Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain

Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly

Cork Institute of Technology
Urja.Pawar@mycit.ie, Donna.OShea@cit.ie, Susan.Rea@cit.ie,
Ruairi.OReilly@cit.ie

Abstract. In the healthcare domain, Artificial Intelligence (AI) based systems are being increasingly adopted with applications ranging from surgical robots to automated medical diagnostics. While a Machine Learning (ML) engineer might be interested in the parameters related to the performance and accuracy of these AI-based systems, it is postulated that a medical practitioner would be more concerned with the applicability, and utility of these systems in the medical setting. However, medical practitioners are unlikely to have the prerequisite skills to enable reasonable interpretation of an AI-based system. This is a concern for two reasons. Firstly, it inhibits the adoption of systems capable of automating routine analysis work and prevents the associated productivity gains. Secondly, and perhaps more importantly, it reduces the scope of expertise available to assist in the validation, iteration, and improvement of AI-based systems in providing healthcare solutions. Explainable Artificial Intelligence (XAI) is a domain focused on techniques and approaches that facilitate the understanding and interpretation of the operation of ML models. Research interest in the domain of XAI is becoming more widespread due to the increasing adoption of AI-based solutions and the associated regulatory requirements [1]. Providing an understanding of ML models is typically approached from a Computer Science (CS) perspective [2] with a limited research emphasis being placed on supporting alternate domains [3]. In this paper, a simple, yet powerful solution for increasing the explainability of AI-based solutions to individuals from non-CS domains (such as medical practitioners), is presented. The proposed solution enables the explainability of ML models and the underlying workflows to be readily integrated into a standard ML workflow. Central to this solution are feature importance techniques that measure the impact of individual features on the outcomes of AI-based systems. It is envisaged that feature importance can enable a high-level understanding of a ML model and the workflow used to train the model. This could aid medical practitioners in comprehending AI-based systems and enhance their understanding of ML models' applicability and utility.

Keywords: Explainable Artificial Intelligence · Healthcare · Feature Importance · Decision trees · Explainable Underlying Workflow

1   Introduction

Interpretability is the degree to which the rationale of a decision can be observed within a system [4]. If a ML model's operation is readily understood then the model is interpretable. Explainability is the extent to which the internal operation of a system can be explained in human terms. XAI is comprised of methodologies for making AI systems interpretable and explainable [5].

The context of interpretability and explainability is generally considered domain-specific in an applied setting. For instance, a ML engineer and a medical practitioner would have a different perspective on what is "explainable" when viewing the same system. Interpretability from the perspective of the ML engineer relates to understanding the internal working of a system so that the technical parameters can be tuned to improve the overall performance. Interpretability from the medical practitioner's perspective would relate to a higher-level understanding of the internal operation of a system as it relates to the medical function it provides. Explainability for a ML engineer may relate to presenting technical information in an understandable format that enables effective evaluation of a system while explainability for medical practitioners may be more related to the rationale as to *why* a course of action is prescribed for a patient.

It is postulated that AI-based systems need to accommodate a medical practitioner's perspective to be considered explainable in a healthcare setting. This presents several challenges which are highlighted and addressed as part of this work:

Designing domain-agnostic systems with XAI and simultaneously accommodating multiple perspectives is a complex problem because explanations require a context of the domain (engineering, medicine, or healthcare) and can be useful for a targeted perspective but trivial for others. For instance, presenting interactive visualisations to explain layers of a neural network is beneficial for ML engineers but of less importance to the radiologists who use the neural network for analysing MRI scans.

The scope of interpretability and explainability for AI-based solutions is broader than the operation of a ML model. It also concerns the workflow adopted to train these models. The workflow can provide technical knowledge regarding the pre-processing steps, the ML models used, and the evaluation criteria (e.g. accuracy, precision) to the ML engineer. It can benefit medical practitioners with an overview of the underlying data, the model's interpretation of the data, and the performance metrics pertinent to medical diagnostics. For instance, the ML models that are used to predict based on a patient's medical record might be inappropriate if the underlying training data does not include records from similar demographics.

The subjective nature of XAI in medical setting presents challenges such as as the association of a trained model's knowledge with the medical features, the provisioning of explanations with regard to the underlying medical dataset [6], and an understanding of how the presence or absence of some medical features' information affects a model's performance and its interpretation of features.

