Saint Petersburg State University

Department of Mathematical Game Theory and Statistical Decisions

Xu Feiran

Master's Thesis

Using Shapley Value for Interpretive Artificial Intelligence based on Feature Interactions Graph

Specialization 01.04.02

Applied Mathematics and Informatics

Master's Program Game Theory and Operations Research

Science Supervisor:

Professor of Department of Mathematical

Modelling of Energetic Systems

Dr. Sc. Ovanes Petrosian

Saint Petersburg

2023

Contents

Abstract

With the increase of computer computing power, AI has been widely used in various fields of life, but at the same time, more and more cases show us the uninterpretability of AI algorithms. Therefore, the interpretability of AI is particularly important. In recent years, more scholars have proposed to work on the interpretability of machine learning models by Shapley values in cooperative game theory, but when there are more features in the dataset, calculating Shapley values becomes a challenge. Some authors have introduced approximate Shapley calculation techniques. However, as the number of players increases, it remains a challenge to strike a balance between sample size and time cost. The sampling method to calculate shapley values is a method that samples the features themselves by simulating random permutation disease and then estimates the contribution of each feature to the prediction result based on each sampling result, but in use, we found that the random permutation process is still long when the number of participants is large. Therefore, we propose a new approach by using a coalition of "high impact" participants. Shapley values are calculated in less time and a more meaningful way to measure the plausibility of the interpretation of the results is proposed.

Introduction

Background and related work

Anomaly detection is a technique used to identify anomalous patterns that do not conform to expected behavior [16] and is used to identify outliers or anomalous patterns in a data set. These anomalies usually manifest as data points or behaviors that are significantly different from normal data and may be due to measurement errors, system failures, malicious activity, or other causes.In numerous practical applications, such as fraud detection, industrial quality control, and intrusion detection, anomaly detection is indispensable[3]. In medical physiological data analysis [27], anomalies this means that if a patient's physiological metric data deviates from other observations, the probability of him having a certain disease is significantly increased. The use of machine learning models in anomaly detection has become quite common[30, 21]. Although artificial intelligence algorithms are widely used in the medical field, there is still a trust [4] problem. Therefore, it is necessary to establish interpretable AI for the sake of trustworthiness of diagnostic results, correctness.

Explainable AI emphasizes the understandability and transparency of models and aims to help people more easily understand how AI systems make decisions[14]. With the widespread use of machine learning techniques such as deep learning, AI models have become increasingly complex. However, this complexity can make the decision-making process of models difficult to interpret, even for researchers in the field of machine learning, who many times are unable to

give reasonable explanations for the models' judgments [11]. Some scholars have even called for using models with interpretability rather than machine learning black box models in high-stakes decision making scenarios to better understand the model outputs and decision making process[20].Explainable AI has become an important research area in order to improve trust and meet regulatory requirements.Ihe goal of interpretability is to describe the internals of a system in a way that is understandable to humans. The success of this goal is tied to the cognition, knowledge, and biases of the user: for a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user.[6]

Currently, researchers in interpretable AI have proposed various solutions for interpretable AI from different perspectives. Using feature maps of convolutional neural networks and calculating their gradients on target output categories, Selvaraju et al. (2017) proposed the new visualisation method Grad-CAM for the visual interpretation of images[24].Grad-CAM's efficacy in tasks such as image classification and target detection has made it a significant research result in the field of XAI. Among the most significant research findings in the discipline of XAI. David Bau et al. proposed in 2018 a method known as GAN Dissection that can analyse the image characteristics learnt by GAN by visualising and analysing the hidden layers in generative adversarial networks (GANs)[1]. This method is applicable to image generation, speech synthesis, and other disciplines, and can help users better comprehend GAN-generated images.

Depending on the approach and application domain, interpretable AI can be

divided into the following two main categories:

- **Intrinsic Interpretability:** This class of approaches focuses on the direct use of models that are inherently more explanatory. These models naturally produce easily understandable results during training without additional post-processing. Examples include: logistic regression, decision trees, linear regression, etc. 2.
- **Post-hoc Interpretability** Such methods aim to explain complex and hard-to-explain black-box models. They are usually applied after the model training is completed to reveal the decision process or feature importance of the model. For example: SHAP [11], LIME [18], etc.2.

Shapley value is a concept from game theory for assigning the contributions of different participants (also called players) in a cooperative game. It was first introduced by Lloyd S. Shapley in 1953 to find a fair and consistent way to measure the contribution of each player to the outcome of a game.Shapley values have been widely used in fields such as economics, operations research, and social sciences, and have also been introduced to interpretable artificial intelligence in recent years. Applying Shapley values to interpretable AI can effectively quantify the contribution of each feature to the predicted outcome of a model. In the context of interpretable AI, Shapley values provide a fair and consistent contribution measure for features that is not limited to a specific type of model, making it possible to apply it to a wide range of machine learning tasks so that we can analyze the Shapley values of individual features to explain the decision process of a model and provide valuable insights to domain experts.

Cancer detection is a crucial medical task that can detect cancer lesions at an early stage, enhance treatment efficacy, and lengthen patient survival. As artificial intelligence technologies continue to advance, a growing body of research investigates the use of machine learning algorithms to accomplish more precise and rapid cancer detection[5].Cancer detection based on machine learning typically involves extracting features from medical images or biomarker data (e.g., blood samples or tissue sections) and training machine learning models to automatically identify and classify tumour types[9]. In recent years, deep learning algorithms have emerged as an essential research direction in the field of cancer detection, as they can automatically learn more complex features and achieve greater classification accuracy.

Early research concentrated on conventional machine learning algorithms[23], including support vector machines, decision trees, and random forests. These models necessitate the manual extraction of features from biomarker data or medical images for use in training classifiers[5]. Due to the intricacy and unpredictability of medical images and biomarker data, these methods have limited generalizability and interpretability, despite their high diagnostic precision.

In recent years, deep learning algorithms have become the predominant method for detecting malignancy[7]. Deep learning algorithms have demonstrated superior performance in cancer detection tasks due to the benefits of automatic learning of high-level features and end-to-end training. This research aims to enhance the precision, dependability, and interpretability of cancer detection algorithms in order to provide clinicians with more effective diagnostic tools.

Aims and objectives of work

When there are a large number of participants, the main challenges of Shapley value calculation include computational complexity, result stability and reliability, and redundant information handling. The amount of computation required increases exponentially with the number of participants due to the need to traverse all possible combinations of collaborations, resulting in computation times that become long and even in some cases results that cannot be obtained in a limited time. In practical applications, the results of each calculation may vary slightly due to random factors and incomplete data, etc. Such differences may be magnified when the number of participants is large, thus affecting the stability and reliability of the Shapley values. Second, a large number of participants is prone to overlapping functional participants, introducing redundant information among multiple combinations, making the calculation of Shapley values more complex and susceptible to interference. Based on these problems, we conducted a study on the application of shapley values to explain artificial intelligence models when there are many participants.

Tree Explainer [12], Deep Explainer (DeepLIFT + Shapley values) [26], and Kernel Explainer (Linear LIME + Shapley values) [11] have emerged as methods for estimating the Shapley value for different algorithms within the field of explainable AI. These elucidable methods have proven to be effective at estimating the Shapley value. However, the majority of them are only appropriate for particular categories of algorithms.

In our previous paper "High-dimensional explainable AI for cancer detection"[31], we used the sampling method [2] for the calculation of the shapley value when there were more participants. The sampling method is a way to approximate the shapley value by random sampling. By randomly sampling the combination space, we can effectively reduce the computational complexity and thus obtain a more accurate estimation result in a limited time.

In the sampling method, we first randomly select a certain number of subsets from all possible combinations and calculate the contributions of each participant in these subsets. Then, based on these sample data, we estimate the Shapley value of each participant using statistical methods. In order to improve the accuracy and reliability of the estimation results, the sample size can be increased appropriately or other optimization strategies can be used. However, during our experiments, we found that the use of the sampling method still required a lot of time, and although it was much faster than calculating the Shapley values directly through the metric, the calculation time of the sampling method was still long.

Consequently, we propose a novel method to increase efficiency based on conventional sampling techniques that can be applied to all high-dimensional AI problems and algorithms: Graph Based Sampling Methodology for Shapley value.

In this paper, we will describe our method, which we refer to as the graph-based sampling method. Utilising a graph of the relationships between all participants, this method helps to increase computational speed. In Section 2, we will discuss briefly the context and related work, including the isolated forest algorithm,

Shapley values, the sampling Shapley method, and our prior work. In Section 3, we describe our most important methods, including the construction of relational graphs, the stochastic search algorithm, and the convergence measure. Our method is theoretically applicable to all applications of Shapley values. This study is novel in that we propose to use the graph method to characterise the relationships between all participants and to use this information to assist users in gathering more useful data to generate samples.

Academic Achievements

I have been working on combining Shapley values with interpretable AI, and have achieved some results, published a few papers, and attended a number of conferences, which are listed below.

Article for Journal

- High-dimensional explainable AI for cancer detection; 2021. Published[31]
- Explainable AI: Graph Based Sampling Approach for High Dimensional AI System; Preparing.

Article for Conference

- Explainable AI: using Shapley value to explain the anomaly detection system based on machine learning approaches; 2020 Control Processes and Stability.[32]

- XAI evaluation: evaluating black-box model explanations for prediction; 2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT)[29].

Structure

This section is an introduction to the paper, which will be described in the following way: Chapter 1 will introduce in detail the anomaly detection and ANN model and show the detection results of the model. Chapter 2 introduces the interpretable algorithm and our proposed new technique to reduce the time of computing shapley values by sampling method. Chapter 3 presents our experimental results and analysis. Chapter 4 presents the conclusions and concludes with future work.

1 ANN Models and Cancer Detection

The concept of artificial neural networks (ANNs) dates back to the 1940s and 1950s, when Warren McCulloch and Walter Pitts first proposed a simplified model of biological neurons, the M-P model[13], which laid the foundation for subsequent work on neural network development. After decades of accumulation, artificial neural networks have become a mature model and are widely used in various fields.

