Penerbit Akademia Baru

Journal of Advanced Research in Computing and Applications

Journal homepage: www.akademiabaru.com/arca.html
ISSN: 2462-1927

Categorization of Early Detection Classifiers for Gastric Carcinoma through Data Mining Approaches

Shanmuga Pillai Murutha Muthu[1,*], Sellappan Palaniappan[1]

[1] Department of Information Technology, School of Science and Engineering, Malaysia University of Science and Technology, 47810 Petaling Jaya, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 20 July 2023
Received in revised form 22 August 2023
Accepted 9 September 2023
Available online 29 September 2023

Keywords:
Gastric carcinoma; logistic regression; prediction

ABSTRACT

Gastric carcinoma, a prevalent and potentially fatal malignancy, underscores the urgency of early detection for effective treatment. Data mining methods, renowned for their ability to extract meaningful insights from large and complex datasets, offer a promising avenue for improving early detection accuracy. In this study, we embarked on a comprehensive exploration of the identification and evaluation of classifiers using data mining techniques to enhance the early detection of gastric carcinoma. The primary objective of our study was to leverage advanced data mining techniques to predict the early diagnosis of stomach cancer, a critical factor in improving patient outcomes. Early detection can significantly impact treatment success rates and patient survival. To achieve this goal, we employed a range of classification algorithms, including Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Decision Tree (DT), and Logistic Regression. These algorithms were selected for their established efficacy in handling diverse datasets and their potential to uncover intricate patterns that may contribute to the early identification of gastric carcinoma. The preliminary evaluation of these classifiers involved the use of key performance metrics such as accuracy, precision, F1 score, and confusion metrics. These metrics are crucial for assessing the reliability and effectiveness of the classification models in distinguishing between individuals with stomach cancer and those without. The results of this preliminary analysis provide valuable insights into the strengths and limitations of each algorithm in the context of early gastric carcinoma detection. The detailed findings and classifier comparisons are presented in this paper, offering a comprehensive overview of the performance of each algorithm. This comparative analysis allows us to discern the most promising approach for early detection based on the specific characteristics of the dataset under consideration. The significance of this research lies in its potential to contribute to the development of robust and accurate screening tools for gastric carcinoma, ultimately improving the prognosis and treatment outcomes for individuals at risk. In conclusion, our study highlights the potential of data mining techniques in enhancing the early detection of gastric carcinoma. The results obtained through the evaluation of classification algorithms provide a foundation for further research and the development of practical tools that can aid in the timely diagnosis of stomach cancer, ultimately saving lives and improving overall healthcare outcomes.

[*] *Corresponding author.*
E-mail address: shanmuga@must.edu.my (Shanmuga Pillai Murutha Muthu)

1. Introduction

Gastric carcinoma, commonly referred to as stomach cancer, is a formidable global health concern characterized by its prevalence, aggressive nature, and impact on patient outcomes. This malignancy originates in the lining of the stomach and can exhibit various subtypes, each with distinct clinical features and prognoses. The growth of cancer cells in the stomach is shown in Figure 1.

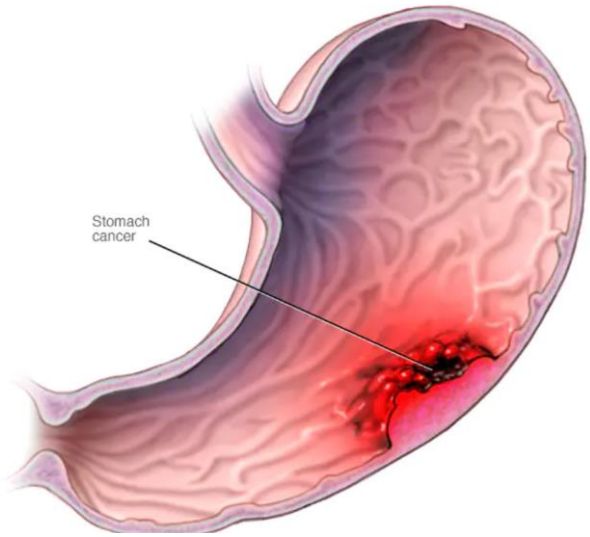

Fig. 1. The growth of cancer cell

Gastric carcinoma is associated with significant morbidity and mortality rates worldwide, making it a pressing challenge for healthcare systems and oncology research [1]. Gastric carcinoma is one of the most diagnosed cancers worldwide. According to the latest available data, it ranks as the fifth most common cancer and the fourth leading cause of cancer-related deaths globally [2]. Its occurrence varies geographically, with higher incidence rates reported in certain regions, including Eastern Asia, parts of South America, and Eastern Europe.