There are several nuanced issues related to the challenges articulated. These include: (a) lack of explainability in underlying feature engineering processes to incorporate clinical expertise; (b) complexity in the integration of XAI approaches with existing ML workflows [1]; (c) a lack of a high-level explainability of the data and the ML model [7]; (d) and a lack of explainability of a model's operation in different medical settings.

A standard ML workflow consists of several stages: data collection, data pre-processing, modeling, training, evaluation, tuning, and deployment. XAI approaches should endeavor to integrate interpretability and explainability into the standard ML workflow. Feature Importance (FI) is a set of techniques that assign weightings(scores) to each feature indicating their relative importance in making a prediction or classification by a ML model [8]. FI techniques are typically used as part of the data pre-processing to enhance feature selection.

Moving towards a solution: While addressing the challenges articulated in their totality is beyond the scope of this paper, addressing the nuanced issues outlined will provide the initial steps for a more complete solution to be derived and is the primary contribution of this work. In this paper, FI techniques are utilised as a means of enabling XAI. It is envisaged that FI will provide a simple but powerful means of integrating XAI into the standard ML workflow in a domain-agnostic manner. The approach can enable the explainability of a ML model as well as the underlying workflow whilst accommodating multiple perspectives. This is realised by three proposed approaches that utilise the associations between FI scores, FI techniques, the inclusion/exclusion of features, data augmentation techniques, and performance metrics. In doing so, it enables multiple levels of explainability encapsulating the operation of the ML model with different underlying datasets in different medical settings. The explainability derived is expected to enable the clinical validation of AI-based systems as discussed in the following sections.

The remainder of the paper is organised as follows: Section 2 presents related work with regards XAI, FI, and its utilisation in an applied ML setting. Section 3 outlines the proposed methodology for enabling XAI in a standard ML workflow. Section 4 outlines the results of the experimental work of the approaches proposed. Section 5 presents a discussion and concluding remarks arising from the work carried out to date.

2   Related Work

Preliminary work for making ML models used in clinical domains increasingly interpretable and explainable has been initiated in [1, 5, 9, 10]. The interpretability and explainability of ML models enable ML engineers to understand, and evaluate, a model's parameters (weights/coefficients) and hyper-parameters (input-size, number of layers) with the model's outcomes (predictions/classifications). It can also enable medical practitioners to effectively comprehend and validate the output derived from ML models as per their medical expertise [1, 5].

There exists a variety of XAI methods that are applicable to the medical domain. Ante-hoc XAI methods achieve interpretability without an additional

step that makes them easier to adopt in existing ML workflows. They include inherently interpretable ML models such as Decision trees [11], Random Forests [11] and Generalised Additive Models (GAMs) [9]. They are typically used to achieve interpretability at the cost of lower performance scores as compared to complex ML models. However, their contribution towards enabling the explainability to non-CS perspectives in different domains is not extensively discussed in the literature [1].

In [12], an XAI-enabled framework to include clinical expertise in AI-based systems is proposed in an abstract format. This work also discusses the use of FI to enable the inclusion of clinical expertise when building AI-based solutions. In [11] FI scores based on Decision trees were used to analyse the importance of features in classifying cervical cancer and achieving interpretability in the model. However, the interpretability in relation to the underlying dataset was not discussed. Also, the utilisation of the FI scores to enable explainability from the perspective of medical practitioners was not addressed. In [13] Random forests were used to classify arrhythmia from time-series ECG data and FI scores were presented as a means of achieving interpretability. However, as the time-series ECG data has numeric values for each sampled record, the FI scores assigned to each time-stamped value were not useful as effective conclusions cannot be drawn by associating a FI score with a single amplitude value in a time-series ECG wave.