1.1 Artificial Neurons

Artificial neurons are derived from the neuron theory, which was realized in the late 19th century that the complex nervous system is made up of a large number of neurons combined. Neurons consist of cells and the many protrusions they send out. The cell has a nucleus inside the cell and the role of the protrusions is to transmit information. Several protrusions are called "dendrites" that introduce the input signal, while only one protrusion is called "axon" that is the output. Such a basic unit with multiple inputs and a single output can be considered from an information processing point of view. The following figure 1.1 introduces the MeCulloch-Pitts model, which is shown in a schematic structure.

For the j neuron, the input signal $_i$ is received from multiple other neurons. The strength of each synapse is expressed as a coefficient w_{ij}, which is the weighted value of the action of the i neuron on the j neuron[19]. The combined effect of the input signals using some operation to give their total effect is called the

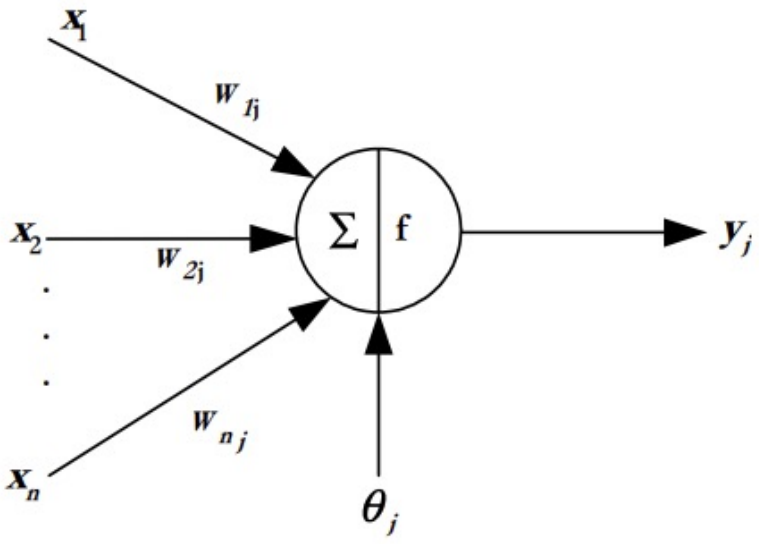

Figure 1.1: M-P model

"net input" and is denoted by I_j. The simplest expression (1.1) of the net input is a linear weighted summation:

$$I_j = \sum w_{ij} x_i \tag{1.1}$$

This action causes a change in the state of neuron j, and the output of neuron j, y_j is a function of its current state. Therefore, the mathematical equation (1.2) of the M-P model is shown below:

$$y_j = sgn(\sum w_{ij} x_i - \theta_j) \tag{1.2}$$

where θ_j is the threshold value and sgn is the sigh function. When the input exceeds the threshold value y_j takes $+1$ as output, conversely, -1 as output.

An artificial neural network composed by using a large number of interconnected neurons will show several features of the human brain, and the artificial neural

network also has preliminary self-adaptation and self-organization capabilities. The value of the weights w_{ij} is changed during the training process to adapt to the requirements of the surrounding environment.

1.2 Basic Elements of Artificial Neural Networks

In the design and application of artificial neural networks, it is usually necessary to consider three most basic elements, namely, the **activation function**[22], the **form of connections** between neurons and **the training of the network**.

Activation function consists of the process from the input signal to the input activation value, which eventually produces the output signal. Activation functions come in various forms, and different features can be used to form artificial neural networks with different functions. In our cancer recognition problem, we choose the *Sigmoid* function as the activation function of the artificial neural network.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1.3}$$

Sigmoid function has smooth and continuous characteristics, and its gradient can be directly expressed by the value of *Sigmoid* function. This advantage makes it possible to have the gradient derived from the output of *Sigmoid* function directly in the subsequent training process, which greatly reduces the computational effort in the training process. At the same time, the *Sigmoid* function maps any real number input to between $(0, 1)$, a feature that makes this function very popular in binary classification problems.

In the network, several neurons form the layers of the network, and the signals are transmitted from input to output in the order of the layers. The neurons in layer i only accept the signals given in layer $i-1$, and there is no feedback between neurons. The figure 1.2 shows the structural features of the forward network. It can be seen that the input node is not involved in the computation, each layer has neurons with computational functions, and each neuron involved in the computation has any number of inputs but only one output. The input node layer is layer 0, and the computational node layers form layer 1 to layer n one at a time from the bottom to the top, and we call this structure an n-layer forward network. In the forward network, the input and output layers are called visible layers and the intermediate layers are called hidden layers.

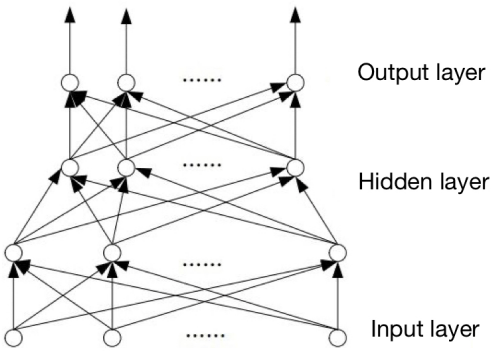

Figure 1.2: Forward Neural Network

Rumelhant and Mcllelland proposed the Back Propagation (BP) algorithm in 1986 [22] , making it one of the most extensively utilised algorithms. According to its fundamental principle, this network's learning process consists of two processes: forward propagation of the signal and backward propagation of the error. In forward propagation, input samples are transmitted from the input

layer to the concealed layer, where they are processed layer by layer, before being transmitted to the output layer. When the output of the feeble output layer does not match the desired output, the error is propagated backwards. The back propagation of error is to reverse the output error back to the input layer, and the error is apportioned to all units in each layer, so as to obtain the error signal of the units in each layer, and this error signal several bits to correct the weighting basis for each unit. Forward propagation and error back propagation are performed continuously, so that the unit weights are continuously adjusted; this is the neural network training process.

1.3 Mathematical principles for training phase

The structure of an artificial neural network, which is used as an example to derive the bp algorithm, is depicted in the following schematic 1.3.

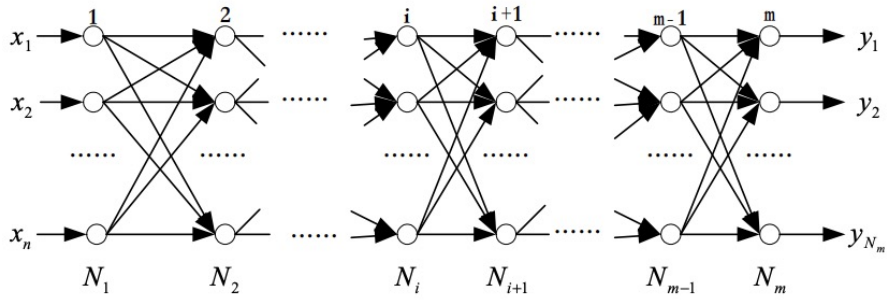

Figure 1.3: Structure of ANN

x_i:input of ANN; y_j:the actual output of the neural network; d_i: expected output of the neural network; W_{ijk}: the j neuron in layer i to the k neuron in layer $i+1$ connection weight; O_{ij}: the output of the j neuron in i layer; θ_{ij} : threshold of the j neuron in i layer; net_{ij}: total input of the j neuron of the i

layer.

Forward propagation process:

$$net_{ij} = \sum_{k=1}^{N_{i-1}} O_{(i-1)k} * W_{(i-1)kj} \tag{1.4}$$

$$O_{ij} = f_s(net_{ij}) = \frac{1}{1 + exp[-(net_{ij} - \theta_{ij}]} \tag{1.5}$$

Error between actual output and desired output:

$$e_j = d_j - y_j \tag{1.6}$$

The objective function to be optimized:

$$E = \frac{1}{2} \sum_j (d_j - y_j)^2 \tag{1.7}$$

Update the weights of the network along the gradient descent direction of the function E, η is the learning rate:

$$\nabla W_{ijk} = -\eta \frac{\partial E}{\partial w_{ijk}} = -\eta \frac{\partial E}{\partial net_{(i+1)k}} * \frac{\partial net_{(i+1)k}}{\partial w_{ijk}} = \eta \delta_{ik} \frac{\partial net_{(i+1)k}}{\partial w_{ijk}} \tag{1.8}$$

Here δ_{ik} :

$$\delta_{ik} = -\frac{\partial E}{\partial net_{(i+1)k}} \tag{1.9}$$

1) Calculate the $\frac{\partial net_{(i+1)k}}{\partial w_{ijk}}$ at first :

$$\frac{\partial net_{(i+1)k}}{\partial w_j ik} = \frac{\partial}{\partial w_{ijk}}(\sum_{h=1}^{N_i} O_{ih} * W_{ihk}) = O_{ij} \tag{1.10}$$

$$\nabla W_{ijk} = -\eta \frac{\partial E}{\partial w_{ijk}} = -\eta \delta_{ij} O_{ij} \tag{1.11}$$

2) Calculate the δ_{ik}:

$$\delta_{ik} = -\frac{\partial E}{\partial net_{(i+1)k}} = -\frac{\partial E}{\partial O_{(i+1)k}} * \frac{\partial O_{(i+1)k}}{\partial net_{(i+1)k}} \tag{1.12}$$

A.

$$\frac{\partial O_{(i+1)k}}{\partial net_{(i+1)k}} = f'(net_{(i+1)k}) = f(net_{(i+1)k})(1 - f(net_{(i+1)k})) \tag{1.13}$$

$$= O_{(i+1)k}(1 - O_{(i+1)k}) \tag{1.14}$$

B.

