Interpretable Machine Learning in Healthcare

Jessie Lee (chiahsua@andrew.cmu.edu),
Yirun Wang (yirunw@andrew.cmu.edu)
Akihiko Sangawa (asangawa@andrew.cmu.edu)
95-729 – E-Commerce Tech, Machine Learning, Analytics, & Bots

1. Introduction

Machine Learning has transformed industries in many different ways with its ability to make accurate predictions. However, in the healthcare setting, achieving high accuracy is far from enough because, most of the time, a single metric can only give a partial answer to a question. Metrics only tell you "what" but not tell you why. Explainability and interpretability will ultimately drive the adoption of machine learning models in the healthcare setting. The current machine-learning approaches to diagnosis are purely associative. With current machine learning methods, a disease is identified by its strong association with certain symptoms rather than by its causal relationship with certain factors. The inability to derive causal relationships might lead to suboptimal decisions or unintended consequences.

Thus, through this research, we aimed at leveraging interpretable machine learning (IML) methods, SHAP and Skater, to extract the interpretation for each classification model to make the model more comprehensible to support physicians in diagnosing cervical cancer. We also wanted to explain how each algorithm works, identify the most critical risk factors relevant to malignant cervical formation, extract causal relationships between features and outcomes, and discuss what trade-offs need to be considered while deciding which algorithm to be implemented. The reason that we chose to tackle cervical cancer diagnosis is that cervical cancer impacts many women's lives. It is the fourth leading cause of death for women around the globe (Mehmood et al., 2021). Even though early screening tests, such as Pap test and HPV DNA test (Mayo Clinic, n.d.), made cervical cancer a preventable disease that minimizes the global burden of cervical cancer, the screening resources are not easily affordable and accessible for women in the most developing countries. Suppose we were able to build an algorithm that guarantees a decent level of accuracy in identifying cervical cancer patients and extracting the most critical factors or causal relationships from the model. In that case, we might be able to create a new digital screening solution to make the screening more readily available for women in areas with scarce healthcare resources.

2. Background

Interpretability is not equivalent to explainability (Molnar, 2022; Miller, 2018). Interpretability is the associations a machine learning model identifies between features and output, in other words, "extraction of relevant knowledge from the model" (Molnar, 2022). It is the extent to which you can predict what will happen, given a change in input or algorithmic parameters. By contrast, explainability is about the internal mechanisms of machine learning models. It explains why a machine learning model makes a certain prediction (Gall, n.d.). For certain problems or tasks, it is not enough to get the prediction (the what). The model must also explain how it came to the prediction (the why), especially when a model will affect human life.

In addition, machine learning models can only be debugged and audited when they can be interpreted. An interpretable machine learning model can ensure fairness (a model does not discriminate against underrepresented groups to make sure an algorithm works as planned), reliability (small changes in the input do not lead to large changes in the predictions), and adoptability (easier for humans to understand and adopt) of a model.

Even though the development of interpretable machine learning (IML) methods has increased and has addressed some of the major challenges complex machines and algorithms face, for example, complex machine learning algorithms usually do not reveal details about why and how a certain prediction was made, existing IML methods do have some major pitfalls. Molnar points out in his book (Interpretable Machine Learning) that feature dependence makes attribution and extrapolation problematic (Molnar, 2022). For example, while evaluating the impact of a certain feature on the target variable, the partial dependence plots could generate fictitious data points that didn't really exist in the real-world data distribution. This leads to a major concern; the IML methods might not be able to capture the true association or causal relationships in the models that we try to interpret. Another major concern about the existing IML methods mentioned by Molnar in his book is the lack of statistical rigor. Unlike traditional statistical methods, most IML methods do not provide confidence estimates. There is no universal standard to evaluate the different interpretations of the same machine learning model using different IML methods.

2.1. Why is Causal Inference Important in Healthcare?

Given that machine learning algorithms can make accurate predictions, why do we need causal inference in healthcare? The main reasons are as follows.

Reason 1. Causal structures will affect healthcare decisions

Simpson's paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears, or reverses when the population is divided into subpopulations (Simpson, 1951). A real-life example of Simpson's Paradox comes from a medical study that examined two kidney stone treatments and how effective they were for stones of various sizes (Charig et al., 1986). One of the treatments was a less invasive new treatment; the other was the current treatment.

Table 2.1.1. The recovery rate of current treatment versus new treatment.

	Treatment A (Current Treatment)	Treatment B (New Treatment)
Small Kidney Stones	93% (81/ 87)	87% (234/270)
Large Kidney Stones	73% (192/263)	69% (55/ 80)
Aggregated	78% (273/350)	83% (289/350)

The percentage indicates the success rate. The number in the parenthesis is the ratio of number of recoveries to total cases. Adapted from "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy" by Charig et al., 1986, British medical journal (Clinical research ed.).

Comparing the effectiveness of two treatments regardless of the size of the stones by looking at the aggregated result, it is clear that treatment B is more effective. However, when the population is subset based on the size of the stones, the conclusion is reversed. Treatment A works better for both small kidney stone groups and large kidney stone groups. Since the two conclusions derived from the aggregated result and the subset result contradicts, in considering which conclusion to trust, we need to understand the underlying causal structure of the treatment and the outcome.

The contradictory results led the research team to delve deeper to understand what caused the success rate to reverse. The research team found that the probability of treatment choice varied according to the diameter of the stones between the two treatment groups. From the above table, we see that the number of patients with large kidney stones in the treatment A group is much higher than that in the treatment B group given that the sample size of the two treatment groups is the same, which indicates that the outcome of treating patients with a more severe condition has a bigger impact on the overall treatment outcome of treatment A group while the treatment outcome of patients with a mild condition largely determined the outcome of the treatment B group.

