Employment of Artificial Intelligence Based on Routine Laboratory Results for the Early Diagnosis of Multiple Myeloma

Wei Yan[1], Hua Shi[1], Tao He[2], Jian Chen[2], Chen Wang[2], Aijun Liao[1], Wei Yang[1] and Huihan Wang[1]*

[1] Haematology Department of Shengjing Hospital, China Medical University, Shenyang, China, [2] Neusoft Research Institute, Northeastern University, Shenyang, China

Objective: In order to enhance the detection rate of multiple myeloma and execute an early and more precise disease management, an artificial intelligence assistant diagnosis system is developed.

Methods: 4,187 routine blood and biochemical examination records were collected from Shengjing Hospital affiliated to China Medical University from January 2010 to January 2020, which include 1,741 records of multiple myeloma (MM) and 2,446 records of non-myeloma (infectious diseases, rheumatic immune system diseases, hepatic diseases and renal diseases). The data set was split into training and test subsets with the ratio of 4:1 while connecting hemoglobin, serum creatinine, serum calcium, immunoglobulin (A, G and M), albumin, total protein, and the ratio of albumin to globulin data. An early assistant diagnostic model of MM was established by Gradient Boosting Decision Tree (GBDT), Support Vector Machine (SVM), Deep Neural Networks (DNN), and Random Forest (RF). Out team calculated the precision and recall of the system. The performance of the diagnostic model was evaluated by using the receiver operating characteristic (ROC) curve.

Results: By designing the features properly, the typical machine learning algorithms SVM, DNN, RF and GBDT all performed well. GBDT had the highest precision (92.9%), recall (90.0%) and F1 score (0.915) for the myeloma group. The maximized area under the ROC (AUROC) was calculated, and the results of GBDT (AUC: 0.975; 95% confidence interval (CI): 0.963–0.986) outperformed that of SVM, DNN and RF.

Conclusion: The model established by artificial intelligence derived from routine laboratory results can accurately diagnose MM, which can boost the rate of early diagnosis.

Keywords: multiple myeloma, artificial intelligence, early diagnosis, gradient boosting decision tree, machine learning

INTRODUCTION

As a hematological malignancy, multiple myeloma (MM) accounts for 1% of all cancer and 13% of hematological tumors with the characteristics of proliferation of malignant plasma cells in the bone marrow (BM), presence of anemia, renal dysfunction, hypercalcemia, and lytic lesions (1). The involvement of other disciplines such as orthopedics, nephrology and hematology often cause misdiagnoses (2). In addition, due to the sub-par distribution of state-of-the-art medical and diagnostic equipment, rural health centers and primary care providers register a high rate of misdiagnoses and missed diagnoses. Howell et al. reported that the time from the appearance of first symptoms to the first instance of seeking medical help ranged from 1 to 7 months, and the time from help-seeking to diagnosis ranged from 2 weeks to 17 months in MM patients. Patients reported between one and ten primary care consultations with what they considered (in hindsight) to be myeloma symptoms, before the referral leading to diagnosis (3). The delay in diagnosis will deprive the patient of the optimal opportunity for treatment and can lead to the development of complications which can only at times be reversed. Increased tumor burden, symptoms, and organ damage all affect the treatment results and the capacity for myeloma patients to receive treatment (4). Improving the time to diagnosis of MM is a sine qua non condition to fulfill to give patients a fair chance of recovery especially in community hospitals and primary care clinics.

Being human-made, artificial intelligence (AI) can simulate intellectual work such as humans' thoughts and judgments and has thus revolutionized the medical field (5). Hence, there is an increasing attention on the application of AI for the diagnosis and treatment of cancer (6, 7). In terms of settling the problems of classification and regression, the gradient boosting decision tree (GBDT) is regarded as a powerful ensemble learning technique (8). This model outperforms other models as the direction of the negative gradient is followed in order to train the residuals of each iteration, which can avoid the over-fitting problem. Furthermore, the GBDT model has performed well for knowledge discovery in various fields (9, 10). The current study is the first time that the artificial intelligence technology, including GBDT, has been used constructing a multiple myeloma early screening model based on a large amount of clinical conventional examination data. With the help of AI technology, the knowledge and experience of authoritative experts will better benefit the public, and effectively improve the current diagnosis rate of myeloma in the areas short of experience, which is of very important clinical significance.

PATIENTS AND METHODS

The Medical Ethics Committee at the Shengjing Hospital of China Medical University approved the present study (2020PS055J) according to the principles of the Declaration of Helsinki. The requirement for personal informed consent has been waived by the Ethics Committee as electronic medical records are researched retrospectively.

