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 been a game-changer across various industries, thanks to its ability to provide accurate predictions. However, in the healthcare sector, high accuracy alone is not sufficient as a single metric can only offer a partial answer to a question. Metrics can only tell us "what," but not "why," making interpretability and explainability crucial for the adoption of machine learning models in healthcare. Current machine learning approaches to diagnosis are purely associative, with diseases being identified by their association with particular symptoms, rather than a causal relationship with specific factors. Failure to establish causal relationships may result in suboptimal decisions or unintended consequences.

To address this issue, we aimed to leverage interpretable machine learning (IML) methods, namely SHAP and Skater, in our research to extract the interpretation for each classification model. The goal was to make the model more understandable, supporting physicians in diagnosing cervical cancer. We sought 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 the trade-offs to consider when deciding which algorithm to implement. We chose to focus on cervical cancer diagnosis due to its significant impact on women's lives. It is the fourth leading cause of death among women globally. Although early screening tests such as the Pap test and HPV DNA test have made cervical cancer a preventable disease, screening resources remain inaccessible and unaffordable to women in developing countries. If we can 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, we may be able to develop a new digital screening solution to make the screening more accessible for women in areas with scarce healthcare resources.

2. Background

Interpretability is not the same as explainability, as noted by Molnar (2022) and Miller (2018). Molnar (2022) defines interpretability as the associations a machine learning model identifies between features and output, or the "extraction of relevant knowledge from the model." It describes the extent to which one can predict what will happen given a change in input or algorithmic parameters. Explainability, on the other hand, pertains to the internal mechanisms of machine learning models and explains why a certain prediction was made (Gall, n.d.). In some cases, simply obtaining the prediction is insufficient. The model must also explain how it arrived at the prediction, particularly when the model affects human life.

Furthermore, machine learning models can only be debugged and audited when they are interpretable. An interpretable machine learning model ensures the fairness of the algorithm and prevents discrimination against underrepresented groups. It also ensures the model's reliability, where small changes in input do not lead to significant changes in predictions, and the model's adoptability, making it easier for humans to understand and implement.

Despite the progress made in developing interpretable machine learning (IML) methods, they still have some major limitations, as Molnar (2022) notes in his book on the subject. For instance, feature dependence can create issues with attribution and extrapolation, causing partial dependence plots to generate fictitious data points not found in the actual data distribution. As a result, IML methods may be unable to capture the true association or causal relationships in the models they aim to interpret. Another significant issue with current IML methods, as Molnar points out, is the lack of statistical rigor. Unlike traditional statistical methods, most IML methods do not provide confidence estimates, and there is no universal standard for evaluating 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.

	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)

Table 2.1.1. The recovery rate of current treatment versus new treatment.
The success rate, expressed as a percentage, is accompanied by the ratio of the number of recoveries to the total cases in parentheses. This information is adapted from the study "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy" by Charig et al. (1986) in the British Medical Journal (Clinical research ed.).

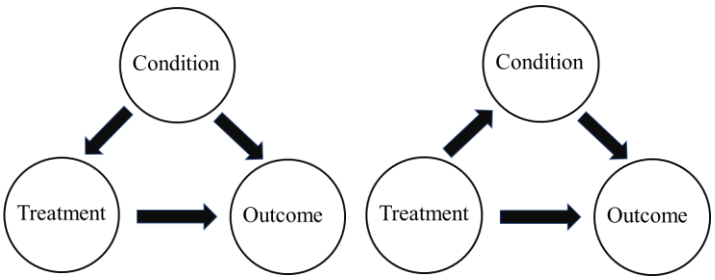
Comparing the effectiveness of two treatments regardless of the size of the stones by looking at the aggregated result, it is clear that treatment B is more effective. However, when the population is subset based on the size of the stones, the conclusion is reversed. Treatment A works better for both 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 treatment on the recovery to make the final decision. Different causal structures will lead to completely different healthcare decisions.



Causal Structure 1 **Causal Structure 2**

Figure 2.1.1 Two distinct causal structures

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. Methods

3.1 Classification Models

Classification models such as Machine Learning, Neural Networks, and Causal Inference can be utilized for various applications. In our study, we employed several machine learning models, including KNN, Decision Tree, Random Forest, and Ensemble Model. We began by using KNN, Decision Tree, and Random Forest to predict classifications. Afterwards, we combined these three models through the implementation of an ensemble model. For this experiment, we leveraged Scikit-Learn, a Python library specialized in tensor computation. Those interested in the code implementation can refer to /ml-research/codes/01_KNN_DT_RF_V2.ipynb for further details.

