Publications

Available Online

J. Sci. Res. **14** (3), 943-958 (2022)

JOURNAL OF
SCIENTIFIC RESEARCH

www.banglajol.info/index.php/JSR

Lifestyle and Dietary Factors Causing Microsatellite Instability Gastric Cancer Detected using Ensemble Modeling

S. K. Brindha[1], P. Charkarborthy[2], S. Chenkual[3], J. Zohmingthanga[4], J. L. Pautu[5], P. Nath[1], A. Maitra[6], N. S. Kumar[2], L. Hmingliana[1*]

[1]Department of Computer Engineering, Mizoram University, Aizawl - 796004, Mizoram, India

[2]Deparment of Biotechnology, Mizoram University, Aizawl – 796004, Mizoram, India

[3]Department of Surgery, Civil Hospital, Aizawl – 796001, Mizoram, India

[4]Department of Pathology, Civil Hospital, Aizawl – 796001, Mizoram, India

[5]Department of Oncology, Mizoram State Cancer Institute, Zemabawk, Aizawl, Mizoram, India

[6]National Institute of Biomedical Genomics, Kalyani - 741251, West Bengal, India

Received 19 February 2022, accepted in final revised form 2 June 2022

Abstract

Aim of this study is to identify diet and lifestyle patterns that cause microsatellite instability gastric cancer (MSI-GC) using supervised machine learning algorithms. There were 142 genetic variants acquired *via* targeted resequencing of 60 biomarker genes from gastric tumor samples and tabulated with respect to MSI status, diet and lifestyle characteristics. Four classifiers (logistic regression, random forest, logistic regression, multilayer perceptron) were used to train the data and evaluated based on their classification efficiency. Data analysis revealed features extracted using ridge regression: extra salt, smoked food, smokeless tobacco products (Khaini /sadha), alcohol and betel nut leaf with lime (khuva) were the core factors for causing MSI-GC. The extracted features were exploited using random forest and multilayer perceptron classifiers, which has produced accuracy, precision, recall, F1 score, and Receiver operating characteristics (ROC) curve of 96 %. The brier score was 0.04 and Matthews correlation coefficient (MCC) was +0.91. Linear regression results revealed khuva was main driving factor and extra salt, smoked food, khaini/sadha and alcohol were confounding factors to cause MSI-GC. This is a first-time report that integrates mutation and diet-lifestyle data using machine learning, to precisely identify the driving and confounding factors for causing MSI-GC.

Keywords: Microsatellite instability gastric cancer (MSI-GC); Matthews correlation coefficient (MCC); Brier score; Exome data; Diet; Supervised learning algorithms.

doi: http://dx.doi.org/10.3329/jsr.v14i3.58331 J. Sci. Res. **14** (3), 943-958 (2022)

1. Introduction

Microsatellites are short repetitive DNA regions spread randomly in the genome. Microsatellite Instability (MSI) is a mechanism that is caused by mismatch repair system

[*] *Corresponding author*: lalhmingliana@mzu.edu.in

deficiency (MMRD) leading to abnormal insertion or deletion of the nucleotide bases in microsatellite regions during DNA replication [1]. Gastric cancer (GC) is third leading cancer in the world with high mortality rate. MSI-GC is one of the subtypes of GC which is reported high in western population than in the asian population [1]. MSI is caused by hyper mutation after MMRD phase in tumor cells, which causes genetic instability in cancer. The prognosis after chemotherapy is slow in MSI subtypes when compared other gastric cancer subtypes such as Epstein-Bar virus (EBV), Chromosomal instability (CIN) and Genomically stable [2]. Much research has been carried out in identifying the GC with MSI, speed of prognosis after chemotherapy and survival period after the surgery using molecular biology techniques [3].

Salattery *et al*. had developed case-control based statistical model in which excessive alcohol drinking and smoking were significant lifestyle factors to cause MSI in colon cancer [4]. Support vector machine classifier had been used to predict MSI status in GC cases using relief-based forward selection algorithm for feature selection from long-noncoding RNAs from The Cancer Genome Atlas [5]. A statistical case-control observational study reported to have high association between red meat, high protein and nitrites intake with MSI in GC subtypes, and negatively correlated in white meat consumption, protective effects with consumption of fruits, vegetables, antioxidants and legumes [6]. As gastric cancer is mainly caused by diet-lifestyle factors, we are interested to associate the lifestyle-dietary patterns with gene data to identify the risk factors for MSI-GC using machine learning algorithms. Not much data mining work had been done to identify the diet and lifestyle risk factors causing MSI-GC by integrating the exome data. The case-control study detects the etiology of the disease by comparing the characteristics of patients (cases) and healthy individuals (controls). This type of study has information bias with regard to the subtype characteristics due to absence of MSI and pathogens in control samples [7].

The aim of the present study is to find the lifestyle and dietary risk factors by retaining the MSI and pathogen as targets. Hence, retrospective cohort study was done by considering the case samples only. The present study was designed to identify the diet and lifestyle factors by integrating both the synonymous and non-synonymous mutations to produce precise and accurate results for MSI-GC.