The biased treatment assignment unveiled a hidden confounding factor, condition, in the underlying causal structure (Figure 2.1.1) of the treatment choice problem. With this causal structure, to evaluate the direct impact of treatment A and treatment B on the recovery rate, we need to break off the causal relationship between condition and treatment by comparing how two treatments perform in treating patients with similar levels of illness. This leads us to the conclusion; treatment A is more effective than treatment B in a scenario with this causal structure (refer to causal structure 1 in Figure 2.1.1).

With a clear underlying causal structure of a problem, the optimal decision seems to be intuitive. However, healthcare problems are much more complicated than the example we mentioned above. Sometimes, there could be more than one causal structure underneath a problem. Imagine another layer of complexity to consider when making the aforementioned treatment decision. Will we still make the same decision if the availability of treatment A is much scarcer than treatment B and the timing of receiving treatment will largely affect the treatment outcome (refer to the causal structure 2 in Figure 1)? With an additional layer of causal structure, the decision is more complicated as we need to weigh the level of impact of the treatment and the timing of receiving

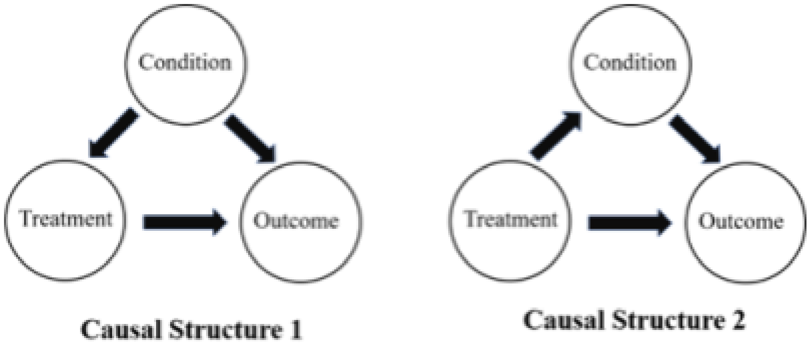


Figure 2.1.1 Two different causal structures of treatment outcome.

treatment on the recovery to make the final decision. Different causal structures will lead to completely different healthcare decisions.

Reason 2. Poor model generalizability

Machine learning can predict accurately if models are trained on a large amount of data. However, insufficient data is one of the biggest challenges faced by healthcare since the healthcare data usually include Personal Identifiable Information and are strictly regulated. Training machine learning models with insufficient amounts of data will lead to poor generalizability of models and cause inconsistent prediction results, making a machine learning model unreliable.

When applying machine learning to address medical problems, there are a few things we need to look after. First, healthcare data tends to have selection biases. Only those who get healthcare services are included in healthcare datasets. Thus, datasets usually cannot represent the overall population. Secondly, even if we eliminate selection biases with randomized controlled trials (RCTs), the imbalance between samples with positive outcomes and negative outcomes remains to be an issue. The main challenge with the imbalance problem is that smaller classes are often more informative, but classification models tend to focus heavily on huge subgroups and ignore smaller subgroups (Ramyachitra & Manikandan, 2014). In this case, causal inference can help extract the true associations between features and outcomes.

3. Experimentation

3.1. Classification Models

Classification models include Machine Learning, Neural Networks, and Causal Inference. We used KNN, Decision Tree, Random Forest, and Ensemble Model for machine learning models. First, we predicted classification using KNN, Decision Tree, and Random Forest. After that, we combined the three models by implementing an ensemble model. In the experiment, the model was implemented by ScikitLearn, one of Python's libraries specialized for tensor computation. The relevant code is implemented in /ml-research/codes/01_KNN_DT_RF_V2.ipynb and should be referred to as necessary.

K-Nearest Neighbor — K-nearest neighbor (KNN) is a supervised learning algorithm that can be used for classification and regression. KNN models classify data points based on feature similarity. KNN finds the K closest points of a point and chooses K points' majority voting result as the point's classification. For each test sample, it will calculate the distance between the test data and each row of training data. It will find the K closest points based on the distance-calculating method selected. Finally, it assigns a class to the test point based on the most frequent class of these rows (Taunk et al., 2019).

Decision Tree — A decision tree is one of the supervised machine learning algorithms. It can be employed for classification or regression tasks (Machine Learning With Python n.d.). Decision trees classify the target data into tree-like hierarchical segments with decision boundaries. The algorithm has an advantage as non-parametric: it efficiently uses large, complicated datasets without imposing complicated parameters (Song & Lu,2015).

Random Forest — Random Forest is a supervised algorithm and can be used as a classifier and regressor (Machine Learning With Python, n.d.c). Random forest is an ensemble machine learning method that samples data, creates multiple-decision trees, and then vote on which machine solution best fits. This ensemble method, by means of sampling and voting, enables the result to be averaged and to mitigate overfitting.

Ensemble Model — The basic idea of Ensemble Modeling is that a single model may predict a specific dataset ideally. Still, the combination of different models has the chance to improve the overall accuracy. Here we used a Voting Classifier from sklearn (Scikit Learn, n.d.). VotingClassifier is to combine conceptually different machine learning classifiers and uses a majority vote, or the average predicted probabilities (soft vote) to predict the class labels. Such a classifier can be useful for a set of equally well-performing models to balance out their weaknesses.

