25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Visualising instance selection for improved explainability using feature extraction

G. F. A. Yeo [a], I. Hudson [a], D. Akman [a] and J. Chan [b]

[a] School of Science (Mathematical Sciences), STEM College, Royal Melbourne Institute of Technology, 124 La Trobe St, Melbourne, VIC, Australia
[b] School of Computing Technologies (Computer Science), STEM College, Royal Melbourne Institute of Technology, 124 La Trobe St, Melbourne, VIC, Australia
Email: anders.yeo@rmit.edu.au

Abstract: Advancement in the collection and storage of data, alongside the modern emphasis of automated decision making have lead to datasets growing exponentially in size and complexity over the last four decades. Lever- aging excessively large data through traditional machine learning can lead to exorbitant run times, storage and general computational bloat, with the trained model potentially being sub-optimal (Kohavi & John 1997). Dimensionality reduction through selection and extraction are common methods of mitigating these issues. Extraction methods map the existing data to lower dimensional space whilst attempting to maintain the characteristics of the original dataset, whilst selection methods attempt to take a representative subset of the data. Alongside elevating the technical computational bloat, data reduction provides a parsimonious representation of the dataset, resulting in comparably simpler models which are more intrinsically interpretable. Therefore, data reduction techniques are included within Explainable Artificial Intelligence (XAI) (Barredo Arrieta et al. 2020). With the increasing reliance on automated decision making, the number of publications related to XAI has increased rapidly over the last ten years. A machine learning model should be not only accurate, but also transparent with an interpretable logic for proposed predictions. Therefore, machine learning models are used for both predictive purposes and retrospective data exploration and analysis.

SpFixedIS is a fixed wrapper instance selection algorithm (Yeo et al. 2023), wherein the number of instances to select are user defined. The algorithm attempts to find the most representative instances of set cardinality with respect to a machine learning model and corresponding performance metric. Within a predictive context, the instances selected through SpFixedIS are able to accurately predict unforeseen observations. Within this paper, we examine the instances selected through SpFixedIS within the application of retrospective explainable data exploration through visualisation. The proposed method will reduce the instance space using SpFixedIS and then map the feature space to two dimensional space using a feature extraction method for ease of visualisation. The feature extraction methods presented are Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018).

SpFixedIS uses a 1-Nearest Neighbours classifier with prediction accuracy as the wrapper and performance metric respectively on the Primate Splice-Junction Gene Sequences DNA dataset. The stability of the instances selected through SpFixedIS across a changing number of instances is examined through the lens of the instances themselves, the unselected instances and through the feature extraction projections in the form of cluster centres. The expected behaviour is exhibited as the number of instances increases with the inherent noise within the dataset being transferred from the unselected instances to the selected instances subset. The reduced dataset cluster centres gravitate away from the cluster centres of the full dataset when the number of instances decreases, therein providing more separation between different classes, however this difference is not reflected uniformly through the feature projections. The projections reveal insight into the structure of the data that metrics alone do not. The UMAP representation provided insight into repeated selection of outlier instances which were repeated identical instances within the dataset. The juxtaposition of instances selected and ease of visualisation provided through feature extraction allow for a holistic view of the instances selected. Therefore, the instance could be visualised and identified within the grand structure of the full dataset. This study is motivated by future work in relation to structure-based kinase peptide interactions in order to locate outliers and unique structures for future drug discovery (Liu et al. 2020).

Keywords: *Instance selection, data visualisation, data reduction, SpFixedIS, UMAP*

1 INTRODUCTION

Modern datasets are growing in size and complexity, which subsequently leads to new challenges and issues for conventional machine learning and data mining methods. These problems can be broadly summarised as two distinct issues. The first issue is technical and pertains to computational bloat, including excessive storage space and run times. Whereas the second and potentially more serious issue is sub-optimal models, which become present in the form of over-fitting, fitting noisy entries or redundant features. Reduced model performance occurs when the dataset is excessively large in both the feature space (Kohavi & John 1997) and instance space (Yeo et al. 2023). Dimensionality reduction is a common method for mitigating the aforementioned issues. Data reduction techniques can be differentiated into two distinct categories, namely selection and extraction techniques. Selection algorithms aim to select a subset of the available data, whereas extraction methods generate new entries in place of the existing dataset. Selection and extraction algorithms can be applied to both features and instances.