i) If $O_{(i+1)k}$ is a node in output layer:

$$\frac{\partial E}{\partial O_{(i+1)k}} = y_k - d_k; \tag{1.15}$$

$$E = \frac{1}{2}\sum_j (d_j - y_j)^2 \tag{1.16}$$

Including the preceding result in the formula:

$$\delta_{ik} = -\frac{\partial E}{\partial net_{(i+1)k}} = -\frac{\partial E}{\partial O_{(i+1)k}} * \frac{\partial O_{(i+1)k}}{\partial net_{(i+1)k}} = (d_k - y_k)O_{(i+1)k}(1 - O_{(i+1)k}) \tag{1.17}$$

$$= (d_k - y_k)y_k(1 - y_k) = (d_k - O_{mk})O_{mk}(1 - O_{mk}) \tag{1.18}$$

ii) If $O_{(i+1)k}$ is a node in hidden layer:

$$\frac{\partial E}{\partial O_{(i+1)k}} = \sum_{h=1}^{N_{i+2}} \frac{\partial E}{\partial net_{(i+2)h}} * \frac{\partial net_{(i+2)h}}{\partial O_{(i+2)k}} = -\sum_{h=1}^{N_{i+2}} \delta_{(i+1)h} w_{(i+1)kh} \tag{1.19}$$

$$net_{ij} = \sum_{k=1}^{N_{i-1}} O_{(i-1)k} * w_{(i-1)kj} \tag{1.20}$$

If $O_{(i+1)k}$ is a hidden node, its actual output is known, but its correct output cannot be determined in advance, with the exception that the total error is related to the output of the hidden layer, while the output of the hidden layer must influence the input of each node in the next hidden layer.

$$\delta_{ik} = O_{(i+1)k}(1 - O_{(i+1)k}) \sum_{h=1} N_{i+2} \delta_{(i+1)h} w_{(i+1)kh} \tag{1.21}$$

Therefore, we derive the formula for updating the weights of the BP algorithm:

$$\nabla W_{ijk} = \begin{cases} \eta(d_k - y_k)y_k(1-y_k)O_{ij}, & \text{if i+1 is output layer} \\ \eta O_{(i+1)k}(1-O_{(i+1)k}) * (\sum_{h=1}^{N_{i+2}} \delta_{(i+1)h} w_{(i+1)kh})O_{ij}, & \\ & \text{if j+1 is hidden layer} \end{cases} \tag{1.22}$$

The neurons in i layer:

$$\delta_{ik} = \begin{cases} (d_k - y_k)y_k(1-y_k), & \text{if i+1 is output layer} \\ O_{(i+1)k}(1-O_{(i+1)k}) * (\sum_{h=1}^{N_{i+2}} \delta_{(i+1)h} w_{(i+1)kh}), & \text{if j+1 is hidden layer} \end{cases} \tag{1.23}$$

$$W_{ijk}(t+1) = W_{ijk}(t) + \nabla W_{ijk} = W_{ijk} + \eta \delta_{ik} O_{ij} \tag{1.24}$$

1.4 Anomaly Detection with ANN model

We use the dataset on breast cancer to detect anomalies. In our case, the value B in the dataset's "diagnosis" column corresponds to benign breast cancer and can be considered a normal sample, whereas the value M corresponds to malignant breast cancer and can be considered an aberrant sample. Moreover, the dataset contains 569 patient cases with 30 characteristics, including hemifaciality, symmetry, concavity, and compactness. The tame samples accounted for 70% of the overall number of samples, while the test samples accounted for

30%. On the basis of the theory outlined in the preceding sections, we devised an ANN model for identifying patients with breast cancer that is malignant. These are the outcomes of our experiments:

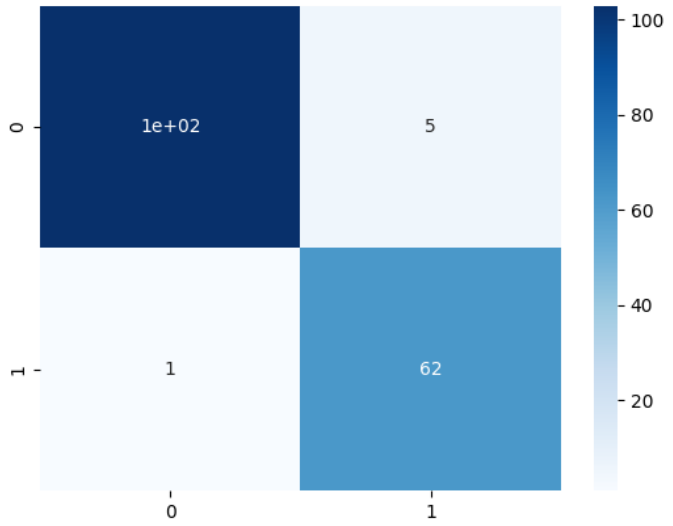

Figure 1.4: confusion matrix of the

Table 1.1: Evalution of ANN model

Accuacy	Precision	Recall	F1-score
96.49%	[99.03%,92.54%]	[95.37%,98.41%]	[97.17%,95.38%]

Precision defines the proportion of predicted positive class (abnormal) samples that are, in fact, positive class. In this case, the precision is 0.99 (normal) and 0.92 (abnormal), indicating that the prediction is highly accurate. Recall identifies the proportion of actual positive classes among samples that were predicted to be positive classes. Recall values of 0.95 (non-anomalous) and 0.98 (anomalous) indicate that the model has a high sensitivity for detecting anomalies. The fscore's cumulative mean is determined by combining accuracy

and recall. In this instance, the F1 scores of 0.9717 (non-anomalous) and 0.9538 (anomalous) indicate that the model performs better overall.

The test results indicate that the neural network model is relatively accurate at detecting breast cancer. This demonstrates that the artificial neural network is capable of learning and capturing breast cancer-related data features. It also demonstrates that breast cancer detection models with a high degree of accuracy can provide supplementary information for clinical diagnosis and aid physicians in determining the nature of lesions with greater speed and precision, allowing for the creation of individualised treatment plans for patients.

In the previous section, the fundamentals of ann and the BP algorithm were described, and breast cancer detection with ann produced more favourable results. However, when implementing these algorithms in the actual world, we must provide the most believable and accessible explanation possible. Therefore, the remainder of this paper will concentrate on the interpretable AI approach, the problems we encountered during our research, and the solutions we developed.

2 Interpretive artificial intelligence for AI models with high-dimensional input

In 1951, Lloyd Shapley invented the idea of Shapley's value[25], for which he was awarded the 2012 Nobel Prize in Economics. Shapley's value is a notion in cooperative game theory. It gives each cooperative game's total surplus produced by the alliance of all players a different allocation. A number of desired characteristics define the Shapley value.

Scott and his colleagues proposed in 2017 to adapt the concept of Shapley values to the field of interpretable artificial intelligence in order to explain predictive models [11]. SHAP (SHapley Additive exPlanations) is an interpretable method that can explain the output of the majority of machine learning models. In analysing their results, we discovered that SHAP does not precisely compute the true Shapley value to guarantee computational efficiency, but instead provides distinct computational strategies for various types of machine learning. However, the properties of Efficiency, Symmetry, Dummy, and Additivity enable an accurate interpretation of the Shapley values for the solutions. Therefore, we wish to reduce the amount of time and memory required to compute Shapley values.

In their 2021 paper[28], Zhanghao and his research team investigated methods for interpreting multivariate Shapley interactions utilising deep neural networks. We propose a new method to compute and explain multivariate Shapley interactions using locally sensitive hashing (LSH) and approximate nearest neighbour

search, thereby enhancing the model's interpretability. The method can divide the input feature space into multiple bins, thereby reducing the complexity of Shapley value computation and estimating Shapley values through the selection of a representative sample set.

The way it works is that a coalition of players works together to achieve a specific overall gain. What should the final allocation of the ensuing surplus be among the players in any given game given that certain players may contribute more to the alliance than others or may have differing negotiating power? Or, to put it another way, how significant is each participant to the total effort, and what benefits can each realistically anticipate? Shapley values offer a potential response to this query.

2.1 Shapley Value in Game Theory

The Shapley value [25] is a concept from game theory for distributing total payoffs in a cooperative game to individual players in an equitable manner. It was introduced in 1953 by Nobel laureate Lloyd Shapley and is regarded as a stable and equitable allocation method.The Shapley value establishes the proportion of each participant's allocation by calculating the average marginal contribution of each participant to the total compensation.The explicit equation of the Shapley value is given in Eq. (3.1):

$$\varphi_i = \sum_{S|i\in S\subseteq N} \frac{(|S|-1)!(|N|-1)!}{|N|!}[v(S) - v(S \setminus \{i\})], \tag{2.1}$$

- i: The number of a player;
- $|S|$: The number of players in coalition S;
- $v(S)$: The characteristic functions of coalition S ;
- N: The set of all players;

In game theory, the player contribution to cooperative games is measured by the Shapley value. Due to the following characteristics, Shapley values are theoretically significant and are often applied in various fields.

- **Efficiency**: The sum of the Shapley values of all agents equals the value of the grand coalition, so that all the gain is distributed among the agents:

$$\sum_{i=1}^{n} \phi_i(v) = v(N) \tag{2.2}$$

- **Symmetry**: If two players have the same impact on all subsets, their Shapley values are equal. This property shows that the contribution of each player is equal and they should receive fair rewards regardless of their role in the game.

$$\phi_i(v) = \phi_j(v), \forall v \tag{2.3}$$

- **Linearity**: The distributed gains should match the gains derived from v and the gains derived from w if two coalition games defined by gain functions v and w are combined:

$$\phi_i(v + w) = \phi_i(v) + \phi_i(w)\phi_i(av) = a\phi_i(v) \tag{2.4}$$

for every i in N, and $a \in R$.

- **Null Player**: The Shapley value for a null player i in game v is zero. A player i is null in v if $v(S \cup i) = v(S)$ for all coalition S that do not contain i.

To increase the accuracy and readability of models in the field of machine learning, Shapley values are frequently utilised in feature importance evaluation and model interpretation. We can comprehend the decision-making process of model prediction and identify areas that want improvement by assessing the degree of contribution of each feature to the model output. The Shapley value can also assist us in locating the model's issues with unfairness and offering appropriate options for improvement. In conclusion, Shapley value is a crucial tool that may help us better understand and analyse the model's prediction outcomes as well as increase the model's dependability and interpretability.