2. Experimental

2.1. *Data Source*

The lifestyle and dietary data were collected using well-structured questionnaires. There were 79 gastric cancer patients who had been followed from 2016-2019 (3 years) at the Civil Hospital, Aizawl, Mizoram, India. The cases were confirmed via. both endoscopy and biopsy. The dataset comprised of 11 significant risk factors causing gastric cancer, has no missing data and noise, as the questionnaires had been answered via. in-person interaction with the patient’s and their health records precision were carefully observed for good feature engineering and better decision making. Targeted resequencing of 60 gene panel specific to gastric cancer was done using next generation sequencing (NGS), 142 variants (both synonymous and non-synonymous) were identified by base-by-base variant calling tool (BBB) [8]. For this retrospective case study, we integrated somatic mutations with their respective case's diet and lifestyle data. For further analysis, 70 datapoints (somatic mutations) positive for MSI and 72 negative for MSI were obtained showing a balanced dataset.

2.2. *Feature engineering*

Feature extraction and learning curves were the significant steps before applying machine learning algorithms on biological data. Learning curves were generated to rule out underfitting and feature selection was done to rule out overfitting, these both were significant problems in data mining methods. The feature selection helps the model to train faster and it becomes less computationally intensive by eliminating the redundant features from the dataset [9]. The main idea was to choose a model that can perform well on unknown data in the future and to extract key features that cause MSI-GC. The learning curves were generated for group of five algorithms (logistic regression, Naive Bayes, multilayer perceptron, support vector machine and random forest) [10]. Features selection and redundant data eliminations were carried using three popular methods: ridge regression [11], extra trees classifier [12] and recursive feature elimination methods [13].

In the present study, the core lifestyle and food habits were chosen as attributes and were evaluated for their risk towards MSI- GC. Hence, we had selected three powerful feature selection algorithms due to their following advantages: a) Ridge regression Method- In this method, ridge penalty avoids nullification of positive and negative co-related features, thereby this plays a significant role in identifying the features of importance. In addition, our dataset has multi-collinearity features which can be efficiently handled by Ridge regression method [14], b) Extra trees classifier- it appropriately ranks the correlated features due to high amount of stochastic-ness in splitting the node during the construction of decision trees and has very low chances for over-fitting [12], and c) Recursive Elimination Method- It is a significant method for feature extraction for a small biological dataset. It removes the feature with least importance during every iteration and ranks them [15].

2.3. *Packages*

Python Jupyter Notebook Version 3 platform was used to build data models using learning packages from scikit-learn (0.20.4) [16]. Numpy (1.16.6), pandas (0.24.2), seaborn (0.9.1), scipy (1.2.3) and matplotlib (2.2.5) had been imported for interpretation and classification of the datapoints [17].

2.4. *Data Labeling*

There were 11 attributes in the dataset which comprised of both diet and lifestyle data, as these were the potential risk factors of gastric cancer based on the past works. Feature set included: sex, age, extra salt, saum, smoked food, Khaini/ sadha, tuibur, alcohol, smoking, khuva and class (MSI - 1 or non MSI - 0), as labeled in Table 1. These gastric cancer risk features had been extensively studied for more than ten years in different populations using conventional statistical methods.

Table 1. Diet and Life Style features analyzed for Gastric Cancer.

Features	Data Labelling
Age (years)	(30-85)
Sex	(1 - Male, 2- Female)
Extra salt	(0 - None, 1 - Little to Average, 2 - High/Excess)
Sa-um	(0 - None, 1 - Little to Average, 2 - High/Excess)
Smoked Food	(0 - None, 1 - Little to Average, 2 - High/Excess)
Khaini / Sadha	(0 - None, 1 - Little to Average, 2 - High/Excess)
Tuibur	(0 - None, 1 - Little to Average, 2 - High/Excess)
Alcohol	(0 - Non drinker, 1 - once or twice a month, 2 - more than once a week, 3 - Daily drinker
Smoking	(0 - No, 1 - Yes)
Paan	(0 - No, 1 - Yes)
MSI	(0 – Negative, 1- Positive)

Sa-um- Fermented pork fat; Smoked Food- smoked vegetables and meat; Khaini / Sadha- smokeless tobacco products; Tuibur- smoke filled tobacco water; Paan- betel leaf with lime and raw arecanut; MSI- Microsatellite Instability

2.5. *Data models*

The lifestyle-diet dataset was randomly divided in 70:30 ratio; 70 (95 instances) was taken for training and 30 (47 instances) was reserved for testing the model and random state=0. Logistic regression, random forest, multilayer perceptron, and naïve bayes models were trained using training dataset. The trained models were tested on the test dataset and evaluated based on: accuracy, precision, recall, F1 score, brier score, Matthews correlation coefficient (MCC) and Receiver operating characteristics (ROC) [18-20] were taken to evaluate the performance of the models.