Patient and Data Selection

In this retrospective research, our institutional databases were screened to investigate the patients admitted to our hospital for the first routine blood checks, hepatic function panel, renal function tests and immunoglobulin tests from January 2010 to January 2020. These included 1,741 records of multiple myeloma (MM) and 2,446 records of non-myeloma (infectious disease, rheumatic immune system disease, hepatic disease and renal disease). We also collected the data for these laboratory items from January 2020 to November 2020, including 68 records of newly diagnosed multiple myeloma (MM) and 70 records of non-myeloma aimed at testing the theory of generalizability. The diagnosis was made based on the 2014 International MM Working Group criteria (IMWG) (11). Nine variables (hemoglobin, serum creatinine, serum calcium, immunoglobulin (A, G and M), albumin, total protein, and ratio of albumin to globulin) were collected based on the use of current diagnostic criteria and medical judgment. Because immunoglobulin assays are not part of routine laboratory tests, we have therefore used six variables to try the alternative, cheaper model.

Data Processing

Based on the diagnostic criteria and doctor-assisted judgments, the related factors for Multiple Myeloma risk prediction have been determined, and the original data related to the prediction have been extracted from the HIS and LIS databases. After extracting the correlating factors, the original sample set could not be directly applied in training machine learning models as the sample set still required further preprocessing of data.

Handling the Missing Value in the Sample Set

The presence of empty values in the extracted raw data is first confirmed. A patient is eliminated from the pool when the number of missing values is larger than the designated threshold. We initially tested the number of missing values from 0–8. The results are depicted in **Supplementary Figure 1**. It was concluded that with the increase of the threshold value, F1 score increases followed by a decrease. This is due to the fact that fewer samples remain after data cleaning through the reduction of the number of the threshold, which results in the weakening of the generalization ability. However, when setting the threshold value as a sufficiently large value, more lost features are filled with normal values, resulting in the decrease of the F1 score. Considering these two factors, we finally set the threshold as 3 in our algorithm. If more than three factors are empty, the sample will be deleted, and if the missing value is below or equal to 3, this sample is deemed as fit to be retained. Through application of the same method, when six variables are considered, the threshold value is set to 2. Due to the fact that reserved data still contain missing values, its missing item will be filled with a normal value. With this missing data processing, we thus can reduce the possible deviation caused by using an abnormal value, and ensure there are no missing values in this data either for training or testing. We have implemented this

method as a specific data processing module in our designed learning model.

Expanding the Number of Positive Classes

Based on the real data extracted from the system, the number of positives is much lower than the negatives. The Synthetic Minority Oversampling Technique (SMOTE) algorithm can be used to fix this unbalanced classification problem, by producing synthetic examples to increase the number of positive classes (12). The SMOTE algorithm analyzes and simulates the minority samples, uses the k-nearest neighbor (KNN) algorithm to synthesize new minority samples and adds the synthesized new samples into the training data, which can expand the sample size (13). The following steps are carried out by the SMOTE algorithm to synthesize new samples: the nearest neighbor algorithm is used and the number of the nearest neighbors for each minority is calculated; random number of samples are selected to randomly implement linear interpolation, and construct new minority samples; finally, new samples are synthesized with original data to produce new training sets (**Figure 1**). Here we obtained 580 synthetic samples based on 1,741 myeloma samples with SMOTE algorithm. These 580 synthetic samples were further integrated with original positive samples for model construction and testing.

Expanding the Number of Correlation Features

Since the newly generated features can reflect the deviation degree of the detection items to its normal range, we thus utilize this feature as the part of the features to build the model. It should be noted that each index feature has a normal range in our system. Subsequently, the data standardization

method was expanded by subtracting the upper and lower limits with the normal value. Moreover, we applied the outlier detection method after standardization method to reduce some outlier values in the sample data set. The relationship between the testing instance and its normal reference range is used to carry out feature correlation and expansion. For example, we assume the detection value of this feature index i as d_i. Each index feature has a normal range $[L_d^i, U_d^i]$. By comparing the differences between d_i and L_d^i and between U_d^i, two novel features l_d^i and u_d^i are generated by
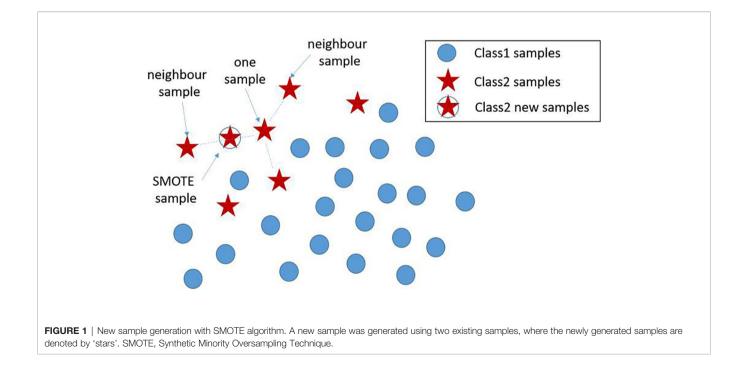
$$l_d^i = \{0, \ d_i > L_d^i d_i - l_d^i, \text{ otherwise} \tag{1}$$

