Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 4, 2023, 505-515.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A New Hybrid Method to Detect Risk of Gastric Cancer using Machine Learning Techniques

Ali Zahmatkesh Zakariaee[1], Hossein Sadr[2*] and Mohammad Reza Yamaghani[3]

1. *Department of Computer Engineering, Rahbord Shomal Institute of Higher Education, Rasht, Iran.*
2. *Department of health Informatics, Guilan Road Trauma Research Center, Trauma institute, Guilan University of medical sciences, Rasht, Iran.*
3. *Department of Computer Engineering, Faculty of Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran.*

Article Info

Article History:
Received 18 July 2023
Revised 05 September 2023
Accepted 15 October 2023

DOI:10.22044/jadm.2023.13377.2464

Keywords:
Artificial Intelligence, Machine Learning, Gastric cancer, Hybrid Method, Neural Network.

**Corresponding author: Sadr@qiau.ac.ir (H. Sdar).*

Abstract

Machine learning (ML) is a popular tool in healthcare while it can help to analyze large amounts of patient data such as medical records, predict diseases, and identify early signs of cancer. Gastric cancer starts in the cells lining the stomach, and is known as the 5th most common cancer worldwide. Therefore, predicting the survival of patients, checking their health status, and detecting their risk of gastric cancer in the early stages can be very beneficial. Surprisingly, with the help of machine learning methods, this can be possible without the need for any invasive methods that can be useful for both patients and physicians in making informed decisions. Accordingly, a new hybrid machine learning-based method for detecting the risk of gastric cancer is proposed in this paper. The proposed model is compared with the traditional methods, and based on the empirical results, not only the proposed method outperform existing methods with an accuracy of 98% but also gastric cancer can be one of the most important consequences of *H. pylori* infection. Additionally, it can be concluded that lifestyle and dietary factors can heighten the risk of gastric cancer, especially among individuals who frequently consume fried foods and suffer from chronic atrophic gastritis and stomach ulcers. This risk is further exacerbated in individuals with limited fruit and vegetable intake and high salt consumption.

1. Introduction

Many diseases have affected humanity throughout history and have taken many lives. Gastric cancer is a prevalent malignancy with a high incidence and mortality rate worldwide. The gastric cancer risk factors vary by country, and are associated with urbanization and economic development. Diagnosing gastric cancer is difficult, with only about 10% of people diagnosed while still in the early stages. Studies indicate that gastric cancer (GC) ranks fifth among the most common cancers worldwide, and is considered a multifactorial and dangerous disease [1]. This factor is responsible for one-third of cancer-related deaths, and is considered the third leading cause of cancer-related fatalities [2]. In Iran, cancer is the second leading cause of death after heart disease [3]. Moreover, the 5-year survival rate in Iran is estimated at less than 25% [4].

Surgery is considered as the primary treatment of gastric cancer. However, due to the lack of clear symptoms in the early stages, and because many of the initial symptoms mimic indigestion, patients often receive treatment in the advanced stages of cancer. This significantly impacts the survival rate, reducing it by up to 50% [5-7]. Hence, the utilization of artificial intelligence and machine learning mining methods is crucial for investigating the characteristics of gastric cancer risk factors and enabling its early prediction and diagnosis [8, 9].

Accordingly, the aim of this study is to develop a new hybrid machine learning-based method for assessing the risk of gastric cancer in the early stage. The proposed method uses the combination of Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) to enhance the prediction accuracy. MLP is a type of artificial neural network that can learn and recognize complex patterns in data. It is useful for processing large amounts of data and identifying hidden relationships between variables. SVM, on the other hand, is a machine learning algorithm that can classify data into different categories based on their features. By combining MLP and SVM, the prediction model can benefit from the strengths of both algorithms. MLP can extract features from the data and identify complex patterns, while SVM can classify the data into different categories based on these features. This can result in a more accurate and reliable prediction model for gastric cancer.

The proposed method was implemented on a dataset collected from gastric patients with various symptoms including indigestion as well as gastric cancer who were referred to Shahid Dr. Fayaz Bakhsh Hospital in Tehran in 1397. The obtained dataset includes demographic information, family and previous disease records, lifestyle and eating habits, disease symptoms, and serological and hematological characteristics, and is categorized into two classes of low-risk and high-risk, aiming to predict the probability of gastric cancer.

According to the empirical findings, the proposed method not only surpasses the existing methods with a 98% accuracy rate but also highlights the significant link between *H. pylori* infection and gastric cancer. The study also suggests that lifestyle and dietary factors can increase the likelihood of developing gastric cancer, particularly in individuals who frequently consume fried foods and suffer from chronic atrophic gastritis and stomach ulcers.

The remainder of this paper is categorized as what follows. The summary of related studies is provided in Section 2. The details of the proposed method are explained in Section 3. Section 4 includes the results of the experiments. Conclusion and the direction of future research are mentioned in Section 5.

2. Related Works

Predicting gastric cancer using machine learning is important because it can help in early detection and treatment of the disease. Gastric cancer is a serious and potentially life-threatening condition, and early diagnosis is crucial for successful treatment outcomes. Machine learning algorithms can analyze large amounts of data and identify patterns that may not be apparent to human experts. By using machine learning to predict gastric cancer, healthcare professionals can potentially identify patients who are at high risk of developing the disease and provide them with appropriate preventive measures or early treatment. This can ultimately lead to better health outcomes and improved quality of life for patients. Accordingly, numerous studies have been conducted in the recent years to use various machine learning methods for gastric cancer prediction that are introduced in the following.

Md. Rejaul Islam Royel *et al*. (2021) investigated the efficiency of machine learning and data mining methods in early detection of gastric cancer risk. They utilized a dataset with 300 samples and extracted 18 important gastric cancer risk factors. They also designed a gastric cancer risk level prediction tool [10].

Yunmei Li *et al*. (2022) explored the predictions of 5-year survival in patients with gastric cancer, and concluded that survival rate of patients with gastric cancer showed different degrees of improvement in each subgroup. However, the overall relative survival rate of patients with gastric cancer remains low. Based on the result of their experiments, analyzing the changes of patients with gastric cancer in the last 10 years will be helpful in predicting the changing trend of cancer in the future. It also provided a scientific basis for relevant departments to formulate effective tumor prevention and control measures [11].