2.6. *Evaluation metrics*

A trade-off was taken between the ROC, accuracy, precision, recall, brier score, and MCC to achieve a precise-decision-making system, because focusing only on accuracy, precision and recall might misguide our perception on the outcome. MCC calculates the classifier's performance based on independent majority of predictions for both positive and negative classes, which an appropriate evaluation metric for binary classifiers. Furthermore, MCC score ranges from -1 to +1, the value will be high if and only if the classifier can classify majority of both positive and negative classes [20], so this parameter had been given a great weightage in determining our model's performance. As our sample size was small, we had randomly split the dataset into two parts training and testing before building the data models, the testing data was used only to validate the performance of the data models and to acquire unbiased outcomes [21]. The data pre-processing, model building, validation, optimization, and result visualization were carried out in Python 3 Jupyter notebook using scikit-learn modules (Fig. 1).

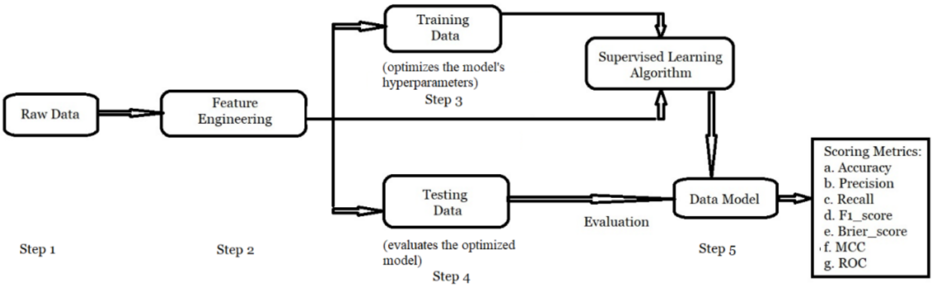


Fig. 1. Flow chart of Feature selection, modelling, optimization and performance measures.

Step 1: Raw diet-lifestyle and mutation data were collected and labeled based on MSI status

Step 2: Feature engineering (Feature selection) was done using Ridge regression, Recursive elimination method and Extra trees classifier.

Step 3: The dataset was split into two-third for training and optimize with different hyperparameter settings to fit and select the model.

Step 4: One-third of dataset used for testing and to evaluate the optimized model.

Step 5: Evaluate the models by comparing the selected performance measures (accuracy, precision, recall, F1 score, ROC, MCC and brier score).

Though the sample size was handful, we had not oversampled the data, as this will hamper our final outcome. So the original collected data from the patients were only used for this classification and interpretation. The presented was balanced dataset, so ROC curve was taken as one of performance measure which showed significant trade-off between false positives and true positives. The ROC was considered to be the gold standard for evaluating biological binary classifiers, especially for the positive class.

Matthews correlation coefficient (MCC) was considered as another very significant evaluation parameter to compare the performance among the classifiers, because the highest value MCC of +1 will be only achieved when majority of positive and negative cases were predicted. Brier score was considered to more significant than the accuracy, as accuracy can mislead lead the model's true potential on an unknown data, the brier score is mean squared error between expected and predicted values, they range between 0 to 1, the smaller the value, the better the performance of the model [22].

The present work focuses to determine the MSI-GC risk factors from diet-lifestyle and mutation characteristics and to determine appropriate classification algorithm which can perform better on them. These extracted features can evidently be the core cause of microsatellite instability gastric cancer.

3. Results and Discussion

Learning curves were constructed using all eleven attributes for a set of 5 supervised machine learning algorithms: random forest, multilayer perceptron, support vector machine, logistic regression and Naive Bayes to rule out underfitting. The learning curves showed all above algorithms were best fit to classify our dataset, except for support vector machine. The error rate between the training score and validation score was wider in learning curve of support vector machine, whereas other four algorithms showed very smooth convergence between the training score curve and validation score curve clearly showed reduced error gaps (Fig. 2).

Feature extraction was done using three powerful methods: ridge regression, recursive elimination method and extra trees classifier to remove irrelevant attributes, to extract strongly correlation features and to rule out overfitting. Ridge regression method showed extra salt, smoked food, Khaini/sadha, alcohol and khuva were top five potential features for causing MSI-GC (Fig. 3). Recursive elimination method had extracted extra salt, Khaini/ sadha, tuibur, alcohol and smoking as top five significant features (Table 2). Extra trees classifier's top five correlated features were extra salt, saum, smoked food, alcohol and khuva (Fig. 4). The selected four algorithms: random forest, multilayer perceptron, logistic regression and Naive Bayes were devised to create data models on above-mentioned three groups of feature sets. Feature selection and learning curves had ruled out no underfitting and overfitting in our dataset, these two were very crucial steps in data preprocessing to enhance accuracy and to reduce the misclassification of datapoints, before building the data models.