The significant geographic disparities in gastric carcinoma incidence can be attributed to a combination of genetic, environmental, dietary, and lifestyle factors. Regions with a high prevalence of Helicobacter pylori infection, a bacterium linked to the development of gastric cancer, often exhibit elevated rates of this malignancy [3].

One of the challenges posed by gastric carcinoma is its often-insidious development and tendency to remain asymptomatic in its early stages. This delayed onset of noticeable symptoms contributes to cases frequently being diagnosed at advanced stages when treatment options are more limited and less effective. As a result, the five-year survival rates for patients diagnosed with advanced gastric carcinoma are notably lower compared to those diagnosed at earlier stages [4].

1.1 Data Mining Techniques in the Early Detection of Gastric Carcinoma

Early detection of diseases, including gastric carcinoma, holds immense significance in healthcare. When combined with advanced data mining prediction methods, early detection has the potential to profoundly impact patient outcomes and significantly improve the effectiveness of treatments. This is particularly true for gastric carcinoma, where timely identification of the disease using data-driven approaches can lead to enhanced patient survival rates and more favorable therapeutic options.

Early-stage detection of gastric carcinoma provides patients with a broader range of treatment choices, including less invasive and more targeted interventions [5]. When cancer is identified at an advanced stage, it tends to have already spread to nearby lymph nodes or distant organs, limiting the curative options available [6]. In contrast, when detected early, the tumor may be localized, enabling more conservative surgical procedures, organ-sparing techniques, and a higher likelihood of complete tumor removal [7].

Traditional diagnostic methods for gastric carcinoma, while valuable, are associated with several challenges and limitations that can impact their accuracy, reliability, and effectiveness in detecting the disease [8]. The main challenges and limitations of traditional diagnostic methods for gastric carcinoma such as Late-Stage Diagnosis, Non-Specific Symptoms, Invasive Procedures, Limited Accessibility, Limited Sensitivity and Specificity, Lack of Surveillance Techniques, Variability in Disease Presentation and Cost and Time Factors [9]. These challenges highlight the need for more advanced and innovative approaches, such as those involving data mining techniques.

The main objective of this research is to leverage data mining techniques to develop and evaluate effective classification models for the early detection of Gastric Carcinoma. The higher result obtained from the performance metrics under cross validation 10 folds will be selected as the best models to be used in the prediction of gastric carcinoma at the early detection. Cross-validation is a fundamental technique in evaluating prediction modules. It reduces biases, provides insights into the model's generalization ability, assists in hyperparameter tuning, and offers a more reliable assessment of how well the model will perform on new and unseen data [10].

2. Methodology

2.1 Overview

Data mining is a multidisciplinary field that involves the exploration and analysis of large datasets to discover hidden patterns, trends, correlations, and actionable insights that might not be immediately evident through traditional methods. It employs a combination of statistical techniques, machine learning algorithms, and domain knowledge to sift through vast amounts of data and extract valuable information [11]. The primary goal of data mining is to transform raw data into useful knowledge, enabling informed decision-making and predictive modeling across various domains.

Machine learning algorithms learn from data by identifying patterns, trends, and relationships within the information provided. The goal is to generalize from the data to perform well on new, unseen data. Machine learning algorithms excel at making predictions and decisions based on the patterns they have learned. They can predict future outcomes, classify data into categories, and recommend actions.

This research begins with a study on the background and understanding of data mining and relevant past investigations around topics like cancers. The outcome of this phase will influence the next phase where data can be collected. Once the data has been collected, we will process the data using pre-processing methods such as data missing and identifying data duplicates. The following phase will feature extraction where the correlation will be collected among specific attributes from the dataset. The next stage will be the classification method. For the classification algorithms, Python is selected to produce the desired models and perform the evaluation. Users may use Python to evaluate different machine learning methods on a dataset that includes a variety of visualization tools and techniques. It includes graphical user interfaces for quick access to this capability, as well as predictive modelling and data analysis. We will perform the 10 folds cross validation based on the obtained data, which separates the training and testing data. Finally, we will use supervised learning

classifier results such as accuracy, precision, recall and f-measure to compare the outcomes of the cross validation.