$$u_d^i = \{0, \ d_i > L_d^i \ U_d^i - d_i, \text{ otherwise} \tag{2}$$

where l_d^i and u_d^i reflect the deviation of the detection value from the normal range. The greater the detection value from the normal range, the larger l_d^i or u_d^i will be obtained. By utilizing the l_d^i, u_d^i, l_d^i as the input features, more detailed and expanded features are thus obtained. Experiments show that this method can reduce the prediction variance of the model.

Building the Prediction Model

Ensemble learning is a popular paradigm employed to leverage the strength of individual classifiers and mitigate their weaknesses. Ensemble techniques consist of combining more than one single classifier under a specific combination rule to solve the same task (14). As a common algorithm for ensemble learning, GBDT is composed of Decision Tree and Gradient Boosting (15) Because this tree model is characterized by high bias, low variance and small depth, highly pruned version of CART tress is thus utilized as the base classifiers for GBDT in each iteration (16).

FIGURE 1 | New sample generation with SMOTE algorithm. A new sample was generated using two existing samples, where the newly generated samples are denoted by 'stars'. SMOTE, Synthetic Minority Oversampling Technique.

The aggregated classifier using the additive modeling structure is as follows (17):

$$\hat{y}_i = F(x_i) = \Sigma_{k=1}^{K} \gamma_k f_k(x_i) \qquad (3)$$

where $\hat{y}_i$ and $F(x_i)$ represent the predicted value of ith sample, γ_k represents the weight of the kth tree, $f_x(x_i)$ represents the prediction result of the kth regression tree for samples, x_i represents the independent values used in fitting each regression tree, K is the number of CART model trees.

For Binary Classification problems, logarithmic loss function, which is also called as the log-likelihood loss function is utilized as the loss function:
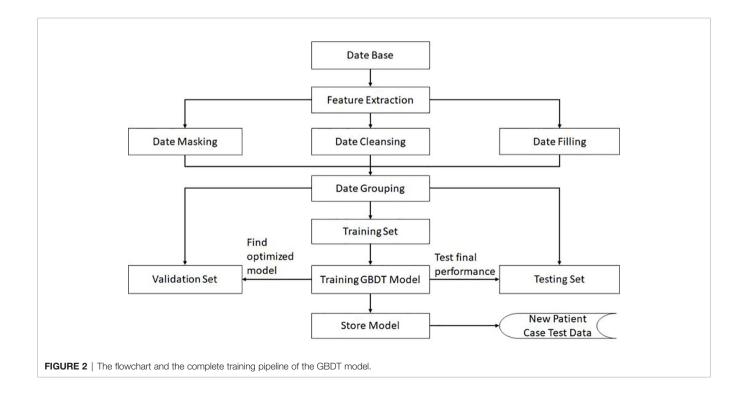
$$L(y, F(x)) = log(1 + exp(-2yF(x))) \qquad (4)$$

where $F(x)$ is given by Equation (3), and x is a generalization of x_i, y represents the true value of the sample.

With this loss function, the common applied gradient descent method is applied to find the optimal model. By calculating the negative gradient, we can obtain the moving direction brings has the steepest decline in the value of the loss function. The optimal model, through an iterative manner, can be found in this moving direction. With each iteration, the gradient descent method first calculates the negative gradient of the current model on all samples, and then trains a new base classifier with the value as the target for quasi merging, thus to calculate the weight of the base classifier. By utilizing this method iteratively, we finally realize the updating of the model.

In order to ensure the generalization ability of the model, the negative data and positive data are first mixed and shuffled, thus changing the original order. Then, using random extraction we obtain the training set and test set, which can ensure the independence of these two data sets. In our algorithm, the data volume ratio of these two data sets is 4:1. For the GBDT algorithm, the important super parameters include the maximum depth of the decision tree and the number of decision trees. The grid search method is then applied on the validation set, and the calculated optimal number of decision trees is 81, while the maximum depth for the decision trees is 6. All these performance results are obtained in the test data set. The complete training pipeline is demonstrated in **Figure 2**.