Mohammadreza Arash *et al*. (2023) established machine learning models to predict the early risk of gastric cancer based on lifestyle factors from six ML methods including multilayer perceptron, support vector machine, k-nearest neighbors, random forest, and XGBoost that were used to build predictive models. This study found 11 important influence factors for the risk of gastric cancer such as Helicobacter pylori infection, high salt intake, and chronic atrophic gastritis, among other factors. Comparisons indicated that the XGBoost had the best performance for the risk prediction of gastric cancer [12].

Meysam Roostaee *et al*. (2023) proposed an approach based on data mining techniques with the aim of minimizing the need for redundant blood tests in diagnosing common diseases by leveraging unsupervised data mining techniques on a large-scale dataset. They used unsupervised methods including pre-processing, clustering, and association rule mining. This study highlights the importance of big data analytics and unsupervised

learning techniques in increasing efficiency in healthcare centers [13].

3. Materials and Methods

3.1. Datasets

Selecting objective data not only ensures a fair evaluation of the prediction model's features but also simplifies the comparison of prediction results and accuracy measurement. The dataset utilized in this research comprises information from 618 patients with stomach diseases and gastric cancer who visited Shahid Dr. Fayaz Bakhsh Hospital in Tehran in 1397. The collected dataset includes demographic details, family and past medical records, lifestyle and dietary habits, disease symptoms, and serological and hematological characteristics. The dataset is categorized into two classes of low-risk and high-risk, and the distribution of data based on target class is depicted in Figure 1.

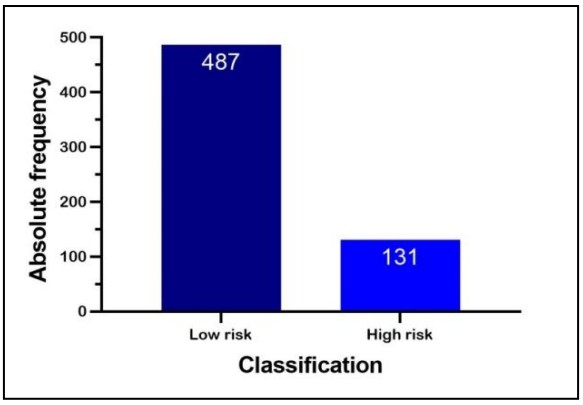

Figure 1. Distribution of the target class.

3.2. Data pre-processing

Pre-processing is a crucial step in data mining because it helps to clean, transform, and prepare data for analysis. The flowchart indicating the pre-processing steps of this study is depicted in Figure 2.

Identifying and managing missing values is the first step in data pre-processing. Deleting data can introduce additional bias and lead to incorrect results. Accordingly, missing values are replaced with the mean of that particular feature in our study [14]. To normalize and standardize the data, Min-Max technique was used to convert the data scale to 0 and 1 (Eq. 1), where X_{min} and X_{max} are the minimum and the maximum values of the feature.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Through data standardization, the values are transformed to have a common mean and scaled by a standard deviation. In standardization, each feature is scaled by subtracting the mean and dividing by the standard deviation, resulting in a mean of zero and a standard deviation of one for each feature. The formula for the standardization method used in this research work is described in (Eq. 2).

$$x' = \frac{x - \mu}{\sigma} \quad (2)$$

Here, μ is the mean of the feature values, and σ is the standard deviation of the feature values. It is worth mentioning that normalization is applied on the whole records of the dataset.

Feature scaling is the next pre-processing step, where the range of values for the independent variables in a dataset is standardized to a specific range [15]. This method allows for the comparison of independent variables within a common range. In our dataset, variables such as 'age', 'weight', 'BMI', 'platelet-count', 'pepsinogen-i,' and 'plr' do not share the same scale.

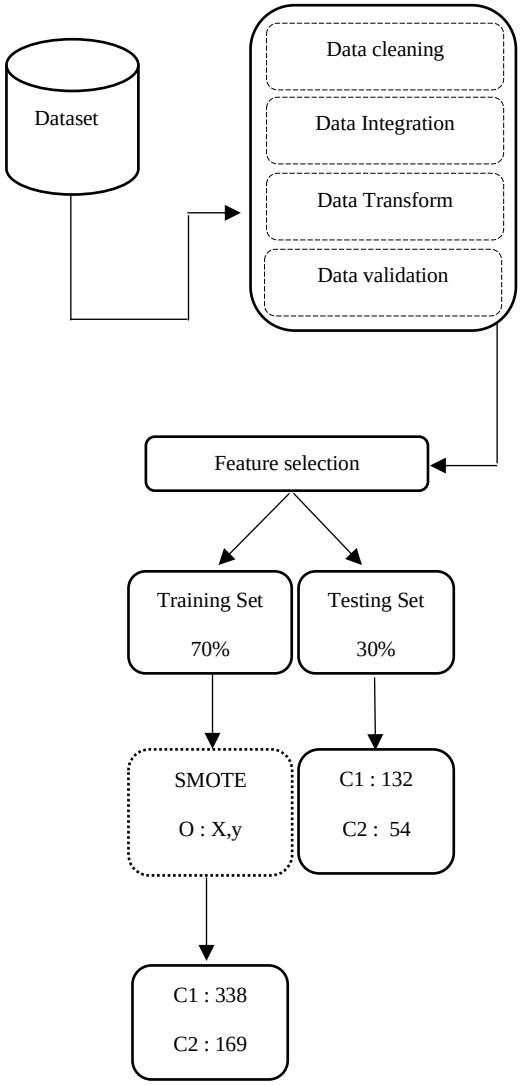

Figure 2. Data pre-processing flowchart.

Therefore, to address this issue, we need to perform feature scaling.

Feature scaling is particularly crucial in the MLP neural network algorithm, which employs gradient descent for data optimization. Additionally, algorithms like SVM, which focus on the range of feature variations and determine data point distances and similarities, are highly influenced by feature scaling.

In the following, the SMOTE method was used to mitigate issues caused by simple oversampling through a combination of replacement and undersampling. In this technique, synthetic data points are generated based on a minority sample and its nearest neighbors, determined by standard Euclidean distance. These synthetic data points are inserted between the existing minority data points [16]. Previous research has demonstrated a significant improvement in classifier accuracy when using the SMOTE method [17-19]. Tt is important to note that the synthetic data generated by SMOTE is only used within the training set and not included in the test set.