Dimensionality reduction, specifically selection methods, are within the Explainable Artificial Intelligence (XAI) toolbox albeit for different purposes within the instance and feature space (Barredo Arrieta et al. 2020). The XAI field of research is becoming increasingly important as more reliance is placed on automated decision making. A decision produced by artificial intelligence must not only be accurate but also provide an interpretable logic allowing for transparent decision making. Broadly, there are two categories of XAI methods, namely post-hoc and intrinsically interpretable methods. Post-hoc methods perform an additional procedure after the model is established in order to provide insight to a prediction or model logic, well-known methods include Local Interpretable Model Agnostic Explanation (LIME) (Ribeiro et al. 2016). Whilst verifying the evaluation of post-hoc methods can be done through external means and domain knowledge (Adadi & Berrada 2018), post-hoc methods, specifically post-hoc feature importance methods, have inconsistent and mixed efficacy representing either the models or dataset (Yeo et al. 2022). On the other hand, the intrinsically interpretable instance selection purpose is two-fold in XAI, the instances themselves are a simpler representation of the dataset and the models based on a smaller subset tend to be simpler and therefore more interpretable. That is, the improved explainability may be viewed through the lens of the data itself or through the model.

Recently we proposed SpFixedIS (Yeo et al. 2023), a fixed instance selection algorithm which allowed for a user specified number of instances to select. Through a stochastic optimisation framework, SpFixedIS attempts to select a subset of instances which optimise a given performance with regards to a particular machine learning model. SpFixedIS is able to train models which were able to outperform or maintain a statistically equivalent predictive performance once a sufficient number of instances were selected. Therein, a representative mapping of the feature space was obtained through the intelligent selection of instances with regards to a classification model. The algorithm was explored from the perspective of model performance, specifically accurately predicting previously unseen data.

In this study we propose the novel procedure of coupling instance selection and feature extraction in order to analyse and examine the selected instance subset of instance selection from the lens of the data itself and retrospective data exploration. The reduced dimensionality through feature extraction enables the ease of visualisation and provides a simple means to verify the selected instances are representative of the full dataset.

2 BACKGROUND

2.1 Fixed Instance Selection

SpFixedIS is a fixed, wrapper instance selection method. A fixed instances selection method is an algorithm which seeks to select a subset of instances with a predefined cardinality, such that each iteration possesses a real valued solution with this pre-set cardinality. Generally, instance selection algorithms are seeking to optimise two criteria, namely performance and compression. Different instance algorithms will prioritise performance preservation over selecting a minimal subset. Fixed instance selection methods remove the compression constraint by setting the number of instances to select and seeking an optimal subset of a given cardinality. On the other hand, the distinction between a filter or wrapper method dictates whether a model needs to be supplied to the algorithm, specifically wrapper methods require additional models to be fitted during the learning process whereas filter methods do not. Therefore, SpFixedIS requires a model with a corresponding performance metric and the number of instances to select alongside the data.

The fixed instance selection wrapper method can be formalised as follows. Given a dataset, T, with n instances and p descriptive features, a machine learning model, C, with a performance metric y_C and the number of

instances to select, k. The dataset can be represented as $T = \{X, Y\}$, where X is the $n \times p$ data matrix and Y is the corresponding response feature of length n. Let $S := S \subset T$ represent a subset of the instances to select of fixed cardinality where $|S| = k$ and conversely denote the remaining unselected instances as $S' := T \backslash S$. Therefore we seek to find the optimal set of instances, X^* which maximises y_C with regard to the unselected instances S' such that

$$X^* := \arg\max_S y_C(S').$$

The instances selected seek to generalise the remaining instances through correctly predicting these unselected instances. However, changing k results in a different number of instances being selected and therefore potentially different behaviours depending on the wrapper, C.