Moreover, Support Vector Machine (SVM) (13), Deep Neural Networks (DNN) (18), and Random forest (RF) (19) were also applied for performance comparisons. These three algorithms used the same training set and test set. For SVM algorithm, the Gaussian kernel function was utilized and the "gamma" parameter was set to 1. Gamma is a regularized super parameter. A larger gamma value denotes a more irregular decision boundary. Concurrently, a smaller gamma denotes a smoother boundary. Consequentially, we found that the gamma value needs to be adjusted when the model is over-fitted or otherwise (13). For DNN, we found that the model with more than four hidden layers seemed to be over fitted, and two hidden layers would not fit well. Therefore, a network with three hidden layers was constructed, where each layer contained 256 neurons and the ReLU activation function was applied. For random forest algorithm, we had tried the number of trees within set {50,100,300,500,600,800,1,000} and the depth of trees within set {5,10,15,20,30,50}. By testing all the combination on the validation set, we set the number of trees as 500 with its depth as 15.

FIGURE 2 | The flowchart and the complete training pipeline of the GBDT model.

Precision P, recall R, and F1 score are three common metrics used to evaluate the performance of a model in Machine Learning. Their formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

From the formula, P stands for the proportion between the number of correctively predicted positives myeloma and the total number of positives in both myeloma and non-myeloma classes, P therefore represents the prediction of myeloma in our model; R stands for the proportion between the correctively predicted myeloma and the total number of actual myeloma patients; and F1-score is the harmonic mean of precision and recall [20]. High F1-score only occurs when both recall and precision are high.

Among them, the clinical interpretation for evaluation criteria is as followed: For the category of myeloma, true positives (TP) indicate that the myeloma patient is correctly predicted to be in this myeloma class, a false positive (FP) indicates that it is not a myeloma patient and it has been incorrectly predicted as a myeloma class, and false negatives (FN) is the total number of incorrect predictions for a certain true myeloma class. For the non-myeloma category, TP represents the non-myeloma is correctly predicted in the non-myeloma class, FP represents the number of samples predicted as non-myeloma but actually myeloma, FN represents the number of non-myeloma incorrect predicted to be myeloma. The threshold for calculating TP and FN is the default value of 0.5.

The balance between the positive and negative samples can potentially cause the model to be impartial to positive and negative cases. Conversely, if the number of negative samples is larger than the number of positive samples, it will result in the deviation to the negative direction due to the over-exposure to negative samples. In our algorithm, the enhanced data is treated as normal data for model construction and testing without special treatment.

The curve of the receiver operator characteristics (ROC) is another important evaluation metric with regard with binary classification problems, and is a probability curve that plots the true positive rate (TPR) against false positive rate (FPR) at various threshold values.

All the experimental programs in this paper have been written in the Python language, and Python version 3.6 was applied as the interpreter. The machine learning development kit using in this paper is scikit learn version 0.20. The random forest applied in this paper is based on sklearn. ensemble class. The support vector machine is based on the SVM class of sklearn, and DBDT is based on Gradient Boosting Classifier class of sklearn. ensemble. The deep learning is operated on the tensorflow 1.12, with numpy 1.15.4 used to process and transform arrays, and Matplotlib 3.0.2 used to draw ROC curves.

RESULTS

Some 1,741 records of multiple myeloma and 2,446 records of non-myeloma (infectious diseases, rheumatic immune system diseases, hepatic disease and renal disease) were analyzed. The basic assay indicators are shown in **Table 1**.

Moreover, we compared performance with or without data standardization and outlier detection. The results are shown in **Supplementary Table 1**. The contents of the table indicate performance comparison of utilizing three features u_d^i, d_i and l_d^i or utilizing only feature d_i with multiple tests, where both the mean value and standard deviation are compared. As can be seen from the statistics, using u_d^i, d_i and l_d^i can reduce the standard deviation. It can be concluded that our method increases the reliability of prediction performance. By taking the F1 scores of positive samples as example, a smaller standard deviation is yielded, indicating the estimation is more stable.

For comparison, SVM, DNN, RF and GBDT models were trained and tested on the same dataset using nine variables as hemoglobin, serum creatinine, serum calcium, immunoglobulin (A, G and M), albumin, total protein, and ratio of albumin to globulin. Among the four machine learning algorithms, GBDT yielded the highest precision 0.929 and 0.899 for the myeloma and non-myeloma respectively. GBDT also has the highest recall (0.900) and F1 score (0.915) for myeloma. The value of P, R, and F1 of the four machine learning algorithms are shown in **Table 2**, and can be calculated according to Reference [21]. The influence weight of each variable on classification calculated by GBDT is shown in **Supplementary Table 2**. The weight of each feature in the table is automatically calculated by machine learning algorithm, which shows the importance of each feature for disease prediction. A larger value indicates a bigger influence on classification by this variable.

