Identification Tool for Gastric Cancer Based on Integration of 33 Clinical Available Blood Indices through Deep Learning

BOWEN ZHANG1,2,#, LONG CHENG3,#, YUZHEN NIU4, AMING WANG1, PENGYI ZHANG2, TIAN TIAN SHEN5, LILI XI6, DEKUI ZHANG3, AND SHUYAN LI1

1School of Medical Information and Engineering, Xuzhou Medical University, Xuzhou, 221004, China
2Department of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou, 730000, China
3Department of gastroenterology, Lanzhou University Second Hospital, Lanzhou, 730000, China
4School of Life Science, Shandong University of Technology, Zibo, 250100, China
5The First School of Clinical Medicine, Lanzhou University, Lanzhou, 730000, China
6Institution of Drug Clinical Trial, The First Hospital of Lanzhou University, Lanzhou, 730000, China

Corresponding author: Shuyan Li (email: li_shuyan@xzhmu.edu.cn) and Dekui Zhang (email: sczd1972@163.com)

#These two authors contributed equally to this work.

This work was supported by the National Natural Science Foundation of China (No.81770525 to DZ), the Industrial Support and Guidance Project of Colleges and Universities in Gansu Province (No. 2019C-21 to DZ ), the Cuiying Science and Technology Project of the Second hospital of Lanzhou University (No.CY2018-ZD01 to DZ) and Key Research and Development Project of Gansu Province (No. 20JR5RA370 to LX).

ABSTRACT Gastric cancer (GC) is one of the most common cancers in the world. In cancer detection, liquid biopsy, as a noninvasive and rapid method, is growing in importance. Different from traditional liquid biopsy using a single biomarker, this study integrated a variety of blood biochemical indices and established an identification system by means of deep learning under the H2O framework method. Based on data from 2951 samples, 58 routine blood biochemical indices, age and gender were collected as comprehensive indices to establish the identification model. Then, the number of indices was reduced to simplify the model, and 33 indices were utilized to build the final identification tool. A tenfold cross-validation technique was used to evaluate the performance of the proposed method. The sensitivity, specificity, accuracy, and area under the ROC curve on the cross-validation set were 85.44%, 83.82%, 84.54% and 0.9165, respectively. The identification tool is built free online at http://www.cppdd.cn/GC2.

The proposed system provides a new approach to identify GC with advantages of being efficient, noninvasive and economical. The deep learning of the integration of these blood biochemical indices will bring insights into the comprehensive understanding of GC pathology, as well as the prevention, screening, diagnosis, and prognosis of GC.

INDEX TERMS Clinical available blood indices, deep learning, gastric cancer, identification, online server

I. INTRODUCTION

Gastric cancer (GC) was responsible for over 1.08 million new cancer cases in 2020 and more than 768,000 deaths worldwide [1]. Although the incidence and mortality of GC have declined, it is still the fifth most commonly diagnosed cancer and ranks fourth in the mortality rate globally [2]. The high prevalence of GC, poor prognosis [3], and limited treatment options have resulted in a heavy medical burden [4].

The development of GC is a multistage, multistep, and multimechanism process that is highly heterogeneous in terms of structural growth, cell differentiation, molecular pathogenesis, and stages [4]. Therefore, the diagnosis of GC is difficult, especially early GC. Meanwhile, the 5-year survival rate of patients with early GC after undergoing surgery exceeds 90%. In contrast, the 5-year survival rate of patients with stage IV GC is less than 5% [5]. The earlier a patient with GC is diagnosed, the easier it can be cured [6]. However, due to nonspecific symptoms of early patients, less than 25% of patients with GC are detected at an early stage [3]. Therefore, a fast and low-cost detection tool is urgently needed to detect GC.

Endoscopy and image examination are the gold standard diagnostic methods for GC [7]. However, the detection accuracy of conventional endoscopy is only 69% to 79% [8], and image examination has difficulty identifying early
lesions with high accuracy [9]. In addition, endoscopy may cause gastric bleeding, gastric perforation, and bacterial infections [10]. Imaging examination should be combined with pathological examination to make a final diagnosis [11]. Doctors are required to have experience in endoscopy and image examination [8]. There is a severe shortage of endoscopists and large-scale endoscopy centers in poverty-stricken areas [12]. Therefore, endoscopy and imaging examinations are not suitable for large-scale screening.

As a minimally invasive method, liquid biopsy can provide information for cancer diagnosis, tumor monitoring, and clinical prognosis [13]. Here, blood is the most important body fluid [14], and plasma is excellent at reflecting gastrointestinal diseases [15]. Experts believe that GC tissue releases tumor markers into the peripheral blood, such as circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), tumor-associated RNA, protein, and circulating tumor cells (CTCs) [16]. With the progress of genomics and proteomics and the development of PCR detection technology, an increasing number of blood biomarkers have been found and applied to the diagnosis and prognosis of cancer [17]. Traditional serum markers, such as carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9), have been applied to diagnose GC [18].

