Cover Page

Student Name: Zhe Zheng

School: Princeton International School of Mathematics and Science

City/State: New Jersey, USA

Country: United States of America

Advisor: Dr. RuoXiu Xiao, Associate Professor, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Title of Project: Artificial Intelligence Diagnostic Approach of Infertility in Chinese Traditional Medicine

Title: Artificial Intelligence Diagnostic Approach of Infertility in Chinese Traditional Medicine

Abstract

Introduction: Infertility is a kind of reproductive disorder that frequently occurs among young couples. It is the failure to pregnant after one year of sexual behaviors. As it gradually drew more public attention, it was addressed infertility as nonnegligible since it induces severe social conflicts, like family violence, divorce, social stigma, emotional stress, depression, anxiety, and low self-esteem. According to Traditional Chinese Medicine (TCM) theory, infertility is mainly attributed to six inducements: Liver Qi Stagnation, Stagnation of Uterine Cell, Kidney Yang Deficiency, Kidney Yin Deficiency, Kidney Qi Deficiency, Internal Obstruction of Phlegm, and Dampness. From the TCM perspective, the disease can be diagnosed by observing a patient's body features like skin spots and feces. The effectiveness of TCM and systematic theoretical knowledge has been recognized widely, yet due to its complexity and massive work of diagnosis implying the underlying conceptual foundations, it is very challenging to diagnose disease manually to consider and weigh multiple factors simultaneously. Therefore, our research took advantage of Artificial Intelligence (AI) in medical use that could help us accurately dealing massive numerical data.

Objective: This topic focused on finding the relationship between syndrome and syndrome type of infertility and extracts features from many symptom data to predict syndrome type via an artificial intelligence approach. Therefore, we construct a model that most effectively diagnoses a patient's syndrome type of infertility using body features. The model should have practical application value and be potentially used in a clinical trial.

Results: This paper used six classifiers-Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor, Support Vector Machine, Artificial Neural Network, and Random Forest. In the beginning, the accuracy of the K-nearest neighbor was only 0.66, and the classification accuracy of the other five models was around 0.8. Subsequently, this paper adopts the methods of hyperparameter tuning and feature selection to optimize the model, and the classification accuracy of the artificial neural network model and random forest model is improved to about 0.85. Finally, we used SMOTE algorithm for data enhancement; the data set capacity was expanded to two times and five times the original. The classification accuracy was also increased by 16.6% and 22%, respectively, based on the original.

Conclusions: The models have practical intelligence for diagnosing a patient's syndrome type of infertility using body features and have received acknowledgment from Dr. Qin, an obstetrician and gynecologist of Beijing Boai Hospital, China Rehabilitation Research Center, from which we get data sources.

Keywords:

Artificial Intelligence, Infertility, Chinese Traditional Medicine, K Nearest Neighbors, Artificial Neural Network, Random Forest.

Declaration:

I state that the submitted manuscript was the research work and the research results obtained under the guidance of my supervisor, Dr. RuoXiu Xiao, Associate Professor at the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. As far as I am aware, except for the references, the manuscript does not contain research results that others have published or written. If there is anything wrong, I am willing to bear all relevant responsibilities.

*Please note that this article is based on the existing scientific basis of Traditional Chinese Medicine diagnosis; infertility diagnosis has recognized accuracy in the medical field. In this study, models were trained based on the above-mentioned diagnostic methods, and finally, each model achieved an accuracy rate of about 85%.

Signature:

Date: August 2022

Table of Contents

I. INTRODUCTION:

Infertility is a kind of reproductive disorder that frequently occurs among young couples. It is the failure to pregnant after one year of sexual behaviors. [1]. According to World Health Organization (WHO) data, forty-eight million couples and 186 million people worldwide will be infertile by 2022 [1]. Addressing infertility is essential because it can induce severe social conflicts, such as family violence, divorce, social stigma, emotional stress, depression, anxiety, and low self-esteem [2].

According to the theory of Traditional Chinese Medicine (TCM), infertility is because of the disruption of life balance and the obstruction of blood circulation [3]. Specifically, it attributes to six inducements: Liver Qi Stagnation, Stagnation of Uterine Cell, Kidney Yang Deficiency, Kidney Yin Deficiency, Kidney Qi Deficiency, Internal Obstruction of Phlegm, and Dampness [4]. Women with Liver Qi stagnation often have an imbalance at ovulation (bloating, irritability, breast tenderness) and menstruation (premenstrual breast tenderness, irritability, anger, painful periods). People with Kidney Qi deficiency typically have low energy and cravings for sugar or bread. People with Kidney Yang deficiency experience symptoms of coldness: cold feet or cold hands or intolerance to cold. People with Kidney Yin deficiency may experience night sweats and hot flashes.

Unlike the Western diagnosis system's convention, of looking for well-defined, well-tested causes to explain a disease state, from the TCM perspective, factors that determine people's health are specific to individuals. Therefore, Traditional Chinese Medicine (TCM) diagnoses disease by observing a patient's body feature, some potential diseases have to correlate with skin expression and temperature [5]. In diagnosing the root causation of infertility, the doctor generalizes, sorts out, and excludes the 52 features of the patient's body and confirms a specific cause of infertility [6]. The effectiveness of TCM and systematic theoretical knowledge has been recognized widely, yet due to its complexity and massive diagnosis means underlying conceptual foundations, it is very challenging to diagnose disease manually considering and weighing multiple factors simultaneously.

As modern medicine developed, countless examples of Artificial Intelligence (AI) in medical uses emerged since AI techniques satisfy the requirements of accurately dealing massive numerical data. Currently, AI commonly appears in ultrasound, magnetic resonance imaging, mammography, genomics, and computed tomography [7]. Besides, numerous precedent

examples of the combination of Chinese medicine and artificial intelligence inspire us and give us confidence in artificial intelligence in the TCM area and assist doctors in judging the cause of infertility. Another advantage of AI is the continuous learning by processing a large amount of data [8][9][10].

This paper used the mainstream algorithm of machine learning to establish the classification and prediction model of sterility syndrome and the type. The advantages and disadvantages of different algorithm models were analyzed and compared, and a series of methods, such as feature selection, were used to optimize the model. The aim is to improve the objectivity of syndrome-type judgment and provide a reference for further research and development of TCM intelligent diagnosis and treatment decision support system.

II. MATERIALS

2.1 Data Sources

We collected 600 valid cases from female infertility case records in the Gynecological Clinic of Traditional Chinese Medicine in China Rehabilitation Research Center Beijing Boai Hospital from October 2018 to March 2022 in the form of Excel.