An immunoglobulin assay is not part of routine laboratory queries, we have therefore used six variables namely: hemoglobin, serum creatinine, serum calcium, albumin, total protein, and ratio of albumin to globulin) to train the model. If the immunoglobulin were considered unwarranted, the model would eventually have 0.797 precision, 0.726 recall and 0.760 F1 score, which are lower than the nine variable model fit with immunoglobulin based on the GBDT model. The value of P, R, and F1 related to six variable data were shown in **Supplementary Table 3** and **Supplementary Figure 2**. For the clinics and hospitals where immunoglobulin (A, G and

TABLE 1 | Subject characteristics.

Variable	Multiple myeloma dataset Mean (SD)	Control dataset Mean (SD)
Creatinine (umol/L)	137.97 (4.62)	119.85 (2.83)
Serum β_2 microglobulin (mg/L)	7.51 (0.30)	6.12 (0.66)
Urine β_2 microglobulin (mg/L)	22.35 (1.24)	16.75 (5.16)
IgA (g/L)	4.43 (0.37)	2.89 (0.04)
IgG (g/L)	14.45 (0.48)	12.11 (0.14)
IgM (g/L)	0.67 (0.80)	1.23 (0.02)
Albumin (g/L)	35.91 (0.26)	29.70 (0.30)
Total protein (g/L)	68.60 (0.46)	58.93 (0.46)
Serum calcium (mmol/L)	2.20 (0.08)	2.04 (0.00)
Hemoglobin (g/L)	107.38 (0.70)	114.59 (0.47)

TABLE 2 | Results of Testing Group based on 9 variables.

Method	Class	P	R	F₁
GBDT	Non-myeloma	0.899	0.928	0.913
	Myeloma	0.929	0.900	0.915
RF	Non-myeloma	0.884	0.903	0.908
	Myeloma	0.901	0.90	0.906
SVM	Non-myeloma	0.830	0.827	0.829
	Myeloma	0.836	0.839	0.837
DNN	Non-myeloma	0.850	0.783	0.815
	Myeloma	0.772	0.842	0.805

M) can be routinely measured, the 9-variable model should be used to achieve higher accuracy. For health centers where immunoglobulin examination is not commonly ordered, the 6-variable model should be applied as a precautionary measure should indications of further immunoglobulin investigations appear.

With the ROC curve, the area under the curve (AUC) can be calculated to measure a classifier's ability to distinguish classes. The higher the AUC, the better the model will be at classifying. **Figure 3** illustrates the performance of comparing the AUC with these four algorithms. We can observe that the classifier with GBDT obtains an AUC of 0.975 [95% confidence interval (CI): 0.963–0.986], and has the best performance when compared to the other three algorithms.
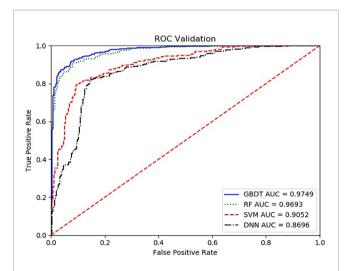
The model we trained was also tested on the new data set, and the performance index obtained is similar with the result on the original data set, but for January to November in 2020. GBDT once more showed the highest recall (0.909), precision (0.952) and F1 score (0.930) for the myeloma set. Also, GBDT showed recall (0.954), precision (0.912) and F1 score (0.932) for the non-myeloma set with the threshold 0.5. The values of P, R, F1 of the 68 newly diagnosed MM and 70 non-myelomas by machine learning algorithms are shown in **Supplementary Table 4**. We can observe that the classifier with GBDT obtained an AUC of 0.974 in the new data shown in **Supplementary Figure 3**. From this analysis, it can be inferred that the model trained by the original data set has a strong generalization ability.
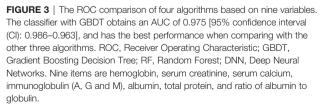
DISCUSSION

Multiple myeloma has an incidence rate of about 28,000 new cases per year in China (22) and due to various clinical presentations, diagnosis is a challenge. Data shows that more than 90% of patients suffer from bone pain and fractures in the early stages of the disease or during the course of disease progression, while about 50% of patients have renal impairment (23–25). The public do not understand MM and with a slow onset of the disease and a lack of typical symptoms in China, MM can hardly be distinguished from other diseases in other departments. This thus leads to delayed treatment and the poor prognosis of patients (26). In order to improve early diagnosis of MM, IMWG recommends the application of a percentage of clonal plasma cell into the bone marrow, serum-free light chain ratios and MRI focal lesions as additional biomarkers for the disease (27). However, in China, these programs are not considered routine inspection programs, and some primary medical care centers do not even carry out the

relevant tests. We hope that based on the integration of data from routine laboratory tests in clinics, an early warning system for MM can be designed to make it easier to proceed.