Studies have shown that the comprehensive sensitivity of blood DNA methylation as a diagnostic marker of GC is 57% (95% CI, 50-63%) [19]. CtDNA is affected by tumor type and stage, which needs further study [13]. CfDNA is more sensitive than conventional tumor markers, but it is insufficient to distinguish other diseases, such as inflammatory diseases and infection [16]. MicroRNA (miRNA) [20], long noncoding RNA (lncRNA) [6], and circular RNA (circRNA) [21] play a key role in the occurrence and development of tumors. However, the chemical instability of RNA prevents it from becoming a good biomarker [19]. The count of circulating tumor cells is considered to be less than 5 in 7.5 ml of blood, which makes detection difficult [13]. The combined detection of multiple biomarkers can avoid the interference of blood mutation template molecules and reduce sampling deviation and individual differences [17], [22]. Various organs, tissues, blood exchange substances, and blood indices are in a state of dynamic balance. This balance is likely to be disrupted due to body diseases, resulting in abnormal blood indices [23]. Wu et al. showed that the neutrophil-lymphocyte ratio (NLR) and platelet-lymphocyte ratio (PLR) were significantly different between GC patients with different stages. The combination of PLR and CEA is better than CEA alone (AUC=0.671) for the diagnosis of gastric cancer (AUC=0.780) [10]. The deep learning method with excellent self-learning ability can deal with multidimensional nonlinear statistical relationships better than traditional methods [24]. The Clinical available blood indices based on deep learning provide a new idea for the differentiation of GC.

In this work, an identification model of GC was established based on blood indices through a deep learning algorithm. GC and other diseases could be effectively distinguished in this model.

II. Materials and Methods

A. Source of Materials

The total number of samples was 2951, including 1836 negative samples and 1115 positive samples. A total of 2682 samples were collected for cross-validation, and 269 samples were collected for external validation. The positive samples were from GC patients, while the negative samples were from 559 healthy people, 229 gastric ulcer patients, 158 gastric polyps patients, 174 gastritis patients, 124 liver cancer patients, 203 lung cancer patients, 124 pancreatic cancer patients, 136 breast cancer patients, and 129 colon cancer patients. To avoid diagnostic errors, the diagnostic results of the above samples were confirmed by several experts. Each sample contains age, gender, and 26 routine blood indices from Sysmex XE-5000 and 32 biochemical

### Table I. Details of the cross-validation sets and external validation sets

| Sample Number of CV set | Sample Number of EV set | Age range of CV set (mean ±SD) | Age range of EV set (mean ±SD) | Male/Female of CV set | Male/Female of EV set |
|-------------------------|------------------------|-------------------------------|--------------------------------|-----------------------|-----------------------|
| GC                      | 1014                   | 16-86(56.2±11.5)              | 27-82(55.5±11.1)               | 739/275               | 78/23                 |
| Healthy                 | 508                    | 0-88(46.0±13.7)               | 24-76(46.4±13.4)               | 267/241               | 29/22                 |
| Gastric ulcer           | 208                    | 2-83(55.7±14.9)               | 16-73(52.4±13.8)               | 140/68                | 18/3                  |
| Gastric polyps          | 144                    | 15-81(56.7±12.2)              | 34-79(49.9±15.6)               | 54/90                 | 8/6                   |
| Gastritis               | 158                    | 10-90(55.5±14.7)              | 38-76(58.0±12.1)               | 72/86                 | 4/12                  |
| Liver cancer            | 113                    | 19-86(56.0±10.9)              | 51-67(56.6±5.6)                | 78/35                 | 9/2                   |
| Lung cancer             | 185                    | 27-81(59.5±11.0)              | 30-78(59.3±11.9)               | 137/48                | 11/7                  |
| Pancreatic cancer       | 112                    | 26-87(60.3±12.1)              | 35-73(56.7±11.6)               | 59/53                 | 9/3                   |
| Breast cancer           | 123                    | 24-80(49.4±11.3)              | 36-52(46.7±4.8)                | 3/120                 | 0/13                  |
| Gastric ulcer           | 208                    | 2-83(55.7±14.9)               | 16-73(52.4±13.8)               | 140/68                | 18/3                  |
TABLE II. NEURAL NETWORK HYPERPARAMETER SETTINGS

| Hyperparameters          | Adjustment method                                      | Hyperparameter settings |
|--------------------------|--------------------------------------------------------|-------------------------|
| input dropout ratio      | adjusted from 0 to 0.5 with a step of 0.1              | 0.2                     |
| activation               | Tanh; TanhWithDropout; Rectifier; RectifierWithDropout; Maxout; MaxoutWithDropout |                        |
| initial weight distribution | UniformAdaptive; Uniform; Normal                      | UniformAdaptive         |
| loss                     | Automatic; CrossEntropy; Quadratic                    | Quadratic               |
| distribution             | AUTO; Bernoulli; Multinomial                           | Bernoulli               |
| epochs                   | adjusted from 500 to 10000 with a step of 100          | 1000                    |
| hidden layer 1           | adjusted from 30 to 120 with a step of 10              | 80                      |
| hidden layer 2           | adjusted from 30 to 120 with a step of 10              | 80                      |
| hidden layer 3           | adjusted from 30 to 120 with a step of 10              | 110                     |