2.2 Data Collection and Processing

In this study, data was collected using a quantitative technique from people who had stomach cancer and without cancer and qualitative techniques by having some conversations with Chief doctors or cancer specialists about the relevant components. The information was collected from hospital patients with the help of the National Cancer Institute (NCI), Malaysia. Overall, 170 data have been collected. The dataset contains 114 Males and 56 Females. For the pre-processing and applying a few data mining algorithms, categorical data is converted to numerical data. The outcome for the actual class (YES) became as (1) and (NO) became as (0). Figure 2 displays a thorough description produced by JupyterNote [12] of the genders impacted by cancer or not.

```
# Choose the column for which you want to count the class labels
class_label_column = 'Cancer Results'  # Replace with the actual column name

# Count the occurrences of each class label
class_label_counts = data[class_label_column].value_counts()

# Display the counts of class labels
print("Class Label Counts:")
print(class_label_counts)
```

```
Class Label Counts:
0    116
1     54
Name: Cancer Results, dtype: int64
```

Fig. 2. Number of cancer and non-cancer

The dataset has 29 variables in total. Age and gender are the only other factors; the rest are signs and symptoms of stomach cancer. These variables are pieces of information or attributes related to individuals within the dataset. Each variable serves a specific purpose in understanding and analysing the data. Age represents the age of the individuals in the dataset, which is a continuous numerical variable. Gender is a categorical variable, typically represented as "Male" or "Female," that indicates the gender of everyone. Most of the variables in the dataset (beyond age and gender) are related to signs and symptoms of stomach cancer. These variables likely represent various medical indicators or observations that can be associated with stomach cancer. Examples of such variables could include abdominal pain, nausea, vomiting, weight loss, and others. The last variable in the dataset is the class label. This variable serves as the target or outcome variable for the dataset. It indicates whether an individual has been affected by stomach cancer or not. This variable is typically binary, with two possible values: "Yes" (indicating the presence of stomach cancer) or "No" (indicating the absence of stomach cancer). Figure 3 displays a description of the data.

	Count	Mean	STD	Min	25%	50%	75%	Max
Abdominal Pain	170	0.941176	0.235989	0	1	1	1	1
Nausea	170	1	0	0	1	1	1	1
Skin Color Turn into Plae	170	0.105882	0.308596	0	1	1	1	1
Eat Yellow Foods	170	0.235294	0.425436	0	1	1	1	1
Frequent Vomiting	170	0.641176	0.481072	0	1	1	1	1
Family History	170	0.317647	0.466937	0	1	1	1	1
Helicobacter Pylori Infection	170	0.217647	0.413865	0	1	1	1	1
Smoking	170	0.682353	0.466937	0	1	1	1	1
Alcohol Consumption	170	0.682353	0.466937	0	1	1	1	1
NSAID(Long-term use of certain medication)	170	0.411765	0.493607	0	1	1	1	1
Stomach Surgery	170	0.105882	0.308596	0	1	1	1	1
High Salt	170	0.811765	0.392055	0	1	1	1	1
Pickled Food	170	0.7	0.459611	0	1	1	1	1
Processed Food	170	0.911765	0.284475	0	1	1	1	1
Carbonated Drinks	170	0.911765	0.284475	0	1	1	1	1
Low Diet Fruits and Vegetables	170	0.7	0.459611	0	1	1	1	1
Obesity or Overweight	170	0.652941	0.477441	0	1	1	1	1
Factory Worker	170	0.329412	0.471388	0	1	1	1	1
Blood Group A	170	0.317647	0.466937	0	1	1	1	1
Exposure to certain chemicals or subtances	170	0.8	0.401182	0	1	1	1	1
Pernicious Anmeia	170	0.370588	0.484389	0	1	1	1	1
Frequent Heartburn	170	0.970588	0.169457	0	1	1	1	1
Gastric	170	0.870588	0.336647	0	1	1	1	1
Missed Meals till stomach pain	170	0.676471	0.469205	0	1	1	1	1
GERD	170	0.5	0.501477	0	1	1	1	1
Black Vomit Colour	170	0.541176	0.499774	0	1	1	1	1
Black Stool Colour	170	0.394118	0.490104	0	1	1	1	1

Fig. 3. Data description of dataset

The dataset contains 114 Males and 56 Females. The detailed description of the gender who got affected with Cancer or not shown in Figure 4.

