Machine Learning-Driven Prediction of TCM Herb Properties Using Chemical Component Analysis

Hanxiao Chang, Jiayi Liu, Xiaoyu Zhai

Abstract—**Traditional Chinese Medicine (TCM) is an ancient medical system that has been practiced for thousands of years, focusing on holistic approaches to health using natural herbs, acupuncture, and other therapeutic techniques. It emphasizes maintaining balance within the body and restoring health by regulating the flow of energy. Although TCM has a long history and widespread utilization, TCM remains controversial in the context of modern medicine because of lacking of empirical validation and significant differences from contemporary scientific methodologies. However, TCM's holistic principles and extensive pharmacological knowledge make it a valuable area for scientific exploration, with potential applications in modern healthcare. This study attempts to bridge the gap between TCM concepts and modern science. Machine learning techniques are well-suited for analyzing the complex chemical properties of TCM herbs, providing an opportunity to identify patterns and relationships that are otherwise difficult to discern. We explore several machine learning models, including neural networks, LightGBM, and logistic regression. These models are used to predict the properties of TCM herbs based on their chemical components. Our approach incorporates mean pooling techniques to aggregate chemical features, addressing compositional complexity and sample imbalance issues. We also employ cross-validation for hyperparameter optimization, ensuring robust model performance despite the limited dataset. The results indicate that Neural Network (NN)provided the best performance in terms of Hamming Loss. Our findings demonstrate that machine learning can effectively contribute to the validation of TCM. Ultimately we expect this approach could support its broader acceptance within the modern healthcare system.**

Index Terms—**ECE 1513, Introduction to Machine Learning**

I. Introduction

Traditional Chinese Medicine (TCM) is an ancient and conventional system of medicine that has been practiced for thousands of years. It mainly focuses on natural herbs, acupuncture, and holistic treatment approaches. Compared with the modern medical system, TCM emphasizes balance within the body and restoring health by regulating the flow of energy, or Qi, through various therapeutic techniques.

Even though TCM has a long history, it often faces skepticism and criticism for its differences from modern scientific methods. It is also limited by empirical validation. There were existing efforts for validating TCM and primarily implemented experimental studies to isolate active compounds, but these methods were resource-intensive and inconclusive. The greatest challenge is integrating TCM concepts with modern science. It affects the understanding of TCM effectiveness and makes it much more difficult to operationalize evidence-based medicine. Another significant challenge is the lack of standardization in TCM practices and materials. For instance, the presence of heavy metals and pesticides in herbal products hinders its international acceptance and credibility [1].

To tackle these challenges, modern technologies like AI and bioinformatics present promising solutions. We can analyze large datasets to discover patterns and validate TCM's efficacy more effectively through some machine learning methods. Therefore, it is crucial to develop international standards for quality control and establish rigorous clinical trials. This is a fundamental step for integrating TCM with modern healthcare systems.

We are suggesting an approach to study the properties of TCM herbs by analyzing their chemical components. Our study will analyze the properties of each important chemical component of TCM, such as flavour property, thermal nature(warm or cold), and potential Herb Meridian Tropism. Furthermore, we will use binary classification techniques to examine the combined effects of multiple chemical components as well. This approach allows us to predict the properties of complex mixtures and understand their contributions to the overall characteristics of the herb. These steps predominantly support an understanding of how different herbs work synergistically in TCM formulations. This approach expects to generate empirical evidence that supports the efficacy of TCM contributing to its acceptance in modern healthcare.

II. System Design

The mapping relationship between the properties of TCM herbs and the chemical features of their constituent compounds was modelled using a neural network for binary classification. This neural network predicts TCM herb properties based on chemical feature representations.

A. Primary Solution

a) Data preparation: The dataset was derived from The Encyclopedia of Traditional Chinese Medicine (ETCM), which catalogues approximately 400 herbs along with their associated properties and chemical components. Additionally, ETCM contains records of over 7,000 compounds, each with 21 chemical features. The data acquisition involved web crawling to extract two datasets: one linking herbs to their components and another associating chemical features with their respective properties. The datasets were subsequently processed using the mean pooling method and combined into the final dataset for training.

TCM herb properties are organized into three primary categories using TCM terminologies: temperature, flavour, and meridian tropism. These categories include 28 classes, with herbs potentially belonging to multiple classes within a category. Due to this overlap, binary classification is preferred over multi-class classification, as the latter requires one-hot encoding, which is incompatible with multi-class membership scenarios. P4 Another challenge is the imbalance in the number of samples across classes, particularly in the temperature category, where some classes have very few samples, which will likely result in the absence of specific classes after splitting the dataset. To mitigate this problem, the original nine temperature classes were generalized into four broader classes based on TCM definitions [2], and the total number of herb property classes was reduced to 24.

