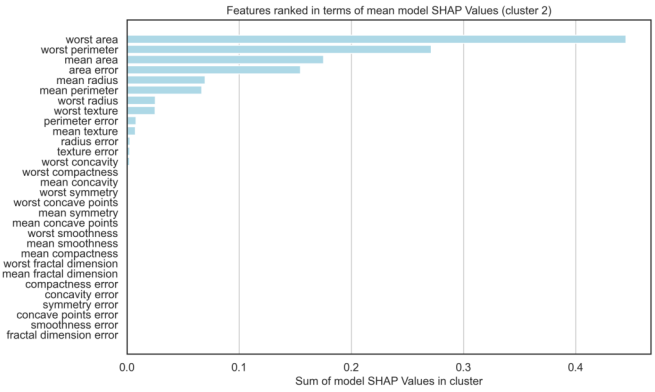
In order to describe this phenomenon across each of the clusters, Kendall Tau correlation was calculated, for each cluster, for each of the following rankings: SHAP values and Feature Presence (Shaps vs feat_pres), LR model coefficients and Feature Presence (Coeff vs feat_pres), LR model coefficients and SHAP values (Coeff vs Shaps). In addition, we seek to explore whether the features that would prove dominant in the LONs could be ascertained from the model coefficients and SHAP values *before* feature selection; in other words, *(c)* application of the metrics to a LR model trained using the full dataset (no feature selection). As seen in Figure 4b, the correlation between the feature rankings according to the SHAP values is typically strong — indicating that, even though the feature subsets change, the feature contributions to the predictions remain relatively stable. This can be contrasted with the correlations between the model coefficients across different clusters and that of the full model which vary from weak to strong, indicating that the model parameters are changing considerably during feature selection (a caveat to this is that model coefficients may also be unstable due to multicollinearity). In terms of feature presence however, the correlations between the full model's coefficients and SHAP scores remained weak (with a maximum of 0.45 between feature presence and coefficients in *Cluster 4*), suggesting that removal of features in advance of performing wrapper-based feature selection may limit valuable regions of the solution space.

6 CONCLUSIONS

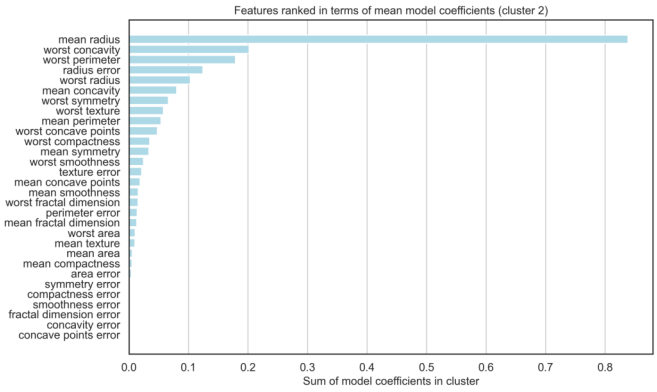
In this work, the use of fitness landscapes associated with evolutionary feature selection towards the aim of explainable artificial intelligence (XAI) has been explored. To this end, we constructed local optima networks (LONs) for feature selection of a well-known machine learning (ML) dataset: the Wisconsin breast cancer dataset. We explored and proposed new ways of mining information from LONs with the purpose of explaining ML Models. Through utilisation of LONs, we demonstrate that high fitness regions of the landscape are not adequately explained using popular *interpretability* (model coefficients) and *explainability* (SHAP values) techniques, calling into question the efficacy of reducing the feature set using



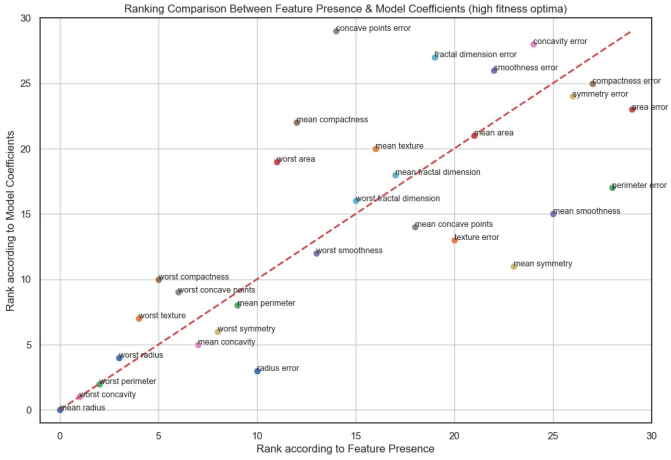
(a) Occurrence of features in high fitness optima in LON