Neural Network — In addition to the above models, a neural network-based classifier was used in the comparison study. Neural networks, also called multi-layer perception (MLP), mimic the mechanism of signal propagation between neurons in the brain by means of several mathematical formulas, e.g., the affine and activation functions (Han et al., 2018). Training of the model is achieved by updating the weights and biases of each function with differential functions derived from forward and backward propagation processes. The model used for this experiment implemented an MLP with three hidden layers. The number of neurons in each layer was 256, 128, and 64, with the final layer using a softmax function to generate a binary discrimination probability. In each layer, batch normalization and dropout functions were added to improve training accuracy (Ioffe & Szegedy, 2015; Srivastava et al., 2014). The training was performed using the minibatch method (Li et al., 2014), with a batch size of 32. The number of training sessions and epochs was set to 30. In this experiment, the model was implemented using PyTorch, one of Python's libraries specialized for tensor computation. The relevant code is implemented in /ml-research/codes/02_NeuralNetwork_CervicalCancerClassification_DI.ipynb and should be referred to as necessary.

3.2. Causal Inference Model

The last model to be experimented with is the causal inference model. The causal inference model called a Bayesian network, developed in the 1980s in the context of exploratory data analysis after its publication by Pearl (1985). Among the many causal inference models in this experiment, we decided to use the fast causal inference (FCI) algorithm (Spirtes, Meek & Richardson, 1995). According to Spirtes (2001), the FCI algorithm is a constraint-based non-parametric algorithm that explores a graphical feature common to all causal-directed acyclic graphs (DAGs) to observationally equivalent sets of statistical tests of conditional independence. In other words, in identifying dependencies between variables, variables purely dependent on each other are connected with a causal edge after excluding the influence of different and unobserved variables. One advantage of the constraint-based algorithm is that it is less likely to suffer on the identified causal relationship due to increases or decreases in other variables. Another advantage of this model is that it can generate correct identification probabilities even in the presence of hidden variables such as covariates or a mixture of factors that can negatively affect the model, such as selection bias. There are often unobserved covariates in an exploratory data analysis such as this one. Instead of implementing this algorithm, we used a Java desktop application called Tetrad, developed by professors in the Philosophy Department at CMU (Cmu-phil/tetrad, 2022). The relevant results are stored in /ml-research/codes/03_causal_inference and should be referred to as necessary[1].

[1] The Causal Analysis_Cervical Cancer.tet file can be downloaded from tetrad-gui-7.1.0-launch.jar located in the same directory, started, and loaded by selecting File > Open Session from the menu. Open Session" menu and load the file.

3.3. Dataset

For this research project, we use the structured dataset according to cervical cancer patients provided by the UCI Machine Learning repository. The dataset has historical medical records of 858 patients related to their demographics and habits. Some patients are determined not to answer particular questions, so the dataset contains missing values; we will complement those values with its mean/median value. The dataset has 36 columns; Examples are as follows:

Explanatory Variable Category (Examples)

a. Age
b. Sexual intercourse information: Number of sexual partners, first sexual intercourse (age)
c. Several pregnancies
d. Smokes information: Smokes or not, years of smoking, and packs of tobacco per year
e. Contraceptive information: Do Hormonal Contraceptives, IUD, and each year
f. Sexually transmitted diseases (STDs) information: Ever had STDs, number of STDs, time since the first diagnosis, and time since the last diagnosis
g. Diagnosis information related to cervical cancer: Cancer, Cervical intraepithelial neoplasia (CIN), Cervical intraepithelial neoplasia (HPV)

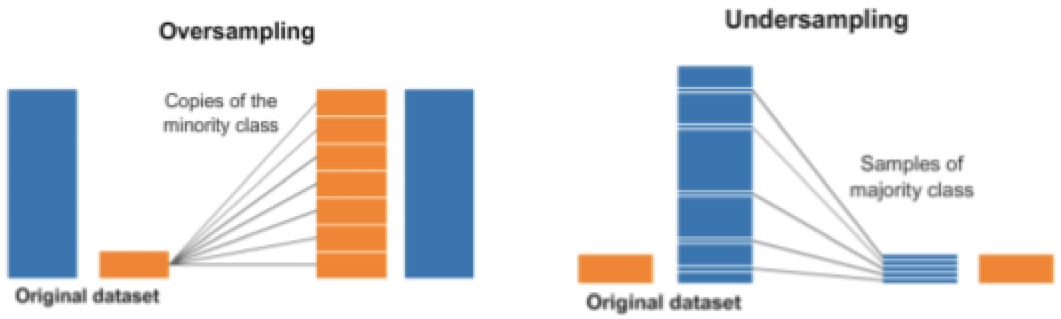
Target Variables

A. Result of Biopsy Diagnosis

After a preliminary review of the dataset, this dataset holds both a discrete variable (1 if an event is applicable, 0 if not) and a continuous variable for the same category information. Since these generally have high correlation coefficients and are likely to have a negative impact on classification results (Schober, Boer & Schwarte, 2018), it was decided to limit the use of continuous variables in this experiment to those that hold more information. After filtering the explanatory variables, we decided to use 13 variables in the discriminant model.

In the current experiment, the goal is to understand the variables that had a significant impact across different classification models. We used the aforementioned variables for general classification models and neural networks. We first checked whether the sample distributions across different features were similar in the positive and negative biopsy groups to implement a causal AI model. Since the distributions across features are similar in both groups, a causal inference could be made[2].

[2] Although a comparison by propensity score is strictly necessary to determine whether or not causal inference can be performed, it was omitted because the distributions are clearly similar. The relevant code is implemented in /ml-research/codes/03_causal_inference/03-01_Feature distribution.ipynb and should be referred to as necessary.



NOTE: These figures refer to post written by Alencar, R. (2017).

Figure 3.4.1. Oversampling and Undersampling Method

3.4. Sampling Methods

Imbalanced-learn (Imblearn) is a library that provides tools to either upsample the minority class or downsample the majority class to balance the proportion of data with different classes (Dwivedi, 2020).

Oversampling means adding more samples from the class with fewer data. On the other hand, undersampling means randomly removing some samples from the majority class to balance the two classes in quantity. Those images are in Figure 3.4.1.