As artificial intelligence develops at an extraordinarily pace, countless applications have been created in the past decade (28–30). Recently, AI has been increasingly adopted to diagnose and predict some diseases, while the medical image analysis community has paid particular attention to the success of machine learning in computer vision (31, 32). Some researchers have been initiated into applying AI techniques to the quantification of early rheumatoid arthritis using Magnetic Resonance Imaging (MRI) data (33). Ni et al. achieved a general accuracy of 84.48% when using radionics analysis based on the LASSO + GBDT method for the noninvasive diagnosis of microvascular invasion in hepatocellular carcinoma [34]. Zhang et al. used LASSO + GBDT to examine the ability of radionics characteristics from MRI in differentiating anaplastic oligodendroglioma (AO) from atypical low-grade oligodendroglioma (35). The majority of researchers have performed quantitative analysis of multi-modality image data

FIGURE 3 | The ROC comparison of four algorithms based on nine variables. The classifier with GBDT obtains an AUC of 0.975 [95% confidence interval (CI): 0.986–0.963], and has the best performance when comparing with the other three algorithms. ROC, Receiver Operating Characteristic; GBDT, Gradient Boosting Decision Tree; RF, Random Forest; DNN, Deep Neural Networks. Nine items are hemoglobin, serum creatinine, serum calcium, immunoglobulin (A, G and M), albumin, total protein, and ratio of albumin to globulin.

for diagnosis and prognosis by using artificial intelligence methods, but few of them have focused on the routine laboratory tests that easily obtained from clinic.

AI techniques have been applied to the treatment of multiple myeloma. Ji et al. constructed a hybrid multi-scale agent-based model (HABM) model to provide new insight into the development of myeloma in a bone marrow micro-environment that is the basis of the immune system, and also build an efficient computational platform for prediction of drug response for discovering the optimal dose combination (36). Zhang et al. built a more efficient approach by combining the standard ordinal logistic regression and the hierarchical modeling. This method can simultaneously analyze numerous variables for detecting important predictors and for predicting multi-level drug response (37). Tang et al. established and validated a novel mathematical model of multiple myeloma cell dynamics. The clinical data compounded with mathematical modeling, suggested that bortezomib-based therapy exerted a selection pressure on myeloma cells (38) Bouchnita et al. developed a hybrid discrete-continuous model to predict the response of MM tumors to treatment with gefitinib and 6-aminonicotinamide (6-AN) (39). There is no published early diagnosing, laboratory results-based AI models as of yet.

According to our preliminary test based on the data of 1,000 myeloma patients and 2,000 non-myeloma patients in our hospital, the predictive value of artificial intelligence can reach more than 90% with the prospect of having a wide application. Our research also indicates that the SVM algorithm is suitable for classifying small-size data, while the DNN algorithm is suitable for classifying large-size data. By efficiently extracting the sample features, the GBDT algorithm can simultaneously train some decision trees on the ability to sort out the features based on their importance, so as to obtain the best performance when comparing with the other three algorithms (40).

Taking the integration of test data as the breakthrough point, this project adopts the methods of big data analysis and artificial intelligence, so as to propose the automatic integration of routine test reports, establish multiple myeloma screening models, give early warnings for multiple myeloma, and improve the diagnosis rate. The research contents are innovative in the medical, information and business fields.

CONCLUSION

In this study, routine exam results obtained from general hospitals are utilized to train machines to realize automatic screening, identify patients at a high risk of diagnosed multiple myeloma and provide early warnings through the big data platform, artificial intelligence and other technologies. This technology can be widely used in general hospitals and primary medical care to improve the early diagnosis rate of myeloma and prevent the occurrence of missed diagnosis and misdiagnosis. At the end, an early warning and screening system for myeloma based on artificial intelligence will be formed.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Study conception and design: WYan and HW. Literature review and data extraction: HS. Quality control: WYan and AL. Model construction and statistical analysis: JC and TH. Manuscript preparation: WYan and HW. Manuscript review: WYan, AL, WYang, and HW. Editing and revisions: WYan, HW, JC, TH, and CW. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.608191/full#supplementary-material

Supplementary Figure 1 | Performance comparison with different threshold in the four algorithms.

Supplementary Figure 2 | The ROC comparison of four algorithms based on 6 variables. Six variables are hemoglobin, serum creatinine, serum calcium, albumin, total protein, and ratio of albumin to globulin.