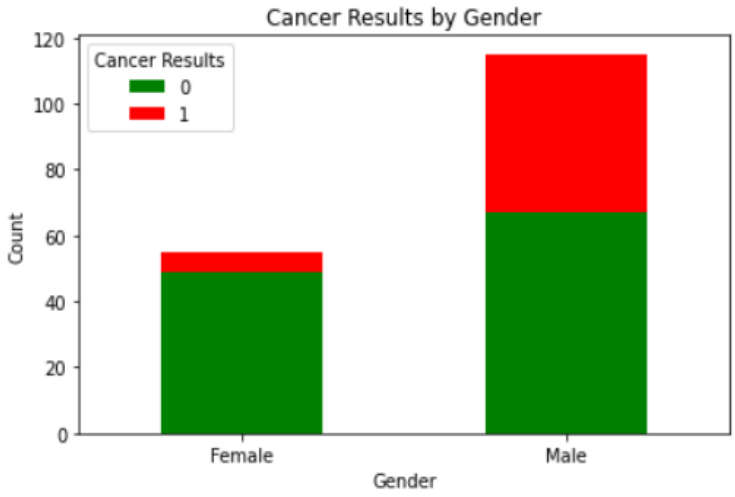

Fig. 4. Gender who got affected Cancer or Non-Cancer

Apart from that, age been considered for gastric carcinoma as well. Figure 5 shows the age range who got affected with the gastric carcinoma.

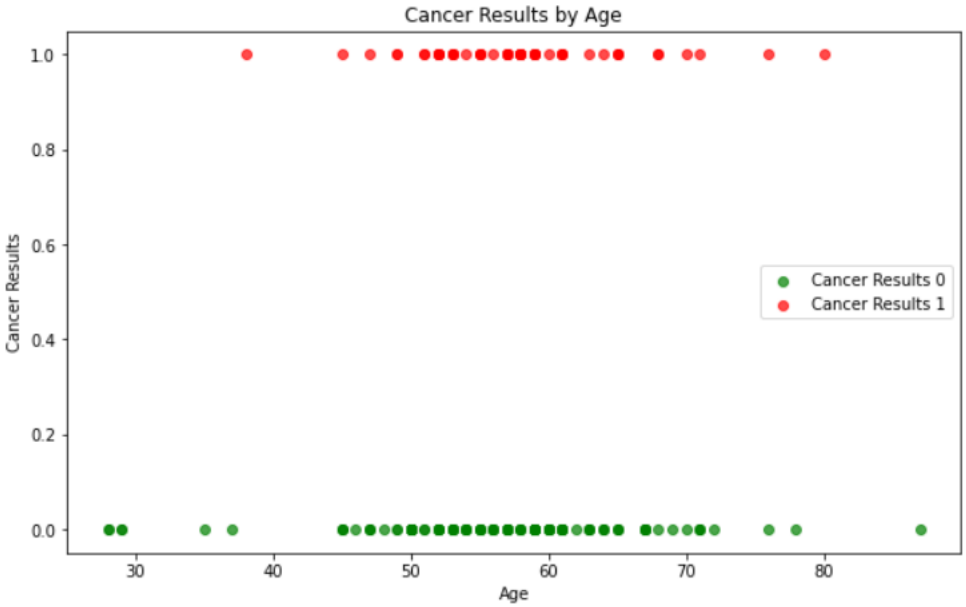

Fig. 5. Age who got cancer

2.3 Model Selection

Four different supervised machines learning i.e. Support Vector Machine, K-NN, Decision Tree and Logistic Regression have been used for analysis the dataset. Python tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. Python's simplicity, rich ecosystem of libraries, and GUI development capabilities make it an ideal choice. Data processing and analysis are simplified with libraries like pandas and NumPy, enabling efficient preprocessing and feature engineering [13]. The availability of machine learning libraries like scikit-learn allows easy implementation, tuning, and evaluation of classification algorithms. Visualizations with Matplotlib and Seaborn aid in result interpretation [14]. The interactivity of a GUI tool enhances user engagement, allowing input, parameter adjustment, and visualized predictions.