SpFixedIS treats the fixed instance selection as a combinatorial optimisation problem and follows a pseudo-gradient descent framework to approach X^*. There is no closed form mathematical representation for y_C, such that the search direction is defined as the secant from two randomly perturbed noisy measurements to the current solution. Whereas the gain sequence is defined by an approximation of the Hessian matrix utilising the Barzilai and Borwein method (Barzilai & Borwein 1988). For more details refer to Yeo et al. (2023).

2.2 Feature Extraction

Feature extraction methods involve projecting the existing features into a lower dimensional space whilst attempting to maintain the characteristics of the original features. Two well-known feature extraction methods were considered in this study, Principal Component Analysis (PCA) and the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018). PCA provides an orthogonal linear transformation to the data whilst UMAP uses applied Riemannian geometry and algebraic topology to provide a non-parametric transformation.

A recent study related to a visualisation framework which uses UMAP to produce explainable outcomes is XMAP (Nguyen & Tran 2021). XMAP is a procedure which attempts to capture the distributions and topological structures of data, define contexts, and build representations for classification tasks. The four steps during the XMAP procedure are: data pre-processing, applying mapping techniques, learning topology and extracting interpretable contexts. The mapping technique used in XMAP is UMAP in order to reduce the dataset to two dimensions, then the topological learning is an instance extraction method based on nodes which behave like cluster centres is then performed on the UMAP projections. The nodes are then evaluated using information theoretics to produce explanations similar to the information provided by a cluster analysis. XMAP provides an interesting framework which essentially performs two extraction techniques, first in the feature space then in the instance space, followed by unpacking the information provided by the extraction methods. The general drawback of extraction techniques is the interpretability of the reduced dataset can be hindered by the projections, conversely selection methods are taking a subset of the available data rendering the interpretability intact.

3 METHODOLOGY

3.1 DATASETS

The dataset we used is the Primate Splice-Junction Gene Sequences (DNA) dataset from the OpenML Suite (Bischl et al. 2019) with the raw data sourced from the UCI machine learning repository (Dua & Graff 2017). The DNA dataset is a real world dataset composed of 3186 observations, 180 descriptive binary features and 3 classes for the target feature. Although not delved into during this study, it is interesting to note the problem description of this dataset. DNA sequences were taken from primates in order to study splice junctions. During the process of protein creation, splice junctions are the points on a DNA sequence in which superfluous DNA is removed. The parts of a DNA sequence retained after splicing are called exons and the discarded sequence are called introns. The classification problem involves recognising the boundaries between exons and introns, introns and exons and the third state which is neither.

3.2 Experimental Design

This study is focused on investigating the subset of instances selected with the SpFixedIS method through the lens of feature extraction techniques. SpFixedIS selected a range of instances in increments of 50 whilst using the original descriptive features. A 1-Nearest Neighbours model, with classification accuracy as the performance metric, was used as the wrapper for SpFixedIS. 1-Nearest Neighbours was chosen due to being the

most sensitive to instance selection. Classification accuracy was selected as there were no highly imbalanced classes within the target feature.

The projections provided by PCA and UMAP were obtained for the full set of instances such that the full structure of the dataset may be visualised and presented as a ground truth baseline. The selected instances were displayed on the projected axis in order to give a clear indication of where the instances appear within the latent structure of the dataset. It should be noted that the target feature is not included in the projections. The cluster centres of the projections were also calculated for each class in order to examine the representativeness of the selected subset of instances.

4 RESULTS AND DISCUSSION

Figure 1 presents the unselected accuracy and the mean accuracy from a leave one out (LOO) cross validation over the selected instances. Figure 2 and Figure 3 present the instances selected by SpFixedIS projected into two dimensional space by UMAP and PCA, respectively. The number of instances selected begin at 3,000 instances and halved each time, with the floor taken if it is not a multiple of 50. It is worth noting that the 50 instances displayed in Figure 2 are the same 50 instances displayed in Figure 3, the instances are the same selected using SpFixedIS however projected into two dimensions using different methods. On the other hand, Figure 4 shows the mean Euclidean distance between each cluster centres of the visualised reduced datasets.

Figure 1. Unselected Accuracy and Mean Leave One Out (LOO) Cross Validation Accuracy across selected instances on the DNA dataset.