Feature selection is the final pre-processing step. Due to the high volume of features in the dataset and the dispersion of some cases, feature selection methods were used to reduce the risk of overfitting and reducing the dimensions of the data [20]. In this regard, Pearson's correlation coefficient was used to show the highest correlation between two specific features; all features are compared with other features in the dataset to determine the best features for building. In this study, after the correlation analysis on the dataset variables, 16 features that had a weak correlation with the risk of gastric cancer were removed, and finally, 36 features were selected. The correlation matrix of the selected features is shown in Figure 3.

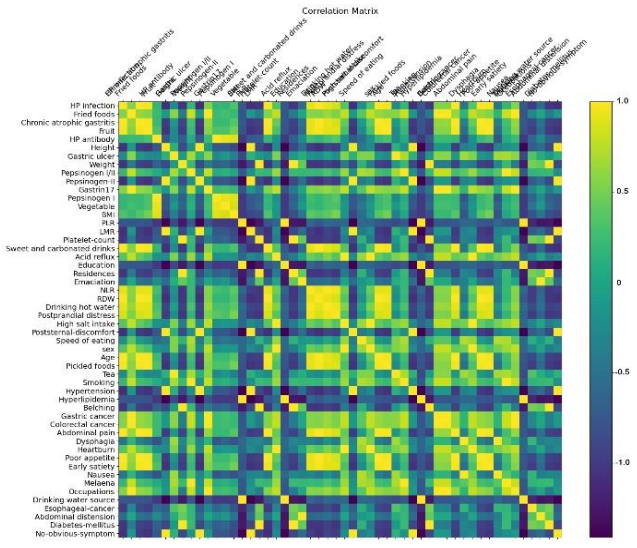

Figure 3. Correlation matrix of selected features.

Furthermore, Relief-F technique [21] was employed to learn feature weights and obtain a good set of features. Relief-F is a popular choice for feature selection due to its ability to capture feature relevance, handle redundancy, accommodate non-linear relationships, and exhibit robustness to noise. By evaluating data correlation, the number of features was reduced from 36 to 20 features. Relief is one of the most widely used feature selection methods in labeled data, which belongs to the category of supervised filtering methods. This method selects a sample completely randomly from among the samples in the dataset and repeats this until the last step. Then the degree of relevance of each feature is updated based on the difference between the selected sample and two nearby samples.

3.3. Proposed hybrid method

The use of data mining in gastric cancer prediction holds significant importance in the field of healthcare. By harnessing the power of advanced algorithms and analyzing vast amounts of patient data, data mining enables the identification of patterns and associations that might otherwise go unnoticed. This early detection capability is crucial in improving patient outcomes as it facilitates timely intervention and treatment. Moreover, data mining helps in assessing an individual's risk for developing gastric cancer by identifying relevant risk factors. This knowledge allows healthcare professionals to implement targeted screening programs and preventive measures, leading to proactive management of the disease. Accordingly, in this paper, a hybrid method is proposed for predicting the risk of gastric cancer. The proposed method uses the combination of MLP and SVM based on the Bagging method. The flowchart of the proposed model is depicted in Figure 4.

It must be noted that the combination of MLP SVM using the bagging method is an effective approach in machine learning ensemble techniques. Bagging, short for bootstrap aggregating, aims to improve the predictive performance and robustness of individual models by aggregating their outputs. In the context of MLP and SVM, bagging involves training multiple MLP and SVM models, each on a different subset of the original dataset created through random sampling with replacement [22-24]. This technique introduces diversity among the individual models, leading to a more accurate and generalized ensemble model.

MLP is a type of artificial neural network widely used for pattern recognition and classification tasks. It is known for its ability to capture complex relationships and non-linearities in data. By

combining multiple MLP models trained on different subsets, bagging helps to mitigate overfitting, improve generalization, and enhance the overall accuracy of the ensemble predictions.
SVM, on the other hand, is a powerful supervised learning algorithm that constructs a hyperplane to separate data points into different classes. SVM excels at handling high-dimensional feature spaces and can effectively handle non-linear classification problems by employing the kernel trick.

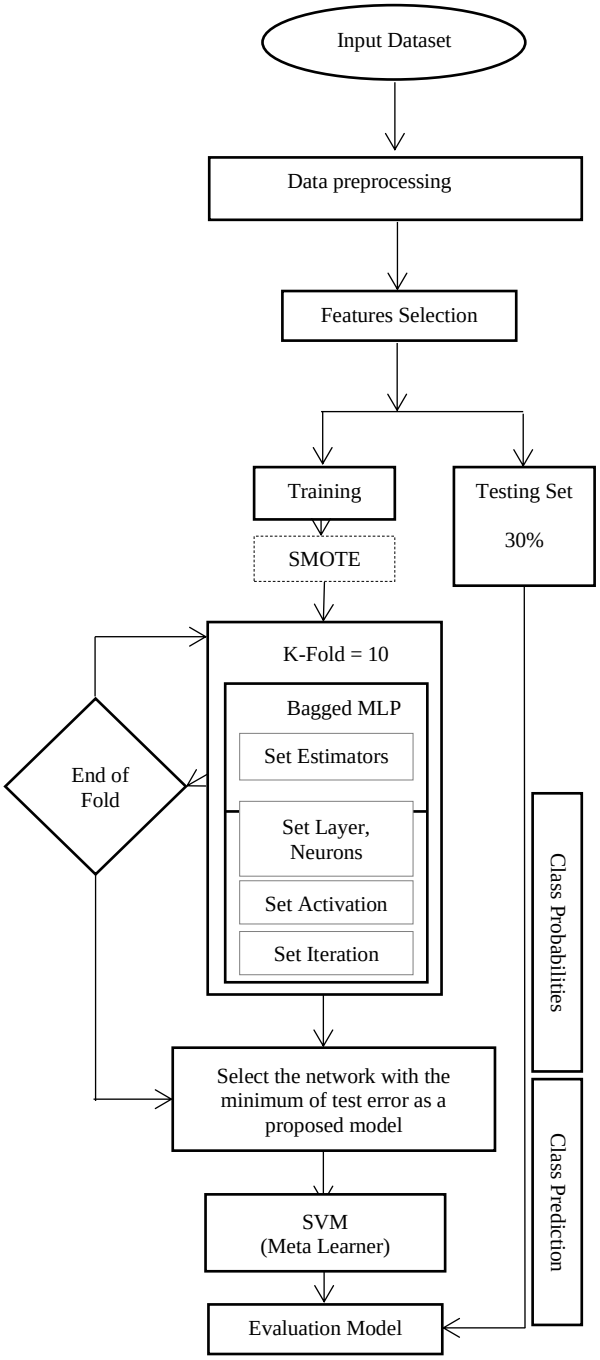

Figure 4. Flowchart of the proposed model.