2.3.1 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification and regression tasks. It works by finding a hyperplane in a high-dimensional space that best separates data points belonging to different classes [15]. SVMs are widely used in various domains due to their effectiveness in handling both linear and non-linear classification problems. It aims to find a hyperplane in a high-dimensional space that best separates data points of different classes, while maximizing the margin between them. Here's a summary of SVM with its key formula: Formula:

Hyperplane equation: In a binary classification problem, the hyperplane equation can be represented as Eq. (1)

$$f(x) = W^T X + b \tag{1}$$

$f(x)$ represents the decision function or decision boundary. It takes an input vector X and produces an output.

W^T represents the transpose of the weight vector W. The weight vector contains the coefficients corresponding to each feature in the input vector X.

X This is the input vector, representing the features of the data point being classified.

b the bias term, also known as the intercept. It is an additional parameter that helps shift the decision boundary.

The distance between the hyperplane and the nearest support vector is the margin. The objective of SVM is to maximize the margin while correctly classifying data points. In the case of non-linearly separable data, SVM introduces slack variables ξᵢ to allow for misclassification or overlapping. SVM's effectiveness stems from finding a hyperplane that maximizes the margin between classes, resulting in improved generalization to new data. It is flexible through kernel functions, which map data to higher-dimensional spaces, allowing non-linear separation [16]. SVMs are employed in various domains, such as image classification, bioinformatics, and finance, due to their robustness and versatility.

2.3.2 K-Nearest Neighbors algorithm (k-NN)

The K-Nearest Neighbors (k-NN) algorithm is a simple yet effective machine learning method used for classification and regression tasks. It makes predictions based on the majority class of the k-nearest data points in the feature space [17]. The k-NN algorithm relies on a distance metric, often the Euclidean distance, to measure the similarity between data points.

For two data points, A and B, with n features (dimensions), the Euclidean distance is calculated Eq. (2)

$$Distance(A,B) = \sqrt{\left(x_{A1} - x_{B_1}\right)^2 + \left(x_{A2} - x_{B_2}\right)^2} + \cdots + \quad (2)$$

Once distances are computed, the k-NN algorithm selects the k-nearest neighbors with the smallest distances to the test data point [18]. In the case of classification, the class labels of these k neighbors are considered, and the class that occurs most frequently becomes the prediction for the test data point.

2.3.2 Decision Tree (DT)

A Decision Tree is a widely used machine learning algorithm for classification and regression tasks. It works by partitioning the feature space into segments and making decisions based on the values of input features. Decision Trees are easy to understand, interpret, and visualize, making them useful for both explanatory and predictive tasks. Decision Trees create a tree-like structure where each internal node represents a decision based on a feature, each branch represents an outcome of that decision, and each leaf node represents a class label (for classification) or a prediction (for regression) [19]. The goal is to construct a tree that maximizes information gain or minimizes impurity at each split, resulting in a tree that accurately classifies or predicts unseen data. For classification, the formula for information gain (used in decision tree splits) is Eq. (3)

Information Gain = Entropy(parent) - ∑ [(count(child) / count(parent)) * Entropy(child)] (3)

where:

- Entropy(parent) is the entropy of the parent node
- count(child) is the number of samples in the child node
- count(parent) is the number of samples in the parent node
- Entropy(child) is the entropy of the child node

The goal is to choose splits that maximize the information gain, leading to more homogeneous child nodes. For regression, Decision Trees use other measures like mean squared error (MSE) to evaluate the quality of splits and make predictions [20]. In summary, Decision Trees are intuitive, interpretable models that create a tree-like structure to make decisions based on input features. The algorithm aims to split the feature space in a way that maximizes information gain (or minimizes impurity) at each step. This process results in a tree that can be used for classification or regression tasks.