The professional attending physicians and chief physicians divided the 600 cases into six major categories: Liver Qi Stagnation, Stagnation of Uterine Cell, Kidney Yang Deficiency, Kidney Yin Deficiency, Kidney Qi Deficiency, and Internal Obstruction of Phlegm and Dampness.

Among the 600 cases, patients were 25-45 years old, and 100 cases in each of the above six syndromes. There are 52 kinds of syndrome information, later turned into numerical variables, where 1 represents the occurrence of such symptoms and 0 represents the absence of such symptoms. Those syndromes include descriptions of bodily form, perspiration amount, and feces.

2.2 Data pre-processing

The quality of data sets ensures the upper limit of model and algorithm performance. Therefore, data preprocessing is a crucial step in constructing a perfect model. Data pre-processing usually includes missing value processing, classified data processing, feature selection, and extraction.

We first inspected the 600 cases with 52 kinds of features. We found out there were five samples has information lost-No.99(Kidney Qi Deficiency), No.299(Kidney Yin Deficiency), No.399(Stagnation of Liver Qi), and No. 499(Stagnation of Uterine Cell), and No.516(Internal Obstruction of Phlegm and dampness). Another circumstance is that some outliers might occur, such as numbers over 1 existing in the form (only 0 or 1 supposed to appear in the data set); hopefully, no outliers have been sifted in the 600 samples. Usually, the above data are invalid and should be discarded strategically. However, since the original data set of this study was not significant, we adopted mean interpolation to replace the missing data with the average of the whole feature column to autofill the missing data.

III. METHODS

3.1 Methods overview

Machine Learning is solving a real problem by collecting a dataset and selecting an appropriate algorithm to train a statistical model based on that dataset. Generally, *Machine learning* can be divided into supervised, unsupervised, and reinforcement learning [11][12]. Supervised learning uses the dataset to generate a model that takes an n-dimensional feature vector as X input and can infer the feature vector's label as Y output. For example, people's job title, salary, credit status, and other information can be used as input to determine whether a bank will approve a loan. Supervised learning mainly includes two applications: classification and regression. Essentially, their purpose is to make predictions on unknown data. Classification is a prediction problem from the feature vectors of samples to class markers, and the prediction results can be two or more classes. In this paper, six classical algorithms in supervised learning are selected to build classification models respectively. The algorithms and experimental results of these six classifiers will be introduced in the following sections.

a. ***Logistic Regression (LR)*:** Despite the word "Regression" included in its name, Logistic Regression is essentially a linear model designed to handle classification tasks. In the logistic regression classifier, regularization is the default. One could also use L1 regularization, L2 regularization, and Elastic-Net regularization to improve numerical stability. Logistic regression is a generalization of linear regression, so it was initially used to solve binary classification problems [13].

However, logistic regression can be naturally extended to the field of multiple classifications.

In the mathematical formula, the linear approach could be described by:

Original linear model: $y = b_0 + b_1 x$

Sigmond Function: $p = sigmoid(y) = \frac{1}{1+e^{-y}}$

Logistic Function: $\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$

P, in this case, represents a possibility ranging from 0 to 1 if the outcome is 0 in this attribute and vice versa.

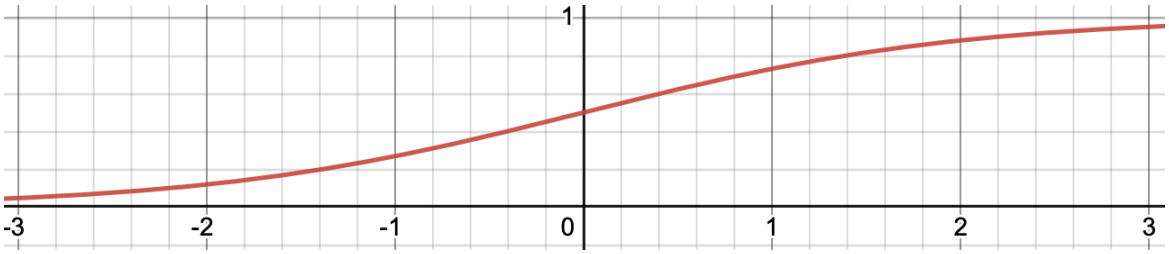

Figure 1 Logistic Regression mathematical model

b. ***Linear Discriminant Analysis (LDA)***: This is a classifier with Linear decision boundaries, generated by fitting class conditional densities to the data and using Bayesian rules. The model assumes that all classes have the same covariance matrix, and the Gaussian density is fitted to each class. In the case of binary classification, the principle of linear discriminant analysis classifier is to project all samples onto a straight line so that the projection points of similar samples are as close as possible, and the projection points of different samples are as far apart as possible. The judgment is based on the fact that the covariance of the former is less than a specific value, and the distance between the class centers of the latter is more significant than a specific value. Linear discriminant analysis can also use a transformation method to project the input to the direction with the absolute power to achieve the effect of dimension reduction [14].

In the figure below, we can see that the two curves of LD1 and LD2 and several points respectively in their projection; When comparing the projection of LD1 and LD2, LD1 projection distance between the two classes is more significant, and in LD2 projection, spacing is minimal, almost overlap. We can also find that in the

projection on the LD1 density, relatively LD2 is more extensive; that maximizes the *Between-Class Variance* and the *Within-Class Variance*.

We will call an optimal vector W (d-dimensional), so the projection of x (d-dimensional) onto W can be calculated using the following equation:

$$\mu_i = \frac{1}{N_i} \sum_{x \epsilon \omega_i} x$$

Projecting from X to W, the mean value of sample points is:

$$\widetilde{\mu_i} = \frac{1}{N_i} \sum_{y \epsilon \omega_i} y = \frac{1}{N_i} \sum_{y \epsilon \omega_i} w^T x = w^T \mu_i$$

To maximize the Between-Class Variance, we need to maximize J(w):

$$J(w) = |\widetilde{\mu_1} - \widetilde{\mu_2}| = |w^T(\mu_1 - \mu_2)|$$

Meanwhile, we need to satisfy the second condition-minimizing the Within-class variance:

We then introduce the definition *Scatter* (The sum of the variances between the projected values and the center):

$$\widetilde{s_i}^2 = \sum_{y \epsilon \omega_i} (y - \widetilde{\mu_i})^2$$

Then we derive:

$$J(w) = \frac{|\widetilde{\mu_1} - \widetilde{\mu_2}|}{\widetilde{s_1}^2 + \widetilde{s_2}^2}$$

Which we tried best to maximize the value of J(w).