Figure 1. Changes in AUC and logloss values for the training set and tenfold cross-validation on the training set as the number of hidden layers increased. (a) The AUC of the training set increased first and then flattened, while the AUC of the tenfold cross-validation first increased and then oscillated. (b) The logloss of the training set first decreased and then oscillated, and the logloss of the tenfold cross-validation tended to increase.

indices from Ortho VITROS 5600. The details of the dataset are listed in Table I, and detailed information can be shown in supplementary table S1. This study met the ethical requirements with the consent of all patients and healthy volunteers and was reviewed and approved by the ethics committee of the Second Hospital of Lanzhou University.

B. Machine Learning Method

Deep learning utilizes multiple neural network layers to obtain more important information from multidimensional data [8]. The artificial neural network can effectively reflect the nonlinear process of tumor development and metastasis [24]. The H2O framework includes an advanced artificial neural network [25]. Deep learning algorithms are embedded in the H2O framework. The deep learning algorithm is similar to the classical multilayer perceptron (MLP) [26], which is optimized through continuous iteration. Stochastic gradient descent was used for training the model, and backpropagation was used for optimization. The performance of the identification model can be further optimized by adjusting the hyperparameters of the neural network.

The principle of layered sampling was followed to divide the total data into the training set (2682 samples) and test set (269 samples). The initial H2O classification prediction model was evaluated by tenfold cross-validation on the training set, and the test set was used for external validation and did not participate in the construction of the model. The first was the choice of the number of hidden layers. In the initial experiment, we set the number of neurons in each hidden layer to 50 and then increased the number of hidden layers to obtain the AUC and logloss of the training set and tenfold cross-validation, as shown in Fig. 1. As shown in Fig. 1a, with the increase in the hidden layer, the AUC of the tenfold cross-validation increased first and then oscillated. The logloss of tenfold cross-validation tended to increase as the hidden layer increased, as shown in Fig. 1b. Experiments showed that as the number of hidden layers
increased, the performance of the model improved, but tenfold cross-validation showed that the model tended to fall into overfitting. To avoid overfitting and underfitting, three hidden layers were selected as the basic structure of the neural network. Then, the grid search method was applied to adjust the hyperparameters of the h2o.deeping function, including “input dropout ratio”, “activation”, “initial weight distribution”, “loss”, “distribution”, etc. Multiple hyperparameters were combined to determine the best solution. The specific settings are shown in Table II. Finally, the number of iterations and the size of the hidden layers were adjusted. The final iterations were set to 1000, and the three hidden layers were 80, 80, and 110. At first, all the features, including age, gender, 26 routine blood indices and 32 biochemical indices were used as the input layer of the neural network, constructing model-1. To simplify the blood detection process and reduce the model noise, the built-in function of the H2O package (h2o.varimp) was applied to calculate the importance of indices. The value of the calculated importance were listed in table III as importance percentage. Then, the number of indices was reduced to establish the final model (model-2).

The deep learning neural network was executed by the H2O package (version 3.32.0.1) in R (version 3.2.4). The H2O package is a parallel machine learning package that provides fast, scalable machine learning algorithms.

C. Assessment Method

The generalization capability of the model can be evaluated by cross-validation, thereby avoiding excessive fit [27]. Tenfold cross-validation divided the training set into 10 parts, 9 of which were used to build the model, and the remaining one was used as the internal test set to verify the performance of the model. Ten models were established by repeating 10 times to detect the accuracy and reliability of the model. The external test set did not participate in the establishment of the model and was only for verifying the performance of the model.

The visual effect of the model is presented with a confusion matrix [28]. Sensitivity (Sens, (1)), specificity (Spec, (2)), and accuracy (ACC, (3)) were calculated by true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN).