Using the Shapley value, the joint reward is distributed among the players ($\sum_{i=1}^{n} \varphi_i = V(N)$). Generally speaking, if a player makes a bigger contribution to the cooperation, then his imputation value will be bigger. In machine learning, the Shapley value approach can explain the contribution of each feature value. It can be used for global explanation and for local explanation[31].

2.2 Sampling Shapley Approach

Shapley value is a method used in game theory to measure the contribution of participants to a game, which enables us to comprehend the influence of

each player on the game's outcome. In machine learning, the Shapley value is commonly employed in feature importance assessment, which informs us of the significance of each feature for model prediction. In practise, however, as the number of feature combinations and computational complexity increase, it becomes impractical to explicitly calculate the Shapley value for each subset of features. Consequently, sampling-based algorithms have become a prevalent method for calculating Shapley values.

In 2009 a sampling method was proposed to reduce the computation complexity of the exact formula for the Shapley value [2]. Basic cooperative game theory research shows that the approximate Shapley algorithm is effective for large-scale games. Although researchers have been attempting to identify an algorithm that can accurately compute Shapley values, the time and resources devoted to this endeavour have been considerable, and the results have been limited. Therefore, it makes more sense to approximate the Shapley values using a sampling-based algorithm.

The algorithm is presented below[31]:

1. Model the game: define $n = |N|$ feature player from input data set D. Set sampling size as M.

2. Set sample M: the population of the sampling process P will be the set of all possible orders of N players, i.e., $P = \pi(N)$. Let $O : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\}$ be a permutation that assigns to each position k the player $O(k)$. By $\pi(N)$ denote the set of all possible permutations with player set N.

3. Observe characteristic function: the characteristics observed in each sampling unit, $O \in \pi(N)$, are the marginal contributions of the players in the order o, i.e. $x(o) = (x(o)_1, \ldots, x(o)_n)$, where $x(o)_i = v\left(\text{Pre}^i(o) \cup \{i\}\right) - v\left(\text{Pre}^i(o)\right)$.

4. Estimate the Shapley values: the estimate $\hat{Sh}_i$ of the parameter Sh, will be the mean of the marginal contributions over the sample M, i.e. $\hat{Sh} = \left(\hat{Sh}_1, \ldots, \hat{h}_1\right)$, where $\hat{Sh}_i = \frac{1}{m} \sum_{O \in M} x(o)_i$.

5. Obtain the final result: the selection process used to determine the sample M will take any order $O \in \pi(N)$ with probability $\frac{1}{n!}$.

The following is the pseudocode for this algorithm[31]:

Algorithm 1: *AlgorithmApproShapley*

begin

Determine m ;

$Cont := 0$ and $\hat{Sh}_i := 0 \ \forall i \in N$;

While $Cont < m$;

begin

Take $O \in \pi(N)$ with probability $1/(n!)$;

For all $i \in N$;

begin

Calculate $Pre^i(O)$;

Calculate $x(O)_i := v(Pre^i(O) \cup \{i\}) - v(Pre^i(O))$;

$\hat{Sh}_i := \hat{Sh}_i + x(O)_i$;

end

$Cont := Cont + 1$

end

$\hat{Sh}_i := \frac{\hat{Sh}_i}{m} \ \forall i \in N$

end

In our previous paper, we used the same dataset but performed interpretive work on the Isolation forest algorithm. It is difficult to calculate the Shapley value explanation using the exact approach by enumerating and analyzing the characteristic function for all coalitions. Therefore we proposed to use the Sampling approach to approximate the results for the exact approach[31]. The error was calculated using the formula $\sum(sh_i^{k-1} - sh_i^k)$, where k is the iteration number and i is the index of features. It can be observed that the algorithm

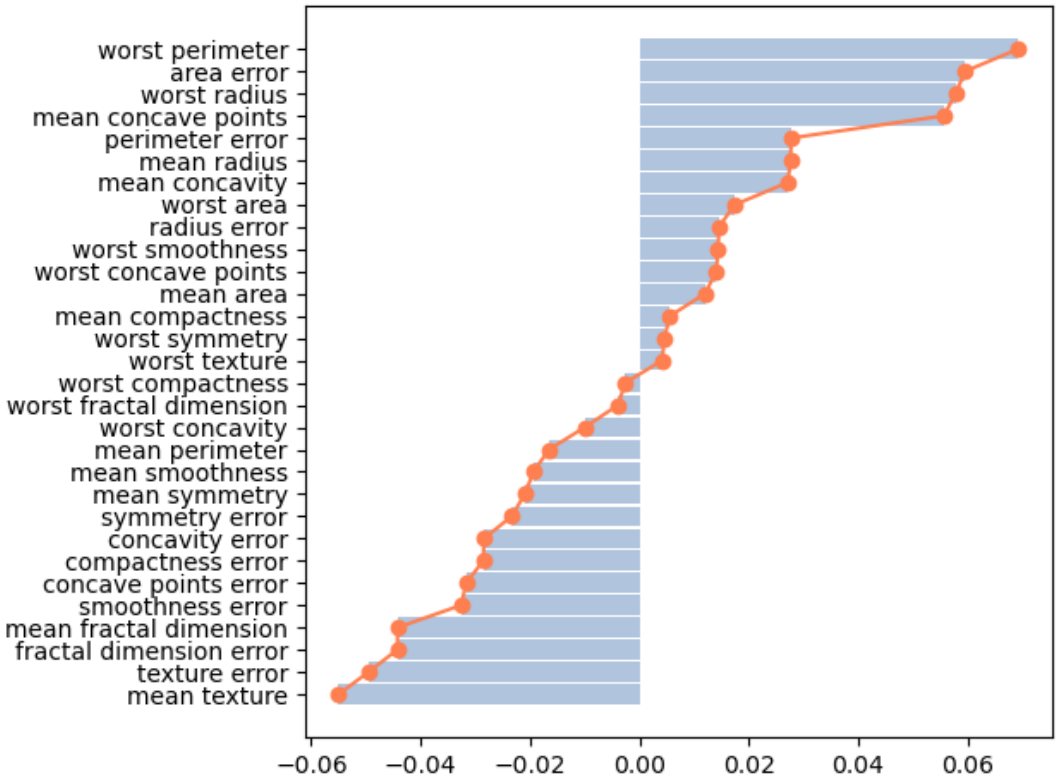

Figure 2.1: Shapley value explanation for Isolation forest model using Sampling approach.

converges faster within 300 iterations (figure 2.1 and figure 2.2).

Despite the fact that the Shapley value is approximated in this manner, the time-intensive character of the sampling method remains significant. Similarly, we did not analyse the information provided by the ranking changes in the previous paper. We have therefore improved the sampling procedure.

2.3 Sampling based on Weighted Graphs

During practical testing, we discovered that the sampling method requires a significant amount of computational time during the sampling phase, which may impose a significant computational burden on the estimation of Shapley

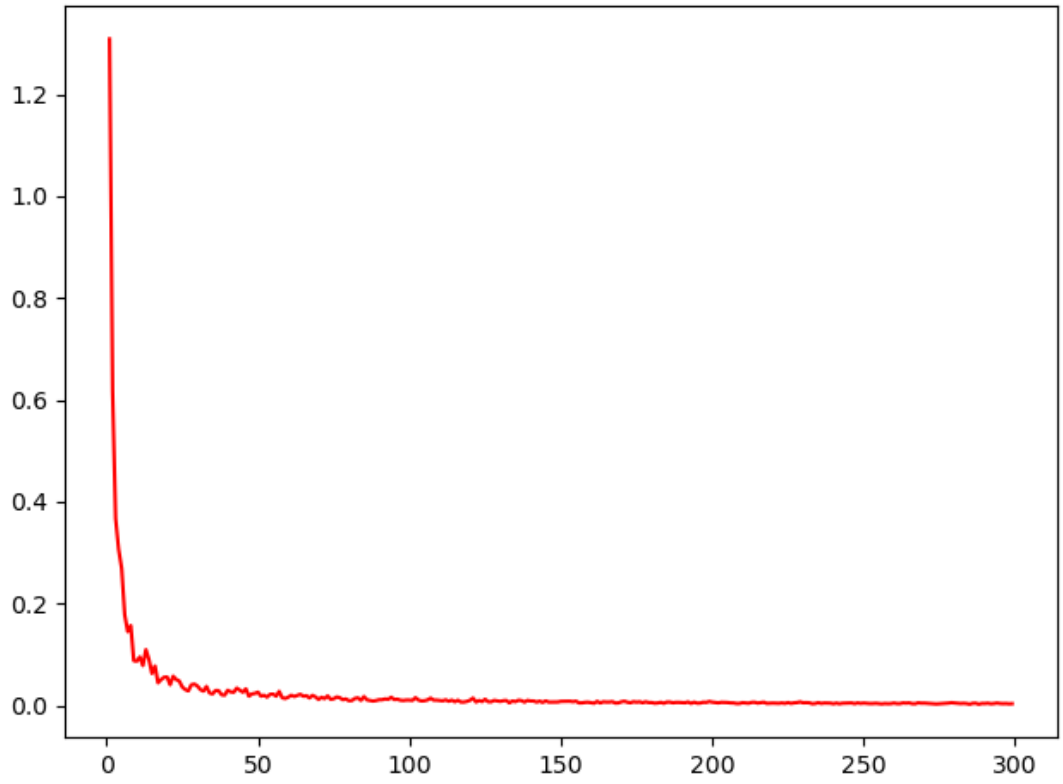

Figure 2.2: Convergence of Samping Shapley algorithm

values for massive datasets and complex models. Therefore, we would like to investigate a method for simplifying the sampling procedure in order to reduce computational complexity and boost efficiency. By enhancing the current sampling technique, we hope to accomplish a faster and more accurate estimation of feature contributions, while preserving its applicability across a variety of scenarios and applications.