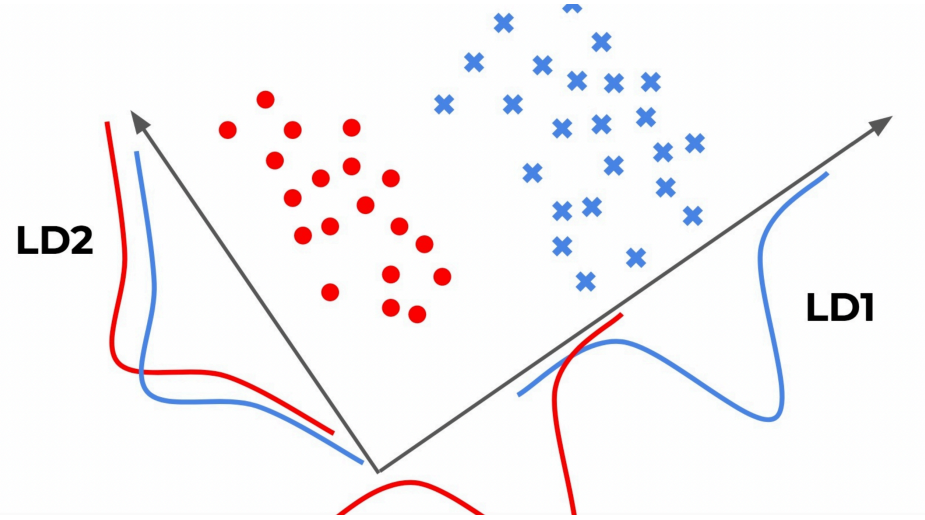

Figure 2 Linear Discriminant Analysis model

c. ***K Nearest Neighbors (KNN)***: *K Nearest Neighbors* is a simple and practical algorithm based on memory. *KNN* accords to the K adjacent samples of each query sample and conducts voting to sift the class with the highest proportion in the adjacent samples category. The class with the most representative ability is classified as the query sample classification [15]. The core of this algorithm is the calculation of variable distance and selection of the K value. In general, a considerable K value will suppress the influence of noise value; however, it will make the classification boundary does not clear enough.

As mentioned, K Nearest Neighbors contains three elements: the choice of K value, distance measurement, and the classification decision rules.

Let the eigenspace χ be an n-dimensional space of real vector numbers R^n,$x_i, x_j \epsilon\ \chi$

$$x_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, \dots, x_i^{(n)})^T, x_j = (x_j^{(1)}, x_j^{(2)}, x_j^{(3)}, \dots, x_j^{(n)})^T$$

The distance $L_p(x_i, x_j) = (\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^p)^{\frac{1}{p}}$ $(p \geq 1)$

When p = 1, $L_1(x_i, x_j)$ is called Manhattan distance:

$$L_1(x_i, x_j) = \sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|$$

When p = 2, $L_2(x_i, x_j)$ is called Euclidean distance:

$$L_2(x_i, x_j) = (\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^2)^{\frac{1}{2}}$$

d. ***Support Vector Machine (SVM)***: This algorithm has been widely used in character recognition, face recognition, pedestrian detection, text classification, and other aspects. SVM has been widely used in psychiatry [16]. Generally, a support vector machine deals with binary classification problems. However, multi-class classification problems like this paper are broken down into several binary classification sub-problems to solve. For the linearly separable feature space, the decision boundary of SVM is the hyperplane with the maximum margin for the learning samples. For nonlinear classification, the original feature space is transformed into a high-dimensional feature space by nonlinear mapping, and then the optimal classification hyperplane is solved by the kernel function.

Define a hyperplane H:

$$g(x) = w^T(x) + b = 0$$

Any vector x can be denoted by:

$$x = x_p + r\frac{w}{||w||}$$

In which

$x_{p:}$ The projection vector of x onto H

r: The vertical distance from x to H

$\frac{w}{||w||}$: Unit vector in the w direction

Then

$$g(x) = w^T\left(x_p + r\frac{w}{||w||}\right) + b = w^T x_p + b + r\frac{w^T w}{||w||} = r||w||$$

$$\therefore r = \frac{g(x)}{||w||} = \frac{|w^T x + b|}{||w||}$$

Assuming the classification is valid, the following inequation

$$|w^T x + b| \geq +1, if\ y_i = +1$$

$$|w^T x + b| \leq -1, if\ y_i = -1$$

refers to the sample points closest to the hyperplane such that the equality sign holds are called *support vectors*.

The following is basic model of Support Vector Machine:

$$max_{w,b}\ \frac{1}{2}||w||^2$$

$$s.t.\, y_i(w^T x_i + b) \geq 1$$

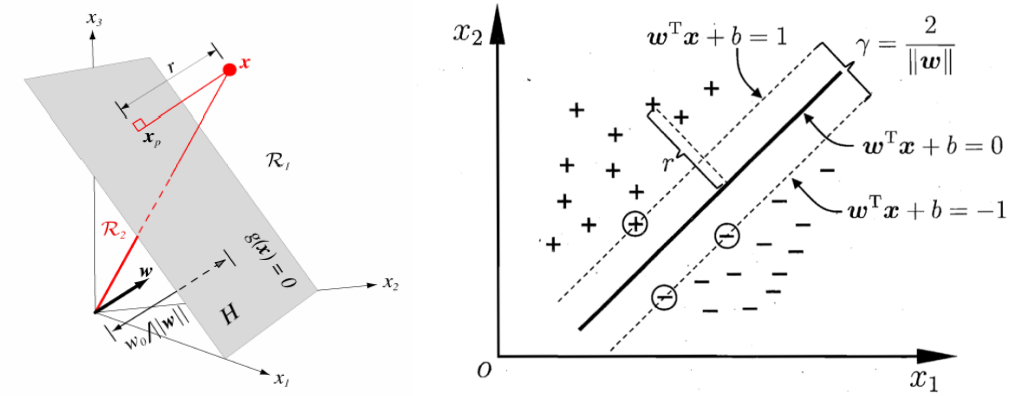

Figure 3 Support Vector Machine concept guide

e. ***Artificial Neural Network (ANN)***: Multilayer Perceptron (MLP) has an input layer, an output layer, and a hidden layer or as a whole [17]. Artificial neural networks simulate the human brain, neurons, synapses, and other substances. Since partial derivatives of the loss function concerning the model parameters are calculated at each time to update the model parameters, the multilayer sensing classifier is trained iteratively. It can also add a regularization term to the loss function, which is used to shrink model parameters to prevent overfitting. Multilayer perceptrons are good at learning nonlinear models and can learn in real-time.

Suppose the number of layers of the neural network is K (K > 1); The number of nodes (excluding bias nodes) at each layer from the input layer to the output layer is, respectively, $m_0,\ m_1,\ m_2,\ m_3,\ m_4, \dots, m_K$.