The primary challenge in this task is the compositional complexity of TCM herbs, as each herb comprises a varying number of chemical components. A mean pooling technique is applied during data preprocessing to address this problem. This method aggregates the chemical features of all components belonging to a herb by taking the average, and denotes this aggregate as a unified representation of the herb's chemical properties.

b) Neural Network training: Given the relatively small dataset consisting of slightly over 400 samples, a compact neural network was constructed with a small number of fully connected layers. Batch Gradient Descent (BGD) was selected as the optimization algorithm for its stability and precision in small datasets, where computational cost is negligible.

The neural network accepts 21 input features corresponding to the chemical features, and yields one binary output for each of the 24 herb property classes. ReLU activation functions were applied to the hidden layers, while a sigmoid activation function was applied to the output layer to accommodate the binary classification requirements. The training process monitors binary cross-entropy loss and accuracy for both training and testing data. The best-performing models and their associated metrics were stored for future use.

To prevent overfitting, an early stopping algorithm was implemented. Initially, both training and testing losses decrease, but a rising trend in testing loss indicates overfitting. This algorithm terminates training when the testing loss increases for ten consecutive steps, as specified by a patience parameter of 10.

c) Hyperparameter tuning: Three key hyperparameters were optimized during model development: the number of layers, the number of neurons per layer, and the learning rate. Considering the limited dataset size, using a separate validation set would reduce the data available for training and testing. The decision of hyperparameters made with limited samples in the validation set would also result in low confidence.

Cross-validation was employed as a more suitable alternative [3]. The dataset was split into a 75% training set and a 25% testing set. The training set was further partitioned into five folds of equal size. During each round of cross-validation, one fold was used as the validation set, while the remaining folds served as the training set.

The average testing loss across all five rounds was used to evaluate the performance of each hyperparameter configuration. Prior experiments determined that most training processes converge within 200 epochs, and this was set as the training period of cross-validation.

For simplicity, all 24 herb property classes shared the same set of hyperparameters, which were evaluated based on the mean of the cross-validation results. The averaged losses for each combination of hyperparameters are presented in the table below. The optimal configuration consisted of a neural network with two hidden layers containing 30 and 60 neurons, respectively, and a learning rate of 0.001.

Network structure	Learning rate 1e-3	Learning rate 5e-4
21-30-30-30-1	0.41036204947158694	0.4143602193022768
21-30-60-30-1	0.40938450093381107	0.40952777800460655
21-30-30-1	0.406978879434367	0.418965047225361
21-30-60-1	0.40664155160387355	0.40722109253207844

TABLE I

CROSS-VALIDATION LOSS WITH DIFFERENT HYPERPARAMETER SETS

B. Alternative Solution

One alternative method for this task is to establish a mapping relationship between the properties of Traditional Chinese Medicine (TCM) herbs and the chemical features of single chemicals instead of compounds. The same dataset is used for the alternative methods, with an additional preprocessing step called reverse mapping. Initially, the dataset links each herb to its corresponding chemical components. In the reverse mapping process, we associate each chemical with the set of herbs it appears in and aggregate the herb properties for that chemical. This transformation is essential because the herb properties are the target variables in our classification tasks.

In the context of the alternative task, we implemented a Neural Network (NN) as the primary model. The NN uses a three-layer architecture optimized for binary classification tasks, with the following layers:

- Input layer: 64 neurons with ReLU activation.
- Hidden layer: 32 neurons with ReLU activation.
- Output layer: 1 neuron with sigmoid activation.

The NN is trained using the Adam optimizer with a learning rate of 0.001 and binary cross entropy as the loss function. Training is conducted for 10 epochs, with 20% of the training data reserved for validation.

C. Solution Evaluation

The proposed approach offers significant advantages over the alternative methods that predict herb properties from individual chemical compounds. The solution directly relates aggregated chemical features of herbs, which is essential in predicting the properties of herbs composed of previously unstudied chemical combinations. Although the approach may have limited data and potentially lower prediction precision, it is well-suited to real-world applications.

We also tested Logistic Regression (LR) and Gradient Boosting (GB) models as alternative models. Logistic Regression, as a linear classifier, provides a baseline for comparison, while Gradient Boosting offers a robust tree-based ensemble method capable of capturing nonlinear relationships.

The Neural Network was chosen as the primary solution because it can model complex, nonlinear relationships between the chemical features and herb properties. While Logistic Regression is simpler and interpretable, it may not capture intricate patterns effectively. Gradient Boosting can be slower to train and more resource-intensive for larger datasets. The NN's ability to generalize well across various targets made it the preferred choice for this task.