Supplementary Figure 3 | The ROC comparison of four algorithms based on 9 variables from new cases set in 2020.

REFERENCES

1. Morandi F, Horenstein AL, Costa F, Giuliani N, Pistoia V, Malavasi F. CD38: A target for immunotherapeutic approaches in multiple myeloma. *Front Immunol* (2018) 9:2722. doi: 10.3389/fimmu.2018.02722

2. Schinke M, Ihorst G, Duyster J, Wäsch R, Schumacher M, Engelhardt M. Risk of disease recurrence and survival in patients with multiple myeloma: A German Study Group analysis using a conditional survival approach with long-term follow-up of 815 patients. *Cancer* (2020) 126:3504–15. doi: 10.1002/cncr.32978

3. Howell DA, Hart RI, Smith AG, Macleod U, Patmore R, Cook G, et al. Myeloma: Patients accounts of their pathways to diagnosis. *PLoS One* (2018) 13:e0194788. doi: 10.1371/journal.pone.0194788

4. Pawlyn C, Cairns D, Kaiser M, Striha A, Jones J, Shah V, et al. The relative importance of factors predicting outcome for myeloma patients at different ages: results from 3894 patients in the Myeloma XI trial. *Leukemia* (2020) 34:604–12. doi: 10.1038/s41375-019-0595-5

5. Das N, Topalovic M, Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. *Curr Opin Pulm Med* (2018) 24:117–23. doi: 10.1097/MCP.0000000000000459

6. Gao F, Zhu Y, Zhang J. Artificial intelligence in computer-aided diagnosis of abdomen diseases. *Sci China Life Sci* (2019) 62:1396–9. doi: 10.1007/s11427-019-1556-7

7. Chen W, Lu Z, You L, Zhou L, Xu J, Chen K. Artificial intelligence–based multimodal risk assessment model for surgical site infection (AMRAMS): development and validation study. *JMIR Med Inform* (2020) 8:e18186. doi: 10.2196/18186

8. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning–data mining, inference, and prediction*. New York: Springer-Verlag (2001).

9. Zhou S, Wang S, Wu Q, Azim R, Li W. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput Biol Chem* (2020) 85:107200. doi: 10.1016/j.compbiolchem.2020.107200

10. Li J, Liu J, Liu P, Qi Y. Analysis of factors contributing to the severity of large truck crashes. *Entropy (Basel)* (2020) 22:E1191. doi: 10.3390/e22111191

11. Rajkumar SV, Dimopoulos MA, Palumbo A, Blade J, Merlini G, Mateos MV, et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol* (2014) 15:e538–48. doi: 10.1016/S1470-2045(14)70442-5

12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* (2002) 16:321–57. doi: 10.1613/jair.953

13. Raschka S. *Python Machine Learning*. UK: Birmingham Press (2015).

14. Schapire RE. Measures of diversity in classifier ensembles. *Mach Learn* (2003) 51:181–207. doi: 10.1049/ic:20010105

15. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Statist* (2001) 29:1189–232. doi: 10.2307/2699986

16. Breiman LI, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees (CART). *Biometrics* (1984) 40:17–23. doi: 10.2307/2530946

17. Freund ,Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* (1997) 55:119–39. doi: 10.1007/3-540-59119-2_166

18. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun ACM* (2017) 60:84–90. doi: 10.1145/3065386

19. Shi ,T, Horvath S. Unsupervised Learning with Random Forest Predictors. *J Comput Graph Stat* (2006) 15:118–38. doi: 10.1198/106186006X94072

20. Chinchor N. *MUC-4 evaluation metrics. Proceedings of the 4th conference on Message understanding*. Virginia, USA: Association for Computational Linguistics (1992) pp. 22–9. doi: 10.3115/1072064.1072067

21. Holz ,HJ, Loew MH. Validation of relative feature importance using natural data. *Pattern Recogn Lett* (2002) 23:367–80. doi: 10.1016/S0167-8655(01)00170-2

22. Du J, Lu J, Gao W, Liu J, He H, Li L, et al. Serum-free light chains combined with the Revised International Staging System could further distinguish the superior and inferior clinical outcome of multiple myeloma patients. *Ann Hematol* (2020) 99:1779–91. doi: 10.1007/s00277-020-04162-8

23. Fiala MA, Wildes TM, Vij R. Racial disparities in the utilization of novel agents for frontline treatment of multiple myeloma. *Clin Lymphoma Myeloma Leuk* (2020) 7:S2152–2650(20)30213-5. doi: 10.1016/j.clml.2020.04.018