The dimension of the input vector is thus defined as m_0, the dimension of the output vector is m_K. The output vectors of each layer of the network are expressed as follows:

$$\textbf{Input Layer}: Y^{(0)} = \left[Y_1^{(0)}, Y_2^{(0)}, Y_3^{(0)}, \dots, Y_{m_0}^{(0)}\right]^T$$

$$\textbf{Hidden Layer 1}: Y^{(1)} = [Y_1^{(1)}, Y_2^{(1)}, Y_3^{(1)}, \dots, Y_{m_1}^{(1)}]^T$$

$$\textbf{Hidden Layer 2}: Y^{(2)} = [Y_1^{(2)}, Y_2^{(2)}, Y_3^{(2)}, \dots, Y_{m_2}^{(2)}]^T$$

$$\dots$$

$$\textbf{Output Layer}: Y^{(K)} = [Y_1^{(K)}, Y_2^{(K)}, Y_3^{(K)}, \dots, Y_{m_K}^{(K)}]^T$$

Next, we define the weight matrix and bias vector for each layer:

$$\boldsymbol{W}^{(1)} \in \mathbb{R}^{m_1 \times m_0} \qquad \boldsymbol{b}^{(1)} \in \mathbb{R}^{m_1 \times 1}$$

$$\boldsymbol{W}^{(2)} \in \mathbb{R}^{m_2 \times m_1} \qquad \boldsymbol{b}^{(2)} \in \mathbb{R}^{m_2 \times 1}$$

$$\dots$$

$$\boldsymbol{W}^{(K)} \in \mathbb{R}^{m_K \times m_{K-1}} \qquad \boldsymbol{b}^{(K)} \in \mathbb{R}^{m_K \times 1}$$

Define the activation function of each layer as $f^{(1)}$, $f^{(2)}$, $f^{(3)}$,…, $f^{(\mathrm{K})}$

Take the first hidden layer as an example:

$$net_i^{(1)} = \sum_{j=1}^{m_0} W_{i,j}^{(1)} Y_j^{(0)} + b_i^{(1)}, (1 \le i \le m_1)$$

$$\boldsymbol{net}^{(1)} = \boldsymbol{W}^{(1)}\boldsymbol{Y}^{(0)} + \boldsymbol{b}^{(1)}$$

$$\boldsymbol{net}^{(1)} = \left[net_1^{(1)}, net_2^{(1)}, \dots, net_{m_1}^{(1)}\right]^T$$

$$\boldsymbol{Y}^{(1)} = f^{(1)}\left(net^{(1)}\right) = [Y_1^{(1)}, Y_2^{(1)}, \dots, Y_{m_1}^{(1)}]^T$$

*Each element in $\boldsymbol{net}^{(1)}$represents the weighted sum of the input layer vector and the bias vector, which can also be called the input vector of hidden layer 1.

Therefore, for the k layer:

$$net_i^{(k)} = \sum_{j=1}^{m_{k-1}} W_{i,j}^{(k)} Y_j^{(k-1)} + b_i^{(k)}, (1 \le i \le m_k)$$

$$\boldsymbol{net}^{(k)} = \boldsymbol{W}^{(k)}\boldsymbol{Y}^{(k-1)} + \boldsymbol{b}^{(k)}$$

$$\boldsymbol{net}^{(k)} = \left[net_1^{(k)}, net_2^{(k)}, \dots, net_{m_k}^{(k)}\right]^T$$

$$\boldsymbol{Y}^{(k)} = f^{(k)}\left(\boldsymbol{net}^{(k)}\right) = [Y_1^{(k)}, Y_2^{(k)}, \dots, Y_{m_k}^{(k)}]^T$$

We obtain the value of $\boldsymbol{net}^{(k)}$ and $\boldsymbol{Y}^{(k)}$ of each layer in the network by forwarding layer-by-layer calculation. Hence, we get the input and the output of each neuron.

f. ***Random Forest (RF)***: Forest is the integration of several decision trees. Among the categories judged by many decision trees, the category with the highest proportion is selected as the final output category of the Random Forest classifier [18]. The random forest algorithm belongs to ensemble learning, for which the principle is to comprehensively use a variety of algorithms to extract features based on a variety of data projections: first, generate relatively weak predictions, then integrate the voting principle, and finally, get better performance results [19]. The key to the random forest is randomness, which uses random sampling by putting back to select samples and randomly selecting features to construct decision trees. This can reduce the correlation between decision trees and improve the model's accuracy. Decision trees use decision criteria derived from data characteristics to predict the target variables. With the increase of tree depth, the decision criteria and model fitting become more complex. Standard decision tree algorithms include *the ID3 algorithm, C4.5 algorithm, C5.0 algorithm, and CART algorithm* [20][21].

Although there are many classification algorithms, they all have a standard set of steps. Step one -- Build a model, describe a predetermined set of categories, and label each tuple with a category label as a training set. After learning, the training set can generate models,

which can be reflected as classification rules, decision trees, mathematical formulas/regressions, etc. Step two -- Use the model. The model's accuracy is evaluated with samples of the test set, and if it meets expectations, the model is applied to new data and tasks.

3.2 Developing Environment

In this paper, we choose to develop web pages based on the Streamlit framework and Python language, using the Jupyter Notebook and Pycharm of the Anaconda platform. Anaconda is a powerful integrated development software for data science and machine learning. Anaconda not only supports Linux, Windows, and Mac systems simultaneously, but also provides environment management capabilities. In order not to conflict with the previous environment and toolkit version, this paper created a new virtual environment, based on which to carry out the subsequent work.

Streamlit is an excellent Python open-source toolkit for rapid web application development. Pages are updated in real time as blocks of code run, making it easy to program and debug. Streamlit development web pages follow a top-down logic, with code running in the same order as the front-end content is displayed.

The specific development environment is shown in the following table:

Table 1 lists the development environments

Toolkit/Environment	Version
python	**3.7.13**
streamlit	**1.8.1**
numpy	**1.19.5**
pandas	**1.3.5**
scikit-learn	**1.0.2**

Anaconda	**4.12.0**
Pycharm	**2021.1.3**

3.3 Model application and evaluation

After the data preprocessing stage in this paper, we divide the original data set into a training set and a test set according to the ratio of 4:1. Firstly, the model is trained and fitted with the training set data. Then the classification ability of the model is verified with the test set data.

In this paper, the model is evaluated using k-fold cross-validation. The training data set is divided into smaller K sets, among which (K-1) is used to form the training set each time, and the left one is considered the training data. The absolute accuracy is calculated as the average value of the values calculated in the cycle [22]. The following figure illustrates how this method works.