2.3.3 Logistic Regression

Logistic Regression is a widely used statistical method for binary classification, which aims to predict the probability that an input belongs to a particular class. Unlike linear regression, which predicts continuous values, logistic regression models the probability of a categorical outcome using the logistic function [20]. Logistic Regression is a statistical technique used for binary classification, where the goal is to predict the probability of an input belonging to one of two classes. The method employs the logistic function to transform the linear combination of input features into a probability value between 0 and 1. This probability is then used to classify the input into one of the two classes based on a predefined threshold. The logistic function, also known as the sigmoid function, is the key component of logistic regression. It transforms the linear combination of input features (represented by the term 'z') into a probability value (represented by 'p'). The logistic function is given by Eq. (4)

$$\rho = \frac{1}{1+e^{-z}} \tag{4}$$

Once the predicted probability ρ is obtained, a threshold (typically 0.5) is used to make the final classification decision. If ρ is greater than or equal to the threshold, the input is classified as the positive class; otherwise, it's classified as the negative class.

Logistic regression is a valuable tool for binary classification tasks, employing the logistic function to model the probability of an input belonging to a particular class. Its simplicity, interpretability, and ability to estimate probabilities make it a popular choice in various fields [21].

3. Results

In this research paper, the results section presented the performance metrics such as Accuracy, Precision, and F1 Measure for four classifiers been mentioned earlier used in the study. The following classification algorithms were applied to the dataset for the early detection of gastric carcinoma using by split the data into training and testing sets. By creating the four models, loop through each model, train it, make predictions, and compute metrics.

In the metrics performance, ratio. `test_size=0.4` means that approximately 40 percent of samples will be assigned to the test data, and the remaining 60 percent will be assigned to the training data. The result for 40% test data and 60% training data shown in Table 1.

Table 1
Metrics Table

Model	Accuracy	Precision	Recall	F1 Score	Confusion Matix
SVM	0.955882	0.9	1	0.947368	[[38 3] [0 27]]
KNN	0.911765	0.818182	1	0.9	[[35 6] [0 27]]
Decision Tree	0.955882	0.9	1	0.947368	[[38 3] [0 27]]
Logistic Regression	0.955882	0.9	1	0.947368	[[38 3] [0 27]]

Using 40 percent testing data and 60 percent training data, the SVM, Decision Tree, and Logistic Regression perform better in terms of accuracy, precision, recall, and F1 Score.

Table 2
Metrics Table

Model	Accuracy	Precision	Recall	F1 Score	Confusion Matix
SVM	0.911765	0.870968	0.84375	0.857143	[[66 4] [5 27]]
KNN	0.882353	1	0.625	0.769231	[[70 0] [12 20]]
Decision Tree	0.882353	1	0.625	0.769231	[[70 0] [12 20]]
Logistic Regression	0.960784	0.888889	1	0.941176	[[66 4] [0 32]]

According to Table 2, while using 60% testing data and 40% training data, Logistic Regression delivers improved accuracy, recall, and F1 Score outcomes. KNN and Decision Tree outperform SVM and KNN in terms of precision.

Table 3
Metrics Table

Model	Accuracy	Precision	Recall	F1 Score	Confusion Matix
SVM	0.933824	0.9	0.878049	0.888889	[[91 4] [5 36]]
KNN	0.970588	0.911111	1	0.953488	[[91 4] [0 41]]
Decision Tree	0.933824	0.9	0.878049	0.888889	[[91 4] [5 36]]
Logistic Regression	0.970588	0.911111	1	0.953488	[[91 4] [0 41]]

According to Table 3, when using 80% testing data and 20% training data, Logistic regression and KNN perform better in terms of accuracy, precision, recall, and F1 Score.

Table 4
Results table

Model	Accuracy	Precision	Recall	F1 Score	Confusion Matix
SVM	0.947059	0.909091	0.925926	0.917431	[[111 5] [4 50]]
KNN	0.929412	0.903846	0.87037	0.886792	[[111 5] [7 47]]
Decision Tree	0.929412	0.903846	0.87037	0.886792	[[111 5] [7 47]]
Logistic Regression	0.947059	0.909091	0.925926	0.917431	[[111 5] [4 50]]

According to Table 4, SVM and Logistic Regression perform better for cross validation 10 folds in terms of accuracy, precision, recall, and F1 Score metrics.