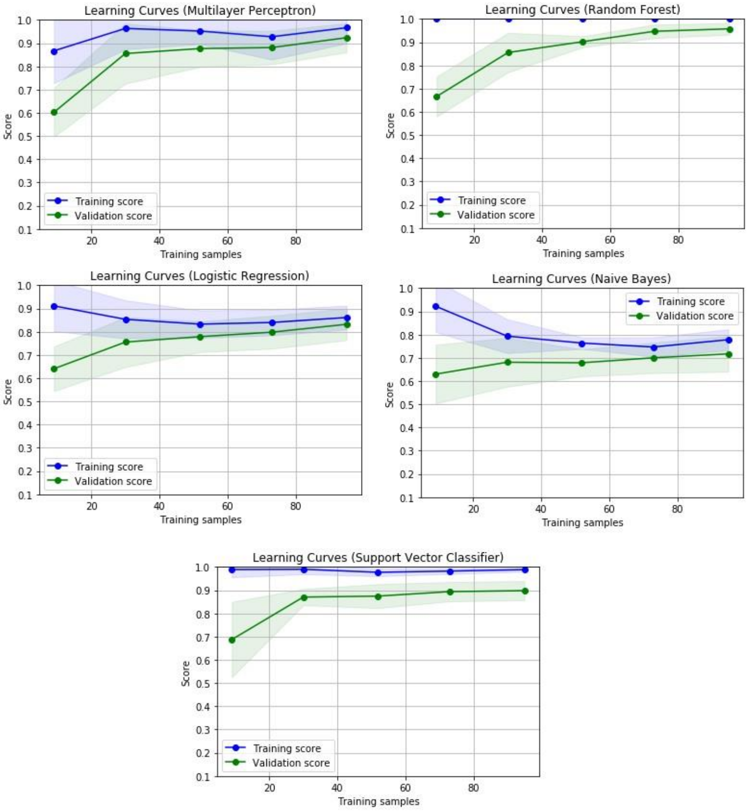


Fig. 2. Learning curves of five supervised learning algorithms.

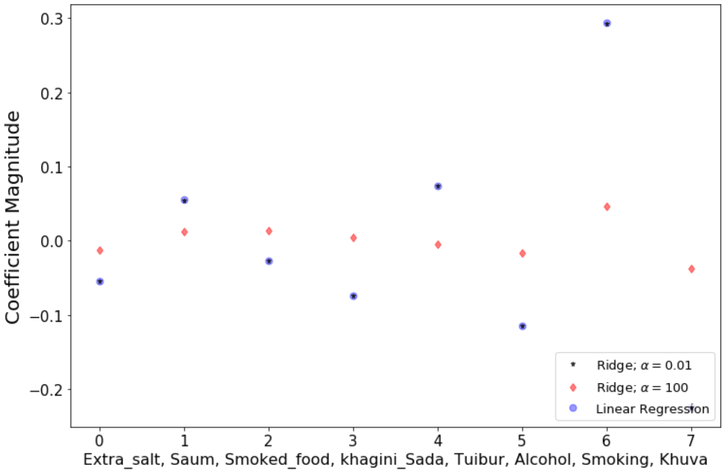


Fig. 3. Feature selection using Ridge regression.

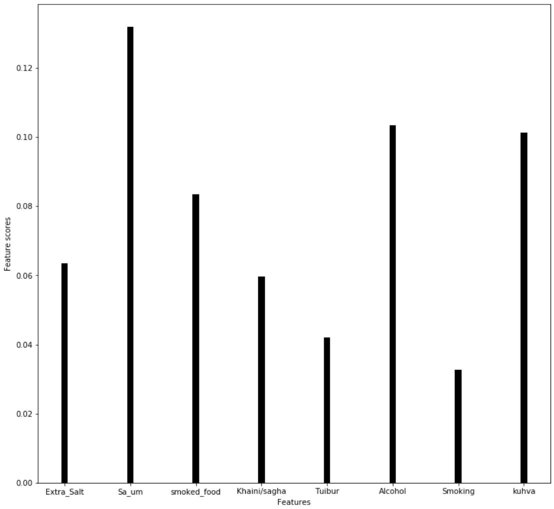


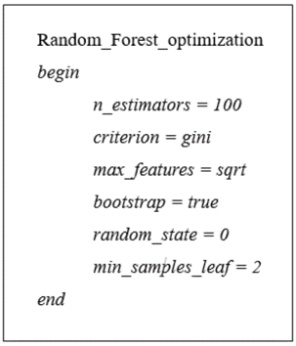
Fig. 4. Feature selection using extra trees classifiers.

Table 2. Feature Selection using Recursive Elimination Method.

Features	Selected Features	Feature Ranking
Extra Salt	True	1
Saum	False	4
Smoked food	False	2
Khaini / Sadha	True	1
Tuibur	True	1
Alcohol	True	1
Smoking	True	1
Khuva	False	3

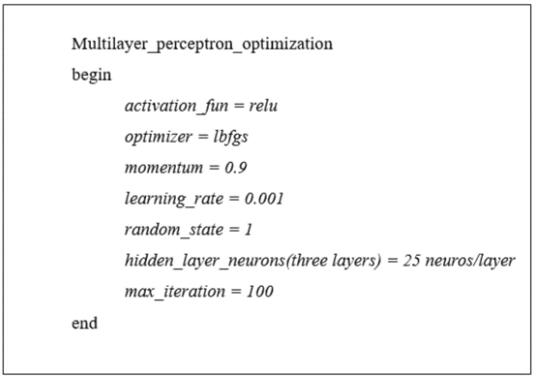
3.1. *Ridge regression feature set and data analysis*