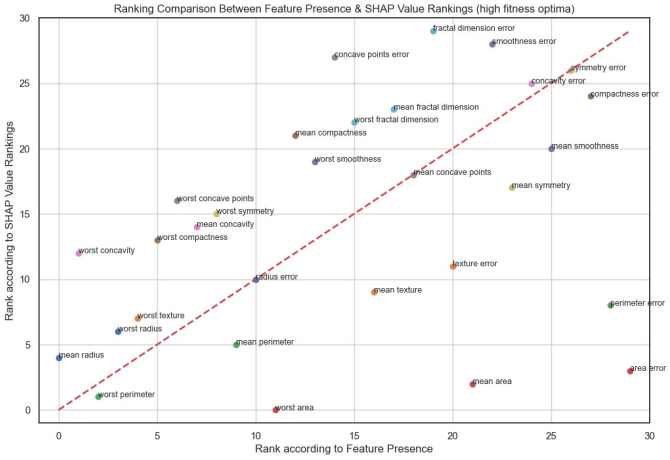
(b) SHAP values for LON nodes in high fitness optima in LON



(c) Logistic coefficients for LON nodes in high fitness optima in LON



(d) SHAP values for LON nodes in high fitness optima



(e) Occurrence of features in LON in high fitness optima

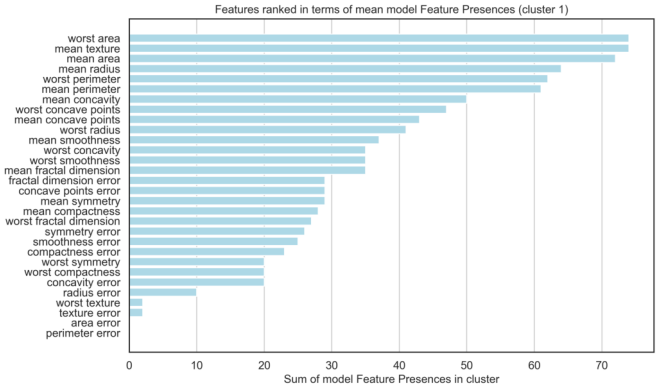
Figure 5: Comparison of feature rankings according to mean model coefficients, SHAP values, and feature presence in high fitness local optima discovered by the ILS sampling technique

these techniques in advance of deploying wrapper-based methods. It is hoped that this work will inspire consideration of fitness landscape analysis for inclusion in the XAI toolbox.

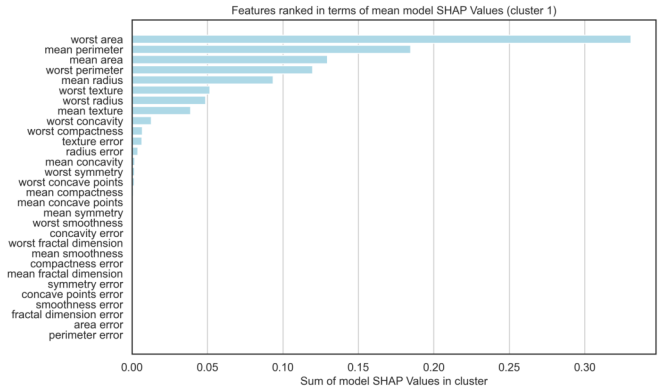
Data Publishing. The data from this work will be made publicly available upon acceptance.

REFERENCES

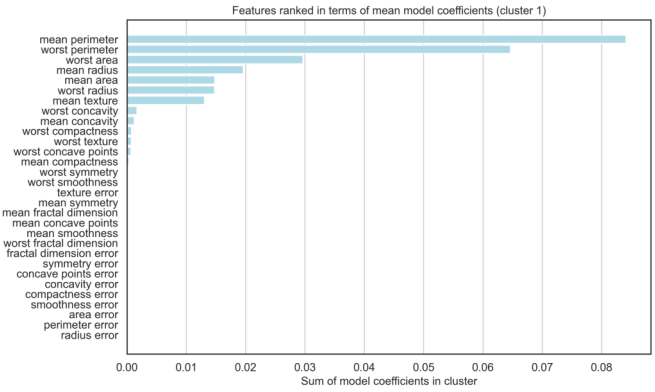
[1] Jason Adair, Alexander Brownlee, and Gabriela Ochoa. 2016. Evolutionary Algorithms with Linkage Information for Feature Selection in Brain Computer