When combined with MLP using the bagging method, SVM models contribute their unique decision boundaries to the ensemble, enhancing its ability to handle diverse and complex data patterns. The final ensemble prediction is typically obtained through majority voting or averaging of the individual model predictions. This ensemble approach significantly reduces the variance and bias that may exist in the individual models, resulting in improved predictive performance and generalization capabilities.
The combination of MLP and SVM using the bagging method provides a powerful and robust framework for solving classification problems, leveraging the strengths of both algorithms. It allows for more accurate predictions, improved model stability, and better handling of complex data patterns.
It is worth mentioning that the dataset was divided into training set (70%) and test set (30%) prior to training. K-fold validation (k = 10) was also used in the training process. K-fold cross-validation is to assess the generalization capability of the model and identify potential issues such as overfitting or under fitting. By averaging the evaluation metrics (across the k iterations, a more reliable estimate of the model's performance can be obtained. In the context of combining MLP and SVM using the bagging method, k-fold cross-validation can be used to evaluate the performance of individual models (MLP and SVM) and tune their hyper-parameters to achieve optimal results.

4. Implementation and Results

4.1. Evaluation metrics

Knowledge generated in the previous step must be carefully examined and interpreted. The objective of knowledge evaluation is to specify its accuracy and suitability for practical applications. Various methods are employed to assess the generated knowledge, which is tied to the used learning models. We employed four standard metrics, namely accuracy, precision, recall, and F-measure, to assess the effectiveness of the proposed model. Additionally, we incorporated the AUC (area under the ROC curve) metric that is commonly used in medical data mining tasks. These metrics can be computed based on the following equations (3, 4, 5, 6). The subsequent section illustrates the results of each separate classifier as well as the combined ensemble classifiers using the following equations.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - SCORE = \frac{2 \times precision \times recall}{precision + recal} \quad (6)$$

4.2. Dataset description

As previously mentioned, the dataset includes 618 patients with stomach diseases and gastric cancer, which are categorized into two classes of low-risk and high-risk. The details of the records obtained from the classification of data into demographic groups, food habits and lifestyle, previous or family history of disease, main symptoms of the disease, and serological and hematological examinations are shown in Tables 1 to 5.

4.3. Training and hyper-parameters

All implementations were conducted using Python programming language based onscikit-learn (sklearn) library, while data manipulation and processing were performed with the pandas and NumPy libraries. Additionally, Matplotlib was used for graph plotting; setting the parameters of the algorithms plays a major role in their accuracy and efficiency. In the proposed model, the training process starts after data preparation and balancing, with model parameters being fine-tuned. The MLP model achieved its highest accuracy, with an r2 value of 0.94, when utilizing either 14 or 20 hidden layers, as illustrated in Figure 5. However, in the interest of model performance, 14 hidden layers were chosen due to the substantial difference in the relative absolute error values, which were 30.65% and 38.23%, respectively.

Table 1. Demographic features.

Features	Low risk (n = 487)	High risk (n = 131)
Sex		
Male	249 (51.13)	66 (50.38)
Female	238 (48.87)	65 (49.62)
Age (year)[a]	48.93 (19.71)	47.81 (18.67)
Weight (kg)[a]	68.46 (16.60)	69.65 (17.30)
Height (cm)	177.35 (9.49)	174.90 (9.61)
BMI[a]	21.89 (5.57)	22.91 (6.15)
Education levels	36 (7.39)	12 (9.16)
Illiterate	28 (5.75)	7 (5.34)
Primary school	107 (21.97)	29 (22.14)
Junior school	122 (25.05)	35 (26.72)
Senior school	194 (39.84)	48 (36.64)
College		
Occupations	78 (16.02)	25 (19.08)
Cadre	81 (16.63)	24 (18.32)
Worker	328 (67.35)	82 (62.60)
Other		
Residences	158 (32.44)	57 (43.51)
City	180 (36.96)	36 (27.48)
Townlet	149 (30.60)	38 (29.01)
Village		

[a] Data are presented as a mean (SD), others are presented as a number (percentage)

Table 2. Lifestyle and eating habits.

Features	Low risk (n = 487)	High risk (n = 131)
High salt intake		
Yes	251 (51.54)	67 (51.15)
No	236 (48.46)	64 (48.85)
Pickled foods		
Often	229 (47.02)	59 (45.04)
Seldom	258 (52.98)	72 (54.96)
Fried foods		
Often	230 (47.22)	117 (89.31)
Seldom	257 (52.78)	14 (10.69)
Fruit		
Often	360 (73.92)	63 (48.09)
Seldom	127 (26.08)	68 (51.91)
Vegetable		
Often	360 (73.92)	65 (49.61)
Seldom	127 (26.08)	66 (50.39)
Tea		
Often	247 (50.72)	60 (49.61)
Seldom	240 (49.28)	71 (54.20)
Smoking		
Yes	234 (48.05)	58 (44.28)
No	253 (51.95)	73 (55.72)
Drinking water source	238 (48.87)	70 (53.44)
Water supply	249 (51.13)	61 (46.56)
Wells water	-	-
Rivers water		
Drinking hot water		
Yes	249 (51.13)	64 (48.85)
No	238 (48.87)	67 (51.15)
Sweet and carbonated drinks	130 (26.69)	72 (54.96)
Often	357 (73.31)	59 (45.04)
Seldom		
Speed of eating		
Fast	254 (52.16)	64 (48.85)
Slow	233 (47.84)	67 (51.15)

All data are presented as a number (percentage).