3.5. Evaluation Method

In the context of healthcare, from patients' and physicians' perspectives, the cost of false positive cases is higher than the cost of false negative cases. By contrast, from hospitals' perspective, high false positive rates might lead to a waste of healthcare resources since patients might get treatments or examinations that could have been avoided. Therefore, accuracy should not be a single evaluation metric for machine learning models implemented in healthcare settings. To ensure the adoptability of models, we need to consider different stakeholders' interests while designing machine learning models. That means we must strike a balance between a false positive rate and a true positive rate. With that being said, for this research, we chose AUC as our evaluation standard when selecting parameters and drawing conclusions.

Common evaluation methods for classification models are Accuracy, Precision, Recall, AUC (Area under the ROC curve), etc. The confusion matrix describes the performance of a classification model. We chose AUC as our evaluation metric.

Table 3.5.1. The components of a confusion matrix

Confusion Matrix		Prediction	
		0	1
Real Results	0	True Negative	False Positive
	1	False Negative	True Positive

Accuracy measure the number of correct predictions / total number of predictions, or

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision describes the correct portion of identification.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall describes the correct identified portion of actual positives. The recall is the same as the ***True Positive Rate* (*TPR*)**.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Receiver Operating Characteristic (ROC) — A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds (Google Machine Learning Education, n.d.). This curve plots two parameters: True Positive Rate (TPR), which is the same as Recall, and False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

A ROC curve plots TPR vs. FPR at different classification thresholds (Google Machine Learning Education, n.d.). The area under the curve (AUC) measures the area under the ROC line. It represents how TPR and FPR trade off. Provides an aggregate performance measure across all possible classification thresholds (Google Machine Learning Education, n.d.).

Permutation Feature Importance / Shaley Value — Then AUC and TPR evaluation, we conduct permutation feature importance for further evaluation. Permutation feature importance (PFI) is an evaluation method for the interpretation of machine learning models. Fisher, Rudin & Dominici state that permutation-based model class reliance (MCR) shows the upper and lower limit on how important a set of variables is as point estimates, and those MCR calculations can perform to any model (2019). In this experiment, the corresponding methods were applied to three models: KNN, decision tree, and random forest. The library is called skater, a library developed by Oracle Open Source (Oracle Open Source., n.d.).

We cannot use the skater framework for neural network interpretation because the library is not compatible with PyTorch. Therefore, we need to find an alternative to interpret the NN model we created. In this experiment, the global model interpretation, i.e., the contribution of the variables to the overall model, was determined by calculating the average Shapley value of each variable. The purpose of this study was not to compare the importance of the variables but to compare the importance of the variables to the model as a whole. The purpose of the present study was not to compare the importance of variables but to derive which variables the model relies on to make a prediction. Even though we used different Interpretable Machine Learning models (Skater and SHAP) to interpret general classification models, the general purpose of these two methods is the same.

4. Experiment Results

The dataset has historical medical records of 858 patients related to their demographics and habits. 55 patients were diagnosed as positive in the results, and 803 of them were negative. The dataset is unbalanced. With the original dataset, we implemented KNN, decision tree, and random forest models. The results are as follows.

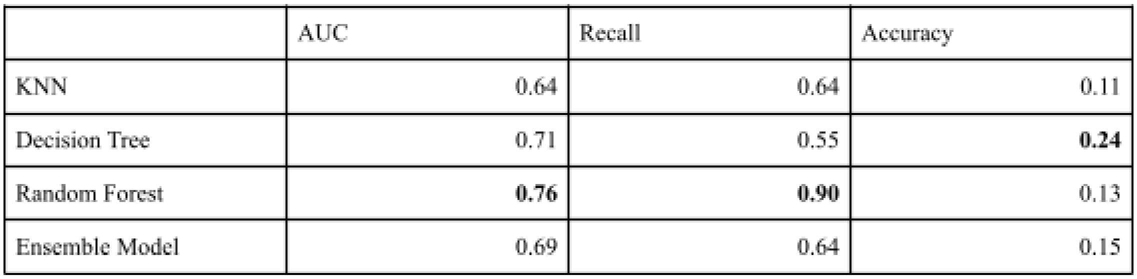
UnderSampling gets the best results in all models. In the training dataset, there are 44 positive samples and 44 negative samples.

Table 4.1. Best performance of three models trained on oversampling and undersampling datasets.

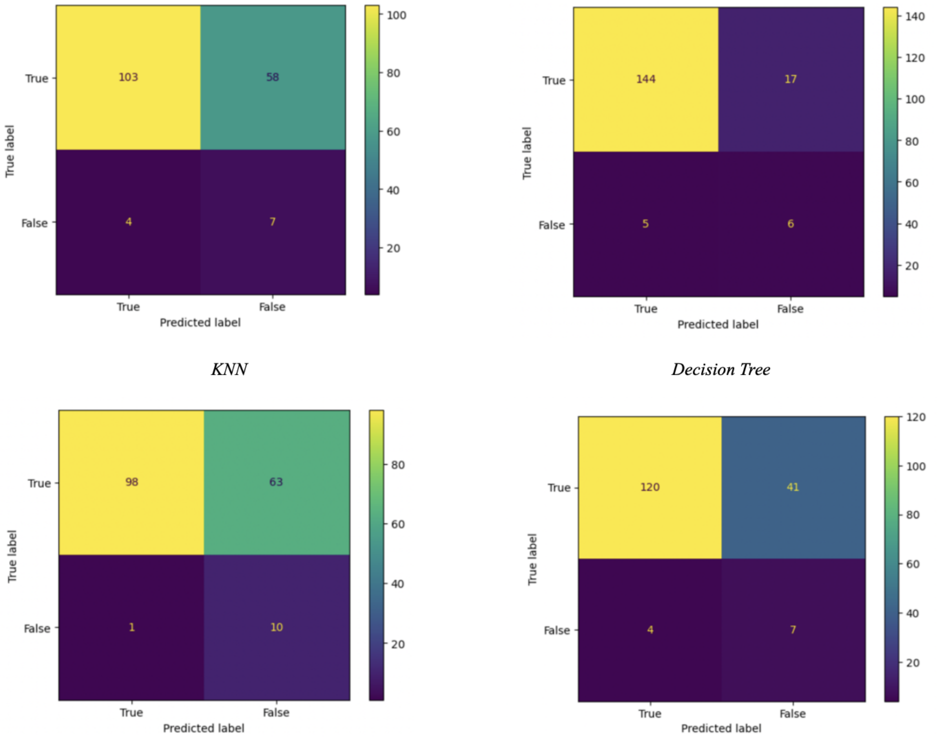
Model selected	Best AUC using oversampling	Best AUC using undersampling
KNN	0.56	**0.64**
Decision Tree	0.57	**0.71**
Random Forest	0.62	**0.76**