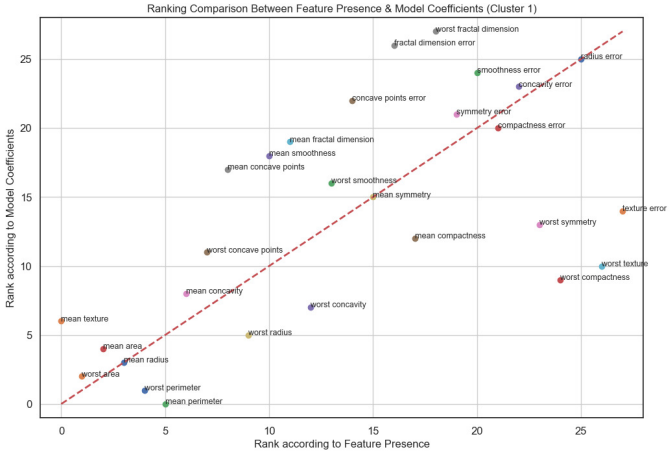
(a) Occurrence of features in *Cluster 1*



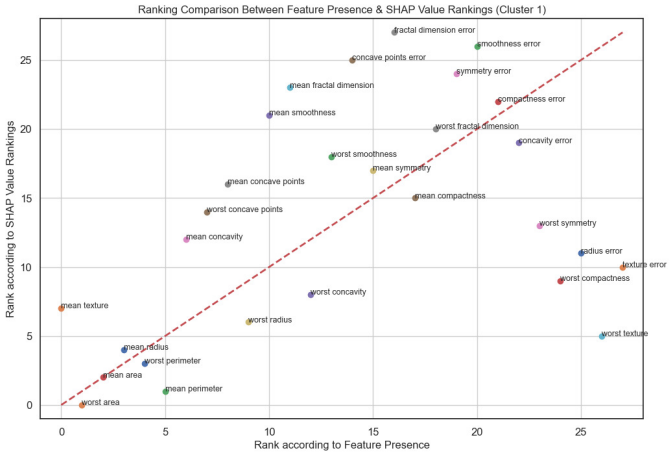
(b) SHAP values for features in *Cluster 1*



(c) Logistic coefficients for features in *Cluster 1*



(d) Feature rankings according to the model coefficients compared with presence in the cluster containing the discovered pseudo-global optima



(e) Occurrence of features in LON in high fitness optima

Figure 6: Exploration of feature rankings according to the SHAP values and model coefficients in comparison with feature presence in the cluster containing the discovered pseudo-global optima

Interfaces. In *Advances in Intelligent Systems and Computing*. Springer Nature, 287–307. https://doi.org/10.1007/978-3-319-46562-3_19

[2] Jason Adair, Sarah L. Thomson, and Alexander Brownlee. 2023. Local Optima Networks for Explainable Artificial Intelligence. (2023).

[3] Miller Janny Ariza-Garzón, Javier Arroyo, Antonio Caparrini, and Maria-Jesus Segovia-Vargas. 2020. Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access* 8 (2020), 64873–64890.

[4] Jaume Bacardit, Alexander EI Brownlee, Stefano Cagnoni, Giovanni Iacca, John McCall, and David Walker. 2022. The intersection of evolutionary computation and explainable AI. In *Proceedings of the Genetic and Evolutionary Computation*