Post-hoc XAI methods are specifically designed for explainability and are applied after a ML model is trained. This makes post-hoc methods difficult to adopt but they are advantageous as they typically support multiple non-interpretable but performant classifiers [5]. Local Interpretable Model-agnostic Explanations (LIME) is one of the commonly used post-hoc XAI methods that was developed to explain the predictions of any ML classifier by calculating FI scores based on some assumptions that don't always hold true across different types of classifiers [14]. Shapley values are another post-hoc XAI technique that was initially introduced in game theory to present the average expected marginal contribution of a player in achieving a *payout* when all possible combinations of players are considered [15]. In XAI, Shapley values are used to assign FI scores to features (players) in achieving predictions (payout) made by a model. In [16] LIME and Shapley's FI scores were compared and it was found that Shapley's FI scores were more consistent when compared to LIME's. This consistency was derived on the basis of objective criteria including similarity, identity, and separability, which are important considerations when generating and providing explanations in a healthcare setting.

3   Methodology

The workflow adopted in this work is depicted in Figure 1. It follows a standard ML workflow with the addition of the FI stage to enable post-hoc explainability. The FI scores are calculated without modifying prior stages of the workflow and are utilised to enable the explainability of the model and the inherent workflow.
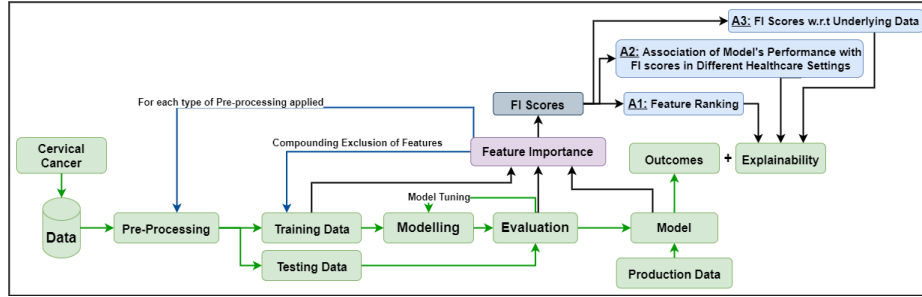
Fig. 1. Integration of FI stage in a standard ML workflow to enable derivation of explainability using three proposed approaches: A1, A2, and A3

Three approaches are proposed that utilise FI scores to enable explainability and interpretability of a ML model and the underlying dataset:

A1. Relative feature ranking: There needs to be a careful validation regarding features that are considered more or less relevant by ML models in healthcare [9]. This approach derives FI scores using two distinct methods. The first is generated using Decision tree FI scores, the FI score of a feature based on its position in the conditional flow of a classification process, and the second is generated using Shapley values, based on weighting the feature's impact on the model's outcome. The derived FI scores are collated and sorted in descending order. This provides a high-level understanding of how a ML model ranks different features to be considered while deriving an outcome.

This enables a comparison between features that are considered important by the classification model (realised by the first approach) and the features that fluctuate the ML model's outcome (realised by the second approach). This enables explainability to be derived as it provides a relative ranking of the features as interpreted by a ML model along with their impact on the outcome. In the applied setting, this can be used by medical practitioners to gain an understanding of the features that are critical in formulating a medical diagnosis, highlighting features whose values cannot be ignored due to their high impact on the model's output.

A2. Feature importance in different medical settings: The availability of medical information in different medical settings is not uniform (e.g. lack of advanced medical tests in small clinics) and therefore, approaches followed by medical practitioners belonging to different medical settings differ. This reduces the associated utility of AI-based solutions. A *gold-standard* solution should be designed to include all the relevant data while providing multiple versions to acknowledge that different healthcare facilities will have different levels of access to this data. This realisation dramatically broadens the applicability and utility of the solution as it acknowledges the inclusion and exclusion of features in different settings.

This approach demonstrates the relative change of FI scores and performance metrics based on the inclusion/exclusion of features. This enables a broader un-

derstanding of a ML model and highlights its suitability to different medical settings (e.g. a general practitioner in a clinic and an emergency room doctor in a hospital will have access to significantly different levels of data regarding an individual's health). If a ML model is trained upon a set of n features, explainability can be derived by training the model on all possible subsets (2^n) of features and can enhance the understanding of how features are re-ranked, and performance is affected, based on the inclusion/exclusion of features.

In an applied setting, this approach is useful to medical practitioners as it aids their understanding based on inclusion/exclusion of clinical test results or medical information with an associated performance score. This enables an informed evaluation regarding the suitability of the AI-based solution on a per actor basis.