K-Nearest Neighbor – K-Nearest Neighbor (KNN) is a supervised learning algorithm that is widely used for classification and regression tasks. This algorithm functions by classifying data points based on their feature similarity. Specifically, KNN identifies the K closest points to a given point and selects the majority vote result of these K points as the classification for that point. During the testing phase, KNN calculates the distance between the test data and each row of training data, identifying the K closest points based on the chosen distance calculation method. The algorithm then assigns a class to the test point based on the most frequently occurring class of these K rows. This approach has been described in detail by Taunk et al. (2019).

Decision Tree – Decision Tree is a supervised machine learning algorithm that can be utilized for both classification and regression tasks. According to Machine Learning with Python (n.d.), decision trees function by segmenting the target data into hierarchical tree-like structures with decision boundaries. A key advantage of decision trees is that they are non-parametric, meaning that they can efficiently analyze complex and large datasets without the need for complicated parameters. As noted by Song & Lu (2015), this feature makes decision trees a highly effective tool for data analysis.

Random Forest – Random Forest is another powerful supervised machine learning algorithm that can function as a classifier or regressor. According to Machine Learning with Python (n.d.c), Random Forest is an ensemble machine learning method that creates multiple decision trees by sampling data and then utilizes a voting approach to determine the best machine solution. This ensemble approach helps to mitigate overfitting and enables the results to be averaged by means of sampling and voting. The effectiveness of this approach makes Random Forest a popular tool in various fields of data analysis.

Ensemble Model – Ensemble Model is based on the concept that while a single model may predict a specific dataset with reasonable accuracy, combining different models can potentially enhance overall accuracy. In our study, we utilized a Voting Classifier from Scikit Learn (n.d.). This classifier combines conceptually distinct machine learning classifiers and utilizes a majority vote or the average predicted probabilities (soft vote) to predict class labels. This approach is particularly useful when dealing with a set of equally well-performing models, as it can help to balance out their weaknesses and achieve better results.

Neural Network – In addition to the models discussed above, we also employed a neural network-based classifier in our comparison study. Neural networks, also known as multi-layer perceptron (MLP), simulate the signal propagation mechanism between neurons in the brain using various mathematical

formulas such as affine and activation functions (Han et al., 2018). The model is trained by updating the weights and biases of each function through forward and backward propagation processes using differential functions. For our experiment, we implemented an MLP with three hidden layers, each with 256, 128, and 64 neurons, respectively. The final layer utilized a softmax function to generate a binary discrimination probability. To enhance the accuracy of the training process, batch normalization and dropout functions were added to each layer (Ioffe & Szegedy, 2015; Srivastava et al., 2014). The model was trained using the minibatch method (Li et al., 2014), with a batch size of 32, and the number of training sessions and epochs were set to 30. We implemented the model using PyTorch, which is one of Python's libraries specialized for tensor computation. The relevant code for this experiment is available in /ml-research/codes/02_NeuralNetwork_CervicalCancerClassification_DI.ipynb and should be referred to as necessary. Overall, the neural network-based classifier proved to be a valuable addition to our comparison study, and its implementation using PyTorch provided robust and reliable results.

3.2 Causal Inference Models

The last model tested was the causal inference model, specifically a Bayesian network developed in the 1980s for exploratory data analysis by Pearl (1985). From several causal inference models, we selected the fast causal inference (FCI) algorithm (Spirtes, Meek & Richardson, 1995). The FCI algorithm is a constraint-based non-parametric approach that explores a graphical feature common to all causal-directed acyclic graphs (DAGs) to observationally equivalent sets of statistical tests of conditional independence. Essentially, it identifies dependencies between variables by connecting variables that are purely dependent on each other with a causal edge, after excluding the influence of different and unobserved variables. This algorithm has several advantages, including being less susceptible to identified causal relationships being affected by increases or decreases in other variables and being able to generate correct identification probabilities even in the presence of hidden variables like covariates or a mixture of factors that could negatively affect the model, such as selection bias. Exploratory data analysis often involves unobserved covariates, which can make causal inference challenging. To implement the FCI algorithm, we used a Java desktop application called Tetrad, developed by professors in the Philosophy Department at CMU (Cmu-phil/tetrad, 2022). The results are stored in /ml-research/codes/03_causal_inference and should be referred to as necessary. To access the "Causal Analysis_Cervical Cancer.tet" file, you can download it from the "tetrad-gui-7.1.0-launch.jar" file in the same directory. Then, start the file and select "File > Open Session" from the menu to load the downloaded file.