Ridge regression selected features were extra salt, smoked food, Khaini/sadha, alcohol and khuva was exploited to build Random Forest, Multilayer perceptron, logistic regression and Naive Bayes classifiers. Random forest and multilayer perceptron classifiers were optimized using hyper-parameters as shown in Fig. 5 and 6, respectively. Hyperparameter tuning was one of the critical steps in the training the model, which produces a big gain during the testing phase [23]. The number of decision trees to build our random forest model was chosen as 100 (Fig 5), as discussed in study by Oshiro *et al*. where the information gain does not increase by doubling the trees, the optimal gain 100% area under curve (AUC) was achieved between 64-128 decision trees [24]. Gini index was used to split node in our decision tree, this hyperparameter optimization had been popularly used in biological data mining [25].



```
Random_Forest_optimization
begin
        n_estimators = 100
        criterion = gini
        max_features = sqrt
        bootstrap = true
        random_state = 0
        min_samples_leaf = 2
end
```

Fig. 5. Random forest optimization using hyper parameters.



```
Multilayer_perceptron_optimization
begin
        activation_fun = relu
        optimizer = lbfgs
        momentum = 0.9
        learning_rate = 0.001
        random_state = 1
        hidden_layer_neurons(three layers) = 25 neuros/layer
        max_iteration = 100
end
```

Fig. 6. Multilayer perceptron optimization using hyper parameters.

Similarly, multilayer perceptron model was also tuned using key hyperparameters: three layers of 25 hidden neurons were created in this neural network, relu activation function was applied to avoid vanishing gradient problem, to reach the global minima momentum was set to 0.9, as our dataset was small, we had chosen ‘lbfgs’ (Limited-memory Broyden Fletcher Goldfarb Shanno) optimizer [26], we were able to optimize the model with 100 iterations (Fig. 6), beyond and below these parameters, we did not get satisfactory results.

The evaluation of Random Forest and multilayer perceptron models had produced accuracy, precision, recall and F1 measure, ROC of 96 % each, these values had very distinctly shown that the-above two algorithms were well-balanced classifier with Matthews correlation coefficient (MCC) of +0.91 and brier score of 0.04 (Table 3 and Fig. 7). MCC value was +0.91 which can be approximated to +1, showing the highly satisfactory performance of the classifier and brier score was 0.04, which was below 0.175

as it was considered to be the very precise model in terms of both performance and accuracy [20,27].

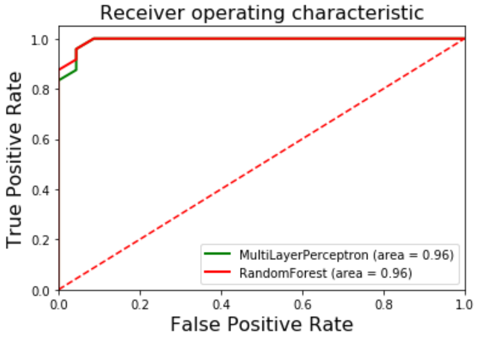


Fig. 7. ROC of random forest and multilayer perceptron models based on Ridge regression feature set.

Table 3. Performance measures of data models with three feature sets.

Feature sets	Classifiers	Accuracy %	Precision %	Recall %	ROC curve %	F1 score %	Brier score	MCC
Ridge regression: Extra salt Smoked food Khaini /Sadha Alcohol Khuva	Random forest	96	96	96	96	96	0.04	+0.91
	Multilayer perceptron	96	96	96	96	96	0.04	+0.91
Recursive elimination method: Extra salt Khaini/Sadha Tuibur Alcohol Smoking	Random forest	94	96	92	94	94	0.06	+0.87
	Multilayer perceptron	94	96	92	74	94	0.06	+0.87
Extra trees classifier: Extra salt Saum Smoked food Alcohol Khuva	Random forest	85	81	92	85	86	0.14	+0.70
	Multilayer perceptron	85	81	91	85	86	0.14	+0.70

3.2. *Recursive elimination feature set and data analysis*

Recursive elimination method extracted feature set was extra salt, Khaini/sadha, tuibur, alcohol and smoking; these features had been utilized to construct data models using random forest, Naive Bayes, logistic regression and multilayer perceptron algorithms. Random forest and multilayer perceptron classifiers were optimized using hyper-parameters settings as shown in Figs. 5 and 6, respectively during the training phase. Random forest and multilayer perceptron models had produced very same results with accuracy 94 %, precision 96 %, recall 92 %, ROC 94 %, F1 score 94 %, brier score 0.06 and MCC of +0.87 (Table 3 and Fig. 8).

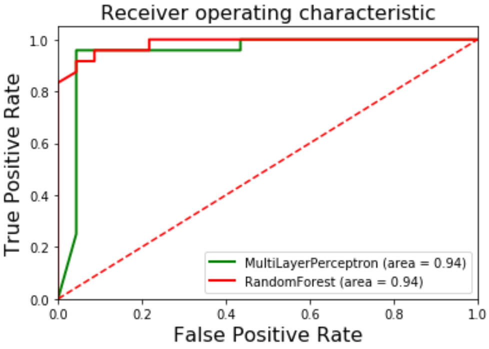


Fig. 8. ROC of Random Forest and multilayer perceptron models based on Recursive Elimination Method feature set.