2.3.1 Pearson correlation coefficient

The Pearson correlation coefficient examines the linear relationship between two continuous variables. This coefficient has been utilised in a number of med-

ical contexts[15]. In biomarker correlation analysis, for instance, it assesses the degree of correlation between various biomarkers, thereby assisting in the identification of potential risk factors or pathological mechanisms[8]. Gene expression correlation studies can also utilise the Pearson correlation coefficient to reveal co-expression patterns, functional similarity, and regulatory mechanisms among genes[17].

In our situation, we use the Pearson product-moment correlation coefficient to investigate the relationship between various features and to construct a graph depicting the relationships between all features. Equation 2.5 displays the formula for the Pearson product-moment correlation coefficient.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.5}$$

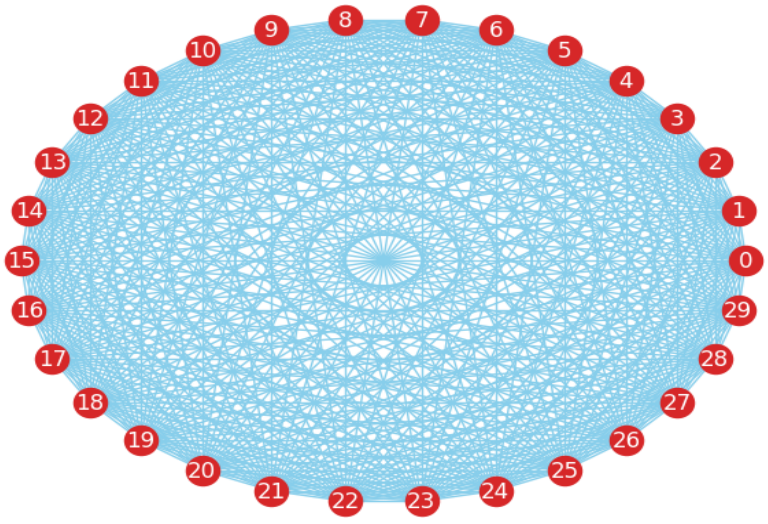

Figure 2.3: Original relationship map

The relationship graph between all features is depicted in figure 2.3. We use

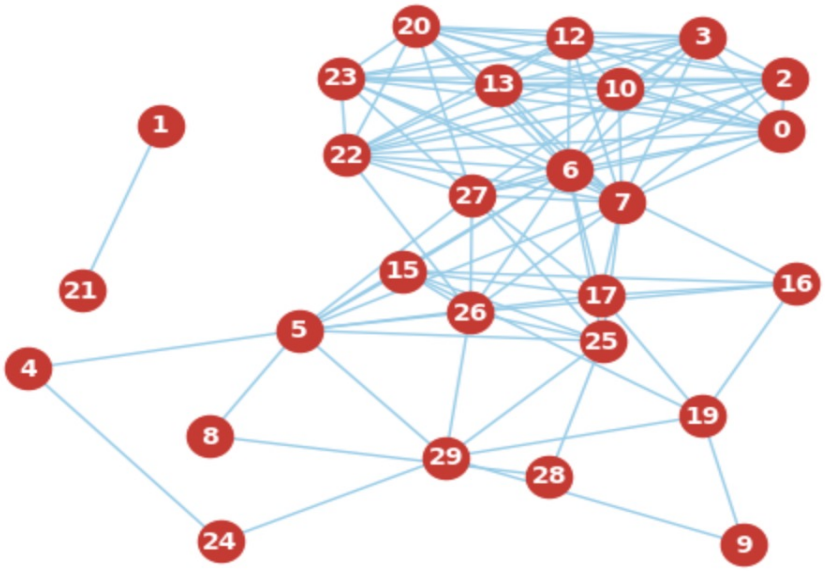

Figure 2.4: Relationship map with most importance link

equation 2.5 to compute the correlation coefficient between every pair of features, and a high correlation coefficient between two features indicates that they share a certain amount of redundant information and can be considered comparable. We argue that the stronger the correlation, the stronger the connection between features. Consequently, based on this concept and in an effort to reduce the number of connections, we propose filtering the weak connections by locating the q% quartiles, which can be altered based on efficiency requirements. As depicted in Figure 2.4, this method enables us to reduce the complexity of the graph while conserving all significant connections between features.

2.3.2 Biased Random Path Searching Method

In the algorithm for the sampling method used to calculate Shapley, the sampling object is the coalition formed by all participants, and during our exper-

iments we ran into the issue of high computational cost. Utilising relational graphs, we endeavoured to create a new method with a lower computational cost. This strategy reduces the number of coalitions processed while retaining the information that the participants in the coalition being processed are highly interconnected. We believe this effect will be realised by generating alliances in a relationship graph filtered by Pearson's correlation coefficient.

The biassed random path search method is a simple and effective heuristic search technique appropriate for complex mathematical model and graph structure issues. The method assigns probability values to neighbouring nodes based on some informative metric beginning at a starting node of the graph. These probabilities reflect the likelihood that the corresponding node will be selected as the next step. Then, based on the designated probability values, a neighbouring node is chosen at random and added to the list of visited nodes. This procedure is repeated until a predetermined termination condition is met. The following is the pseudocode for this algorithm:

Algorithm 2: Random Path Generation

Input: Number of sample paths N, random length L, adjacency matrix A

Output: N generated paths

for $i \leftarrow 1$ **to** N **do**

Select a random starting node $Node_1^{(i)}$

for $t \leftarrow 1$ **to** $L-1$ **do**

Compute the sum of weights for adjacent nodes:

$S_t^{(i)} = \sum_k A_{Node_t^{(i)},k}$

Normalize the weights in the adjacency matrix row:

$P_t^{(i)} = \frac{A_{Node_t^{(i)}}}{S_t^{(i)}}$

Choose the next node based on normalized weights:

$Node_{t+1}^{(i)} \sim P_t^{(i)}$

end

end

In our case, we used this method to help us find affiliates with a high level of interactive information. Our results are shown in the next section.

2.3.3 Convergence Measurements

We evaluate the convergence of the algorithm using Mean Absolute Error (MAE), show in equation (2.6), which is a measure of the average absolute deviation between the predicted and actual values. To ascertain whether the algorithm has converged, the MAE values of two consecutive iterations are compared. At each

iteration stage, we compare the MAE of the Shapley value estimate from the current iteration with the MAE from the previous iteration. By observing the variation between MAE values, we can determine the algorithm's efficacy at various iteration stages. The equation of MAE:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.6}$$

where,

- n represents the number of observations.
- y_i represents the true value of the i-th observation.
- $\hat{y}_i$ represents the predicted value of the i-th observation

As stated in Section 2.2, we do not analyse the information from the ranking changes, but rather only the numerical error in presuming the algorithm has stabilised. As a result, we also incorporated the Spearman rank Correlation Coefficient,show in equation (2.6), as a second metric to assist us in observing the ranking information changes at each iteration.The Spearman correlation coefficient is a commonly used nonparametric statistical method to measure the degree of association between two variables. In feature selection, the Spearman correlation coefficient can be used to assess the correlation between each feature and the target variable and accordingly eliminate features that are uncorrelated or redundant with the target variable[10]. The equation of Spearman Rank

Correlation Coefficient(2.7):

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.7}$$

- ρ represents the Spearman Rank Correlation Coefficient
- d_i represents the difference between the ranks of the paired data points for the i-th observation.
- n represents the number of observations (data pairs)
- $\sum d_i^2$ is the sum of the squared differences between the ranks of the paired data points

2.4 Summary of Sampling Method based on Weighted Graphs

We consolidate the previously mentioned technical details and outline the procedure for implementing the sampling method based on weighted graphs. The novel sampling technique has been meticulously designed to reduce computational complexity. Both the number of samples and the path length influence the convergence speed and accuracy of the algorithm throughout the process. This is the pseudocode for the sampling method based on weighted graphs:

Algorithm 3: Graph Based Sampling Approach for the Shapley value

Data: Data D, filtering parameter p, sample size M, sample length L_j, sample size N

Result: $\hat{Sh}_i$

Initialize $c \leftarrow 0$, $\hat{Sh}_i \leftarrow 0$

$a_{ij} \leftarrow \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$, where $a_{ij} \in A$.

for *all* $a \in A$, *if* $a < p$, *then set it to 0* **do**

end

while $c < M$ **do**

$l \leftarrow 0$

Randomly select a player $player_i$ and set length L, and add $player_i$ to the set O. **while** $l < L$ **do**

Randomly select the next player $player_j$ based on the probability distribution $P \sim \frac{A_{ij}}{\sum A_i}$, and add $player_j$ to the set O. Set $i = j$, $l \leftarrow l + 1$

end

for i *in* O **do**

Calculate $Pre^i(O)$. Calculate $x(O)_i = v(Pre^i(O) \setminus \{i\}) - v(Pre^i(O))$. Set $\hat{Sh}_i = x(O)_i$. Set $count_i \leftarrow count_i + 1$

end

Set $c \leftarrow c + 1$

end

Compute $\hat{Sh}_i = \hat{Sh}_i / counter_i$.

3 Results and Analysis

In this section, we'll show the findings from our experiments in two different ways. The Shapley values derived using the sampling approach based on relational graph information will be displayed first, followed by the corresponding analysis. To demonstrate the advancement we have made, we shall compare our approach to sampling with the original method used in the past.

3.1 Dataset description

The Breast Cancer Wisconsin (Diagnostic) dataset is accessible through the UCI Machine Learning Repository. It has 30 real attributes, one numeric attribute (id field), and one categorical attribute, which is a class label, with a dimension of 569 by 32. Since this is a two-class classification problem, also known as Binary Classification, there are two class values for diagnosis in this data set: M (Malignant) and B (Benign).

The following ten real-valued characteristics are computed for each cell nucleus:

- Radius (mean of distances from the centre to the perimeter's locations).
- Texture (standard deviation of grayscale values).
- Perimeter.
- Area.
- The uniformity (local variation in radius lengths).