3.3 Dataset

For this research project, we utilized a structured dataset provided by the UCI Machine Learning repository that included historical medical records of 858 cervical cancer patients. The dataset encompassed various explanatory variable categories, such as patients' demographics and habits, including age, sexual intercourse habits (including the number of sexual partners and the age at which they had their first sexual intercourse), pregnancy history, smoking habits (including whether or not the patient smokes, the number of years they have smoked, and the number of packs of tobacco they consume per year), contraceptive habits (including whether the patient uses hormonal contraceptives or an intrauterine device (IUD) and how long they have used these contraceptives), and sexually transmitted disease (STD) history (including whether the patient has ever had an STD, the number of STDs they have had, the amount of time since their first STD diagnosis, and the amount of time since their most recent STD diagnosis). Additionally, the dataset contained diagnosis information related to cervical

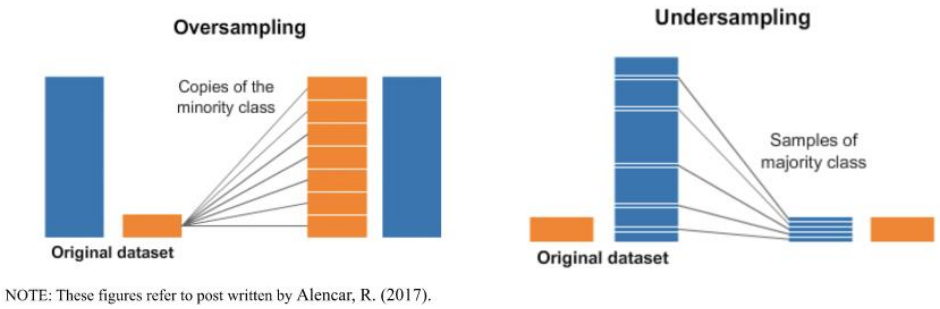
cancer, including whether the patient has cancer, cervical intraepithelial neoplasia (CIN), or cervical intraepithelial neoplasia (HPV).

However, missing values existed in the dataset as some patients did not provide answers to specific questions. To address this, we replaced the missing values with either the mean or median value, depending on the specific variable. We limited the use of continuous variables to those that hold more information to avoid negative impacts on classification results. After filtering the explanatory variables, we selected 13 variables for the discriminant model.

The target variable for this research project was the result of biopsy diagnosis. Our objective was to identify the variables that had a significant impact on biopsy diagnosis by using general classification models and neural networks. To ensure a causal inference, we needed to verify that the sample distributions across different features were similar in both the positive and negative biopsy groups. By doing so, we were able to determine that any observed differences in the outcome were not due to differences in the sample characteristics. This allowed us to make a causal inference and identify the variables that had a significant impact on biopsy results. Even though a propensity score comparison is crucial in determining whether causal inference is feasible or not, we did not perform it in this study as the distributions were found to be comparable. However, for future reference, the relevant code for conducting a propensity score analysis is available in the /ml-research/codes/03_causal_inference/03-01_Feature distribution.ipynb file.

3.4 Sampling Methods

The Imbalanced-learn (Imblearn) library offers a range of tools that can help balance the proportion of data with different classes by either up-sampling the minority class or down-sampling the majority class (Dwivedi, 2020). Oversampling involves adding more samples from the class with fewer data, while under-sampling randomly removes some samples from the majority class to achieve a balance between the two classes in quantity. You can find visual representations of these techniques in Figure 3.4.1.



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

Figure 3.4.1. Oversampling and undersampling Method

3.5 Evaluation Methods

Healthcare decisions are critical and can have significant impacts on patients, physicians, and hospitals. From the patients' and physicians' perspectives, false positive cases can be costly, and from hospitals' perspectives, high false positive rates can result in a waste of healthcare resources. Therefore, when designing machine learning models for healthcare settings, accuracy should not be the sole evaluation metric. Instead, we need to consider different stakeholders' interests and strike a balance between false

positive and true positive rates. For our research, we chose AUC as our evaluation standard while selecting parameters and drawing conclusions.