A3. Understanding the Data: The data on which a model was trained and how it was pre-processed can have significant consequences in a medical setting [6, 7]. As such, this approach demonstrates the association of FI scores and performance metrics with the data augmentation techniques that are used in a dataset. This association provides explainability by enabling an understanding of how differences in underlying data impact the performance metrics and the FI scores.

In an applied setting, the medical practitioners can validate the ranking of features as interpreted by a ML model trained on data augmented using different techniques and be able to associate it with the corresponding performance metrics. This furthers the interpretability of the underlying workflow used for processing the data and can enable better selection of augmentation techniques by incorporating clinical expertise along with the expected performance metrics.

The dataset and the modelling technique utilised for experimental work are discussed in Section 3.1 and 3.2 respectively. In Section 3.3, the two FI techniques: one based on Decision trees and the other based on Shapley values used in this work are discussed.

3.1   Dataset

In this work, the "Cervical Cancer Risk Factors" dataset available from the UCI data repository is used [17]. This dataset was used in [18] to train different ML models to predict the occurrence of cervical cancer based on a person's health record. The performance of different models was compared based on accuracy, precision, recall, and F-score (harmonic mean of precision and recall) values [18]. The work did not address the interpretability and explainability of ML models and the underlying workflow.

The dataset contains 36 feature attributes representing risk factors responsible for causing cervical cancer and the results of some preliminary and advanced medical tests. In the dataset, 803 out of 858 records have a negative Biopsy result while 55 have a positive result. The class-imbalance problem is addressed using Imbalanced-learn that offers many data sampling techniques to balance the number of the majority and minority classes [19]. Table 1 denotes the number of records corresponding to positive and negative biopsy results after a sampling technique is applied.

Resampling Method	Sam.	0:1 Ratio
Random Over sampling (ROS)	1606	803:803
Adaptive Synthetic Over sampling (ASS)	1606	803:803
Random Under sampling (RUS)	110	55:55
Neighbourhood cleaning Under sampling (NCUS)	725	670:55
SMOTEtomek Combination sampling (S-TOM)	1600	800:800
SMOTE edited nearest neighbours Combination sampling (S-ENN)	1429	652:777

Table 1. Number of samples in different data sampling techniques [18]. Legend: Number of Samples (Sam.), Biopsy results ratio - Positive (0): Negative (1).

3.2   ML Model

Decision trees are graphs where nodes represent sets of data samples and edges represent conditions. Each node has an associated *impurity factor* indicating the diversity of classes/labels in that node. A node is *pure* if all the data samples present in it belong to the same class/label. In a classification problem, the conditions in the edges of Decision trees are designed to decrease the impurity. Therefore, from root node to leaf nodes, the impurity factor decreases and each leaf node should contain data samples that are classified under a single class/label.

Decision trees are more interpretable as compared to complex models such as Support Vector Machines or Neural Networks [5]. They also provide sufficient performance scores in the given dataset [18]. In this work, a Decision tree was chosen as it achieves sufficient performance while retaining interpretability [1].

3.3   Feature Importance (FI)

FI identifies the important features as considered by a ML model from a dataset for making a classification or prediction. In this paper, FI using decision trees and Shapley were used. When a decision tree is trained, FI scores can be calculated by measuring how much a feature contributes towards a decrease in the impurity [20]. The FI scores obtained represent features considered important by the Decision tree model. Shapley values can be used to generate the FI score of a feature by first calculating a model's output including and excluding that feature to get the contribution of that feature alone. This contribution is then weighted in presence of all subsets of features. This whole process is summed for all the subsets of features to get a weighted and permuted FI score. The FI scores obtained represent the impact of different features on a model's outcome.

4   Results

The main challenge in achieving explainability using FI was to present the FI scores generated after training the model with different data augmentation techniques and different feature sets in an integrated manner such that their association with the performance metrics (e.g. accuracy, F-scores) and relative ranking of the features can be effectively utilised in a domain-agnostic manner. This integration of information enables explainability of both: the ML model and the underlying data, thereby broadening the scope of explainability.

The three approaches outlined in Section 3 have been implemented. The value of the approaches and the derived explainability is demonstrated in this section. This is considered a contribution towards a long-term generic workflow for simplifying the integration of XAI in an applied setting such as healthcare.