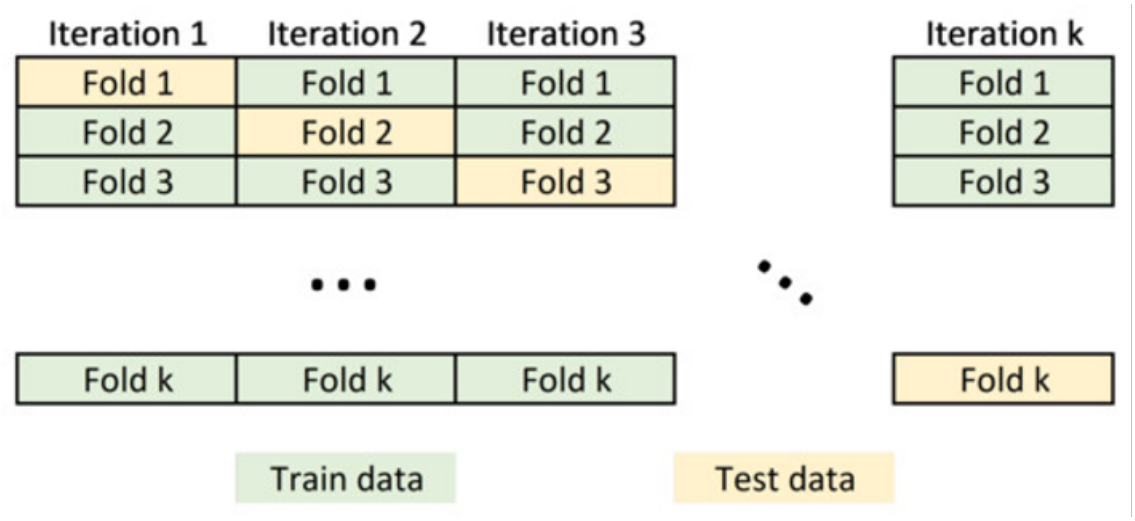

Figure 4 K-Fold Cross-Validation Model

The 10-fold cross-validation fitting experiment of 6 models was carried out on the training dataset containing 480 (600 × 80%) samples, and the remaining test dataset containing 120 (600 × 20%)samples were used for verification. The training and validation accuracy were obtained as shown in the figure below.

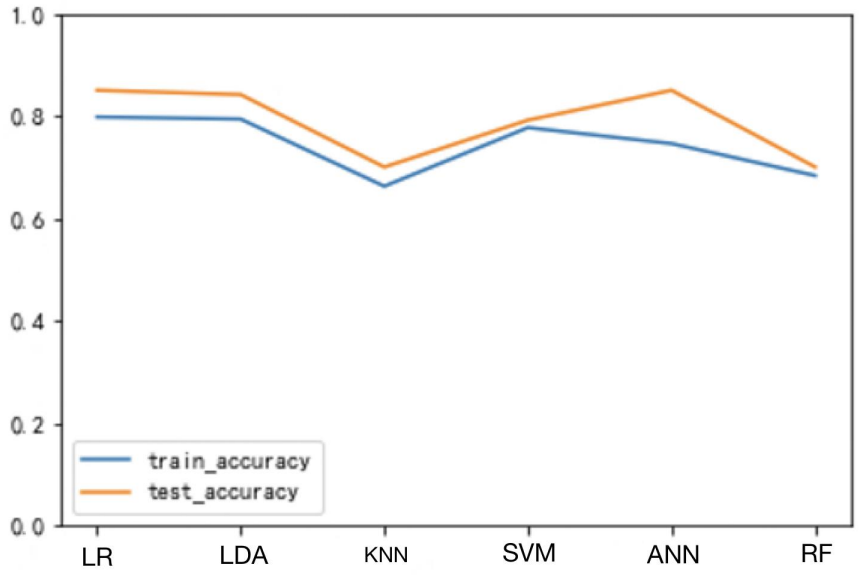

Figure 5 Accuracy line chart of training set and test set

In addition to Accuracy, this paper also uses **Precision**, **Recall**, and their weighted average **F1-score** when evaluating the model's performance. Both the precision and recall rates are calculated using ***The Confusion Matrix***. *The Confusion Matrix* is a table that summarizes the success probability of the classification model in predicting sample instances belonging to different classes [23]. The column index or table header of the confusion matrix is the classification result of the model, and the row index or the other table header is the actual result.

Table 2 Confusion Matrix

	Real Positive	Real Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Precision is the ratio of correct positive predictions to the total number of positive predictions, and the formula is as follows:

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

Recall is the ratio of the correct positive prediction to the total number of positive samples in the data set, and its formula is as follows:

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

**In the following data summary, six major categories of infertility are abbreviated as: Kidney Qi Deficiency (KQD), Kidney Yang Deficiency (KYaD), Kidney Yin Deficiency(KYiD), Liver Qi Stagnation(LSQ), Stagnation of Uterine Cell(SUC), and Internal Obstruction of Phlegm and Dampness(IOPD).*

Table 3 Evaluation index of Logistic Regression (LR) model validation set

Label	Precision	Recall	F1-score	Support
KQD	0.74	0.61	0.67	23
KYaD	0.95	0.95	0.95	20
KYiD	1.00	0.95	0.97	19
LSQ	0.88	0.82	0.85	17
SUC	0.61	0.82	0.70	17
IOPD	0.96	0.96	0.96	24
Accuracy			0.85	120

- For the Logistic Regression model, the accuracy of the training set is 0.798, and the accuracy of the validation set is 0.85. The specific classification evaluation indexes of the validation set are shown as follows.

Table 4 Evaluation index of linear discriminant analysis model validation set

Label	Precision	Recall	F1-score	Support
KQD	0.76	0.70	0.73	23

KYaD	0.94	0.85	0.89	20
KYiD	1.00	0.95	0.97	19
LSQ	0.84	0.94	0.89	17
SUC	0.61	0.82	0.70	17
IOPD	0.95	0.83	0.89	24
Accuracy			0.84	120

- For the Linear Discriminant analysis model, the accuracy of the training set is 0.794, and the accuracy of the validation set is 0.842. The specific evaluation indicators of the validation set are shown as follows.

Table 5 Evaluation indexes of K-nearest neighbor model validation set

Label	Precision	Recall	F1-score	Support
KQD	0.54	0.61	0.57	23
KYaD	1.00	0.70	0.82	20
KYiD	0.89	0.89	0.89	19
LSQ	0.90	0.53	0.67	17
SUC	0.39	0.76	0.52	17
IOPD	0.94	0.71	0.81	24
Accuracy			0.70	120

- For the K-nearest neighbor model, the accuracy of the training set is 0.663, and the accuracy of the validation set is 0.7. The specific evaluation indicators of the validation set are shown as follows.