Various evaluation methods exist for classification models, including Accuracy, Precision, Recall, and AUC (Area under the ROC curve). The confusion matrix is used to describe the performance of a classification model. While choosing our evaluation metric, we considered the importance of balancing the interests of different stakeholders in healthcare decision-making. Therefore, we selected AUC as our evaluation metric, given its ability to evaluate the performance of a model across all possible thresholds.

The accuracy metric calculates the proportion of correct predictions out of the total number of predictions.

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

Table 3.5.1. The components of a confusion matrix

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

Precision refers to the proportion of correct identifications among the total number of positive identifications.

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

Recall measures the proportion of actual positive cases that were correctly identified by the model. It is equivalent to the True Positive Rate (TPR).

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

Receiver Operating Characteristic (ROC) – The Receiver Operating Characteristic (ROC) curve is a graph that shows how well a classification model performs at all possible classification thresholds. It plots two parameters, the True Positive Rate (TPR), which is equivalent to Recall, and the False Positive Rate (FPR). (Google Machine Learning Education, n.d.).

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

The ROC curve plots TPR versus FPR at various classification thresholds, and the area under the curve (AUC) measures the trade-off between TPR and FPR. It provides an overall performance measure across all possible classification thresholds (Google Machine Learning Education, n.d.).

Permutation Feature Importance / Shaley Value – After evaluating the performance of the machine learning models using AUC and TPR, we conducted permutation feature importance to further assess their interpretability. Permutation feature importance (PFI) is a commonly used evaluation method for interpreting machine learning models. According to Fisher, Rudin & Dominici (2019), PFI can estimate the importance of variables in any model by calculating the upper and lower limits of model class reliance (MCR) as point estimates. In this experiment, we applied PFI to three models: KNN, decision tree, and random forest, using the skater library developed by Oracle Open Source.

However, skater is not compatible with PyTorch, which is the neural network library we used for this experiment. Therefore, we employed an alternative method to interpret the neural network model, by calculating the average Shapley value of each variable to determine the global model interpretation, i.e., the contribution of variables to the overall model. Our objective was not to compare the importance of variables but to determine which variables the model relied on to make predictions. Despite using different interpretable machine learning models (Skater and SHAP) to interpret general classification models, the objective of these two methods remained the same.

4. Experiment Results

The dataset consists of 858 patients' historical medical records, including their demographics and habits. Out of these, 55 patients tested positive, while 803 patients tested negative, resulting in an imbalanced dataset. We applied KNN, decision tree, and random forest models to the original dataset. However, all models' performances were not satisfactory. To address this issue, we used under-sampling techniques and trained the models on a balanced dataset consisting of 44 positive samples and 44 negative samples. Table 4.2 presents the performance of the four models, with the Random Forest model showing the highest AUC and Recall. Meanwhile, the Decision Tree model achieved the highest accuracy among all the models.

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

	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

Table 4.2. The performance metrics, including AUC, recall, and accuracy, of the KNN, decision tree, and random forest models.