Table 3. Family and previous disease records.

Features	Low risk (n = 487)	High risk (n = 131)
Esophageal cancer		
Yes	145 (29.77)	81 (61.83)
No	342 (70.23)	50 (38.17)
Gastric cancer		
Yes	75 (15.40)	29 (22.14)
No	412 (84.60)	102 (77.86)
Colorectal cancer		
Yes	23 (4.72)	21 (16.03)
No	464 (95.28)	110 (83.97)
Diabetes mellitus		
Yes	130 (26.69)	72 (54.96)
No	357 (73.31)	59 (45.04)
Hypertension		
Yes	87 (17.86)	69 (52.67)
No	400 (82.14)	62 (47.33)
Hyperlipidemia		
Yes	26 (5.34)	71 (54.20)
No	461 (94.66)	60 (45.80)
HP infection		
Positive	79 (16.22)	48 (36.64)
Negative	316 (64.89)	45 (34.35)
Unidentified	92 (18.89)	38 (29.01)
Chronic atrophic gastritis	123 (25.26)	73 (55.73)
Yes	364 (74.74)	58 (44.27)
No		
Gastric ulcer		
Yes	131 (26.90)	72 (54.96)
No	356 (73.10)	59 (45.04)

All data are presented as a number (percentage).

Table 4. Disease symptoms.

Features	Low risk (n = 487)	High risk (n = 131)
Abdominal pain		
Yes	327 (67.15)	99 (75.57)
No	160 (32.85)	32 (24.43)
Abdominal distension	268 (55.03)	100 (76.34)
Yes	219 (44.97)	31 (23.66)
No		
Acid reflux		
Yes	135 (27.72)	104 (79.39)
No	352 (72.28)	27 (20.61)
Belching		
Yes	254 (52.16)	97 (74.05)
No	233 (47.84)	34 (25.95)
Early satiety		
Yes	239 (49.08)	105 (80.15)
No	248 (50.92)	26 (19.85)
Postprandial distress		
Yes	246 (50.51)	99 (75.57)
No	241 (49.49)	32 (24.43)
Heartburn		
Yes	263 (0.54)	81 (61.83)
No	224 (0.56)	50 (38.17)
Melaena		
Yes	151 (31.01)	84 (64.12)
No	336 (68.99)	47 (35.88)
Emaciation		
Yes	73 (14.99)	73 (55.73)
No	414 (85.01)	58 (44.27)
Poor appetite		
Yes	299 (61.40)	78 (59.54)
No	188 (38.60)	53 (40.46)
Dysphagia		
Yes	291 (59.75)	72 (54.96)
No	196 (40.25)	59 (45.04)
Nausea		
Yes	188 (38.60)	63 (48.09)
No	299 (61.40)	68 (51.91)
Poststernal discomfort	119 (24.44)	81 (61.83)
Yes	368 (75.56)	50 (38.17)
No		
No obvious symptom		
Yes	77 (15.81)	14 (10.69)
No	410 (84.19)	117 (89.31)

All data are presented as a number (percentage).

Table 5. Serological and hematology features.

Features	Low risk (n = 487)	High risk (n = 131)
Pepsinogen-I[a]	119.24 (0.62)	172.38 (110.29)
Pepsinogen-II[a]	22.26 (9.63)	46.43 (30.11)
Gastrin17[a]	40.76 (22.01)	19.38 (12.16)
Pepsinogen-I/II[a]	6.99 (5.99)	10.54 (21.19)
HP-antibody		
Negative	361 (74.13)	38 (29.01)
Weakly positive	47 (9.65)	19 (14.50)
Positive	79 (16.22)	74 (56.49)
NLR[a]	7.18 (4.18)	9.94 (4.05)
PLR[a]	225.14 (128.49)	364.37 (142.49)
LMR[a]	13.86 (5.86)	4.42 (0.81)
Platelet-count[a]	329.94 (72.34)	281.48 (59.57)
RDW[a]	12.42 (0.64)	14.74 (1.58)

[a] Data are presented as a mean (SD), and others are presented as a number (percentage).

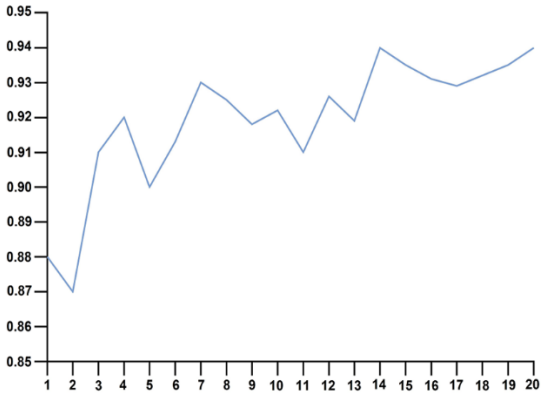

Figure 5. Hidden layer ranking.

The bagged MLP model, employing 50 epochs and 10 estimators, demonstrated the highest accuracy. It employed ReLU activation for hidden layers and sigmoid activation for the output layer. Weight and bias initialization in MLP ranged from -1 to 1, with values being iteratively updated for optimal results. SVM, as a machine learning algorithm, falls under the supervised algorithm category, and is commonly employed for the regression and classification tasks. In our proposed hybrid model, SVM served as a linear meta-learner. The LinearSVC function, which organizes samples in a 'one-versus-all' manner, enhances accuracy in predicting classes and delivers them as outcomes [25]. This approach results in the creation of a composite model, which often outperforms a single model constructed from the original data. Table 1 presents the optimal hyper-parameters for each algorithm [26]. The summary of used hyper-parameters is depicted in Table 6.

Table 6. Hyper-parameters.