The unselected accuracy reflects the selected instance's ability to generalise the remaining instances within the dataset through the lens of the model. Meanwhile, the LOO cross validation of the selected instances with respect to nearest neighbours models provide a measure of purity from within the selected instances. This interpretation changes when using non instance based techniques. These metrics show the trade-off across selecting different number of instances in Figure 1.

Nearest neighbours models provide a clear distinction between useful and redundant instances, that is for any given subset there are instances which constitute the decision boundary which ultimately form the predictions. With the increase in the number of features, these boundary points tend to become more difficult to discern. From 3000 to 1500 instances selected, the unselected accuracy increases to perfectly classifying the unselected whilst the mean LOO cross validation accuracy decreases. That is, the bias-variance trade-off in the selected instances, wherein the decision boundary is captured in its entirety at the expense of a potentially more parsimonious model through reduced instances. Examining how the 3000 to 1500 instances appear through the feature extraction methods, there does not appear to be much of a difference visually. This extends to the 750 instances selected for both UMAP and PCA, in which the subsequent larger subsets appear as a denser variation. This interpretation is supported by the relative lack of movement in cluster centres between 750, 1500 and 3000 instances selected.

Shifting the attention to the smaller number of instances selected, the two accuracy measures in Figure 1 appear to be more erratic. This noise can be mitigated with repeated runs. Regardless, a lower number of instances provide a multitude of combinations to construct a similar decision boundary, such that the mean LOO cross validation is inherently noisy. The instances selected by SpFixedIS not only attempt to reproduce the decision boundary but also provide a representative mapping of the full feature set regardless of the number of instances selected. This mapping is most evident in the projections with fewer instances selected, such that the selected instances are not localised but spread to give an impression of the full projected dataset.

The cluster centres of the dataset when juxtaposed alongside the original cluster centres of the full dataset provide a clear display of the selected instance behaviour with respect to the projected method. The cluster