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

(1)
### Table III. Ranking of Important Features and Results of Correlation Test

| Rank | Features (Abbreviation) | Importance percentage | $P$-value* | $P$ value† | $P$ value‡ | $P$ value§ |
|------|-------------------------|-----------------------|-----------|-----------|-----------|-----------|
| 1    | Creatine kinase isoenzymes (CK-MB) | 0.0494 | 1.89E-14 | 1.70E-01 | 2.39E-02 | 7.99E-47 |
| 2    | Magnesium (Mg) | 0.0469 | 8.59E-01 | 4.63E-05 | 1.02E-06 | 1.98E-01 |
| 3    | Immature granulocytes (IG#) | 0.0452 | 8.89E-19 | 9.04E-29 | 1.67E-92 | 1.91E-11 |
| 4    | Immature granulocyte ratio (IG%) | 0.0431 | 1.20E-31 | 5.21E-01 | 3.37E-76 | 3.82E-08 |
| 5    | Plateletcrit (PCT) | 0.0390 | 1.72E-08 | 8.61E-24 | 3.00E-01 | 2.28E-02 |
| 6    | Potassium (K) | 0.0384 | 5.33E-03 | 5.00E-53 | 2.58E-03 | 4.60E-07 |
| 7    | Gamma Glutamyltransferase (γ-GGT) | 0.0288 | 3.00E-38 | 4.84E-27 | 1.17E-46 | 7.75E-04 |
| 8    | Basophil count (BASO#) | 0.0251 | 9.06E-38 | 1.45E-17 | 4.98E-01 | 3.24E-03 |
| 9    | Calcium (Ca) | 0.0232 | 7.74E-11 | 5.30E-73 | 3.90E-05 | 1.43E-02 |
| 10   | Monocyteratio (MO%) | 0.0223 | 1.08E-26 | 2.65E-189 | 8.42E-01 | 5.52E-04 |
| 11   | Basophil ratio (BASO%) | 0.0197 | 3.07E-03 | 7.08E-194 | 1.16E-04 | 5.92E-02 |
| 12   | Low-density lipoprotein (LDL) | 0.0196 | 1.07E-16 | 1.66E-01 | 4.90E-01 | 1.65E-07 |
| 13   | Red blood cell distribution width-CV (RDW-CV) | 0.0188 | 1.27E-49 | 2.18E-73 | 1.10E-05 | 1.24E-34 |
| 14   | Creatine kinase (CK) | 0.0185 | 5.64E-22 | 7.35E-53 | 2.50E-01 | 9.63E-11 |
| 15   | Blood urea nitrogen/Creatinine (BUN/Cre) | 0.0184 | 2.00E-10 | 4.28E-21 | 3.27E-03 | 1.72E-02 |
| 16   | Total bilirubin (TBil) | 0.0167 | 9.96E-01 | 2.51E-01 | 2.47E-02 | 1.29E-01 |
| 17   | Mean platelet volume (MPV) | 0.0165 | 1.98E-01 | 6.21E-02 | 7.67E-01 | 1.12E-07 |
| 18   | Alkaline phosphatase (ALP) | 0.0161 | 1.80E-01 | 2.72E-04 | 1.00E-14 | 2.21E-02 |
| 19   | Basophil ratio (BASO%) | 0.0151 | 2.97E-40 | 8.43E-194 | 1.04E-04 | 6.44E-02 |
| 20   | Eosinophil transaminase (ALT) | 0.0150 | 9.53E-11 | 2.88E-05 | 9.16E-11 | 2.02E-04 |
| 21   | Triglyceride (TG) | 0.0147 | 1.30E-15 | 4.86E-21 | 3.24E-01 | 5.27E-18 |
| 22   | Gender | 0.0143 | 1.95E-28 | 1.55E-16 | 8.61E-20 | 7.77E-17 |
| 23   | Aspartateaminotransferase (AST) | 0.0140 | 3.39E-01 | 1.11E-03 | 1.42E-08 | 2.51E-01 |
| 24   | High-density lipoprotein (HDL) | 0.0137 | 3.61E-01 | 1.25E-15 | 2.12E-06 | 8.52E-01 |
| 25   | Red blood cell distribution width-SD (RDW-SD) | 0.0133 | 3.17E-62 | 2.16E-112 | 6.91E-15 | 1.16E-16 |
| 26   | Eosinophil cells (EO) | 0.0133 | 1.08E-02 | 7.06E-08 | 1.29E-02 | 2.35E-04 |
| 27   | Large platelet ratio (P-LCR) | 0.0131 | 1.11E-01 | 1.19E-13 | 2.74E-04 | 3.47E-15 |
| 28   | Indirect bilirubin (I-Bil) | 0.0129 | 2.16E-01 | 7.65E-01 | 9.91E-01 | 7.28E-03 |
| 29   | Albumin/Globulin (ALB/GLB) | 0.0126 | 1.48E-05 | 4.76E-19 | 9.41E-48 | 3.14E-02 |

Note: The statistical significant difference of features between gastric cancer (positive sample group) and all negative sample group, including three sub-groups (normal group, other cancer group and other gastric disease group). The statistical significant difference of features between gastric cancer group and normal sub-group. The statistical significant difference of features between gastric cancer group and other cancer sub-group. The statistical significant difference of features between gastric cancer group and other gastric disease sub-group. The detailed statistical method about how the significant differences were calculated was shown in supplementary table S2.

The receiver operating characteristic curve (ROC) was drawn by using sensitivity as the y-axis and 1-specificity as the x-axis. The area under the curve (AUC) was obtained by calculating the area under the ROC curve. The closer the AUC is to 1, the better classification of the model.

### III. Results
A. Good Identification Performance

Model-1 is based on 60 features, with excellent classification and prediction performance. For the external test set, the AUC of model-1 was 0.9152, the sensitivity was 80.20%, the specificity was 91.07%, and the accuracy was 86.99%. The ROC curve of the external test set is shown in Fig. 2a.