Table 4.2 shows the performance of the four machine learning models. Random Forest has the highest AUC and Recall among all the models. The Decision Tree got the highest accuracy.

Table 4.2. The results of KNN, Decision Tree, Random Forest, and Ensemble Model.



	AUC	Recall	Accuracy
KNN	0.64	0.64	0.11
Decision Tree	0.71	0.55	**0.24**
Random Forest	**0.76**	**0.90**	0.13
Ensemble Model	0.69	0.64	0.15



KNN

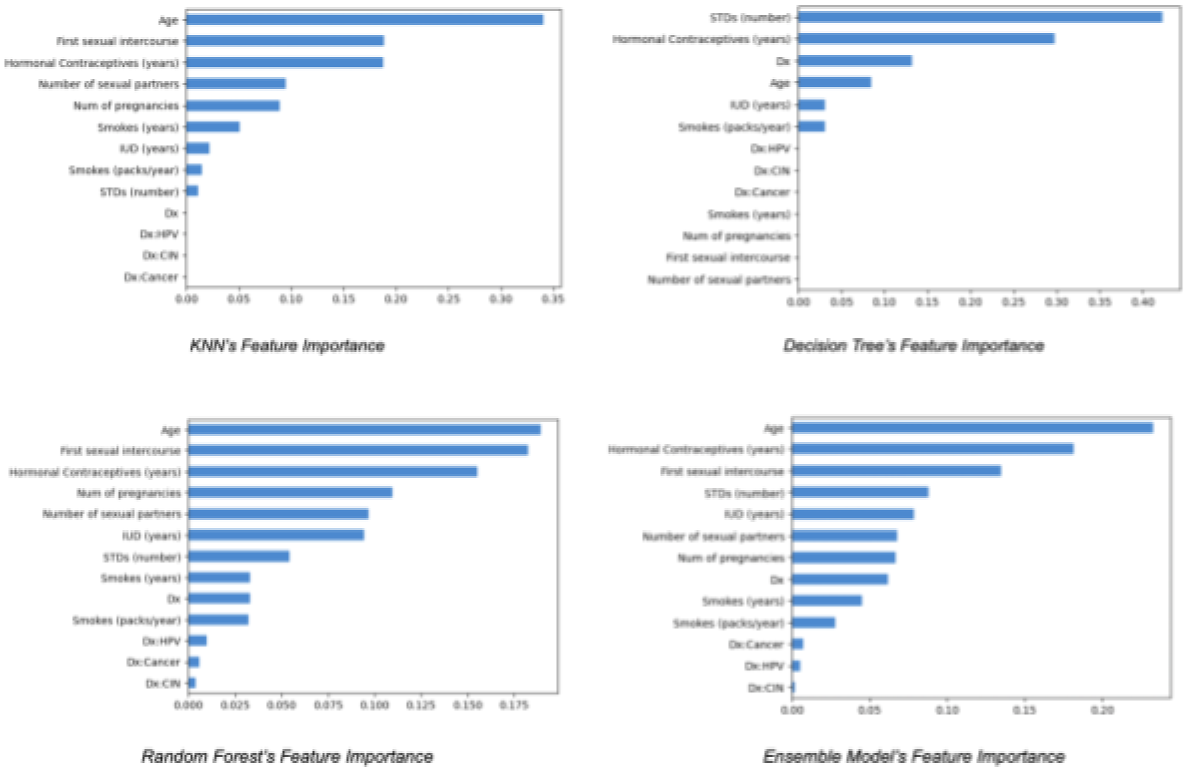
Decision Tree

Random Forest

Ensemble Model

Figure 4.1 Confusion Matrix of the Four Models.

In the experiments, we used the undersampling dataset. For each model, we ran the model multiple times and selected the parameter that led to the highest AUC. In the KNN model, k = 3; Decision Tree model, the max depth is 3. Random Forest's parameters are n_estimators = 40 and max_depth = 34. We then combined the three models above into an Ensemble Model using a soft voting strategy. To create an ensemble model, we used KNN, decision tree, and random forest with the best parameters selected from training every model. The ensemble model was expected to have the best performance among all the models. However, the decision tree got the best performance in AUC and Recall. In the test dataset, only one out of ten patients was misdiagnosed. Though the test dataset size is small, it still proves that the complex model does not always lead to the best result.