- The surface.D compactness ($perimeter^2/area - 1.0$).
- Concavity (severity of the contour's concave portions).
- Concave elements (number of the contour's concave portions).
- Symmetry.
- Fractal dimension ("coastline approximation" - 1).

For each image, the mean, standard error, and "worst" or maximum (mean of the three largest values) of these features were calculated, yielding 30 features. Field 3 is the Mean Radius, field 13 is Radius SE, and field 23 is the Worst Radius.Four significant numerals are used to record every feature value.

This dataset contains no attributes with missing values. The distribution of the class is 357 benign and 212 malignant.

3.2 Simulation Results for Sampling method based on Graph

Based on the Pearson correlation coefficient introduced in 2.3.1, we first plotted a heat map of the thirty features of the dataset. The outcomes are depicted in figure 3.1.

Then, we filtered the weak connections between the features and set the parameter for filtering to 75% of the quartiles. This will help us to select coalitions with higher relevance. The results are depicted in figure 3.2:

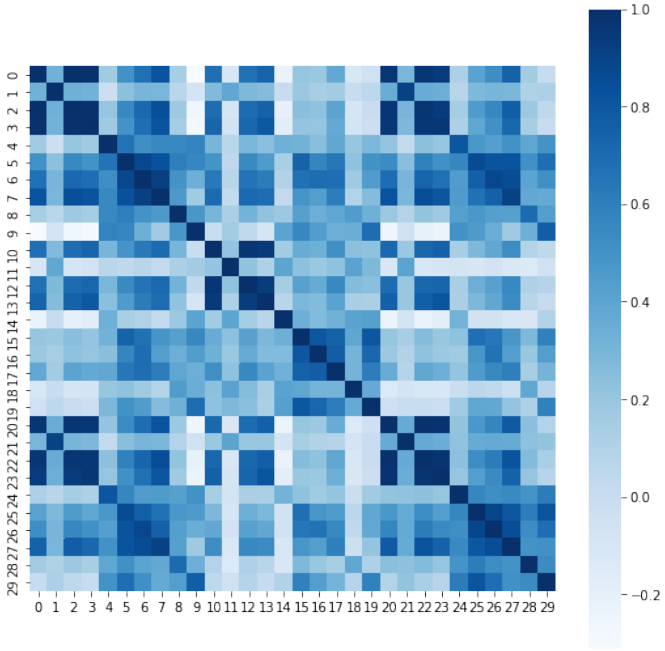

Figure 3.1: Interaction based on Correlation

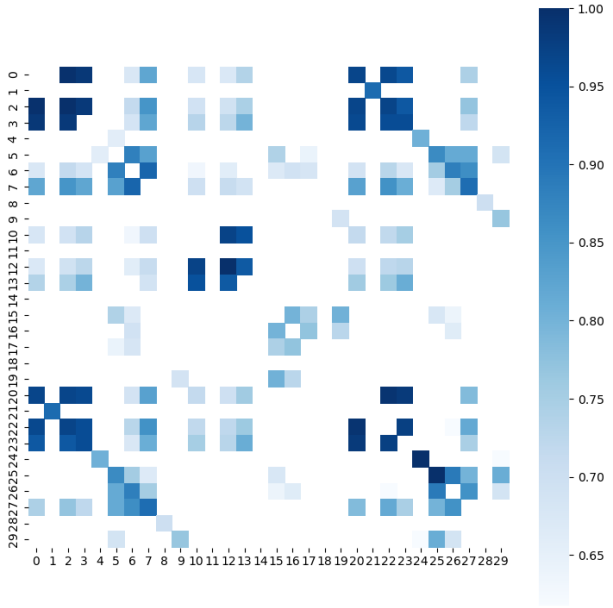

Figure 3.2: After Enhancing

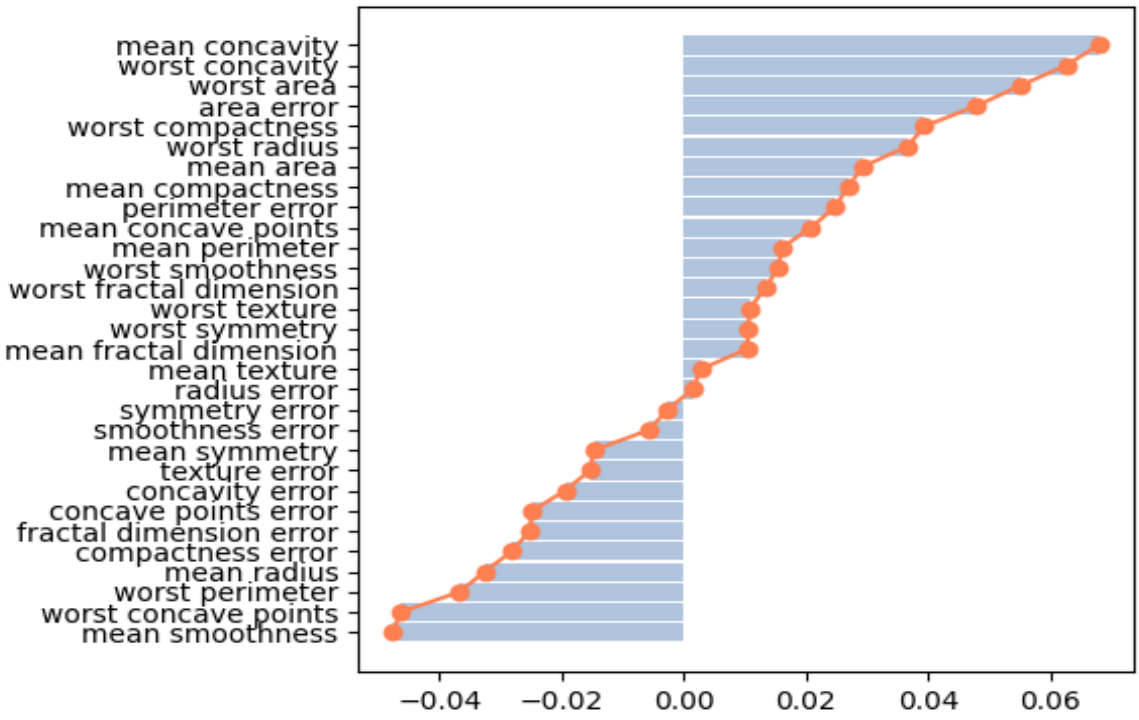

Figure 3.3: Shapley value of Graph sampling

Based on the experimental findings which shows on figure 3.3, this result concluded that four parameters in the breast cancer data-set mean concavity, worst concavity, worst area, and worst compactness were more significant than other features. Particularly, these characteristics helped better to discriminate between malignant and benign tumours.

The graph Shapley sampling method accomplishes a 40% reduction in time cost compared to other methods with the same number of iterations due to the random scale design. This enhancement and efficiency gain gives the graphical sampling technique a significant advantage in managing complex datasets, bringing convenience to related research and applications. Moreover, the graphical Shap-

ley sampling method not only reduces computational cost, but also enhances computational reliability and stability to a certain extent, allowing for more precise results to be obtained in a shorter amount of time. The next subsection will show the detailed results.

3.3 Comparison with Original sampling method

In our earlier work [31], we employed MAE as a metric to assess convergence, however we failed to take into account a disadvantage. MAE is small does not account for changes in ranking; it only indicates that the overall numerical error has been minor throughout the course of two iterations. In order to improve our trials, we included observations for the Spearman rank correlation coefficient.

Figure 3.4 and figure 3.5 demonstrate that the MAE stabilises after 500 iterations, while the Spearman correlation stabilises after approximately 2,000 iterations for Original sampling method. Random samples create a fluctuation near the 2,500 iteration; however, the rank and Shapley values for the original sampling method remain nearly the same between the 2,000 and 3,000 iterations.

The correlation between the two iterations is closer to 1, which means that the change in ranking between the two iterations is smaller, and at the same time, the overall mae of the data is almost constant (< 0.0001) , therefore, for the original sampling method, although the algorithm is more time consuming, it still achieves stable results.