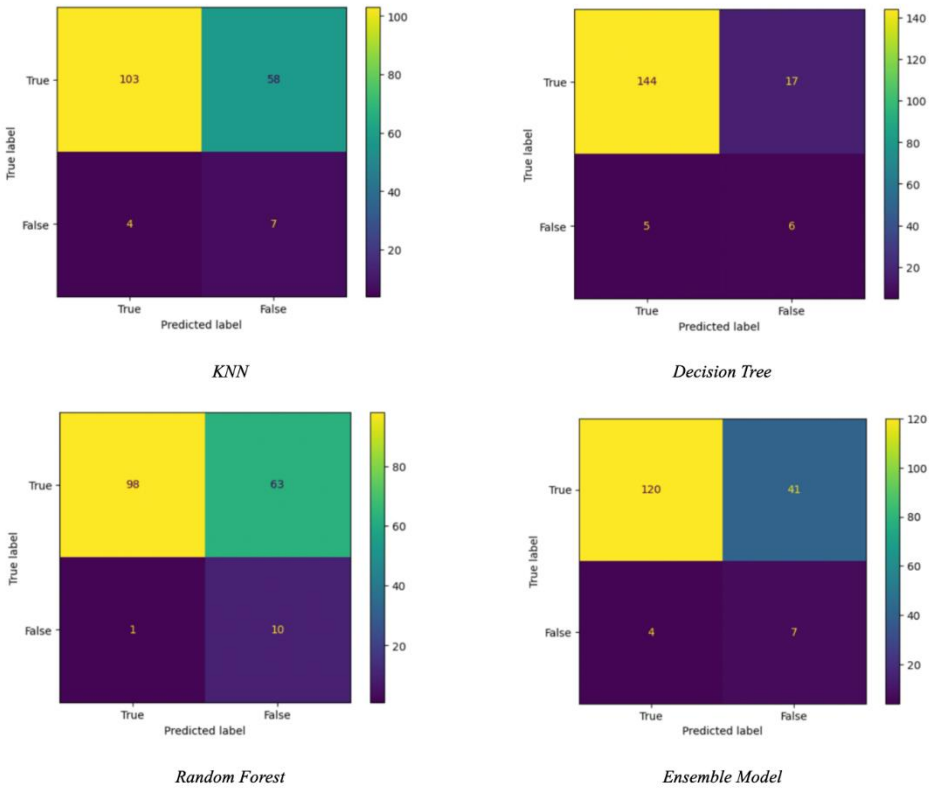


Figure 4.1. Confusion Matrix of the Four Models.

To evaluate the models, we utilized the undersampled dataset and ran each model multiple times to determine the optimal parameters resulting in the highest AUC. Specifically, for KNN, k = 3; for the Decision Tree, the max depth was set to 3, and for Random Forest, n_estimators = 40 and max_depth = 34. After training each model, we combined them into an ensemble model using a soft voting strategy. The goal was to achieve the best performance among all the models. However, surprisingly, the decision tree outperformed the ensemble model in terms of AUC and Recall. When testing on a small dataset, only one out of ten patients was misdiagnosed. This result demonstrates that a more complex model does not always lead to the best outcome.