Algorithm	Parameters
MLP	hidden_layer_size = 14 activation = 'relu,Sigmoid' learning_rate = 'adaptive' momentum = 0.9 optimizer = 'sgd'
SVM	kernel=linear C = 10
Random forest	Bootstrap = True max_features = sqrt (n_features) n_estimators = 10 criterion = 'Gini'
Decision tree	criterion = 'Gini' min_samples_leaf = 1 min_samples_split = 2 max_features = log2 (n_features) splitter = 'best'

4.4. Performance evaluation

The goal of this paper is to introduce a model that utilizes data mining algorithms to predict the risk

of gastric cancer. Accordingly, various data mining algorithms besides the proposed method were implemented on the collected datasets. The results of empirical experiments are presented in Table 7. Their Confusion Matrices and ROC curves are respectively shown in Figures 6 and 7. As can be seen, the proposed method has the highest performance based on all evaluation metrics, which clearly demonstrates the superiority of the proposed method for gastric cancer prediction.

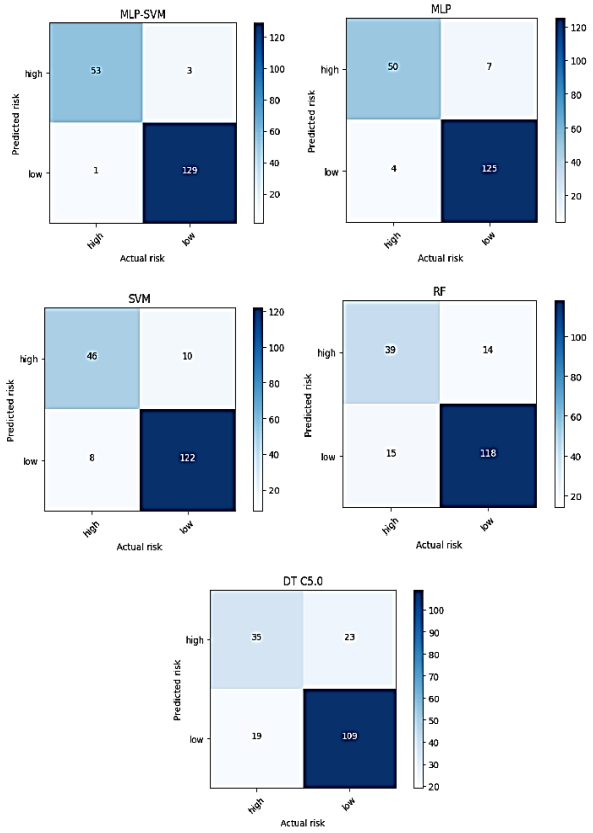

Figure 6. Confusion matrices.

Table 7. Test performance comparison using various performance evaluation metrics.

Algorithm	Accuracy	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	F1-score (F1)	AUC
MLP-SVM	**0.98**	**0.98**	**0.98**	**0.95**	**0.96**	**0.99**
MLP	0.94	0.93	0.95	0.88	0.90	0.97
SVM	0.90	0.85	0.92	0.82	0.84	0.93
RF	0.84	0.72	0.89	0.74	0.73	0.87
DT C5.0	0.77	0.65	0.83	0.60	0.63	0.81

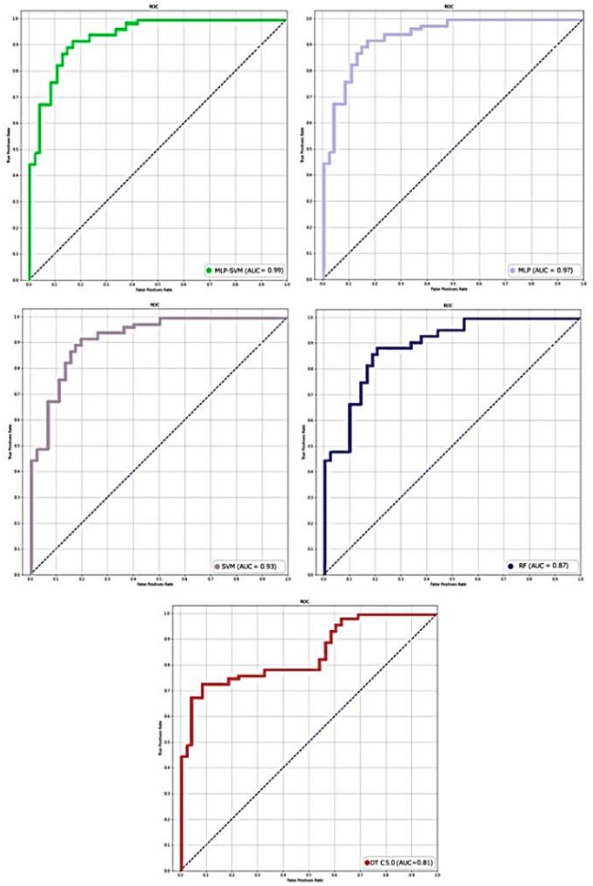

Figure 7. ROC curves.

4.5. Subset selection

Feature selection is the initial step in any data mining task. Therefore, conducting an essential experiment to evaluate the impact of individual features is another crucial aspect of addressing the given problem. Using the mentioned feature selection method, all independent and dependent optimal features were selected. The highest score was obtained with 20 features, which increased the accuracy of the proposed model to 98%. The feature score chart is shown in Figure 8.

The most important features selected by the Relief-F feature selection algorithm are also shown in Figure 9.

The relative importance of each independent variable was also computed using the proposed method and other baselines. Table 8 presents the key factors influencing the risk of gastric cancer, with their total importance exceeding the average total importance of all 36 features.

According to the results, Helicobacter pylori (HP) infection is identified as a risk factor for gastric cancer. HP is involved in the invasion, metastasis, and clinical staging of gastric cancer, thereby promoting its pathogenesis. Therefore, it serves as a potential marker for assessing the clinical progression and prognosis of gastric cancer.