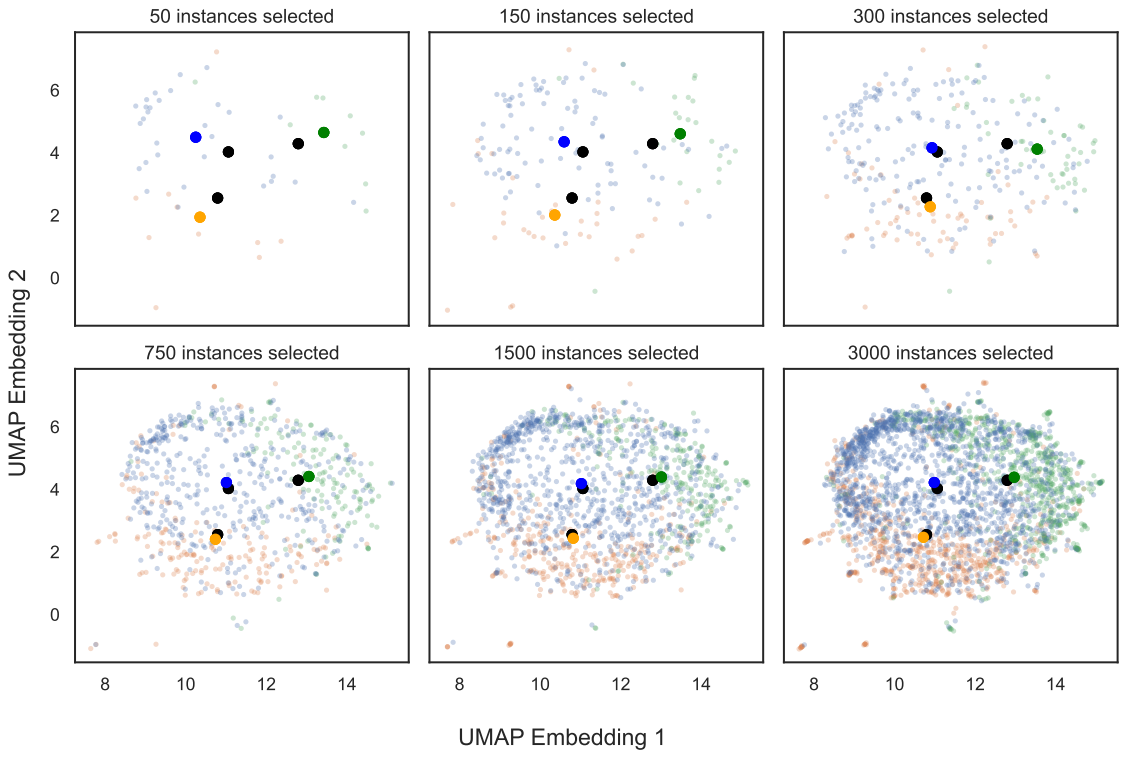

Figure 2. UMAP Projections of SpFixedIS selected instances on the DNA dataset.

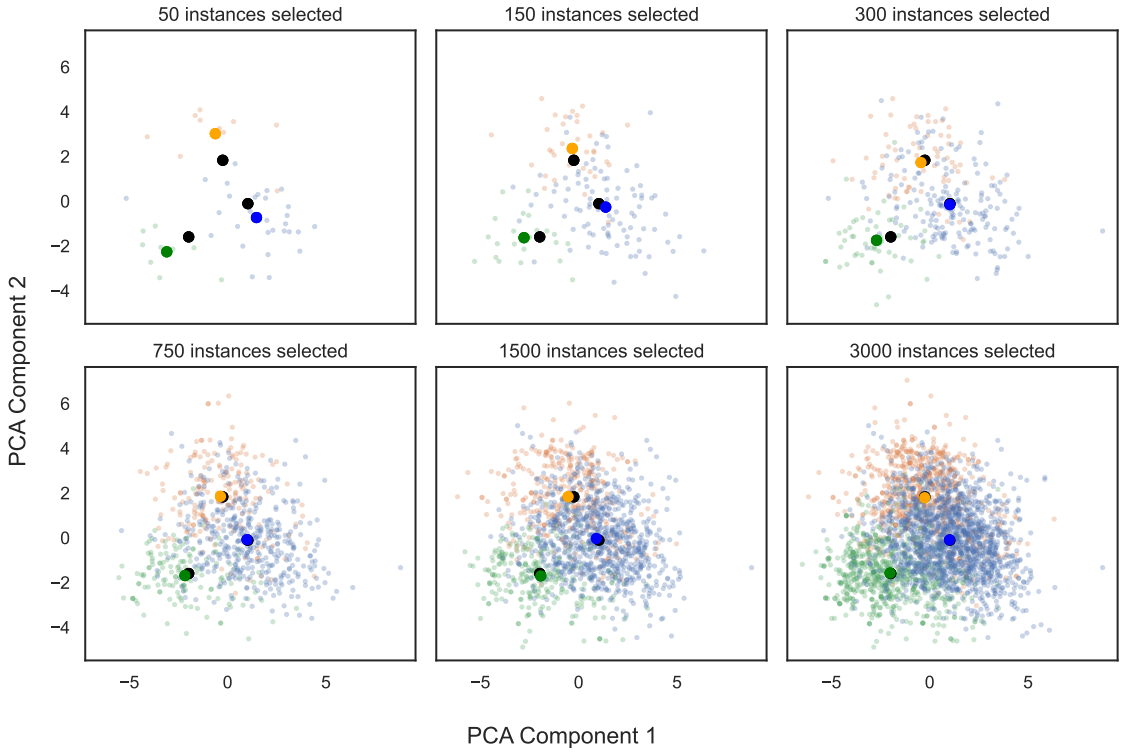

Figure 3. PCA Projections of SpFixedIS selected instances on the DNA dataset.

centres of each class gravitate away from those of opposing classes with the distance greater the fewer instances are selected which occurs for both the non-linear UMAP and linear PCA projections. Relative to the positions of the original cluster centres, the fewer the number of instances selected the further away the cluster centres move. This behaviour is clearly seen in the difference in distance cluster centres presented in Figure 4, specifically the difference in the original features which provide an almost uniform distance between each set of selected instances. This is the expected behaviour, the more instances selected the more the distribution reflects that of the original dataset.