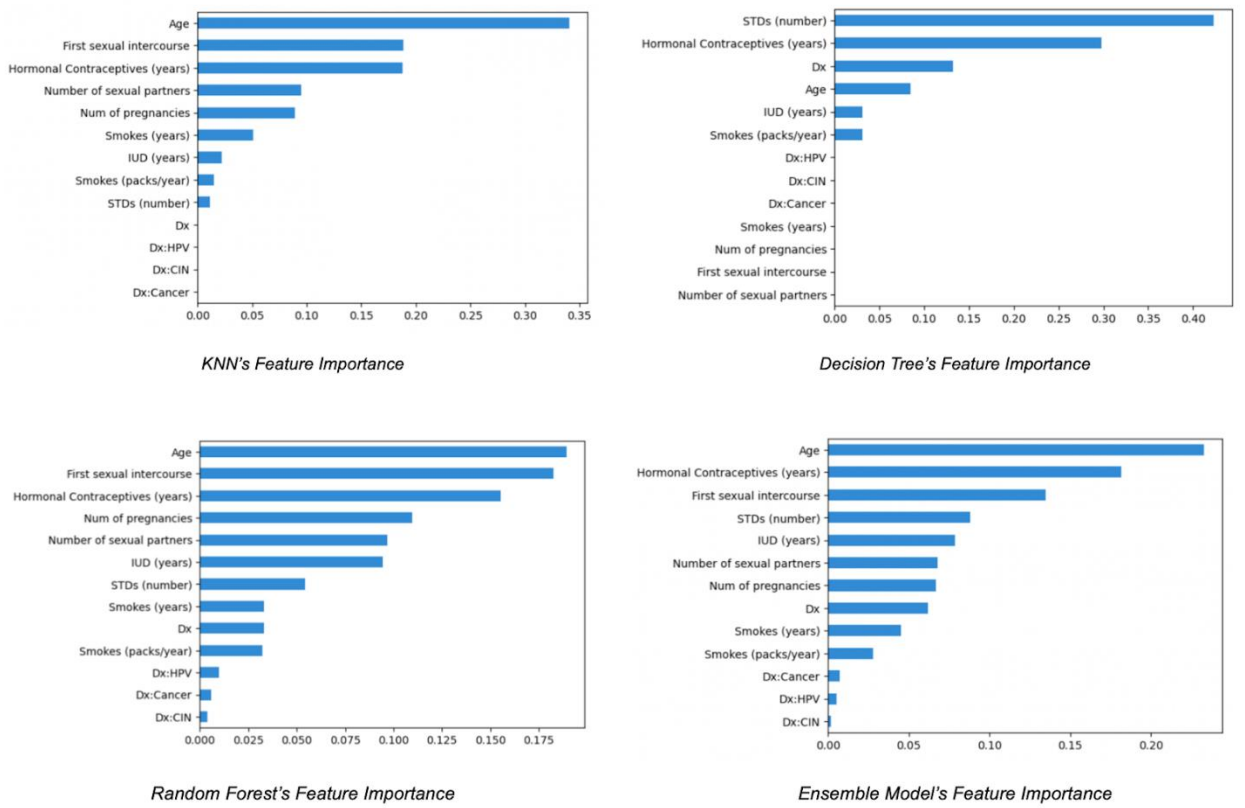


Figure 4.2. Permutation Feature Importance

Figure 4.2 illustrates the feature importance rankings for each model, revealing that age, hormonal contraceptives, and first sexual intercourse are the top three most significant factors in KNN, Random Forest, and the Ensemble Model. Hormonal contraceptives are also the second most important feature in the Decision Tree model. However, our results did not reflect the well-known strong correlation between Dx CIN and cervical cancer, indicating that general classification models may need assistance in capturing essential real-world features.

In Figure 4.3, we present the results of the Neural Network experiment, where the true positive rate is 64.3%. Our primary objective was to identify the model's critical features, rather than creating a high-precision neural network model; thus, we will not delve into the details of the model's detection capability. Nevertheless, we can interpret it as a model with a certain degree of detection capability. In Figure 4.4, we explore the interpretation of the neural network model, which reveals that Hormonal Contraceptives (years), Age, Skomer (years), Number of Sexual Partners, and First Sexual intercourse are the top five crucial features in the model, respectively. These features' direct impact on cervical cancer is not intuitive, suggesting that the neural model can only capture the association between explanatory and target variables and does not capture causal relationships, similar to other models such as KNN, Random Forest, and the ensemble model.