Random forest classifier was tuned for both gini index and entropy for better information gain, where both gave the same evaluation scores. Though the brier score and MCC were below the accepted range 0.175, and +1, respectively and the other parameters were good but not very satisfactory to prove the recursive elimination method feature set to cause the MSI-GC due to the mild deterioration in accuracy, recall, F1 score, ROC, and MCC values (Table 3) when compared to results produced by ridge regression feature set.

3.3. *Extra trees classifiers feature set and data analysis*

Feature set from extra trees classifier was extra salt, saum, smoked food, alcohol and khuva, had been used to develop models. We were able to achieve accuracy 85 %, precision 81 %, recall 91 %, ROC 85 %, F1 score 86 %, brier score 0.14 and MCC +0.70 (Table 3 and Fig. 9) using algorithms: multilayer perceptron tuning was done using hyperparameters as mentioned in Fig. 6 and random forest was tuned for both impurity measures gini index and information-gain entropy, and other parameters were as illustrated in Fig. 5. Extra trees classifier's features of importance were not significantly correlated to MSI-GC, as the evaluation metrics scores were below average performance when compared to other two feature sets derived from ridge regression and recursive elimination methods.

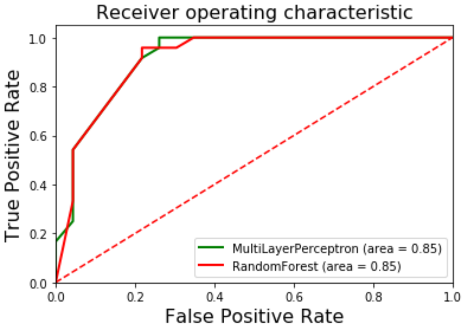


Fig. 9. ROC of random forest and multilayer perceptron models based on extra trees classifier feature set.

Furthermore, random forest is an ensemble modeling, had been proved to be the best models to predict the disease rate from epidemiological features and to reduce the disease occurrence in biological domain with no issues of overfitting [28]. Multilayer perceptron is a non-linear back propagation algorithm, which had predicted heart disease with accuracy, sensitivity and specificity of 98 % each with 4000 epochs from features (age, sex, type of chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, old peak and slope) between 164 healthy and 139 cases [29]. In addition, random forest classifier has produced high accuracy to classify churna names of Siddha medicine (Indian Traditional Medicine) [30]. From the above previous works, random forest and multilayer perceptron had been widely used in bio-data mining, our results also performed highly satisfactory results for these two strong classifiers.

Based on the above discussions, it was evident that extra salt, smoked food, khaini/sadha, alcohol and khuva (features extracted by Ridge regression) were primary lifestyle-diet factors causing microsatellite instability gastric cancer in Mizoram population. Among these predicted diet-lifestyle factors (extra salt, smoked food, khaini/sadha, alcohol and khuva): salt was known to be the strongest risk factors causing gastric cancer with confidence index >95 %, based on study by *Lee and Derakhshan* [31]. Smoked food was significantly contributed to cause of gastric cancer in 78 % of cases in North-eastern states of India [32]. Khaini/sadha (smokeless tobacco products) mixed with lime was also chewed raw or along with betel leaf was also the risk factor for stomach cancer with high mortality rate in Mizoram among the Northeastern states in India [33]. Alcohol had been one of the significant factors to cause to stomach cancer in Mizoram population [34]. Khuva (betel leaf with raw arecanut and lime or paan) is often consumed in empty stomach by people in Mizoram and had been shown to be highly prevalent risk factor for causing stomach cancer [35,37].

4. Identification of Confounders and Driving Factor

Multiple linear regression was used which involved multiple co-variates to identify the confounding factors [36]. The regression results showed features: extra salt, smoked food, khaini / sadha and alcohol (independent variables) vary with the MSI-GC (dependent variable), showing these were potential confounders. The main driving factor for causing MSI-GC is Khuva showing a significant p-value 0.0007 and slope -0.2142 which shows high negative co-relation between Khuva and MSI-GC when compared to rest of the features (Table 4). The regression coefficient -0.20 showed the dependent variable (MSI-GC) was highly dependent variable (Khuva) and mean square error 0.13 both were lowest values among the other attributes where the error rate between the actuals and predicted was very minimal (Table 4). Khuva (betel quid) chewers were prone to have MSI in head and neck cancer patients [37] and it is also one of main driving factor for causing several MSI regions in oral cancer [38]. This was the first report which showed our predicted features such as: smoking tobacco, excess salt, smoked food, and alcohol were confounding factors and khuva was the driving factor for causing MSI in gastric cancer.

Table 4. Estimation of coefficients of linear model.