The hyperparameters for the NN were chosen to balance performance and computational efficiency:

- **Number of layers and neurons:** The three-layer architecture with 64 and 32 neurons was selected after initial experimentation to provide sufficient capacity without overfitting.
- **Activation functions:** ReLU activation was used for hidden layers to handle nonlinearity effectively, and sigmoid activation was used in the output layer for binary classification.
- **Optimizer:** The Adam optimizer was selected for its adaptive learning rate, ensuring faster convergence.
- **Learning rate (0.001):** This value was chosen after testing common ranges, providing stable training and good accuracy.
- **Epochs (10):** Limited to 10 to prevent overfitting and ensure timely training for multiple models.

III. Results

A. Primary Solution

The performance of the model was evaluated using metrics of testing loss and accuracy. For each TCM herb property, the network's output was passed through a sigmoid activation function to compute the cross-entropy loss. Additionally, a sign function was applied to the sigmoid output to convert predictions into binary values, which were then used to assess the accuracy of the predictions.

The loss and accuracy for both the training and testing datasets were visualized in the plots below. The x-axis of these plots represents the number of training iterations, while the y-axis corresponds to logistic loss or accuracy percentage. In the loss plot, lower values indicate better performance, whereas in the accuracy plot, higher values denote improved performance. The best-performing model in each scenario is marked with a red triangle, indicating the epoch at which the testing loss reaches its minimum or the testing accuracy reaches its maximum. This point is also used as the stopping criterion to prevent overfitting.

The trained model achieved varying levels of accuracy across different herb properties. Certain properties, such as ”sour” and ”hot” achieved accuracy rates as high as 90%, while others, such as ”sweet” and ”liver” achieved only 60%, while other properties with accuracies hovering around 70%.

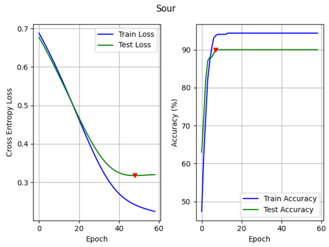

Fig. 1. Target Sour

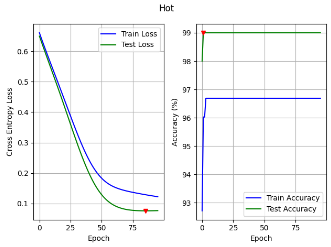

Fig. 2. Target Hot

Three distinct patterns were observed in the loss and accuracy curves as the number of epochs increased. The classic behaviour, (such as ”salty”), showed a testing loss curve consistently above the training loss curve and a testing accuracy curve consistently below the training accuracy curve.

For properties such as ”bitter,” although the loss curve behaviour was similar, both the training and testing accuracy curves exhibited fluctuations throughout the training process.

A more abnormal pattern was observed for properties like ”heart,” where the training loss remained higher than the testing loss. This unexpected behaviour is likely a result of the limited number of samples in the dataset that leads to coincidental predictions.

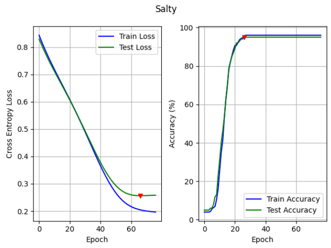

Fig. 3. Target Salty

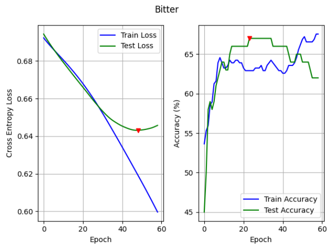

Fig. 4. Target Bitter

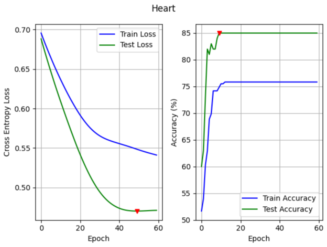

Fig. 5. Target Heart

It is noteworthy that the best-performing models for loss and accuracy did not always align in terms of the optimal number of training epochs. In some cases, the model achieved lower loss at the expense of reduced accuracy, or vice versa. This discrepancy is potentially due to differences in the optimization objectives for loss and accuracy. While minimizing loss focuses on improving the confidence of predictions, maximizing accuracy prioritizes the correct classification of binary outputs.

Several limitations and challenges fall outside the scope of this study. There exist inconsistencies in the documentation of TCM herb properties across various sources. These discrepancies likely stem from differences in historical, geographic, and cultural contexts.

Additionally, the model employed in this study does not account for the dosage or concentration of chemical constituents within the herbs. Variations in these factors can significantly impact the therapeutic effects of TCM herbs. Further research is needed to address these challenges comprehensively.