As shown in Fig. 3, with the increase in features involved in modeling, the performance of the model increases. Finally, we selected 33 features to build the final model (model-2). Model-2 has good performance, with a sensitivity, specificity, total accuracy and area under the curve of 85.44%, 83.82%, 84.54% and 0.9165 for the cross-validation set, respectively. The selected features are shown in Table III. For the external test set, the sensitivity, specificity, and accuracy of model-2 were 85.15%, 81.55%, and 82.90%, respectively. The ROC curve of the test set is shown in Fig. 2b, and the AUC was 0.9126.
TABLE IV. PERFORMANCE COMPARISON OF THIS WORK WITH OTHER METHODS FOR IDENTIFYING GC. (N/A: NOT APPLICABLE)

| Identification method | Sample number | Sample type       | Sens (%) | Spec (%) | AUC   |
|-----------------------|---------------|-------------------|----------|----------|-------|
| This work (cross-validation set) | 2682          | Blood indices     | 85.44    | 83.82    | 0.9165|
| This work (testing set)   | 269           | Blood indices     | 85.15    | 81.55    | 0.9126|
| Our previous work [29]  | 322           | Blood indices     | 82.52    | 86.30    | 0.9092|
| spectral-spatial [9]    | 120           | Gastric tissue    | 96.6     | 96.7     | N/A   |
| hsa_circ_0005758 (circRNA) [30] | 263         | Gastric tissue    | 75.0     | 67.7     | 0.721 |
| LINC00152 (IncRNA) [31] | 142           | Gastric tissue    | 62.5     | 68.1     | 0.645 |
| pepsinogen [32]        | 3785          | Serum             | 69       | 73       | 0.76  |
| SERS [33]              | 98            | Urine             | 90       | 93.8     | 0.981 |
| miR-106a (MiRNA) [20]  | 141           | Gastric juice     | 73.8     | 89.3     | 0.871 |
| DNA methylation [19]   | 4172          | Plasma            | 71       | 89       | 0.87  |
| DNA methylation [19]   | 4172          | Serum             | 50       | 98       | 0.91  |
| hsa_circ_0001017 (circRNA) [15] | 242         | Plasma            | 67.6     | 89.7     | 0.851 |
| hsa_circ_0000181 (circRNA) [18] | 207         | Plasma            | 99.0     | 20.6     | 0.582 |
| Protein P08493 [5]     | 30            | Plasma            | 73.3     | 80.0     | 0.813 |
| MYC (cfDNA) [34]       | 136           | Plasma            | 75.44    | 75.95    | 0.8155|
| miR-20a (MiRNA) [35]   | 120           | Plasma            | 75       | 88       | 0.8593|
| Blood routine, CEA and CA [36] | 213         | Whole blood       | 87.0     | 84.1     | 0.86  |
| mass spectral [37]     | 150           | Serum             | 85.3     | 88.0     | N/A   |

B. User-Oriented Online Diagnostic Tool

We have designed an online identification website for patients and medical workers at http://www.cppdd.cn/GC2. The user can input the corresponding blood test data in the text box according to the prompt and obtain the corresponding identification results. The homepage of this website is shown in Fig. 4. Users should use the same equipment as the blood testing equipment used in this study. The system deviation caused by using other blood testing equipment might affect the accuracy of this identification.

IV. Discussion

A. Advantages of This Study

A comprehensive comparison of previous works and our method is given in Table IV. In general, the identification method for GC presented in this study is suitable for the screening of GC on a large scale compared with previous methods, especially in regions that lack medical resources, with the advantages listed below.

First, the blood indices on which this method were built were all collected from a Clinical available blood biochemical detector. These data are inexpensive and easy to obtain. Traditional methods for the detection of gastric cancer always rely on the detection of characteristics in gastric tissue. For example, Lu et al. detected the content of hsa_circ_0005758 (circRNA) in gastric tissue, and the sensitivity and specificity were 75.0% and 67.7%, respectively [30]. Pang et al. detected the content of LINC00152 (IncRNA) in gastric tissue; the sensitivity was 62.5%, and the specificity was 68.1% [31]. In addition, spectral-based methods on tissue images were also discovered and achieved outstanding results. For example, Li et al. analyzed the spectral characteristics of gastric mucosa tissue to diagnose GC through a deep learning method and achieved good results, with a specificity of 96.7% and a sensitivity of 96.6% [9]. The above studies using gastric tissue require experienced gastroscopic doctors to accurately extract gastric mucosa samples, which is not suitable for screening GC on a large scale. The sensitivity of the traditional GC biomarker pepsinogen was 69% [32]. Lin et al. used urinary surface-enhanced Raman spectroscopy (SERS) to diagnose GC from healthy sample based on gold nanoparticles, with a sensitivity of 90% and specificity of 93.8% [33]. However, Lin's study was not efficient at distinguishing between GC and breast cancer, with specificity of 81.4% and a lower sensitivity of 62.0%. Cui et al. measured the content of mir-106a in gastric juice samples, and the sensitivity and specificity were 73.8% and 89.3%, respectively [20]. The sampling of gastric juice is more complicated than that of blood. Research on blood markers is extensive. Hu et al. conducted a meta-analysis of DNA methylation in serum and plasma, and the results showed high specificity and low sensitivity [19]. Li et al. and Zhao et al. analyzed the content of circular RNA in plasma, and the results were not ideal [15], [18]. The use of plasma markers, such as protein P08493, MYC (cfDNA), and miR-20a (MiRNA), did not achieve good results [5], [34], [35].
Comprehensive multiple indices can achieve better results in the diagnosis of GC. Similar to this work, Zhu et al. used the gradient boosting decision tree (GBDT) method to analyze clinical blood indices, such as hemoglobin, and biomarkers, such as carcinoembryonic antigen (CEA), to establish a gastric cancer identification model and achieved good results [36]. Su et al. used machine learning to analyze the mass spectra of various proteins in serum and obtained good results [37]. Mass spectrometers are costly and have high requirements for operators, so this method is not suitable for large-scale screening.