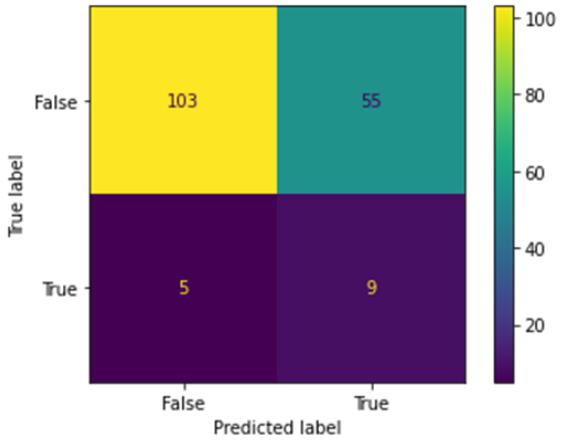


Figure 4.3. Confusion Matrix of Neural Network Model

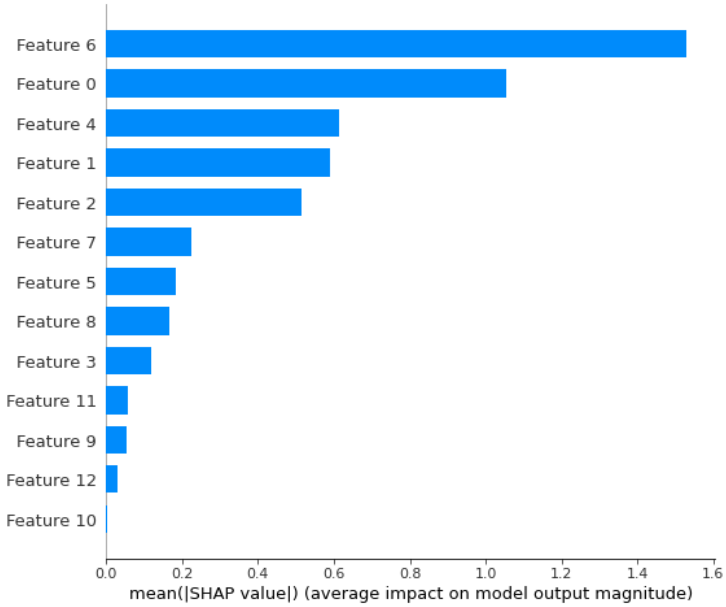


Figure 4.4. Feature Importance in Neural Network Model by Averaged Shapley Value

Moving on, let's dive into the results of the causal AI experiment. Figure 4.5 displays the outcome of the FCI algorithm, which identified the causal relationship between variables. As shown on the right side of the figure, two variables - diagnosis of CIN and number of STDs - are causally related to the biopsy result. CIN refers to cervical intraepithelial neoplasia, an abnormal cell that can lead to cervical cancer, while STDs are a known cause of cervical cancer. It's easy to see why this result makes intuitive sense, but to get a more accurate understanding, we needed to delve deeper into the relationship between STDs and cervical cancer.

To do so, we re-ran the FCI algorithm, this time including STD details as explanatory variables. Figure 4.6 shows that genital herpes is another causal factor detected in positive biopsy results. Studies have shown that the probability of developing cervical cancer increases when genital herpes is transmitted along with other HPVs. Therefore, we can conclude that the risk factors for cervical cancer are a diagnosis of CIN and genital herpes in terms of causal relationships. These findings align with expert knowledge cited from trusted medical webpages, validating the accuracy of our results.

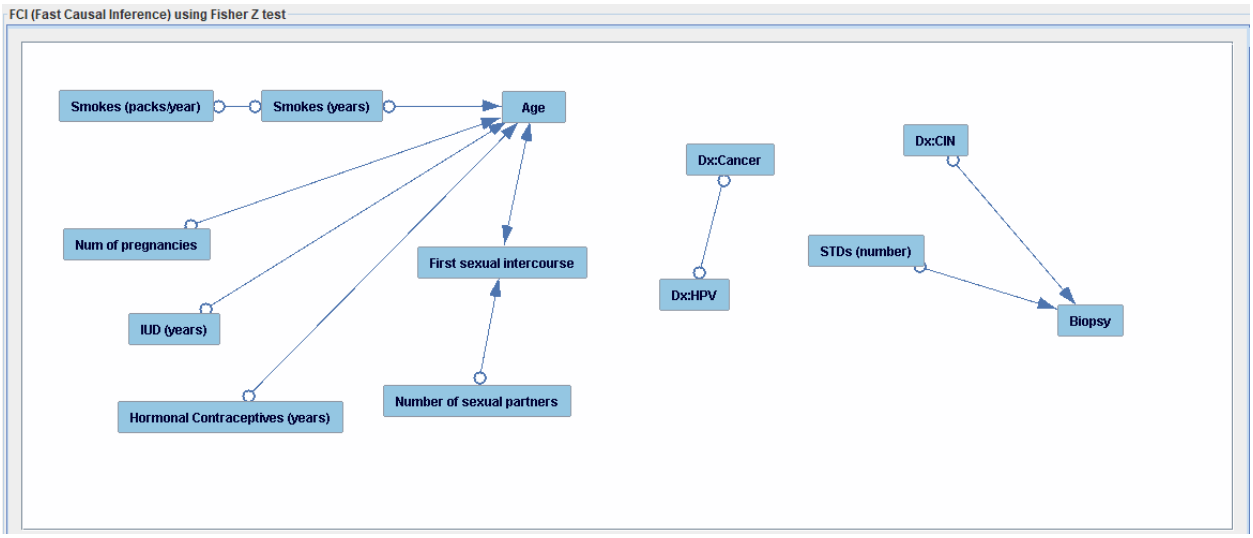
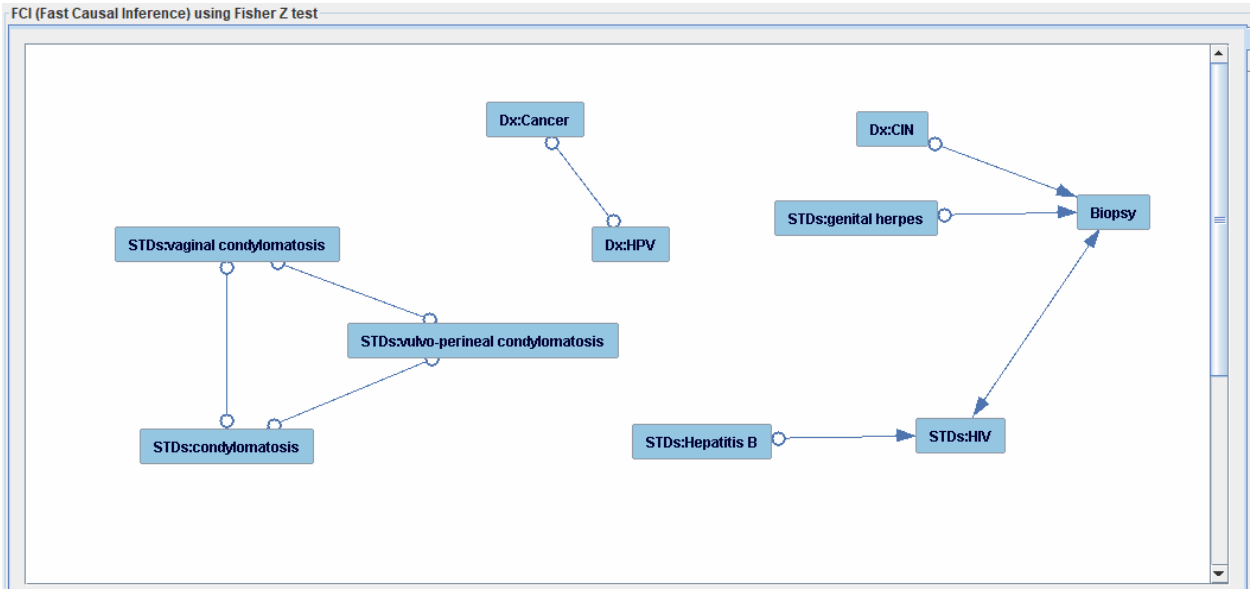