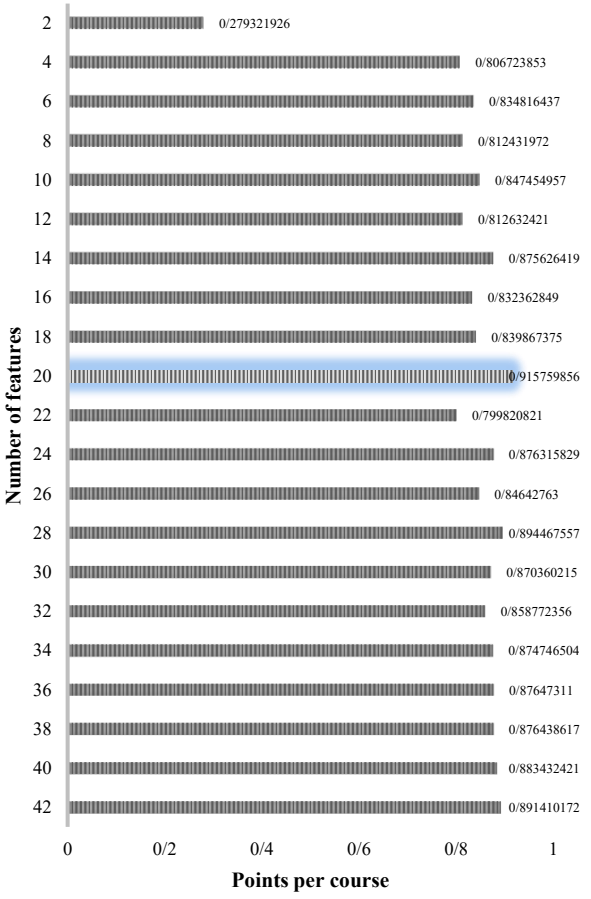

Figure 8. The feature score chart.

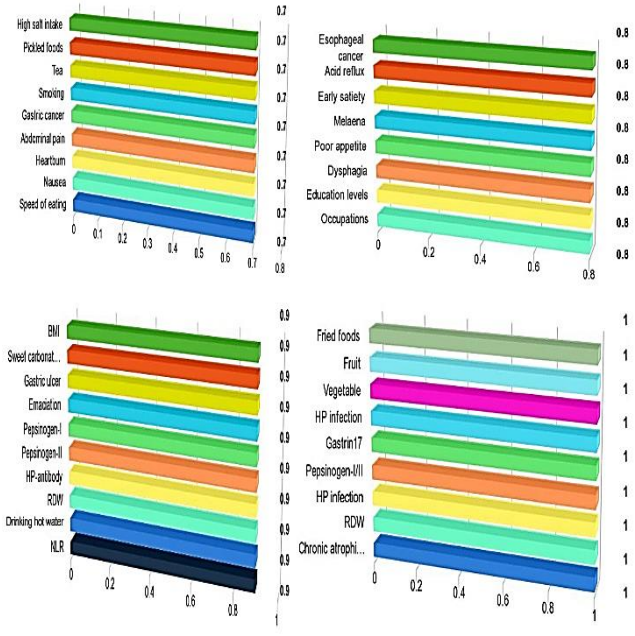

Figure 9. The most important features selected by the Relief-F feature selection algorithm.

Table 8. Important independent variables for the risk of gastric cancer.

Variables	MLP - SVM	MLP	SVM	RF	DT C5.0	Total
HP infection	0.07	0.06	0.07	0.06	0.05	0.31
Fried foods	0.07	0.05	0.05	0.07	0.04	0.28
Chronic atrophic gastritis	0.04	0.07	0.07	0.05	0.04	0.27
Fruit	0.04	0.04	0.05	0.07	0.03	0.23
HP antibody	0.06	0.05	0.04	0.04	0.03	0.22
Gastric ulcer	0.04	0.07	0.06	0.03	0.02	0.22
Pepsinogen-I/II	0.04	0.04	0.04	0.04	0.04	0.20
Gastrin17	0.05	0.04	0.04	0.03	0.03	0.19
Pepsinogen-I	0.05	0.04	0.04	0.03	0.03	0.19
Vegetable	0.04	0.04	0.04	0.03	0.03	0.18
BMI	0.03	0.03	0.03	0.04	0.04	0.17
Sweet and carbonated drinks	0.03	0.03	0.03	0.04	0.04	0.17
Acid reflux	0.02	0.02	0.02	0.05	0.05	0.16
Emaciation	0.03	0.03	0.03	0.03	0.03	0.15
NLR	0.03	0.03	0.03	0.03	0.02	0.14
RDW	0.03	0.03	0.03	0.03	0.02	0.14
Drinking hot water	0.03	0.03	0.03	0.03	0.02	0.14
Postprandial distress	0.03	0.03	0.03	0.02	0.03	0.14
High salt intake	0.03	0.03	0.03	0.02	0.02	0.13
Speed of eating	0.02	0.02	0.02	0.04	0.03	0.13

5. Conclusions

Early and accurate screening for gastric cancer can significantly improve patients' chances of survival. Treating patients in the early stages is also easier and less expensive than in later stages. Non-invasive diagnostic methods, implemented using artificial intelligence and machine learning algorithms, offer substantial benefits compared to traditional and more invasive methods. In this study, our sample population consisted of individuals who presented at the hospital with symptoms of indigestion or other digestive issues. Some of these individuals underwent endoscopy for further evaluation. By collecting various data including medical and disease records, conducting blood tests, and gathering demographic information; we categorized the patients into two groups: low-risk and high-risk, based on their potential for developing gastric cancer.

The presence of numerous predictive factors can pose challenges for both doctors and advanced software systems in analyzing the key factors for diagnosing the risk of gastric cancer. In this research work, we employed a proposed method that utilized cumulative enhancement improvement and featured a feature selection approach to achieve an impressive accuracy rate of 98%. Comparing our proposed method with other machine learning techniques revealed that, through proper training of data mining algorithms and precise selection of independent and dependent features, we can design a model that does not

require invasive methods. Instead, it relies on optimal features, considering patient information such as dietary habits, lifestyle, medical records, serological, and hematological tests. This model accurately examines the factors influencing gastric cancer and diagnoses individuals as low-risk or high-risk with the utmost precision.

There are numerous possibilities for improving this research work and overcoming the limitations of this study. One approach is to expand the scope by conducting the same experiment on larger real-world datasets. Further investigation can explore different combinations of data mining methods for predicting gastric cancer. Additionally, applying new feature selection methods can provide a wider understanding of the important features, thereby enhancing prediction accuracy.

References

[1] F. Hashemi Amin, M. Ghaemi, SM. Mostafavi, L. Goshayeshi, K. Rezaei, M. Vahed, and B. Kiani , "A Geospatial database of gastric cancer patients and associated potential risk factors including lifestyle and air pollution, "*BMC Research Notes*, vol. 14, pp. 1-3, 2021.