Table 6 Evaluation index of support vector machine model validation set

Label	Precision	Recall	F1-score	Support
KQD	0.67	0.43	0.53	23
KYaD	0.90	0.95	0.93	20
KYiD	0.95	0.95	0.95	19
LSQ	0.88	0.82	0.85	17
SUC	0.50	0.82	0.62	17
IOPD	0.95	0.83	0.89	24
Accuracy			0.79	120

- For the support vector machine model, the accuracy of the training set is 0.777, and the accuracy of the validation set is 0.792. The specific evaluation indicators of the validation set are shown as follows.

Table 7 Evaluation index of artificial neural network model validation set

Label	Precision	Recall	F1-score	Support
KQD	0.71	0.65	0.68	23
KYaD	0.95	0.95	0.95	20
KYiD	0.95	0.95	0.95	19

LSQ	1.00	0.76	0.87	17
SUC	0.59	0.76	0.67	17
IOPD	0.96	1.00	0.98	24
Accuracy			0.85	120

- For the artificial neural network model, the accuracy of the training set is 0.746, and the accuracy of the validation set is 0.85. The specific evaluation indicators of the validation set are shown as follows.

Table 8 Evaluation index of validation set of random forest model

Label	Precision	Recall	F1-score	Support
KQD	0.29	0.17	0.22	23
KYaD	0.83	0.95	0.88	20
KYiD	0.95	1.00	0.97	19
LSQ	0.88	0.82	0.85	17
SUC	0.27	0.35	0.31	17
IOPD	0.88	0.92	0.90	24
Accuracy			0.70	120

- For the random forest model, the accuracy of the training set is 0.683, and the accuracy of the validation set is 0.7. The specific evaluation indicators of the validation set are shown as follows.

When the original dataset is not segmented, the average accuracy and standard deviation of 10-fold cross validation of logistic regression, linear discriminant analysis, K-nearest neighbor, support vector machine, artificial neural network and random forest algorithm are shown in the following figure, respectively.

Table 9 Accuracy and standard deviation of each model

Model	Average Accuracy	Standard Deviation
LR	0.833333	0.034960
LDA	0.806667	0.070396
KNN	0.668333	0.034521
SVM	0.768333	0.042459
MLP	0.815000	0.066018
RF	0.788333	0.048333

The line chart of the evaluation results of the six models is shown below. It is not difficult to find that, except for the KNN model, the accuracy of the other models is good, but there is still a large room for improvement. Therefore, the focus of the following research will be on the adjustment and optimization of the model.

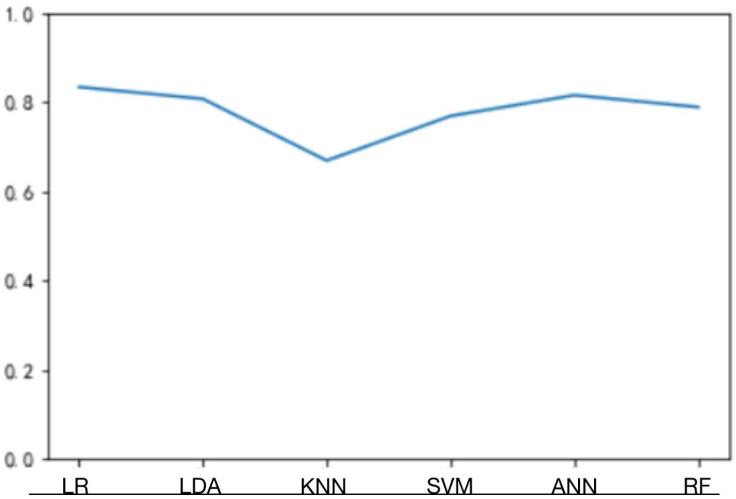

Figure 6 Line chart of accuracy of each model

IV. DATA ANALYSIS–Model Adjustment and Optimization

4.1 Grid Search Hyperparameters

The first tuning strategy adopted in this paper is a grid search for the best combination of hyperparameters, which has been shown to improve the model performance [24]. Hyperparameters are parameters that the classifier cannot learn directly and that need to be passed to the constructor by the programmer. This article uses the GridSearchCV function provided by Scikit-Learn, which considers all parameter combinations for a given value. In order to compare with the performance before parameter adjustment, ten-fold cross-validation is still adopted here to verify the calculation accuracy.

- In the logistic regression model, the hyperparameters include reciprocal of regularization intensity C and solver. The grid search results show that the best accuracy is 0.83, which is not improved compared to before tuning.

	Average	Standard Deviation	Parameter Combination
0	0.828333	0.054288	{'C': 0.1, 'solver': 'newton-cg'}
1	0.828333	0.054288	{'C': 0.1, 'solver': 'lbfgs'}
2	0.828333	0.054288	{'C': 0.1, 'solver': 'sag'}
3	0.831667	0.054493	{'C': 0.1, 'solver': 'saga'}
4	0.805000	0.060116	{'C': 1, 'solver': 'newton-cg'}
5	0.805000	0.060116	{'C': 1, 'solver': 'lbfgs'}

Figure 7 Pattern parameters of Logistic Regression model

- In the linear discriminant analysis model, the default solver is Singular Value Decomposition (SVD), and the grid search results show that when the solver is Least Squares, Shrinkage value shrinkage=0.5 has an accuracy of 0.818, which increases by 1% compared with 0.807 before parameter adjustment

	Average	Standard Deviation	Parameter Combination
0	0.806667	0.059255	{'shrinkage': 0.01, 'solver': 'lsqr'}
1	0.806667	0.059255	{'shrinkage': 0.01, 'solver': 'eigen'}
2	0.818333	0.057951	{'shrinkage': 0.05, 'solver': 'lsqr'}
3	0.818333	0.057951	{'shrinkage': 0.05, 'solver': 'eigen'}
4	0.810000	0.066332	{'shrinkage': 1, 'solver': 'lsqr'}
5	0.810000	0.066332	{'shrinkage': 1, 'solver': 'eigen'}

Figure 8 Pattern parameters of the Linear Discriminant analysis model

- In the K-neighbor model, the critical hyperparameter is the number of neighbors, N_neighbors. When N_neighbors is 1-73 in turn, the accuracy of training and testing sets is shown in the following figure. It can be seen from the figure that when N_neighbors =63, the average accuracy score is 0.733, which is 10% higher than 0.668 before parameter adjustment.