A1: Relative Feature Ranking When the ML model was trained on data sampled using Random Over Sampling, FI scores assigned to different features were plotted as depicted in Figure 2. Random over sampling provided higher accuracy than other sampling techniques [18], as such it was selected for this approach.
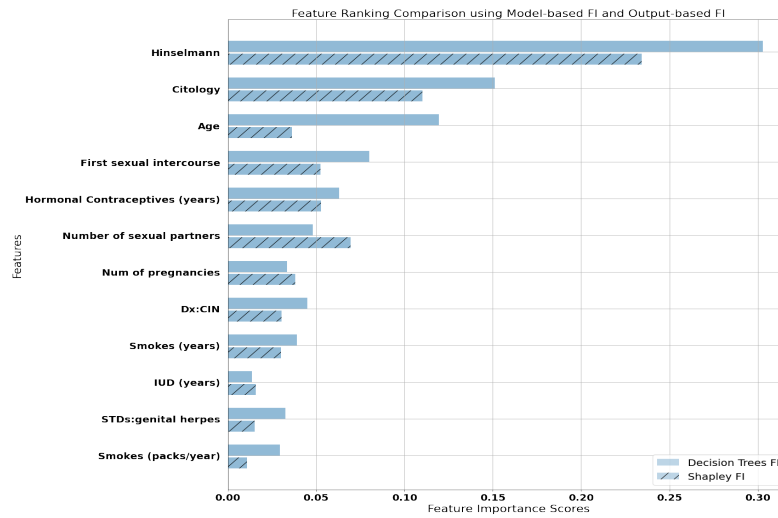
Fig. 2. `A1`: Relative Feature Ranking using FI scores generated by Decision trees and Shapley based FI techniques. The two bars represent the FI scores assigned to individual features by the two FI techniques to provide the basis for the comparison.

In Figure 2, the feature *Schiller Test* was omitted due to its high correlation with the biopsy results as it is an advanced medical test conducted to diagnose cervical cancer [11]. It can be observed that the feature *Hinselmann* is the highest-ranked feature by both FI approaches. There is a similarity between the two sets of FI scores obtained as features considered important by the model (represented by Decision trees based FI) will automatically have a higher impact on its outcome (represented by Shapley based FI). The value derived from these FI scores is that a medical practitioner can understand and validate the ranking of features. This enables the incorporation of clinical expertise to improve feature engineering processes and achieve improved models for future use.

A2: FI in Different medical settings The random over sampled data was used to train multiple instances of the model, each time on a subset of features that excluded the highest-ranked feature from the previous instance. The

resulting FI scores assigned to individual features are indicative of the impact the omission of the highest-ranked feature from a prior instance has on the ML model and is depicted in Figure 3.

The change in the relative ranking of different features can be observed on the omission of the highest-ranked feature. For instance, when all the features were present (All features), *Schiller* (orange segment) was given the highest importance followed by *Age* (yellow segment). When feature *Schiller* is omitted (second bar), it was noticed that *Hinselmann* (grey segment) was given the highest importance instead of *Age*. Thus the inclusion or exclusion of a feature does not behave in an ordered fashion as dictated by a *gold-standard* approach that includes all features.
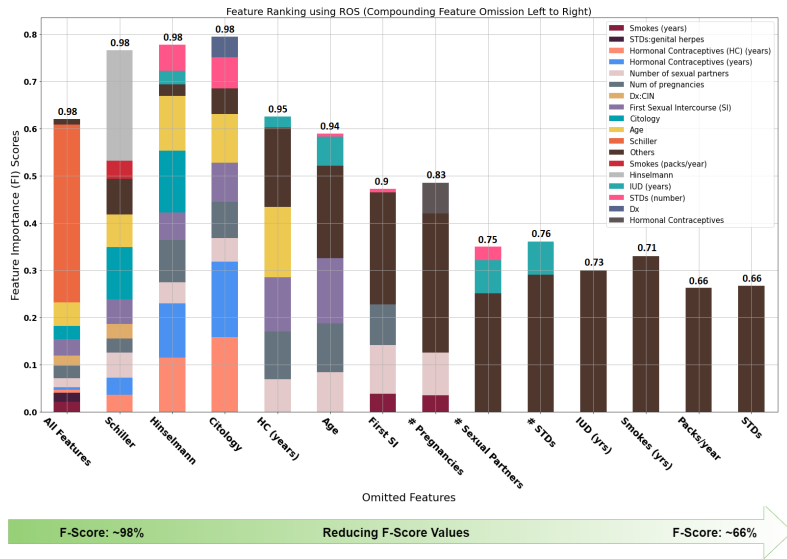
Fig. 3. A2: Impact of excluding the highest ranked features on FI scores, F-scores, and Ordering of features based on FI score