Second, this method was developed based on a large and comprehensive dataset, including not only gastric cancer samples and healthy samples but also other major cancer samples and other gastric diseases with symptoms similar to gastric cancer. These samples included samples that are likely to be confused in mass screening of GC. Most of the
previous work was only deduced on a small sample size [9, [33], [37]. Only a few works were studied on a large dataset with thousands of samples, as in this work, such as in works by Hu et al. [19] and Huang et al. [32].

Third, this method presents a better identification performance than previous works, especially for the test set, which implies that this method has a better generalization ability and stronger robustness, even than our previous work [29], with a better AUC value.

B. Analysis of Key Features

Cancer is associated with dysregulation of multiple biological processes, and the analysis of key features facilitates the discovery and validation of biomarkers that contribute to understanding cancer pathogenesis and developing drugs [38]. To discover insights into key features in different populations, all samples were categorized into four population groups: the GC group, normal group, other cancer group, and other gastric disease group. We selected features with significant differences and drew a boxplot between groups, as shown in Fig. 5. In Fig. 5a, the levels of PCT in the normal group were higher than those in the GC group. Current studies have shown that tumor cells can indirectly promote platelet production and activation. Activated platelets can also protect and even promote the growth and metastasis of tumor cells [39], [40]. Our statistics may show that the hematopoietic function of bone marrow in some patients with gastric cancer is affected by chemotherapy drugs or radiotherapy, resulting in the suppression of hematopoietic function and thrombocytopenia [41], [42].

As shown in Fig. 5b, the levels of potassium in the normal group were higher than those in other groups. Cancer and gastric diseases may lead to disordered homeostasis and potassium ion balance. Wu et al. showed that the potassium content in GC tissue was higher than that in normal tissue [43]. Ding et al. proved that the Eag1 potassium channel was overexpressed in GC tissue [44]. In this experiment, the low levels of potassium in the blood of GC patients may be related to this, which requires further study.

In Fig. 5c, the basophil count of the normal group is shown to be higher and more concentrated than that of the other groups, which is also confirmed by the higher basophil ratio in Fig. 5d. Tumor-infiltrating basophils are considered to be a poor prognostic factor for GC, but there is no significant correlation between blood basophils and tumor-infiltrating basophils [45]. The higher percentage of basophils in the normal group may be due to the increase in white blood cells in the other groups.

As shown in Fig. 5e, the calcium content in the normal group was higher than that in the GC group. Studies have shown that calcium may inhibit the damage of salt to gastric mucosa, thereby reducing the risk of GC [46]. However, the work of Xie et al. indicated that serum calcium concentration was positively correlated with the expression of calcium-sensing receptor (CaSR). Calcium can activate the overexpression of CaSR to promote the proliferation of GC cells [47]. These contradictory conclusions need further study.

Monocytes (Mo) are believed to promote tumor growth and proliferation [48]. High levels of monocytes are thought to confer poor prognosis in GC patients [49]. Figure 5f shows that the Mo% of the normal group is significantly higher than that of the other groups, but there is no significant difference in the Mo content among the groups. The results show that the count of white blood cells in all groups except the normal group increased significantly.

A low concentration of creatine kinase (CK) in GC tissue has been demonstrated [50]. This is similar to the results shown in Fig. 5g. CK is very important for ATP homeostasis in cells. Abnormal CK levels may lead to apoptosis [51].

The increase in eosinophils in the tissue can improve the prognosis of patients with GC. It has been proven that eosinophils have a good anticancer effect on colorectal cancer [52]. In Fig. 5h, the EO% of the normal group is significantly lower than that of the GC group. Eosinophils in cancer patients may increase due to autoimmune responses.

The red blood cell distribution width-SD (RDW-SD) in the normal group was lower than that in the other groups, as shown in Fig. 5i. Wei et al. showed that the red blood cell distribution width of GC patients is significantly higher than that of the normal group and further increases with the development of tumors. This phenomenon may be related to the decrease in hemoglobin caused by malnutrition in GC patients and may be related to inflammation [53]. Chang et al. showed that the ratio of CK-MB to total CK in hematological malignancies was higher in patients with colorectal cancer, lung cancer, and hepatocellular carcinoma [54]. This conclusion is roughly the same as the distribution of CK-MB and CK in this study and provides a direction for distinguishing patients with gastric diseases from patients with gastric cancer.