Figure 4.5. Causal Directional Acyclic Graph with FCI and Continuous Variables



Please note that while other sexually transmitted diseases (STDs) were included in the causal inference analysis, they were not found to have any significant causal relationships with the variables in the figure. Therefore, they have been excluded from the illustration.

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

5. Discussion & Conclusion

In this study, we found that the ensemble model, particularly the Voting Classifier, did not perform better than the best model in the set of implemented models, which was the decision tree. This can be attributed to the mechanism of the Voting Classifier algorithm, which uses weighted average probabilities to infer the result. Although the ensemble model could perform better when each model accurately classifies multiple classes, in this case, where the target variable is binary, the ensemble model's classification results could be easily biased by the model with a bigger weight. As such, for binary classification problems, a simple non-ensemble model with higher accuracy may be more appropriate (Scikit Learn, n.d.).

Furthermore, the interpretation results of all the trained models using interpretable machine learning (IML) methods were similar in terms of their permutation-based important features and averaged Shapley values. However, these important features may not capture the causal relationships inferred by the Causal AI algorithm (FCI algorithm), which considers probabilistic and conditional independence to extract causal dependencies (Spirtes, 2001). For instance, the causal inference algorithm identified "Dx CIN" and "STDs: genital herpes" as causes of positive biopsy, but neither of them was included in the important features of each model. This demonstrates that IML methods may not be able to derive the true causal relationships between features and target variables (Molnar, 2022). On the other hand, while causal inference algorithms have a powerful mechanism to detect causal relationships, discerning true causal relationships without expert knowledge can be challenging. Moreover, causal relationships are sometimes detected without a clear direction. Two approaches to deriving true causal relationships are verifying causal relationships extracted with AI algorithms with expert knowledge and incorporating expert knowledge into a causal AI algorithm to ensure correct causal inference.

In conclusion, we make three key points. Firstly, the lack of statistical rigor and comparison standards are common drawbacks shared by existing IML methods, making it difficult to determine which interpretation to trust when different IML methods provide different interpretations of the same machine learning model. Secondly, identifying associations through interpretability is not equivalent to explaining causations identified by causal inference models. Lastly, to create more adoptable machine learning models, we need to consider different stakeholders' interests and explain prediction results to people from diverse backgrounds and varying levels of knowledge in medicine. To achieve explainability in machine learning models, we need to verify whether the causal inference was made correctly by incorporating expert knowledge in developing an explainable machine-learning model.

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-20352506026236 12

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#