Furthermore, the compounded omission of the highest-ranked features (left to right) significantly reduces the total sum of FI scores assigned to features in each instance (≥ 0.7 to ≤ 0.3). This is accompanied by a reduction in performance metrics such as F-scores (denoted at the top of each bar) and indicates less accurate models due to the absence of more important features or the presence of less important features. This approach warrants the derivation of multiple instances of a single model such that the relationship among features can be fully understood and can be validated with clinical expertise. Based on a threshold value of performance metrics, a medical practitioner can select a ML model that is trained with the features that are accessible to his/her medical setting and assigns appropriate importance scores to the available features while generating an outcome.

A3: Understanding Underlying Data The model was trained on the data sampled using different sampling techniques as discussed in Section 3.1 and FI scores were plotted corresponding to each of the sampled versions as depicted in Figure 4. FI scores relating to a particular type of sampled data are assigned a particular color. Performance metrics corresponding to each of the sampling techniques are noted in the legend.
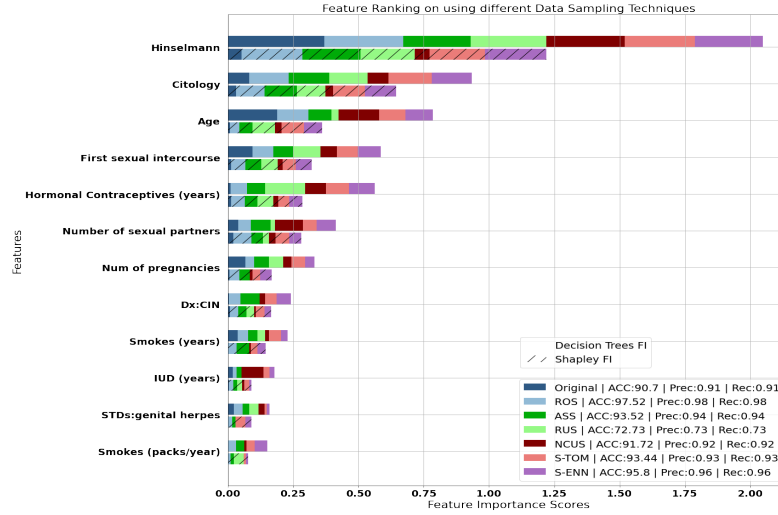
Fig. 4. A3: Ranking of features (by Decision trees and Shapley based FI) on using different data sampling techniques and corresponding performance metrics. Performance metrics are noted as follows Accuracy (*ACC*), Precision (*Prec*) and Recall (*Rec*).

This approach enables the interpretability of the underlying dataset by presenting the difference in FI scores when using data augmented using different techniques. For instance, in Figure 4, there is a lack of similarity in the FI scores associated with under-sampling techniques (e.g. NCUS, RUS) as compared to over/combination-sampling techniques (e.g. S-TOM, ROS). The lesser the volume of data generated using under-sampling techniques the less diverse the values of a feature. This is evident when comparing the sorted ordering of FI scores in under-sampling techniques to over/combination sampling techniques.

As depicted in Figure 4, the under-sampled data provided less accuracy and recall values (70-90%) compared to the over/combination-sampled data (93-97%). The association of performance metrics aligned with FI scores enables the explainability to validate the suitability of datasets from a domain-specific perspective. In contrast to over/combination sampled data, in the under-sampled data augmented using the NCUS technique (dark-red bars), *Age* is assigned a higher FI score than the *Cytology* test which would be considered an invalid approach as a Cytology test is a diagnostic aid with a high level of efficacy when detecting cervical cancer [21]. A medical practitioner should disregard the use of NCUS data due to its invalid FI ranking along with the low-performance scores.