One should note that there were some features that are not statistically significant (with high p-value) between gastric cancer sample group and whole negative sample group, such as Mg and DBil in table III, but they were at least had significant difference (p-value <0.05) with one of the 3 negative sub-groups (normal group, other cancer group and other gastric cancer group). This is just the reason why that deep learning is utilized here. Deep learning methods with excellent self-learning ability can handle multi-dimensional nonlinear statistical relationships better than traditional statistical methods [24]. Therefore, these features were also included as key features in the final model since their relationship with gastric cancer is in a multi-dimensional nonlinear way instead of linear statistical way.

VII. CONCLUSION

We utilized an artificial neural network under the H2O framework method to process routine blood data and blood biochemical data, extracted 33 important indices as
biomarkers for the identification of GC, and established an identification model and user-friendly web testing terminal. These biomarkers are considered to have great value in the physiological research of GC. Compared with conventional diagnostic methods, the diagnostic model established in this study has the advantages of higher accuracy, noninvasiveness, and inexpensive detection. It has the potential to screen a wide range of people and effectively reduces the pressure on the medical system. With the further collection and learning of blood test data, the performance of this diagnostic model can be further improved.

APPENDIX

The original data of the samples for modeling and validation are provided in supplementary table S1. The detailed statistical method and results for difference of selected features between gastric cancer group and other subgroups are provided in supplementary table S2. The box plots for all 33 selected features are provided in supplementary file S3.

REFERENCES

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," (in English), CA-Cancer J. Clin., Article vol. 71, no. 3, pp. 209-249, May 2021.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," (in English), CA-Cancer J. Clin., Article vol. 68, no. 6, pp. 394-424, Nov-Dec 2018.

[3] Z. H. Zhang, M. M. Dou, X. F. Yao, H. Tang, Z. B. Li, and X. Y. Zhao, "Potential Biomarkers in Diagnosis of Human Gastric Cancer," Cancer Invest., vol. 34, no. 3, pp. 115-22, 2016.

[4] E. Van Cutsem, X. Sagarra, B. Topal, K. Haustermans, and H. Prenen, "Gastric cancer," Lancet, vol. 388, no. 10060, pp. 2654-2664, 2016.

[5] B. Zhou et al., "Plasma proteomics-based identification of novel biomarkers in early gastric cancer," Clin Biochem, vol. 76, pp. 5-10, Feb 2020.

[6] E. Delshad, F. T. Shamsabadi, S. Bahramian, F. Mehravar, H. Maghsoudi, and M. Shafiee, "In silico identification of novel IncRNAs with a potential role in diagnosis of gastric cancer," J Biomol Struct Dyn, vol. 38, no. 7, pp. 1954-1962, Apr 2020.

[7] Y. Huang et al., "Serum microRNA panel excavated by machine learning as a potential biomarker for the detection of gastric cancer," Oncol Rep, vol. 39, no. 3, pp. 1338-1346, Mar 2018.

[8] P. H. Niu, L. L. Zhao, H. L. Wu, D. B. Zhao, and Y. T. Chen, "Artificial intelligence in gastric cancer: Application and future perspectives," (in English), World J. Gastroenterol., Review vol. 26, no. 36, pp. 5408-5419, Sep 2020.

[9] Y. P. Li et al., "Early diagnosis of gastric cancer based on deep learning combined with the spectral-spatial classification method," Biomedical Optics Express, vol. 10, no. 10, pp. 4999-5014, Oct 2019.

[10] Y. Y. Wu, M. Q. Jiang, Y. Y. Qin, F. Q. Lin, and M. Y. Lai, "Single and combined use of neutrophil-lymphocyte ratio, platelet-lymphocyte ratio and carcinoembryonic antigen in diagnosing gastric cancer," Clin Chim Acta, vol. 481, pp. 20-24, Jun 2018.

[11] W. Ge, L. M. Zheng, and G. Chen, "The Combination of Seven Preoperative Markers for Predicting Patients with Gastric Cancer to Be Either Stage IV or Non-Stage IV," Gastroenterol Res Pract, vol. 2018, p. 3450981, 2018.

[12] L. M. Zhang, Y. Zhang, L. Wang, J. Y. Wang, and Y. L. Liu, "Diagnosis of gastric lesions through a deep convolutional neural network," Dig Endosc, Sep 22 2020.

[13] C. T. Uchoa Guimaraes et al., "Liquid biopsy provides new insights into gastric cancer," Oncotarget, vol. 9, no. 19, pp. 15144-15156, 2018-Mar-13 2018.

[14] M. Bahreini, A. Hosseinizadegan, A. Rashidi, S. R. Miri, H. R. Mirzaei, and P. Hajian, "A Raman-based serum constituents' analysis for gastric cancer diagnosis: In vitro study," Talanta, vol. 204, pp. 826-832, Nov 1 2019.