The conditions are same for the graph sampling method. The MAE stabilises

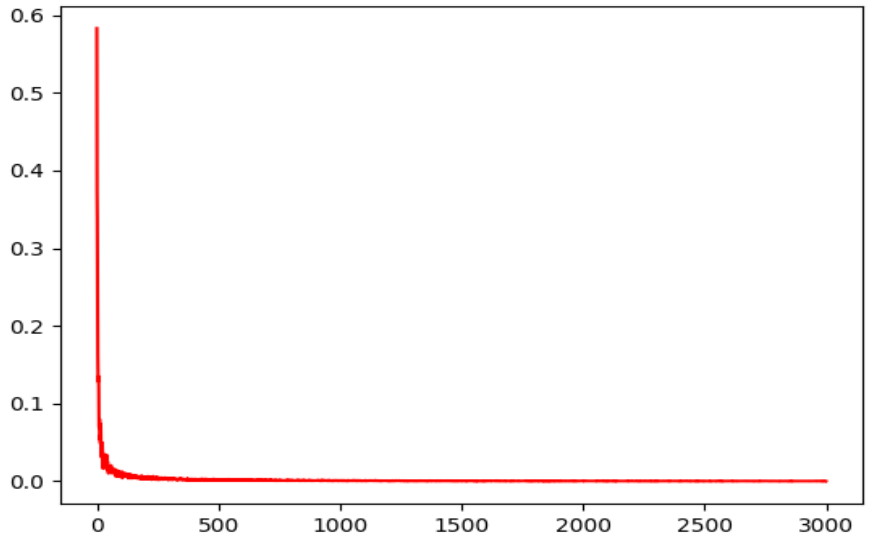

Figure 3.4: MAE for original sampling

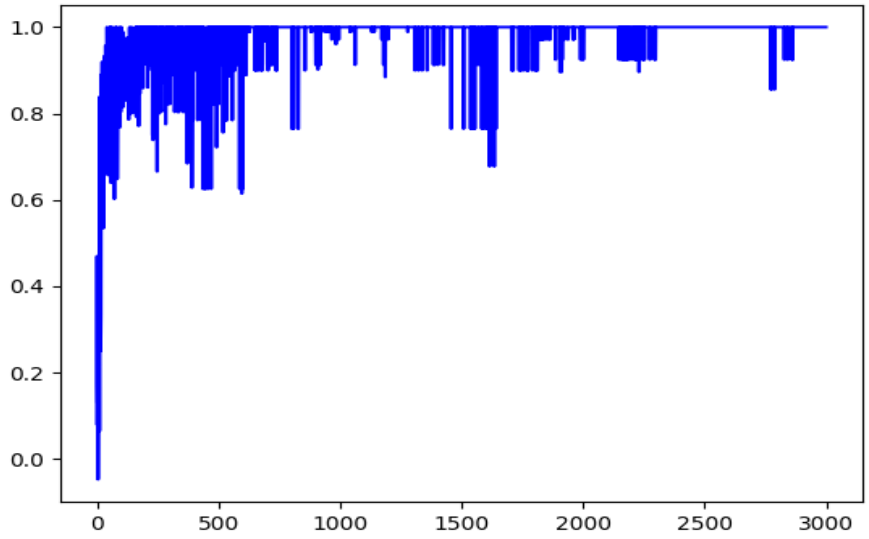

Figure 3.5: Spearman rank for original sampling

after 500 iterations, as seen in figures 3.6 and 3.7. When doing simulations, there is a certain amount of randomness in each random sample, which could cause variations in the results regarding the Spearman Rank Correlation Coefficient.

The size of the sample can also have an impact on how stable the results are. The results fluctuated after 2500 iterations. When the number of iterations reached 3,000, the results stabilised and no longer displayed significant changes as a result of further increasing the number of iterations.

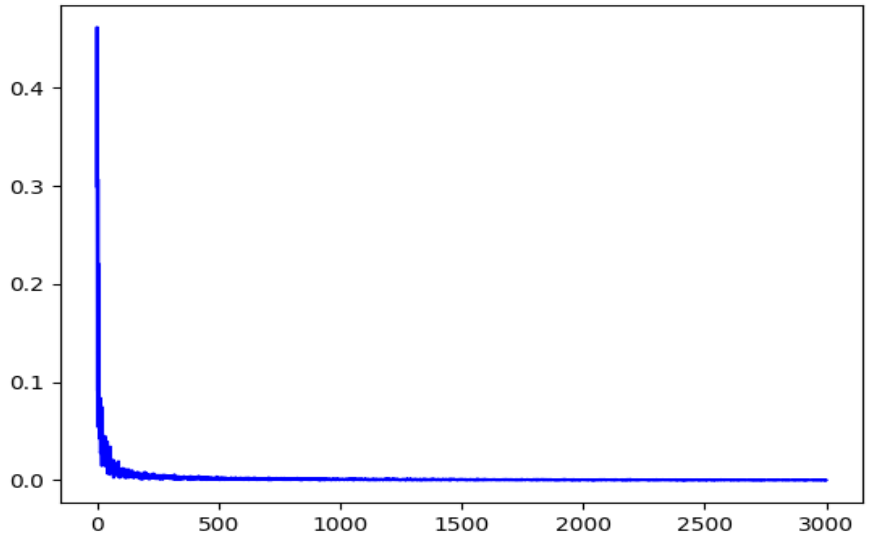

Figure 3.6: MAE for sampling based on graph

Figure 3.8 and figure 3.9 depict the efficacy of the graph-based Shapley sampling method in interpreting the results. Notably, although the Shapley values and feature rankings calculated by the two methods differ slightly, they both follow a similar pattern in identifying significant characteristics. By comparing the two graphs, we can see that 3 of the top 5, 8 of the top 10, and 3 of the bottom 5 features are shared by both approaches, with only minimal variations.

The table 3.3 gives details of the feature rankings given by the two methods. The features that change within two places in the new method's results are represented by the features in the grid highlighted in green in the table. It is

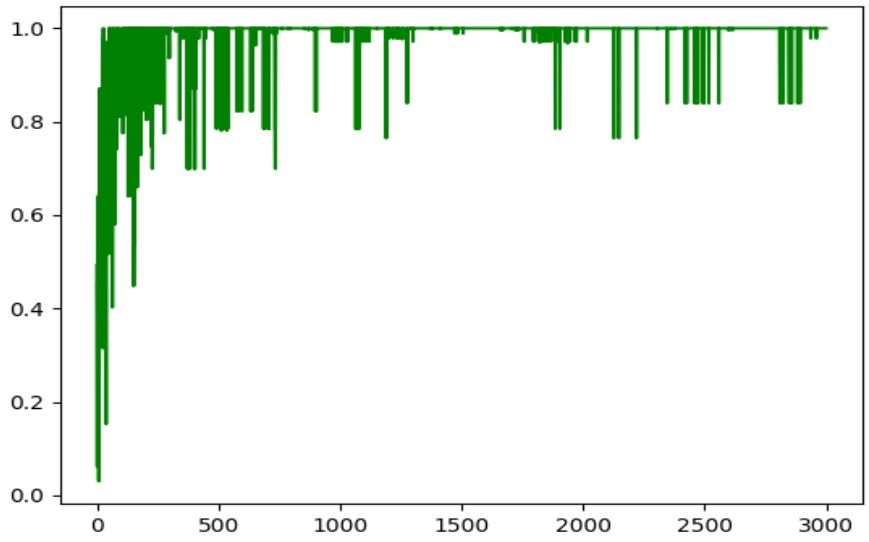

Figure 3.7: Spearman rank correlation coefficient for sampling based on graph

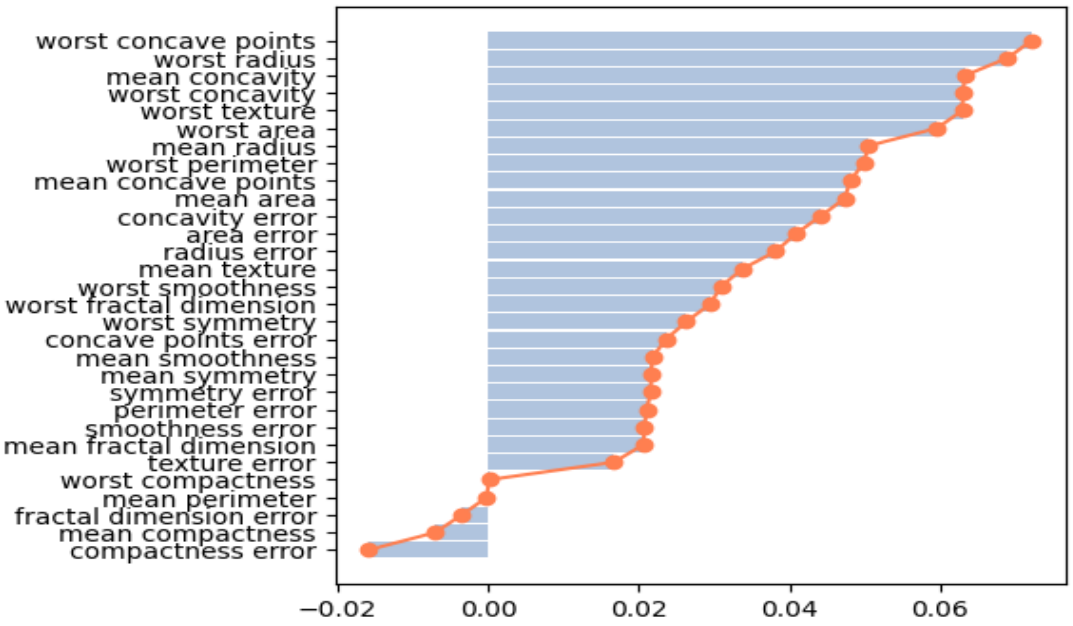

Figure 3.8: Shapley value for sampling method based on graph

evident that the results of the graph-based sampling approach are almost identical to those of the original sampling method for the most and least significant features (Top 10 and Bottom 10), and that the features with the largest changes

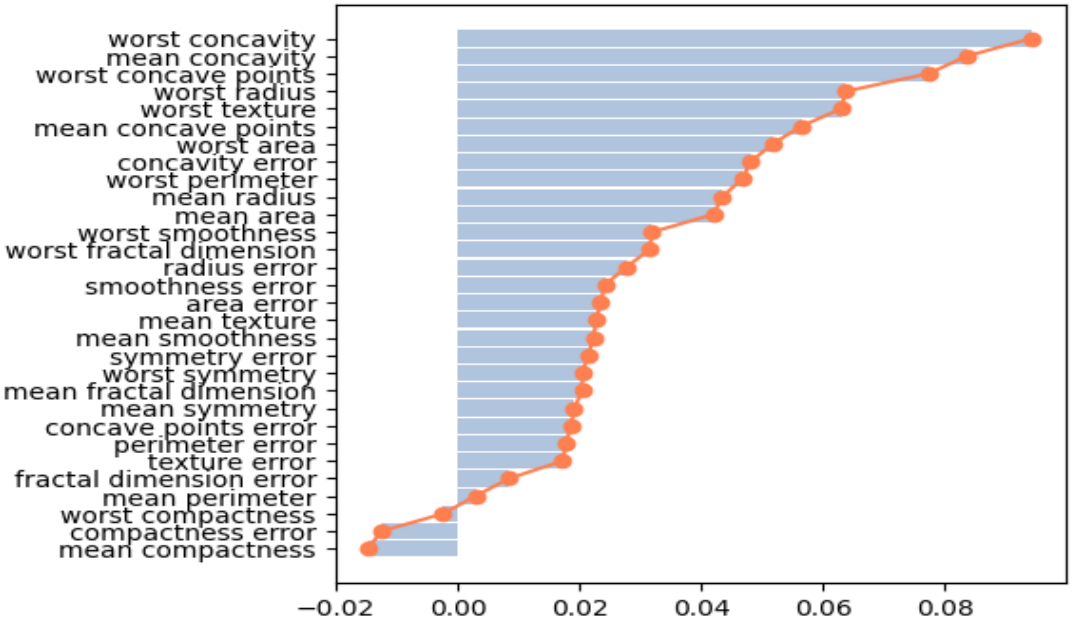

Figure 3.9: Shapley value for original sampling

in feature weight ranking are created in the middle ten features.