KNN's Feature Importance

Decision Tree's Feature Importance

Random Forest's Feature Importance

Ensemble Model's Feature Importance

Figure 4.2. Permutation Feature Importance

Figure 4.2 shows the feature importance understood by each model. Age, hormonal contraceptives, and first sexual intercourse are the three most important features in KNN, Random Forest, and Emsemble Model. Hormonal contraceptives are also an important (the second) feature in the Decision Tree model. Besides, according to medical knowledge, Dx CIN is highly correlated to cervical cancer (National Cancer Institute at the National Institutes of Health, n.d.). However, this strong correlation was not reflected in our results. The general classification models may need help to capture the important features in the real world.

The results of the Neural Network experiment are presented in Figure 4.3. In this model, the true positive rate is 64.3%. Since our main purpose was to extract the important features of the model rather than creating a neural network model with higher precision, we will not discuss the details of the model's detection capability. Still, it can be interpreted as a model with a certain level of detection capability. Next, we discuss the interpretation of the neural network model. The result exhibits in Figure 4.4. Figure 4.4 indicates that the top five important features in the neural network model are Hormonal Contraceptives (years), Age, Skomer (years), Number of Sexual Partners, and First Sexual intercourse, respectively. The interpretation of the neural network model is almost identical to those of other models such as KNN, Random Forest, and the ensemble model. Since these variables do not directly affect cervical cancer intuitively, we can conclude that the neural model can only capture the association between explanatory variables and target variables; the model does not capture causal relationships.

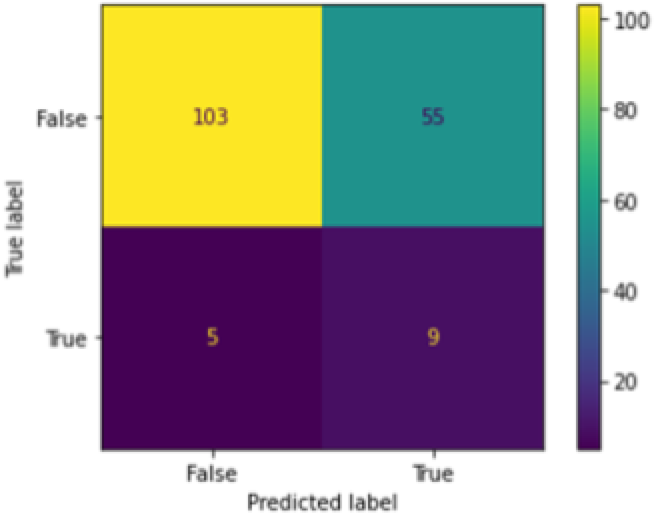


Figure 4.3 Confusion Matrix of Neural Network Model

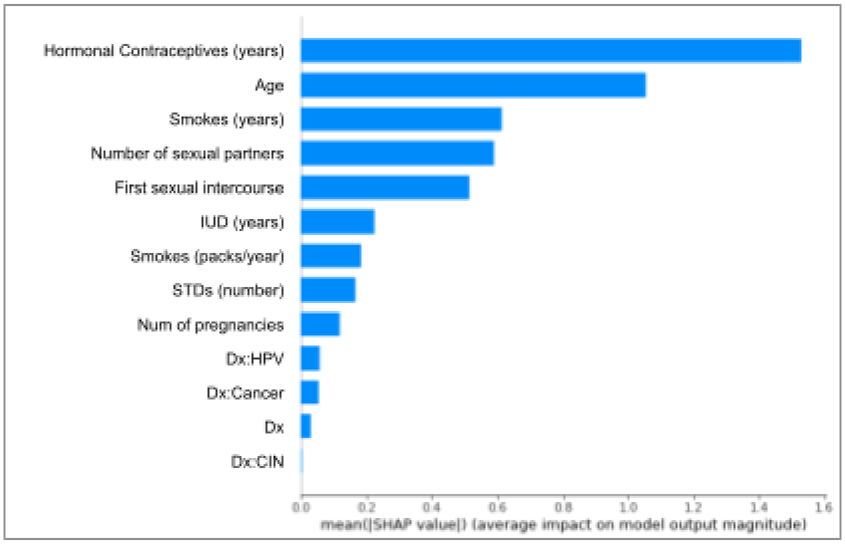


Figure 4.4 Feature Importance in Neural Network Model by Averaged Shapley Value

Next, we explain the experiment results of causal AI. After the run, the FCI algorithm identified the causal relationship, and the results are shown in Figure 4.5. On the right side of Figure 4.5, two variables are shown that are causally related to the biopsy result: diagnosis of CIN and number of STDs. CIN stands for Cervical intraepithelial neoplasia, an abnormal cell in a woman's cervix that becomes malignant and leads to cervical cancer (National Cancer Institute at the National Institutes of Health, n.d.). The number of STDs is a sexually transmitted disease, which is also a cause of cervical cancer. Conversely, for other variables not included in the causal relationship of the biopsy results. Intuitively, we can conclude that this result is correct. However, the number of sexually transmitted diseases has abstracted away the true causal relationship, leaving room for further drill down to which of the sexually transmitted diseases directly indicate a causal relationship. Therefore, we re-ran the FCI algorithm again, including the STD details.

After re-execution, including the STD details as explanatory variables, the algorithm identified genital herpes as another causal factor detecting positive biopsy results in Figure 4.6. Some studies have reported that the probability of cervical cancer is increased when genital herpes is transmitted with other HPVs (Kirchheimer, n.d.). Therefore, we conclude that the risk factor of cervical cancer is a diagnosis of CIN and genital herpes in terms of causal relationships. We can also conclude that this result is correct according to the expert knowledge citing from trusted medical webpages.