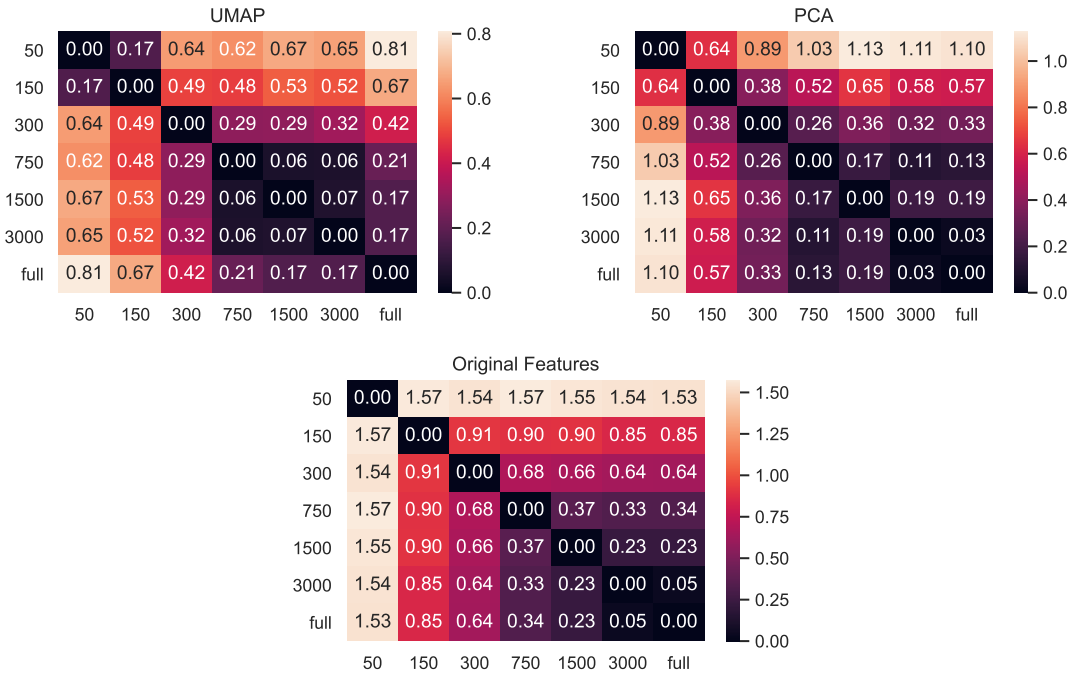

Figure 4. Mean Euclidean distance difference in cluster centres.

The projections of the datasets themselves reveal insight into the structure of the data that metrics alone do not. As a broad statement, there are overlapping clusters of each class with outliers which is evident within both methods of feature extraction. The different instances selected using SpFixedIS are independent of one another, such that the instances in the subset of 50 have no guarantee on being included within the subset of 150. However, it is interesting to note that within the UMAP projection SpFixedIS consistently selects what appear to be displaced outliers away from the main cluster of instances. This is due to many of those perceived outlier points being identical entries within the original dataset. SpFixedIS attempts to select a subset of fixed size which can generalise the remaining instances through the lens of a model, therefore by selecting one of these identical entries, the subsequent copies will be correctly classified. Therefore, the instances provided by SpFixedIS and the ease of visualisation offered by feature extraction return a subset of representative instances which can allow for an organised assortment of unique instances allowing for easier identification of insights.

In terms of comparing the UMAP projections to that of PCA, this is an interesting dataset due to the binary descriptive feature combination. The different cluster centres for the PCA projections line up more closely to the original difference in cluster centres in Figure 4 compared to that of UMAP. However, the UMAP projection in Figure 2 provide insight through the induced structure into displaced outliers which are not present within the PCA projections in Figure 3. Although, PCA has the additional benefit of the projections being a linear combination of the original instances rendering it highly interpretable, the projections themselves need not be inherently interpretable with the instance space reduced and the original features intact.