[15] T. W. Li et al., "Plasma circular RNA profiling of patients with gastric cancer and their droplet digital RT-PCR detection," J Mol Med (Berl), vol. 96, no. 1, pp. 85-96, Jan 2018.

[16] C. T. Liu, C. Q. Hong, X. C. Huang, E. M. Li, Y. W. Xu, and Y. H. Peng, "Blood-based Markers in the Prognostic Prediction of Esophageogastric Junction Cancer," J Cancer, vol. 11, no. 15, pp. 4332-4342, 2020.

[17] Z. B. Huang, H. T. Zhang, B. Yu, and D. H. Yu, "Cell-free DNA as a liquid biopsy for early detection of gastric cancer," Oncol Lett, vol. 21, no. 1, p. 3, Jan 2021.

[18] Q. F. Zhao, S. J. Chen, T. W. Li, B. X. Xiao, and X. J. Zhang, "Clinical values of circular RNA 6000181 in the screening of gastric cancer," J Clin Lab Anal, vol. 32, no. 4, p. e22233, May 2018.

[19] W. L. Hu et al., "Diagnostic accuracy of DNA methylation in detection of gastric cancer: a meta-analysis," (in English), Oncotarget, Article vol. 8, no. 68, pp. 113142-113152, Dec 2017.

[20] L. Cui et al., "Gastric Juice MicroRNAs as Potential Biomarkers for the Screening of Gastric Cancer," (in English), Cancer, Article vol. 119, no. 9, pp. 1618-1626, May 2013.

[21] Y. Xie et al., "Downregulated expression of hsa_circ_0074362 in gastric cancer and its potential diagnostic values," (in English), Biomark Med, vol. 12, no. 1, pp. 11-20, Jan 2018.

[22] H. L. Sun et al., "A nomogram based on serum bilirubin and albumin levels predicts survival in gastric cancer patients," (in English), Oncotarget, vol. 8, no. 25, pp. 41305-41318, Jun 20 2017.

[23] P. A. Weiss and P. A. Weiss, "Can cancer be diagnosed with a blood test during routine examinations?", Clinical journal of oncology nursing, vol. 11, no. 6, pp. 940-1, 2007 2007.

[24] S. J. Que et al., "Application of preoperative artificial neural network based on blood biomarkers and clinicopathological parameters for predicting long-term survival of patients with gastric cancer," World J. Gastroenterol., vol. 25, no. 43, pp. 6451-6464, 2019.

[25] A. S. Elsayad, A. I. El Desouky, M. M. Salem, and M. Badawy, "A Deep Learning H2O Framework for Emergency Prediction in Biomedical Big Data," Ieee Access, vol. 8, pp. 97231-97242, 2020.

[26] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," IEEE transactions on neural networks, vol. 3, no. 5, pp. 683-97, 1992 1992.

[27] G. X. Jiang and W. J. Wang, "Error estimation based on variance analysis of k -fold cross-validation," Pattern Recognition, vol. 69, pp. 94-106, 2017.

[28] I. Duntsch and G. Gediga, "Indices for rough set approximation and the application to confusion matrices," International Journal of Approximate Reasoning, vol. 118, pp. 155-172, Mar 2020.

[29] J. P. Wu et al., "GCdiscrimination: identification of gastric cancer based on a milliliter of blood," Brief Bioinform, vol. 22, no. 1, p. 536-544, Jan 18 2021.

[30] R. D. Lu, Y. F. Shao, X. P. Tao, G. L. Ye, B. X. Xiao, and J. M. Guo, "Clinical significances of hsa_circ_0067582 and hsa_circ_0005758 in gastric cancer tissues," J Clin Lab Anal, vol. 33, no. 9, p. e22984, Nov 2019.

[31] Q. Q. Pang et al., "Increased expression of long intergenic non-coding RNA LINC00152 in gastric cancer and its clinical significance," (in English), Tumor Biol., Article vol. 35, no. 6, pp. 5441-5444, Jun 2014.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3172477, IEEE Access
**Tiantian Shen** is currently pursuing the M.D. degree in the First School of Clinical Medical, Lanzhou University. Her current research interests include virtual drug screening and biological activity verification.

**Lili Xi** received the Ph.D. degrees from the School of Chemistry and Chemical Engineering of Lanzhou University in June 2010. Since July 2010, she has been working in the Institution of Drug Clinical Trial of the First Hospital of Lanzhou University. Her current research interests include computer-aided drug molecular design and computer-aided diagnosis models.

**Dekui Zhang** is currently a Chief with the Department of Gastroenterology, Lanzhou University Second Hospital. His current research interests include diagnosis and treatment of esophagus and gastropathy, inflammatory bowel disease and digestive tract tumors, and endoscopic minimally invasive treatment of digestive diseases.

**Shuyan Li** is currently a Professor in College of Medical Information and Engineering, Xuzhou Medical University. Her current research interests include application of artificial intelligence in clinical diagnosis systems and computer-aided drug design.