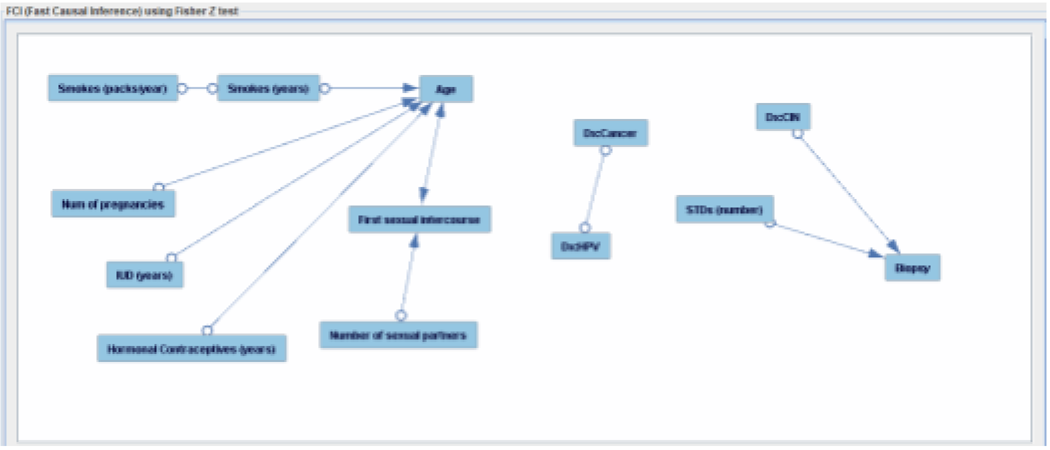
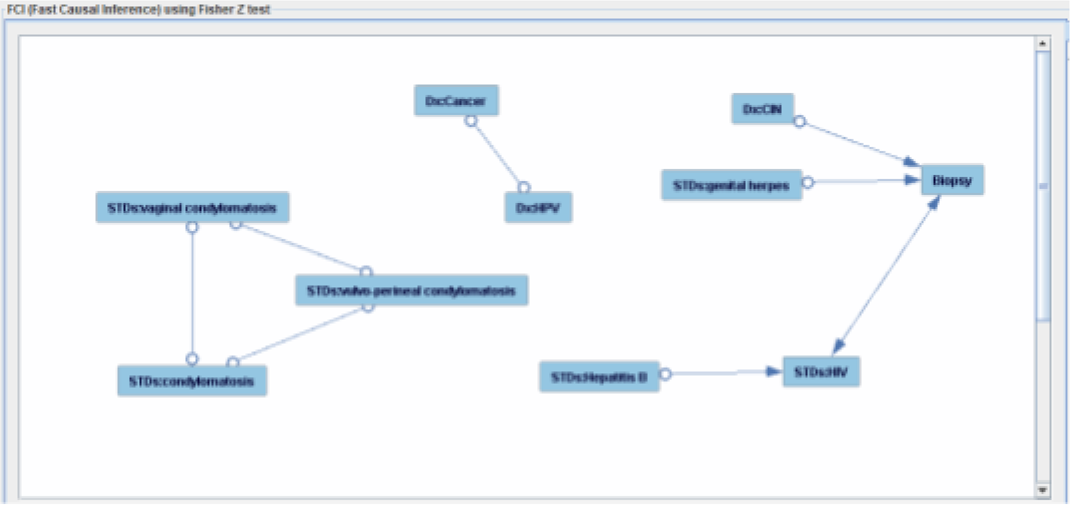


Figure 4.5. Causal Directional Acyclic Graph with FCI and Continuous Variables


FCI (Fast Causal Inference) using Fisher Z test
Dx:Cancer
Dx:CIN
Dx:HPV
STDs:genital herpes
Biopsy
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:condylomatosis
STDs:Hepatitis B
STDs:HIV

NOTE: Other STDs that not shown in the figure are included at causal inference, but they does not shown any causal relationships with each other, so they are excluded from the figure.

Figure 4.6. Causal Directional Acyclic Graph with FCI, Continuous Variables and STDs

5. Discussion & Conclusion

The ensemble model, especially the Voting Classifier, does not outperform the best model in the set of implemented models, in this case, the decision tree. It is easily understandable for us because of the mechanism of the Voting Classifier. The ensemble model could perform better because the ensemble model with the Voting Classifier algorithm uses weighted average probabilities as its parameter to infer the result (Scikit Learn, n.d.). If each model can classify multiple classes accurately, the ensemble model can classify data into different classes. However, in this case, the target variable is binary (two classes: positive biopsy or negative biopsy). In the case of binary classification problems, the classification results of the ensemble model could be easily biased by the model with a bigger weight. Therefore, we can conclude that in the case of such a binary classification problem, it might be more appropriate to apply a simple (non-ensemble) model with higher accuracy.

Next, from the IML interpretation results, we see that the interpretations of all the trained models are similar in terms of their permutation-based important features and averaged Shapley values. However, these important features do not capture the causal relationships inferred by the Causal AI algorithm (FCI algorithm), which extracts causal dependencies by considering probabilistic and conditional independence (Spirtes, 2001). For example, "Dx CIN"" and "STDs: genital herpes" are identified as causes of the positive biopsy, but neither of them was included in the important features of each model. This result shows that IML methods may not be able to derive the true causal relationships between features and target variables (Molnar, 2022). On the other hand, even though the causal inference algorithm has a powerful mechanism to detect causal relationships, it is difficult to discern whether the causal relationships are true without expert knowledge. Moreover, sometimes causal relationships are detected without a clear direction (in this case, we are not sure which node is the cause). There are two approaches to deriving true causal relationships. The first approach is verifying causal relationships extracted with AI algorithms with expert knowledge. The second approach incorporates expert knowledge into a causal AI algorithm to ensure correct causal inference.