Independent variable	Slope	p-value	Regression coefficient	Mean square error
Extra Salt	-0.0349	0.573	-0.032	0.15
Smoked Food	0.1014	0.127	0.096	0.16
Khaini / Sadha	0.0308	0.608	-0.007	0.15
Alcohol	-0.0633	0.303	-0.066	0.15
Khuva	-0.2142	0.0007	-0.200	0.13

5. Logistic Regression and Naïve Bayes Models with Three Feature Sets

Even though, the logistic regression and Naive Bayes algorithm showed no underfitting of our dataset, the models evaluation metrics showed MCC values were below +0.50, accuracy, precision, recall, F1 score, and ROC were below 75 % with brier score above 0.20, for all three feature sets extracted by Ridge regression, recursive elimination method and extra trees classifier method. These score clearly showed very average performance of logistic regression and naïve bayes classifiers, thus they were not discussed further due to their unsatisfactory outcomes.

6. Remarkable Stepes Taken to Develop Precise-Data Models

The significant criteria followed to achieve high precision in predicting MSI-GC in this work are: a) To avoid the precision and recall bias in our results, we had carried out the investigation using case-case study design. The case-case study can significantly key out the underlying diet-lifestyle patterns for causing MSI-GC cancer under high precision. b) It had been well proved in asian and western population-based studies that occurrence of MSI-GC in old-age group and our interest was to find the dietary and lifestyle features that causes MSI-GC, we had not included age during the feature selection phase. c) The mutations data were selected as they alter the protein function, every mutation was integrated with the patient's lifestyle and dietary features to frame the dataset for the present work. This integrated dataset can evidently feature out the diet-lifestyle risk factors for causing MSI-GC as diet and lifestyle changes the genetic code in any disease. d) The presented was a balanced dataset with 142 instances in our dataset (70 instance positive for MSI and 72 instances for non-MSI), so ROC was a significant evaluation parameter which was given a higher weightage to grade the performance of the data models. e) Feature selection was one of main phase in data preprocessing as it avoids overfitting, reduces training time, improves the accuracy and precision by selecting the right subset of features and decreases the complexities in biological data. f) Learning curves were used to select the best classifier to well classify this heterogeneous data. g) Hyperparameter tuning was does before the training phase to optimize the models. h) Besides the accuracy, precision, and recall scores, the error rates were significantly scrutinized using brier score – this value should be below 0.175 and MCC values nearing to +1 were strictly observed to compare the performance between the classifiers. The models qualified for the above two error parameters were only taken into consideration for further data interpretation and discussion in this work.

7. Conclusion

This is the first computationally integrated study of gene-diet characteristics that had filtered out the MSI-GC causing risk factor from lifestyle-diet factors of gastric cancer patients. MSI-GC causing factors were extra salt, smoked food, Khaini/sadha, alcohol and khuva based ridge regression feature selection. These features had shown remarkable performance on random forest and multilayer perceptron models by scoring high and balanced accuracy, precision, recall, f1 score, ROC, MCC values and low error rates via brier scores. This information gain will be very valuable for the physicians and the public to avoid these foods and habits for healthy living and for better prognosis and treatment of microsatellite instability gastric cancer.

Acknowledgments

The authors thank DBT, New Delhi for establishing the Bioinformatics Infrastructural Facility (No. BT/BI/12/060/2012 - BTISNeT), Mizoram University which was utilized for completing this work. The authors also thank the patients who gave consent for the questionnaire data and samples as well as to the Lab technicians for their contribution.