[2] M. Arbyn, E. Weiderpass, L. Bruni, S. De Sanjosé, M. Saraiya, J. Ferlay, F.Bray, "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis,"*The Lancet Global Health*, vol. 9, no. 2, pp. e191-e203, 2020.

[3] T. Samavat and E. Hojatzadeh, "Programs for prevention and control of cardiovascular diseases," *Ministery of Health*. Tehran: Javan, 2012.

[4] K. Kalan Farmanfarma, N. Mahdavifar, S. Hassanipour, and H. Salehiniya, "Epidemiologic study of gastric cancer in Iran: a systematic review,*" Clinical and experimental gastroenterology*, pp. 511-542, 2020.

[5] I.J.Choi, J.H. Lee , Y.I. Kim, C.G. Kim, S.J. Cho, J.Y. Lee, K.W. Ryu, B.H. Nam, M.C.Kook, and Y.W. Kim, "Long-term outcome comparison of endoscopic resection and surgery in early gastric cancer meeting the absolute indication for endoscopic resection," *Gastrointestinal endoscopy*, vol. 81, no. 2, pp. 333-341, 2015.

[6] S. Fukunaga, Y. Nagami, M. Shiba, M.Ominami, T.Tanigawa, H.Yamagami, H.Tanaka, K. Muguruma, T. Watanabe, K. Tominaga, and Y. Fujiwara, "Long-term prognosis of expanded-indication differentiated-type early gastric cancer treated with endoscopic submucosal dissection or surgery using propensity score analys*is,"* *Gastrointestinal endoscopy*, vol. 85, no. 2, pp. 143-152, 2017.

[7] Z. Khodaverdian, H. Sadr, S.A. Edalatpanah, and M. Nazari , "An energy aware resource allocation based on combination of CNN and GRU for virtual machine selection,*" Multimedia Tools and Applications*, pp. 1-28, 2023.

[8] M.P.Kalashami, M.M. Pedram and H. Sadr," EEG feature extraction and data augmentation in emotion recognition,*" Computational Intelligence and Neuroscience*, 2022.

[9] Z. Khodaverdian, H. Sadr, M. Nazari, and S.A. Edalatpanah, " Predicting the workload of virtual machines in order to reduce energy consumption in cloud data centers using the combination of deep learning models," *Journal of Information and Communication Technology*, vol. 55, no. 5, pp. 158- 63, 2023.

[10] M. Rejaul, I. Royel, M.A. Jaman, F. Masud, A. Ahmed, and A. Muyeed, "Machine learning and data mining methods in early detection of stomach cancer risk,*" Journal of Applied Science and Engineering*, vol.24, no.5, pp. 1-8, 2021.

[11] Y. Li, A. Feng, S. Zheng, C. Chen, and J. Lyu, "Recent estimates and predictions of 5-year survival in patients with gastric cancer: A model-based period analysis," *Cancer Control*, vol. 29, 2022.

[12] M.R. Afrash, M. Shafiee, and H. Kazemi-Arpanahi, "Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors," *BMC gastroenterology*, vol. 23, no. 1, pp. 1-13, 2023.

[13] M. Roostaee, and R. Meidanshahi, "Hidden Patterns Discovery on Clinical Data: An Approach based on Data Mining Techniques,*" Journal of AI and Data Mining*, 2023.

[14] J. Brownlee, "Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end," *Machine Learning Mastery*, 2016.

[15] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, and D.N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques,*" TrAC Trends in Analytical Chemistry*, vol. 132, 2020.

[16] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique,*" Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.

[17] M. Nakamura, Y. Kajiwara, A. Otsuka, and H. Kimura, "Lvq-smote–learning vector quantization based synthetic minority over–sampling technique for biomedical data," *BioData mining*, vol. 6, no.1, pp. 1-10, 2013.

[18] H. Sadr, M. Nazari, M.M. Pedram, and M. Teshnehlab, "Exploring the efficiency of topic-based models in computing semantic relatedness of geographic terms.*" International journal of web research*, vol.2 , no. 2, pp. 23-35, 2019.

[19] F. Mohades Deilami, H. Sadr and M. Tarkhan, "Contextualized multidimensional personality recognition using combination of deep neural network and ensemble learning,*" Neural Processing Letters*, vol. 54, no. 5, pp. 3811-3828, 2022

[20] H. Yin and K. Gai. "An empirical study on preprocessing high-dimensional class-imbalanced data for classification," in *2015 IEEE 17th international conference on high performance computing and communications, 2015 IEEE 7th international symposium on cyberspace safety and security, and 2015 IEEE 12th international conference on embedded software and systems*, IEEE, 2015.

[21] A.K. Shukla, S.K. Pippal, S. Gupta, B. Ramachandra Reddy, and D. Tripathi, "Knowledge discovery in medical and biological datasets by integration of Relief-F and correlation feature selection techniques," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 6637-6648, 2020.

[22] D.D. Nguyen, P.C. Roussis, B.T. Pham, M. Ferentinou, A. Mamou, D.Q. Vu, Q.A.T. Bui , D.K. Trong, and P.G. Asteris , "Bagging and multilayer perceptron hybrid intelligence models predicting the swelling potential of soil," *Transportation Geotechnics*, vol. 36, 2022.

[23] Z. Khodaverdian , H. Sadr, and S.A. Edalatpanah. "A shallow deep neural network for selection of migration candidate virtual machines to reduce energy consumption," in *2021 7th International conference on web research (ICWR)*, IEEE, 2021.

[24] A. Bellili, M. Gilloux, and P. Gallinari, "An MLP-SVM combination architecture for offline handwritten digit recognition: Reduction of recognition errors by Support Vector Machines rejection mechanisms," *Document Analysis and Recognition*,vol. 5, no. 4 pp. 244-252, 2003.

[25] M. Awad, R. Khanna, M. Awad, and R. Khanna, "Support vector machines for classification,*" Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designer*s, pp. 39-66, 2015

[26] H. Sadr, M.M. Pedram, and M. Teshnehlab , 2021. "Convolutional neural network equipped with attention mechanism and transfer learning for enhancing performance of sentiment analysis", *Journal of AI and data mining*, vol. 9, no. 2, pp.141-151, 2021.