In conclusion, there are three key points we would like to make at the end of this research paper. First, through this research, we found that lack of statistical rigor and comparison standard are common drawbacks shared by all the existing IML methods. Therefore, when different IML methods give different interpretations of the same machine learning model, we might be clueless about which interpretation to trust. Second, interpretability (identifying associations) is not equivalent to explainability (identifying causations). The associations identified by IML methods could be inconsistent with the causations identified by causal inference models. Lastly, to make a machine learning model more adoptable, we need to consider different stakeholders' interests and think about what the user journey looks like in designing a machine learning model. We need to think about how to explain prediction results to people from diverse backgrounds with different knowledge of medicine since explainability can greatly influence whether a model is adopted or not. To realize the "explainability" of machine learning models, we can not simply rely on the IML methods and causal inference models without verifying whether the causal inference was correctly made. We need to incorporate expert knowledge in developing an explainable machine-learning model. In addition, We need to think about how to explain prediction results to people from diverse backgrounds with different levels of knowledge of medicine.

Reference

Alencar, R. (November 15, 2017). *Resampling strategies for imbalanced datasets | Kaggle*. https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook

Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. (1986, March 29). *Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy*. British medical journal (Clinical research ed.). Retrieved December 14, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339981/

Cmu-phil/tetrad. (2022). [Java]. cmu-phil. https://github.com/cmu-phil/tetrad (Original work published 2015)

Dwivedi, R. (September 20, 2020). *What is Imblearn Technique—Everything To Know For Class Imbalance Issues In Machine Learning*. Analytics India Magazine. https://analyticsindiamag.com/what-is-imblearn-technique-everything-to-know-for-class-imbalance-issues-in-machine-learning/

Fernandes, K., Cardoso, J. S., & Fernandes, J. (n.d.). *UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set*. Retrieved December 14, 2022, from https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#

Fisher, A., Rudin, C., & Dominici, F. (2019). *All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models* (arXiv:1801.01489). arXiv. https://doi.org/10.48550/arXiv.1801.01489

Gall, R. (n.d.). *Machine learning explainability vs. Interpretability: Two concepts that could help restore trust in ai*. KDnuggets. Retrieved December 14, 2022, from https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

Google Machine Learning Education. (n.d.). *Classification: ROC Curve and AUC | Machine Learning*. Google Developers. Retrieved December 15, 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Han, S.-H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dementia and Neurocognitive Disorders*, *17*(3), 83–89. https://doi.org/10.12779/dnd.2018.17.3.83

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* (arXiv:1502.03167). arXiv. https://doi.org/10.48550/arXiv.1502.03167

Kirchheimer, S. (n.d.). *Herpes Virus Linked to Cervical Cancer*. WebMD. Retrieved December 14, 2022, from https://www.webmd.com/genital-herpes/news/20021105/herpes-virus-linked-to-cervical-cancer

Li, M., Zhang, T., Chen, Y., & Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 661–670. https://doi.org/10.1145/262333

Machine Learning With Python. (n.d.). *Classification Algorithms—Decision Tree*. Retrieved December 15, 2022, from https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_decision_tree.htm

Machine Learning With Python. (n.d. b). *KNN Algorithm—Finding Nearest Neighbors*. Retrieved December 15, 2022, from https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm

Machine Learning With Python. (n.d. c). *Classification Algorithms—Random Forest*. Retrieved December 15, 2022, from https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_random_forest.htm

Mayo Foundation for Medical Education and Research. (2022, December 14). *Cervical cancer*. Mayo Clinic. Retrieved December 14, 2022, from https://www.mayoclinic.org/diseases-conditions/cervical-cancer/diagnosis-treatment/drc-203525060 2623612

Mehmood , M., Rizwan, M., Ml, M. G., & Abbas, S. (2021, December). *Machine Learning Assisted Cervical Cancer Detection*. Frontiers in public health. Retrieved from https://pubmed.ncbi.nlm.nih.gov/35004588/

Miller, T. (2018). *Explanation in Artificial Intelligence: Insights from the Social Sciences* (arXiv:1706.07269). arXiv. https://doi.org/10.48550/arXiv.1706.07269

Molnar, C. (December 14, 2022). *Interpretable machine learning*. christophm.github.io. Retrieved December 14, 2022, from https://christophm.github.io/interpretable-ml-book/

National Cancer Institute at the National Institutes of Health. (n.d.). *Definition of CIN 1—NCI Dictionary of Cancer Terms—NCI*. Retrieved December 14, 2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cin-1

Oracle Open Source. (n.d.). *Overview—Skater 0 documentation*. Retrieved December 14, 2022, from https://oracle.github.io/Skater/overview.html

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. *UCLA Computer Science Department Technical Report 850021 (R-43)*, *Proceedings, Cognitive Science Society, UC Irvine*, 329–334

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. https://doi.org/10.48550/arXiv.1912.01703

Ramyachitra, D. D., & Manikandan, P. (2014). IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW. *International Journal of Computing and Business Research*, *5*(4). https://www.semanticscholar.org/paper/IMBALANCED-DATASET-CLASSIFICATION-AND-SOLUTIONS-%3A-A-Ramyachitra-Manikandan/3e8ea23ec779f79c16f8f5402c5be2ef403fe8d3

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Scikit Learn. (n.d.). 1.11. Ensemble methods. Scikit-Learn. Retrieved December 15, 2022, from https://scikit-learn.org/stable/modules/ensemble.html

Simpson, E. H. (1951). *The Interpretation of Interaction in Contingency Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 13(2), 238–241. http://www.jstor.org/stable/2984065

Song, Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Spirtes, P., Meek, C., & Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 499–506.

Spirtes, P. (2001). An Anytime Algorithm for Causal Inference. *International Workshop on Artificial Intelligence and Statistics*, 278–285. https://proceedings.mlr.press/r3/spirtes01a.html

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747

UCI *Machine Learning Repository: Cervical cancer (Risk Factors) Data Set*. (n.d.). Retrieved December 14, 2022, from https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#