RANK	Original sampling	Sampling based on graph
1	worst concavity	worst concave points
2	mean concavity	worst radius
3	worst concave points	mean concavity
4	worst radius	worst concavity
5	worst texture	worst texture
6	mean concave points	worst area
7	worst area	mean radius
8	concavity error	worst perimeter
9	worst perimeter	mean concave points
10	mean radius	mean area
11	mean area	concavity error
12	worst smoothness	area error
13	worst fractal dimension	radius error
14	radius error	mean texture
15	smoothness error	worst smoothness
16	area error	worst fractal dimension
17	mean texture	worst symmetry
18	mean smoothness	concave points error
19	symmetry error	mean smoothness
20	worst symmetry	mean symmetry
21	mean fractal dimension	symmetry error
22	mean symmetry	perimeter error
23	concave points error	smoothness error
24	perimeter error	mean fractal dimension
25	texture error	texture error
26	fractal dimension error	worst compactness
27	mean perimeter	mean perimeter
28	worst compactness	fractal dimension error
29	compactness error	mean compactness
30	mean compactness	compactness error

In addition, we discovered that the Shapley Value determined by the graph-based sampling approach was, to some extent, less than that of the original sampling method. These issues might arise because, in the previous operation, we eliminated the potential of coalition building among weakly related characteristics, which resulted in the loss of some information and such experimental findings.

3.4 Validation of Interpretation Results

Based on the results in the previous subsection, we decided to validate the feature importance. We can see that the three features worst concave points , worst radius, mean concavity are among the features that contribute more to the detection of malignant tumors in both methods. Therefore, we decided to remove these features and retrain the model. At the same time, we also remove the three features that contribute least to this work and retrain them as controls. The following table is the results of the retraining:

Table 3.1: Evalution of ANN model

Features	Accuacy	Precision	Recall
Without Top 3	87.32%	[91.67%,92.53%]	[93.42%,90.94%]
Without Bottom 3	96.31%	[98.03%,92.31%]	[94.25%,98.19%]

We found that when the three features with the largest contributions were removed, the evaluation metrics of the model decreased significantly using the same training process. However, when the three features with the smallest contribution are removed, the metrics of the model are basically unchanged using the same training process.

4 Conclusion and Future works

4.1 Conclusion

This finding indicates that while the original Shapley sampling method and the graph-based Shapley sampling method may differ in terms of specific values, they have a high degree of consistency in identifying essential characteristics. This consistency gives researchers and practitioners greater assurance that similar results can be obtained when analysing data using both methods. This is additional evidence that the graph-based Shapley sampling method can be an effective and efficient alternative for feature selection, model interpretation, and data analysis.

In this research, we try to find a solution to the high-dimensional complicated artificial intelligence systems feature importance assessment challenge.Although the Shapley Sampling algorithm is a frequently used evaluation technique, it has a high processing complexity and is challenging to utilise with high-dimensional datasets. In order to evaluate feature importance, we provide a biassed graph-based Shapley Sampling approach.

In contrast to the original Shapley Sampling technique, our approach creates a relationship graph using Pearson product moment correlation coefficients, and then employs a biassed random path search method to create coalitions as samples. This speeds up computations without sacrificing precision. In particular, our strategy enhances the quality of the sampled coalitions and balances com-

putational complexity and accuracy by regulating the size of the coalitions.

The experimental findings demonstrate that, despite modest variations in feature ranks and Shapley values, the original technique and our suggested method both consistently score positive and negative contributions on a scale of one to ten. Additionally, for the same number of iterations, the computation time of the new approach is 40% faster than the old algorithm.

Additionally, we added the capability to measure outcomes by combining the MAE and Spearman index correlation. These techniques enable us to thoroughly assess the Shapley Sampling method's effectiveness.

Challenges and Future Work

1. Future research must look more closely at how the interpretation results can be used in practical situations. Incorporate with various real business requirements to strengthen or develop the model depending on the findings of the interpretation. Investigate ways to improve the model's ability to generalise depending on the results of the interpretation.

2. Investigate the explanation scheme's metrics to confirm or validate the explanation scheme's correctness. The industry currently lacks a statistic that can accurately gauge the outcomes of interpretation.

3. It is necessary to extend the application of our findings to additional machine learning models, such as reinforcement learning and unsupervised learning. Other categories of machine learning models should take Shapley value-based explanatory approaches into consideration.

4.2 Acknowledgment

At the conclusion of this thesis, I would like to express my gratitude to those who have supported and inspired me throughout this endeavour.

Professor Pertrosyan Ovanes, my supervisor, is the first person I desire to thank. Throughout the course of my research, he has assisted and guided me with patience and consideration. Not only did he provide me with valuable advice and direction, but he also improved my understanding of the pertinent knowledge and skills. His knowledge and experience have inspired my future studies and research and had a profound impact on me.

Second, I'd like to thank my mentor, Zou Jinying, whose guidance and direction in academic research, as well as his care and assistance in daily life, have made me feel like a member of a large family. With his assistance, I have gained many research skills and had many experiences I never anticipated I would have.

In addition, I would like to thank my family. I've always been determined to advance because of their consistent encouragement and support. Their selfless care and unwavering support give me the strength to move forward. I would like to express my heartfelt appreciation to my girlfriend, who has been an unwavering source of support and encouragement throughout my academic journey. Her selfless love and understanding have helped me overcome writing slumps and navigate the challenges of pursuing higher education. Without her constant support, I would not have had the motivation to keep moving forward. I am grateful for her presence in my life and will always cherish her unwavering

support.

Finally, I would like to express my gratitude to everyone who assisted and supported me during my research for this paper.

References

[1] David Bau et al. "Gan dissection: Visualizing and understanding generative adversarial networks". In: *arXiv preprint arXiv:1811.10597* (2018).

[2] Javier Castro, Daniel Gómez, and Juan Tejada. "Polynomial calculation of the Shapley value based on sampling". In: *Computers & Operations Research* 36.5 (2009), pp. 1726–1730.

[3] Raghavendra Chalapathy and Sanjay Chawla. "Deep learning for anomaly detection: A survey". In: *arXiv preprint arXiv:1901.03407* (2019).

[4] Robert Challen et al. "Artificial intelligence, bias and clinical safety". In: *BMJ Quality & Safety* 28.3 (2019), pp. 231–237.

[5] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[6] Leilani H Gilpin et al. "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[7] Varun Gulshan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22 (2016), pp. 2402–2410.

[8] Anthony J.G. Hanley et al. "Prediction of Type 2 Diabetes Using Simple Measures of Insulin Resistance: Combined Results From the San Antonio Heart Study, the Mexico City Diabetes Study, and the Insulin Resistance Atherosclerosis Study ". In: *Diabetes* 52.2 (Feb. 2003), pp. 463–469. ISSN: 0012-1797. DOI: 10.2337/diabetes.52.2.463.

[9] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[10] Ying Liu. "A comparative study on feature selection methods for drug discovery". In: *Journal of chemical information and computer sciences* 44.5 (2004), pp. 1823–1828.

[11] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[12] Scott M Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[13] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[14] Menaka Narayanan et al. "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation". In: *arXiv preprint arXiv:1802.00682* (2018).

[15] Michael C Oldham, Steve Horvath, and Daniel H Geschwind. "Conservation and evolution of gene coexpression networks in human and chimpanzee brains". In: *Proceedings of the National Academy of Sciences* 103.47 (2006), pp. 17973–17978.

[16] Animesh Patcha and Jung-Min Park. "An overview of anomaly detection techniques: Existing solutions and latest technological trends". In: *Computer networks* 51.12 (2007), pp. 3448–3470.

[17] Karl Pearson. "VII. Note on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58.347-352 (Jan. 1895). Publisher: Royal Society, pp. 240–242. DOI: 10.1098/rspl.1895.0041. URL: https://doi.org/10.1098/rspl.1895.0041 (visited on 06/21/2020).

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you? Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[19] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

[20] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[21] Lukas Ruff et al. “Deep one-class classification”. In: *International conference on machine learning*. PMLR. 2018, pp. 4393–4402.

[22] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[23] Gouda I Salama, M Abdelhalim, and Magdy Abd-elghany Zeid. “Breast cancer diagnosis on three different datasets using multi-classifiers”. In: *Breast Cancer (WDBC)* 32.569 (2012), p. 2.

[24] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[25] Lloyd S Shapley et al. “A value for n-person games”. In: (1953).

[26] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning important features through propagating activation differences”. In: *International conference on machine learning*. PMLR. 2017, pp. 3145–3153.

[27] Hongtao Xie et al. “Automated pulmonary nodule detection in CT images using deep convolutional neural networks”. In: *Pattern Recognition* 85 (2019), pp. 109–119.

[28] Hao Zhang et al. “Interpreting multivariate shapley interactions in dnns”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 10877–10886.

[29] Yuyi Zhang et al. “XAI evaluation: evaluating black-box model explanations for prediction”. In: *2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT)*. IEEE. 2021, pp. 13–16.

[30] Bo Zong et al. “Deep autoencoding gaussian mixture model for unsupervised anomaly detection”. In: *International conference on learning representations*. 2018.

[31] J Zou et al. “High-dimensional explainable AI for cancer detection”. In: *International Journal of Artificial Intelligence* 19.2 (2021), p. 195.

[32] Jinying Zou, Feiran Xu, and Ovanes Petrosian. “Explainable AI: using Shapley value to explain the anomaly detection system based on machine learning approaches”. In: *Control Processes and Stability* 7.1 (2020), pp. 355–360.