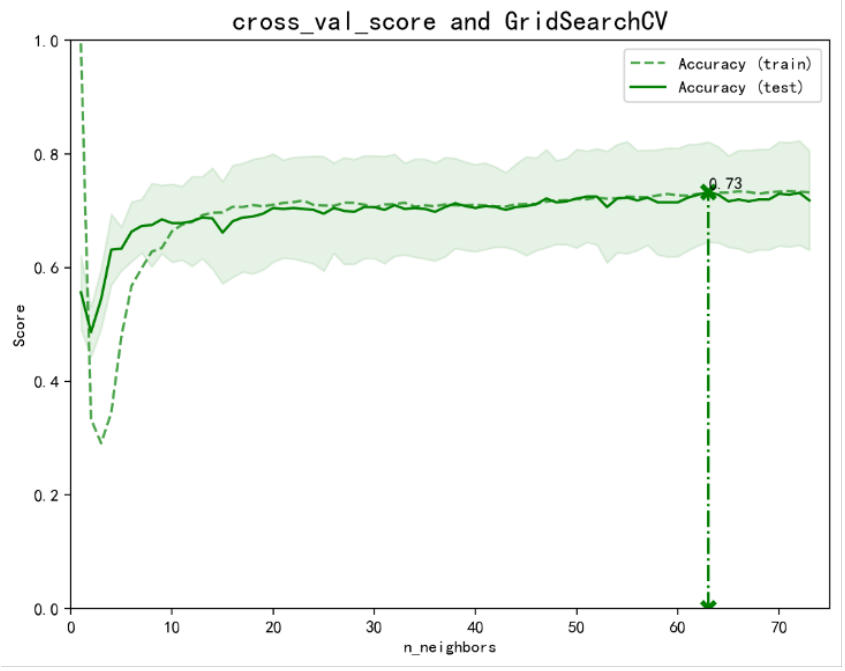

Figure 9 Relationship between the average accuracy of the K-Nearest Neighbor model and the value of hyperparameter

- In the support vector machine model, the hyperparameters that affect the performance are C, Gamma, and kernel. The results of the grid search show that the accuracy of the optimal parameter combination can reach 0.833, which is 8% higher than 0.768 before parameter adjustment.

	Average	Standard Deviation	Parameter Combination
1	0.780000	0.082597	{'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
2	0.771667	0.090077	{'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
3	0.833333	0.067082	{'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
4	0.776667	0.090431	{'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
5	0.788333	0.055802	{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}

Figure 10 Pattern parameters of Support Vector Machine model network

- In the artificial neural network model, the grid search results show that the accuracy of the optimal parameter combination can reach 0.813, which is not improved compared with the value before parameter adjustment.

	Average	Standard Deviation	Parameter Combination
0	0.730000	0.050442	{'hidden_layer_sizes': (100,), 'max_iter': 20,...
1	0.798333	0.059838	{'hidden_layer_sizes': (100,), 'max_iter': 20,...
2	0.813333	0.052599	{'hidden_layer_sizes': (100,), 'max_iter': 50,...
3	0.786667	0.059535	{'hidden_layer_sizes': (100,), 'max_iter': 50,...
4	0.793333	0.056862	{'hidden_layer_sizes': (100,), 'max_iter': 300...
5	0.775000	0.088270	{'hidden_layer_sizes': (100,), 'max_iter': 300...

Figure 11 Pattern parameters of the Artificial Neural Network model

- In the random forest model, the grid search results show that the accuracy of the optimal parameter combination can reach 0.835, which is 6% higher than 0.788 before parameter adjustment.

	Average	Standard Deviation	Parameter Combination
0	0.816667	0.058214	{'max_depth': 15, 'n_estimators': 500}
1	0.835000	0.059372	{'max_depth': 15, 'n_estimators': 800}
2	0.826667	0.056862	{'max_depth': 15, 'n_estimators': 1200}
3	0.825000	0.049018	{'max_depth': 25, 'n_estimators': 500}
4	0.828333	0.056789	{'max_depth': 25, 'n_estimators': 800}
5	0.833333	0.062805	{'max_depth': 25, 'n_estimators': 1200}

Figure 12 Pattern parameters of random forest model network

In summary, the performance improvement of the six models after grid search hyperparameters is indicated in the following table. Among them, the performance of the K-nearest neighbor, support vector machine, and random forest model is significantly improved compared with that before parameter adjustment. However, parameter adjustment has little influence on logistic regression and artificial neural network models.

Table 9 Comparison of average accuracy before and after the mesh parameters of each model

Model	Before Adjustment	After Adjustment
LR	0.833333	0.831667
LDA	0.806667	0.818333
KNN	0.668333	0.733333
SVM	0.768333	0.833333
MLP	0.815000	0.813333
RF	0.788333	0.835000

4.2 Data Enhancement

Data augmentation can be described as any method that makes the original dataset larger. For example, image data can be enhanced by scaling, rotating, and brightening it; Audio data enhancement can change the tone and playback speed; Or, to expand the text dataset, we can substitute synonyms, add noise, or change the sentence structure. Data enhancement has been widely used in computer vision [25], natural language processing [26], and other fields. It has become a strategy to solve the problem of a small number of data sets to optimize the performance of models. Data augmentation brought many benefits to machine learning and deep learning, like reducing the dependence on extensive, accurate data, especially in privacy-sensitive fields, making unbalanced data sets uniform, reducing overfitting, and improving the generalization ability of models.

This paper used the Synthetic Minority Oversampling Technique (SMOTE) of the Imbalanced-Learn library for data enhancement. The SMOTE method was initially designed to solve the problem of uneven numbers of various types of samples by inserting new samples for classes with small numbers of samples [27]. The specific principle is to complete for a minority sample S, use the K-nearest neighbor algorithm to find the nearest K similar samples to S, randomly select one neighbor S-neighbor, and then randomly interpolate on the line between S and S-neighbor as a new sample generation according to the formula. *Distance* is defined as the Euclidean distance of N-dimensional feature space between samples. Data augmentation is classified into supervised and unsupervised types, among which the supervised data augmentation can be divided into a single sample and diverse data. SMOTE method belongs to the supervised diversified data augmentation.

After the expansion of the original dataset, from 600 samples to 1200, 3000, the average accuracy of the 10-fold cross-validation of the six models changed, as shown in the following figure and table. Although the model's generalization ability may decrease, it can be concluded that the model's classification accuracy will be improved when the dataset capacity is expanded, which can achieve the purpose of optimizing the model performance.