5   Conclusions and Future Work

XAI is a crucial tool for enabling medical practitioners to understand and evaluate AI-based solutions effectively in the healthcare domain. It provides additional benefits in the form of increased confidence in solutions being adopted amongst medical practitioners and increased exposure to the operation of the solutions.

In this paper, an alternative perspective regarding how FI scores can be integrated into a ML workflow is adopted. FI scores are used to surface pertinent information relating to associations between features, models, and data to provide explainability. This perspective is realised in three distinct approaches.

A1) A model/output-based perspective with regards to the relative ranking of a feature, this informs the medical practitioner which features the model considers most important and which features fluctuate the model outcome. A2) Relative feature ranking in different medical settings, this incorporates a hierarchical perspective which considers diagnostic capacity in the form of feature inclusion/exclusion aligning it more closely to the real world. This informs the medical practitioner how the model will perform and rank features in different medical settings enabling a more informed interpretation of a model's operation. A3) The impact of data augmentation approaches on the performance of a model and the validity of their FI scores in a medical setting. This informs the medical practitioner how suitable the augmented data is and how valid it is in a medical setting. The simple but powerful nature of FI enables the applicability of the three approaches proposed in a domain-agnostic manner.

It is intended to extend the work by developing a framework that automates the training and validation of models appropriate to the intended level of a hierarchy in order to enable explainability from a multi-level perspective. The workflow comprising that hierarchy will empirically evaluate the applicability of combining XAI and recommendations to increase operational efficacy.

References

1. Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? (Ml):1–28, 2017.
2. Benjamin P. Evans, Bing Xue, and Mengjie Zhang. What's inside the black-box? A genetic programming method for interpreting complex machine learning models. *GECCO 2019 - Proceedings of the 2019 Genetic and Evolutionary Computation Conference*, pages 1012–1020, 2019.
3. Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–15, 2019.

4. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2017.
5. Erico Tjoa and Cuntai Guan. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. 1, 2019.
6. D Douglas Miller. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. *npj Digital Medicine*, 2(1):62, 2019.
7. Namrata Vaswani, Yuejie Chi, and Thierry Bouwmans. Rethinking pca for modern data sets: Theory, algorithms, and applications [scanning the issue]. *Proceedings of the IEEE*, 106(8):1274–1276, 2018.
8. Ben Hoyle, Markus Michael Rau, Roman Zitlau, Stella Seitz, and Jochen Weller. Feature importance for machine learning redshifts applied to sdss galaxies. *Monthly Notices of the Royal Astronomical Society*, 449(2):1275–1283, 2015.
9. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.
10. Devam Dave, Het Naik, Smiti Singhal, and Pankesh Patel. Explainable ai meets healthcare: A study on heart disease dataset, 2020.
11. X. Deng, Y. Luo, and C. Wang. Analysis of risk factors for cervical cancer based on machine learning methods. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 631–635, 2018.
12. U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly. Explainable ai in healthcare. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–2, 2020.
13. P Nisha, Urja Pawar, and Ruairi O'Reilly. Interpretable machine learning models for assisting clinicians in the analysis of physiological data.
14. Marco Tulio Ribeiro and Carlos Guestrin. " Why Should I Trust You ?" Explaining the Predictions of Any Classifier. pages 1135–1144, 2016.
15. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.
16. R. El Shawi, Y. Sherif, M. Al-Mallah, and S. Sakr. Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 275–280, June 2019.
17. Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis*, pages 243–250. Springer, 2017.
18. Sean Quinlan, Haithem Afli, and Ruairi O'Reilly. A comparative analysis of classification techniques for cervical cancer utilising at risk factors and screening test results. In *AICS*, pages 400–411, 2019.
19. Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.
20. J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
21. Anita Ww Lim, Rebecca Landy, Alejandra Castanon, Antony Hollingworth, Willie Hamilton, Nick Dudding, and Peter Sasieni. Cytology in the diagnosis of cervical cancer in symptomatic young women: a retrospective review. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 66(653):e871–e879, dec 2016.