Conference Companion. 1757–1762.
[5] Ben Carterette. 2009. On rank correlation and the distance between rankings. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 436–443.
[6] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).
[7] Tansel Dokeroglu, Ayça Deniz, and Hakan Ezgi Kiziloz. 2022. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing* 494 (2022), 269–296.
[8] Tom Jansen, Gijs Geleijnse, Marissa Van Maaren, Mathijs P Hendriks, Annette Ten Teije, and Arturo Moncada-Torres. 2020. Machine learning explainability in breast cancer survival. In *Digital Personalized Health and Medicine*. IOS Press, 307–311.
[9] Femke M Janssen, Katja KH Aben, Berdine L Heesterman, Quirinus JM Voorham, Paul A Seegers, and Arturo Moncada-Torres. 2022. Using Explainable Machine Learning to Explore the Impact of Synoptic Reporting on Prostate Cancer. *Algorithms* 15, 2 (2022), 49.
[10] Ryota Kitani and Shinya Iwata. 2023. Verification of Interpretability of Phase-resolved Partial Discharge using a CNN with SHAP. *IEEE Access* (2023).
[11] Sang Gyu Kwak and Jong Hae Kim. 2017. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology* 70, 2 (2017), 144–156.
[12] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
[13] Wilson E Marcílio and Danilo M Eler. 2020. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee, 340–347.
[14] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review* 55, 5 (2022), 3503–3568.
[15] Christopher Z Mooney, Christopher F Mooney, Christopher L Mooney, Robert D Duval, and Robert Duvall. 1993. *Bootstrapping: A nonparametric approach to statistical inference*. Number 95. sage.
[16] Werner Mostert, Katherine M Malan, Gabriela Ochoa, and Andries P Engelbrecht. 2019. Insights into the feature selection problem using local optima networks. In *European Conference on Evolutionary Computation in Combinatorial Optimization (Part of EvoStar)*. Springer, 147–162.
[17] Gabriela Ochoa and Sebastian Herrmann. 2018. Perturbation strength and the global structure of QAP fitness landscapes. In *International Conference on Parallel Problem Solving from Nature*. Springer, 245–256.
[18] Gabriela Ochoa, Marco Tomassini, Sebástien Vérel, and Christian Darabos. 2008. A study of NK landscapes' basins and local optima networks. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. 555–562.
[19] Gabriela Ochoa and Nadarajen Veerapen. 2022. Neural Architecture Search: A Visual Analysis. In *International Conference on Parallel Problem Solving from Nature*. Springer, 603–615.
[20] Lucas Marcondes Pavelski, Marie-Éléonore Kessaci, and Myriam Delgado. 2021. Local Optima Network Sampling for Permutation Flowshop. In *2021 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1131–1138.
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[22] Isak Potgieter, Christopher W Cleghorn, and Anna S Bosman. 2022. A Local Optima Network Analysis of the Feedforward Neural Architecture Space. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[23] Michal W Przewozniczek and Marcin M Komarnicki. 2022. Empirical linkage learning for non-binary discrete search spaces in the optimization of a large-scale real-world problem. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 35–36.
[24] Michal Witold Przewozniczek, Renato Tinós, and Marcin Michal Komarnicki. 2023. First Improvement Hill Climber with Linkage Learning–on Introducing Dark Gray-Box Optimization into Statistical Linkage Learning Genetic Algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 946–954.
[25] Sam J Silva, Christoph A Keller, and Joseph Hardin. 2022. Using an Explainable Machine Learning Approach to Characterize Earth System Model Errors: Application of SHAP Analysis to Modeling Lightning Flash Occurrence. *Journal of Advances in Modeling Earth Systems* 14, 4 (2022), e2021MS002881.
[26] Peter F. Stadler. 2002. Fitness landscapes. *Biological Evolution and Statistical Physics. Lecture Notes in Physics* 585 (2002), 183–204.
[27] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.
[28] Matheus Cândido Teixeira and Gisele Lobo Pappa. 2022. Analysis of Neutrality of AutoML Search Spaces with Local Optima Networks. In *Intelligent Systems*, João Carlos Xavier-Junior and Ricardo Araújo Rios (Eds.). Springer International Publishing, Cham, 473–487.
[29] Matheus C Teixeira and Gisele L Pappa. 2022. Understanding AutoML search spaces with local optima networks. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 449–457.
[30] Sarah L Thomson, Jason Adair, Alexander EI Brownlee, and Daan van den Berg. 2023. From fitness landscapes to explainable AI and back. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*. 1663–1667.
[31] Renato Tinós, Michal Przewozniczek, Darrell Whitley, and Francisco Chicano. 2023. Genetic Algorithm with Linkage Learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 981–989.
[32] Risto Trajanov, Stefan Dimeski, Martin Popovski, Peter Korošec, and Tome Eftimov. 2021. Explainable landscape-aware optimization performance prediction. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 01–08.
[33] Sébastien Verel, Fabio Daolio, Gabriela Ochoa, and Marco Tomassini. 2012. Local optima networks with escape edges. In *Artificial Evolution: 10th International Conference, Evolution Artificielle, EA 2011, Angers, France, October 24-26, 2011, Revised Selected Papers 10*. Springer, 49–60.
[34] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 3 (2020), 261–272.
[35] Raymond E Wright. 1995. Logistic regression. (1995).
[36] Christopher Yeung, Ju-Ming Tsai, Brian King, Yusaku Kawagoe, David Ho, Mark W Knight, and Aaswath P Raman. 2020. Elucidating the behavior of nanophotonic structures through explainable machine learning algorithms. *ACS Photonics* 7, 8 (2020), 2309–2318.
[37] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. 2019. Deep neural network or dermatologist?. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9*. Springer, 48–55.
[38] Quan Zheng, Ziwei Wang, Jie Zhou, and Jiwen Lu. 2022. Shap-CAM: Visual Explanations for Convolutional Neural Networks Based on Shapley Value. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*. Springer, 459–474.