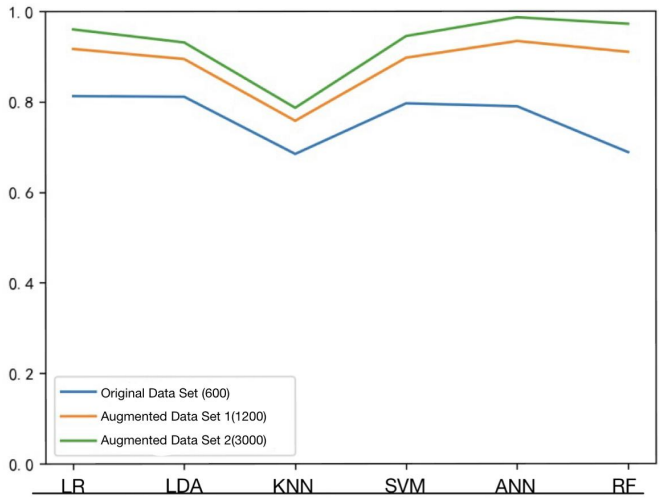

Figure 13 Impact of Data Enhancement on model accuracy

Table 10 Influence of dataset capacity on model accuracy

Model\Sample size	600	1200	3000
LR	0.816667	0.921667	0.960333
LDA	0.808333	0.894167	0.926667
KNN	0.691667	0.761667	0.786667
SVM	0.803333	0.890833	0.941000
MLP	0.785000	0.944167	0.984333
RF	0.675000	0.909167	0.974333

V. CONCLUSION

This paper focuses on a subproblem in traditional Chinese medicine -- syndrome classification. From the point of view of machine learning, the solution of computer

specialty is given. Many factors go into clinical decision-making and are recorded in the patient's electronic information database. The development of artificial intelligence technology makes it possible to mine and make good use of the data to the greatest extent. A complete machine learning process usually includes data processing, model training, evaluation, validation, tuning, and other steps. All kinds of numerical data features in a binary discrete numerical variable, such as over-sampling, under-sampling, standardization, and binarization, are thus omitted. Because there are missing values in the original data set, a random forest regression prediction method is used to fill in the missing values. In addition, four zero-variance features were removed in the preprocessing phase.

The experiments included six models, including logistic regression, linear discriminant analysis, K-nearest neighbor, support vector machine, artificial neural network, and random forest. The model was evaluated by 10-fold cross-validation. In the beginning, the accuracy of the K-nearest neighbor was only 0.66, and the classification accuracy of the other five models was around 0.8. Subsequently, this paper adopts the methods of hyperparameter tuning and other methods to optimize the model, and the classification accuracy of the artificial neural network model and random forest model is improved to about 0.85. Finally, the SMOTE algorithm helped with data enhancement; the data set capacity, therefore, expanded to two times and five times. The classification accuracy was also increased by 16.6% and 22%, respectively, based on the original. Data enhancement is just an attempt. Although the ability to fit existing data sets was significantly improved, more accurate data should be collected to improve the model's generalization ability.

In addition, according to the data from the training, verification, and evaluation of the model in this paper, the classification effect of kidney Qi deficiency syndrome and stasis cell palace syndrome were obviously behind that of other categories without exception. The reason may be that the characteristics of the two types of sample data are not prominent, or the two types of the syndrome are difficult to distinguish.

VI. REFERENCE

1. “Infertility.” *World Health Organization*, World Health Organization,

2. “5 Reasons Infertility Awareness Matters.” *Carolinas Fertility Institute*, 17 Apr. 2020,

3. “Traditional Chinese Medicine for Infertility: West Wimbledon Physio Clinic.” *West Wimbledon Physiotherapy*, 1 Feb. 2022,

4. Zhu, Jihe, et al. “Acupuncture Treatment for Fertility.” *Open Access Macedonian Journal of Medical Sciences*, Republic of Macedonia, 19 Sept. 2018,

5. Cyranoski, David. “Why Chinese Medicine Is Heading for Clinics around the World.” *Nature News*, Nature Publishing Group, 26 Sept. 2018,

6. Jiang, Lijuan. "Knowledge management system and infertility treatment using Traditional Chinese medicine." (2013).

7. Kumar, Yogesh, et al. “Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review, Synthesizing Framework and Future Research Agenda.” *Journal of Ambient Intelligence and Humanized Computing*, Springer Berlin Heidelberg, 13 Jan. 2022

8. Xu, M.B., et al." Construction and application of TCM syndrome Differentiation model of infertility based on artificial intelligence." Chinese Journal of Traditional Chinese Medicine 36.09(2021):5532-5536.

9. Wang Y. Shi X. Li L. Efferth T. Shang D; “The Impact of Artificial Intelligence on Traditional Chinese Medicine.” *The American Journal of Chinese Medicine*, U.S. National Library of Medicine.

10. Wang, Yulin, et al. "The impact of artificial intelligence on traditional Chinese medicine." *The American Journal of Chinese Medicine* 49.06 (2021): 1297-1314.

11. Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

12. Arora, Surbhi. “Supervised vs Unsupervised vs Reinforcement.” *AITUDE*, 29 Jan. 2020,

13. Schober P, Vetter T R. Logistic regression in medical research[J]. Anesthesia and analgesia, 2021, 132(2): 365.

14. Tharwat A, Gaber T, Ibrahim A, et al. Linear discriminant analysis: A detailed tutorial[J]. AI communications, 2017, 30(2): 169-190.

15. Zhang Z. Introduction to machine learning: k-nearest neighbors[J]. Annals of translational medicine, 2016, 4(11): 218.

16. Pisner D A, Schnyer D M. Support vector machine[M]//Machine learning. Academic Press, 2020: 101-121.

17. Rana A, Rawat A S, Bijalwan A, et al. Application of multi-layer (perceptron) artificial neural network in the diagnosis system: a systematic review[C]//2018 International conference on research in intelligence and computing in engineering (RICE). IEEE, 2018:1-6.

18. Biau G, Scornet E. A random forest guided tour[J]. Test, 2016, 25(2): 197-227.

19. Dong X, Yu Z, Cao W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14(2): 241-258.

20. Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.

21. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning[J]. Journal of Applied Science and Technology Trends, 2021, 2(01): 20-28.

22. Wong T T, Yeh P Y. Reliable accuracy estimates from k-fold cross validation[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(8): 1586-1594.

23. Beauxis-Aussalet E, Hardman L. Simplifying the visualization of confusion matrix[C]//26th Benelux Conference on Artificial Intelligence (BNAIC). 2014.

24. Shekar B H, Dagnew G. Grid search-based hyperparameter tuning and classification of microarray cancer data[C]//2019 second international conference on advanced computational and communication paradigms (ICACCP). IEEE, 2019: 1-8.

25. Kaur P, Khehra B S, Mavi E B S. Data augmentation for object detection: A review[C]//2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2021: 537-543.

26. Shorten C, Khoshgoftaar T M, Furht B. Text data augmentation for deep learning[J]. Journal of big Data, 2021, 8(1): 1-34.

27. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.