24. Zhu X, Wang L, Zhu Y, Diao W, Li W, Gao Z, et al. Extramedullary plasmacytoma: long-term clinical outcomes in a single-center in China and literature review. *Ear Nose Throat J* (2020) 17:145561320950587. doi: 10.1177/0145561320950587

25. Joo CW, Quach H, Raje N. Perspectives in the rapidly evolving treatment landscape of multiple myeloma: expert review of new data presentations from ASH 2019. *Clin Lymphoma Myeloma Leuk* (2020) 29:S2152–2650(20)30313-X. doi: 10.1016/j.clml.2020.06.015

26. Li X, Liu J, Chen M, Gu J, Huang B, Zheng D, et al. Health-related quatlity of life of patients with multiple myeloma: a real-world study in China. *Cancer Med* (2020) 9:7896–913. doi: 10.1002/cam4.3391

27. Ishida T. Antibody therapy for multiple myeloma: novel approaches and future perspectives. *Rinsho Ketsueki* (2020) 61:912–21. doi: 10.11406/rinketsu.61.912

28. Lei L, Wang Y, Xue Q, Tong J, Zhou CM, Yang JJ. A comparative study of machine learning algorithms for predicting acute kidney injury after liver cancer resection. *Peer J* (2020) 8:e8583. doi: 10.7717/peerj.8583

29. Ichikawa D, Saito T, Ujita W, Oyama H. How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. *J Biomed Inform* (2016) 64:20–4. doi: 10.1016/j.jbi.2016.09.012

30. Huang CY, Lin CH, Chang CS, Li JL, Lee CC. Predicting gastrointestinal bleeding events from multimodal in hospital electronic health records using deep fusion networks. *Conf Proc IEEE Eng Med Biol Soc* (2019) 2019:2447–50. doi: 10.1109/EMBC.2019.8857244

31. Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of Multi-Modality imaging and artificial intelligence for diagnosis and prognosis of early stages of alzheimer's disease. *Transl Res* (2018) 194:56–67. doi: 10.1016/j.trsl.2018.01.001

32. Li J, Liu S, Qin Y, Zhang Y, Wang N, Liu H. High-order radiomics features based on T2 FLAIR MRI predict multiple glioma immunohistochemical features: A more precise and personalized gliomas management. *PLoS One* (2020) 15:e0227703. doi: 10.1371/journal.pone.0227703

33. Stoel BC. Artificial intelligence in detecting early RA. *Semin Arthritis Rheumatol* (2019) 49:S25–8. doi: 10.1016/j.semarthrit.2019.09.020

34. Ni M, Zhou X, Lv Q, Li Z, Gao Y, Tan Y, et al. Radiomics models for diagnosing microvascular invasion in hepatocellular carcinoma: which model is the best model? *Cancer Imaging* (2019) 19:60. doi: 10.1186/s40644-019-0249-x

35. Zhang Y, Chen C, Cheng Y, Teng Y, Guo W, Xu H, et al. Ability of radiomics in differentiation of anaplastic oligodendroglioma from atypical low-grade oligodendroglioma using machine-learning approach. *Front Oncol* (2019) 9:1371:1371. doi: 10.3389/fonc.2019.01371

36. Ji Z, Su J, Wu D, Peng H, Zhao W, Zhao BN, et al. Predicting the impact of combined therapies on myeloma cell growth using a hybrid multi-scale agent-based model. *Oncotarget* (2017) 8:7647–65. doi: 10.18632/oncotarget.13831

37. Zhang X, Li B, Han H, Song S, Xu H, Hong Y, et al. Predicting multi-level drug response with gene expression profile in multiple myeloma using hierarchical ordinal regression. *BMC Cancer* (2018) 18:551. doi: 10.1186/s12885-018-4483-6

38. Tang M, Zhao R, van de Velde H, Tross JG, Mitsiades C, Viselli S, et al. Myeloma Cell Dynamics in Response to Treatment Supports a Model of Hierarchical Differentiation and Clonal Evolution. *Clin Cancer Res* (2016) 22:4206–14. doi: 10.1158/1078-0432.CCR-15-2793

39. Bouchnita A, Volpert V, Koury MJ, Hellander A. A multiscale model to design therapeutic strategies that overcome drug resistance to TKIs in multiple myeloma. *Math Biosci* (2020) 319:108293. doi: 10.1016/j.mbs.2019.108293

40. Li R, Zhang W, Suk H, Wang L, Li J, Shen D, et al. Deep learning Based imaging data completion for Improved Brain Disease Diagnosis. *Randomized Controlled Trial* (2014) 17:305–12. doi: 10.1007/978-3-319-10443-0_39