References

1. M. Ratti, A. Lampis, J. C. Hahne, R. Passalacqua, and N. Valeri, Cell. Mol. Life Sci. **75**, 4151 (2018). https://doi.org/10.1007/s00018-018-2906-9
2. E. Puliga, S. Corso, F. Pietrantonio, and S. Giordano, Cancer Treat. Rev. **95**, ID 102175 (2021). https://doi.org/10.1016/j.ctrv.2021.102175
3. F. Zhao, X. Yuan, D. Ren, G. Shen, Z. Wang, F. Zheng, R. Ahmad, Z. Ma, and J. Zhao, J. Environ. Pathol. Toxicol. Oncol. **38**, 21 (2019). https://doi.org/10.1615/JEnvironPatholToxicolOncol.2018026876
4. M. L. Slattery, K. Anderson, K. Curtin, K. N. Ma, S. Schaffer, and W. Samowitz, Int. J. Cancer **93**, 601 (2001). https://doi.org/10.1002/ijc.1370
5. T. Chen, C. Zhang, Y. Liu, Y. Zhao, D. Lin, Y. Hu, J. Yu, and G. Li, BMC Genomics **20**, 846 (2019). https://doi.org/10.1186/s12864-019-6135-x
6. D. Palli, A. Russo, L. Ottini, G. Masala, C. Saieva, A. Amorosi, A. Cama, C. D.'Amico, M. Falchetti, R. Palmirotta, A. Decarli, R. M. Costantini, and J. F. Fraumeni, Cancer Res. **61**, 5415 (2001).
7. W. J. Song and K. C. Chung, Plast Reconstr. Surg. **126**, 2234 (2010). https://doi.org/10.1097/PRS.0b013e3181f44abc
8. A. Maitra, N. Biswas, K. Amin, et al. Nat. Commun. **4**, 2873 (2013).
9. S. Gupta and R. R. Sedamkar, J Sci. Res. **13**, 901 (2021). https://doi.org/10.3329/jsr.v13i3.53290
10. J. N. van Rijn, S. M. Abdulrahman, P. Brazdil, and J. Vanschoren, Int Symp. Intelligent Data Anal. 298 (2015). https://doi.org/10.1007/978-3-319-24465-5_26
11. J. Friedman, T. Hastie, and R. Tibshirani, J. Stat. Software **33**, 1 (2010). https://doi.org/10.18637/jss.v033.i01
12. E. Hemphill, J. Lindsay, C. Lee, I. I. Măndoiu, and C. E. Nelson, BMC Bioinformatics **15** ID S4 (2014). https://doi.org/10.1186/1471-2105-15-S13-S4
13. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Machine Learning **46**, 389 (2002). https://doi.org/10.1023/A:1012487302797
14. R. Muthukrishnan and R. Rohini, IEEE Xplore 18 (2016). https://doi.org/10.1109/ICACA.2016.7887916
15. Q. Chen, M. Zhaopeng, L. Xinyi, J. Qiangguo, and S. Ran, Genes **9**, 301 (2018).
16. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Machine Learning Res. **12**, 2825 (2011).
17. J. D. Hunter, Comput. Sci. Eng. **9**, 90 (2007). https://doi.org/10.1109/MCSE.2007.55
18. G. W. Brier, Monthly Weather Rev. **78**, 1 (1950). https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
19. K. H. Tilaki, Caspian J. Int. Med. **4**, 627 (2013).
20. D. Chicco and G. Jurman, BMC Genomics **21**, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7
21. A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, Plos One **14**, ID e0224365 (2019). https://doi.org/10.1371/journal.pone.0224365
22. R. Dinga, B. W. J. H. Penninx, D. J. Veltman, L. Schmaal, and A. F. Marquand, Bio Rxiv (2019). https://doi.org/10.1101/743138
23. N. DeCastro-García, A. Muñoz Castañeda, D. García, and M. V. Carriegosm, Complexity **2019**, ID 6278908 (2019). https://doi.org/10.1155/2019/6278908
24. T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, Lect. Notes Comput. Sci. **7376**, 154 (2012). https://doi.org/10.1007/978-3-642-31537-4_13
25. Y. Qi, I. C. Zhang, Random Forest for Bioinformatics, in Ensemble Machine Learning, ed. Y. Ma (Springer, 2012) pp. 307–323. https://doi.org/10.1007/978-1-4419-9326-7_11
26. M. M. Najafabadi, T. M. Khoshgoftaar, F. Villanustre, and J. Holt, J. Big Data **4**, 22 (2017). https://doi.org/10.1186/s40537-017-0084-5

27. W. K. Michael and A. G. Thomas, Diagn. Progn. Res. **2**, 2 (2018).
28. M. S. Bannick, M. McGaughey, and A. D. Flaxman, Int. J. Epidemiol. **49**, 2065 (2019). https://doi.org/10.1093/ije/dyz223
29. J. S. Sonawane and D. R. Patil, Int. Conf. on Information Communication and Embedded Systems (ICICES2014) (2014) pp. 1-6. https://doi.org/10.1109/ICICES.2014.7033860
30. J. R. Florence, S. S. Priyadharsini, and G. S. Chandran, J Sci. Res. **14**, 189 (2022). https://doi.org/10.3329/jsr.v14i1.54739
31. Y. Y. Lee and H. M. Derakhshan, Arc. Iran Med. **16**, 358 (2013).
32. A. K. Barad, S. K. Mandal, H. S. Harsha, B. M. Sharma, and T. S. Singh, J. Gastroint. Oncol. **5**, 142 (2014).
33. P. D. Rajesh, M. Garima, M. Sharayu, and B. B. Yeole, Ind. J. Med. Paediatr. Oncol. **32**, 1 (2011).
34. M. Mridul, R. K. Devi, K. R. Phukan, T. Kaur, M. Deka, L. Puia, D. Barua, J. Mahanta, and K. Narain, Asian Pac. J. Cancer Prev. **13**, 4725 (2012). https://doi.org/10.7314/APJCP.2012.13.9.4725
35. S. Gupta, R. Gupta, D. N. Sinham and R. Mehrotra, Ind. J. Med. Res. **148**, 56 (2018). https://doi.org/10.4103/ijmr.IJMR_2023_17
36. R. McNamee, Occup. Environ. Med. **62**, 500 (2005). https://doi.org/10.1136/oem.2002.001115
37. J. C. Lin, C. C. Wang, R. S. Jiang, W. Y. Wang, and S. A. Liu, Sci. Rep. **6**, ID 22614 (2016).
38. S. C. Su, L. C. Chang, C. W. Lin, M. K. Chen, C. P. Yu, W. H. Chung, and S. F. Yang, Hum. Genet. **138**, 1379 (2019). https://doi.org/10.1007/s00439-019-02083-9