B. Alternative Solution

The models were evaluated using test accuracy, measuring the proportion of correctly classified samples for each target variable. This metric reflects the effectiveness of each model in handling the binary classification tasks. Validation accuracy was also monitored during training to assess the generalization ability of the Neural Network.

The results indicate strong performance across all models for most target variables, with test accuracies exceeding 90% in most cases. The Neural Network consistently outperformed the other models, achieving the highest accuracy for the majority of target columns. However, all models struggled with specific targets, such as ”isLiverMeridian” and ”isSweet,” where accuracy dropped to around 60-70%, suggesting that these targets may require further feature engineering or more complex models.

Despite the strong performance of the model, TCM herbs are inherently complex, often comprising diverse combinations of chemical constituents which complicates the establishment of a one-to-one mapping between individual herbs and their chemical profiles. Moreover, multiple herbs may share similar chemical compositions, potentially introducing ambiguity into the predictions.

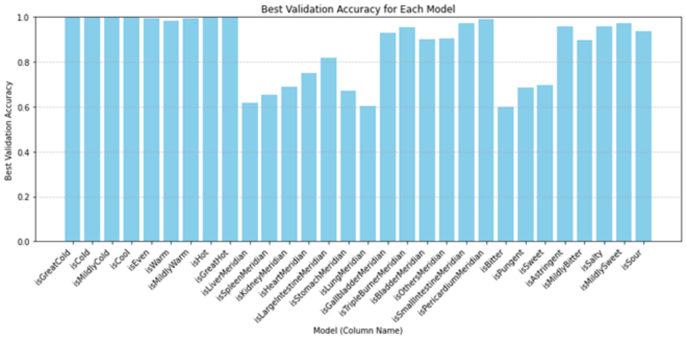

Fig. 6. Best validation accuracy for each model

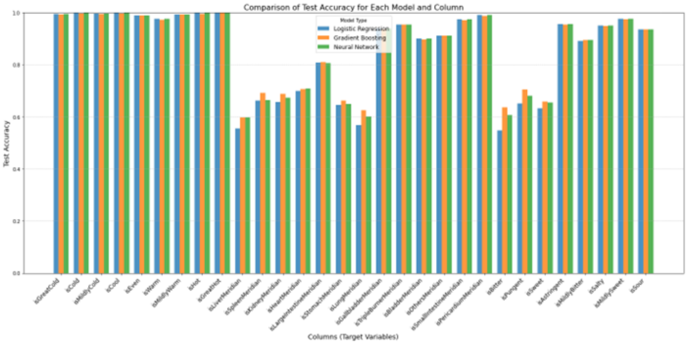

Fig. 7. Best validation accuracy for each model

IV. Concluding Remarks

This study highlights the potential of using modern machine-learning techniques to bridge the division between traditional TCM knowledge and empirical validation. We can conclude that aggregated chemical features offer a more realistic approach to understanding the properties of complex herbal combinations.

Our model used batch gradient descent for training and employed ReLU activation functions for hidden layers, along with a sigmoid activation function for the output layer. The binary cross-entropy loss allowed us to effectively train the model for binary classification. Cross-validation was crucial in hyperparameter tuning, optimizing parameters such as the number of layers, neurons, and learning rates. Among the models evaluated, Neural Networks(NN) achieved the best performance in terms of Hamming Loss, demonstrating their ability to effectively capture key relationships despite the challenge of a relatively small dataset. This result highlights the potential of using modern machine-learning techniques to bridge the division between traditional TCM knowledge and empirical validation. Although the accuracy of predictions remains a challenge, we can conclude that aggregated chemical features offer a more realistic approach to understanding the properties of TCM. Our findings may pave the way for future research about improving model precision and integrating TCM into evidence-based medical practices. We aspire our approach to facilitate greater acceptance of TCM within the modern healthcare system.

References

[1] N. Huang, W. Huang, J. Wu, S. Long, Y. Luo, and J. Huang, "Possible opportunities and challenges for traditional chinese medicine research in 2035," *Frontiers in Pharmacology*, vol. 15, 2024. [Online]. Available: https://doi.org/10.3389/fphar.2024.1426300

[2] Z.-M. L. T. C. C.-Y. L. S.-H. T. X.-B. Z. W. Z. Z.-Y. L. R.-R. Z. H.-J. Y. X.-J. W. L.-Q. H. Hai-Yu Xu, Yan-Qiong Zhang, "Etcm: an encyclopaedia of traditional chinese medicine," *Nucleic Acids Research*, vol. 47, 2019. [Online]. Available: https://doi.org/10.1093/nar/gky987

[3] C. SCHAFFER, "Selecting a classification method by cross-validation," *Machine Learning*, vol. 13, 1993. [Online]. Available: https://doi.org/10.1007/BF00993106