5 CONCLUSIONS AND FUTURE WORK

This paper presents the fixed instances selected by SpFixedIS projected into two dimensional space using the feature extraction techniques UMAP and PCA. The projections provide a means to visualise the instances selected with respect to their structure within the full dataset, allowing for a holistic view of the instances selected. The instances provided by SpFixedIS coupled with the ease of visualisation through feature extraction,

allow a subset of representative instances to be arranged as an organised assortment in order to produce rapid identification of insights.

This study and the dataset selected for this study were motivated by future work related to Kinase Protein interactions (KPIs) (Bradley & Beltrao 2019). Kinase proteins are involved in signalling and maintaining cell behaviour (Cunningham et al. 2017), such that any deviation may result in a range of diseases including cancer (Gross et al. 2015). Our future study will analyse KPIs using machine learning, specifically structure-based kinase peptide interactions, from a dataset constructed from protein kinase structures curated from the RCSB Protein Data Bank (Basse et al. 2016). The dataset is composed of binary descriptive features with 6 classes of the target feature corresponding to the 6 most populous human kinase groups (Manning et al. 2002). The method proposed in this paper will be applied and extended to allow for rapid identification of unique protein structures, thereby potentially helping to identify protein structures within different Kinase groupings by a domain expert. This research is in progress to ultimately identify so called outliers and unique KPI structures for future drug discovery (Liu et al. 2020). Our results will be compared to deep learning mixture analytics and classification performed on a reduced data set of the KPI data (using only peptide interaction information) in collaboration with Marseilles University (Hudson et al, in prep).

REFERENCES

Adadi, A. & Berrada, M. (2018), 'Peeking inside the black-box: A survey on explainable artificial intelligence (xai)', *IEEE Access* **6**, 52138–52160.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020), 'Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai', *Information Fusion* **58**, 82–115.

Barzilai, J. & Borwein, J. (1988), 'Two-point step size gradient methods', *IMA Journal of Numerical Analysis* **8**, 141–148.

Basse, M.-J., Betzi, S., Morelli, X. & Roche, P. (2016), '2p2idb v2: update of a structural database dedicated to orthosteric modulation of protein–protein interactions', *Database* **2016**.

Bischl, B., Casalicchio, G., Feurer, M., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N. & Vanschoren, J. (2019), 'Openml benchmarking suites'.

Bradley, D. & Beltrao, P. (2019), 'Evolution of protein kinase substrate recognition at the active site', *PLoS biology* **17**(6), e3000341.

Cunningham, A. D., Qvit, N. & Mochly-Rosen, D. (2017), 'Peptides and peptidomimetics as regulators of protein–protein interactions', *Current opinion in structural biology* **44**, 59–66.

Dua, D. & Graff, C. (2017), 'UCI machine learning repository'.

Gross, S., Rahal, R., Stransky, N., Lengauer, C., Hoeflich, K. P. et al. (2015), 'Targeting cancer with kinase inhibitors', *The Journal of clinical investigation* **125**(5), 1780–1789.

Kohavi, R. & John, G. H. (1997), 'Wrappers for feature subset selection', *Artificial Intelligence* **97**(1), 273–324.

Liu, C., Ke, P., Zhang, J., Zhang, X. & Chen, X. (2020), 'Protein kinase inhibitor peptide as a tool to specifically inhibit protein kinase a', *Frontiers in Physiology* **11**, 574030.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002), 'The protein kinase complement of the human genome', *Science* **298**(5600), 1912–1934.

McInnes, L., Healy, J. & Melville, J. (2018), 'Umap: Uniform manifold approximation and projection for dimension reduction'.

Nguyen, S. & Tran, B. (2021), 'Xmap: explainable mapping analytical process', *Complex & Intelligent Systems* pp. 1–18.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), "why should i trust you?": Explaining the predictions of any classifier, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 1135–1144.

Yeo, A., Hudson, I., Akman, D. & Chan, J. (2022), A simple framework for xai comparisons with a case study, *in* '5th International Conference on Artificial Intelligence and Big Data (ICAIBD)', pp. 501–508.

Yeo, G. F. A., Akman, D., Hudson, I. & Chan, J. (2023), 'A stochastic approximation approach to fixed instance selection', *Information Sciences* **628**, 558–579.
URL: *https://doi.org/10.1016/j.ins.2023.01.090*