The evaluation results suggest that Logistic Regression outperforms Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Decision Tree classifiers across various testing scenarios, including data splits of 40%, 60%, and 80%. Furthermore, during cross-validation, both Logistic Regression and SVM exhibit superior performance, with Logistic Regression consistently demonstrating better results in terms of accuracy, precision, recall, and F1 Score.

The accuracy of 94.7% achieved by Logistic Regression in cross-validation indicates a high overall correctness in predicting stomach cancer cases. Precision, measuring the ratio of correctly predicted positive observations to the total predicted positives, is impressive at 90.9%, indicating a low rate of false positives. Recall, capturing the ratio of correctly predicted positive observations to all actual positives, is also commendable at 92.59%, signifying the model's ability to effectively capture instances of stomach cancer. The F1 Score, which balances precision and recall, is at a strong 91.7%.

The consistent superior performance of Logistic Regression across different data splits and cross-validation underscores its robustness and reliability in predicting stomach cancer. The decision to favour Logistic Regression over SVM, K-NN, and Decision Tree classifiers can be attributed to its ability to effectively model the underlying patterns in the data, achieving a balance between precision and recall.

In the context of gastric cancer prediction, where the objective is to identify potential cases with high accuracy and minimize false positives, Logistic Regression's superior performance is particularly valuable. The high precision implies that when the model predicts a positive case, it is highly likely to be accurate. Similarly, the high recall suggests that the model can successfully identify a significant proportion of actual positive cases.

4. Conclusions

According to the investigation, using machine learning algorithms can provide more precise and quantitative methods for detecting stomach cancer early on. A useful tool for medical professionals in the diagnosis of cancer, the Logistic Regression Classifier and SVM Classifier model displays notable precision in detecting early detection. These models can be used to diagnose gastric cancer early, saving the lives of many people who still have no idea what is happening within their stomachs but only experience the most rudimentary symptoms. The application of these classifiers in early detection is particularly crucial as it addresses the challenge of identifying cases where patients may not be aware of the underlying pathology. By providing an efficient and accurate means of diagnosis, these models empower medical professionals to intervene at an early stage, potentially saving lives and improving overall patient outcomes.

In summary, the main objective of this study is to leverage the power of data mining methods to develop accurate and efficient classifiers for the early detection of Gastric Carcinoma. By achieving this objective, the study aims to contribute to the advancement of medical diagnosis and treatment strategies, ultimately improving patient outcomes in the fight against stomach cancer.

Acknowledgement
This research was not funded by a grant.

References

[1] Sexton, Rachel E., Mohammed Najeeb Al Hallak, Maria Diab, and Asfar S. Azmi. "Gastric cancer: a comprehensive review of current and future treatment strategies." *Cancer and Metastasis Reviews* 39 (2020): 1179-1203. https://doi.org/10.1007/s10555-020-09925-3

[2] Rawla, Prashanth, and Adam Barsouk. "Epidemiology of gastric cancer: global trends, risk factors and prevention." *Gastroenterology Review/Przegląd Gastroenterologiczny* 14, no. 1 (2019): 26-38. https://doi.org/10.5114/pg.2018.80001

[3] Mentis, Alexios-Fotios A., Marina Boziki, Nikolaos Grigoriadis, and Athanasios G. Papavassiliou. "Helicobacter pylori infection and gastric cancer biology: tempering a double-edged sword." *Cellular and Molecular Life Sciences* 76 (2019): 2477-2486. https://doi.org/10.1007/s00018-019-03044-1

[4] Li, Yunmei, Aozi Feng, Shuai Zheng, Chong Chen, and Jun Lyu. "Recent estimates and predictions of 5-year survival in patients with gastric cancer: A model-based period analysis." *Cancer Control* 29 (2022): 10732748221099227. https://doi.org/10.1177/10732748221099227

[5] Chen, Zhi-da, Peng-Fei Zhang, Hong-Qing Xi, Bo Wei, Lin Chen, and Yun Tang. "Recent advances in the diagnosis, staging, treatment, and prognosis of advanced gastric cancer: a literature review." *Frontiers in Medicine* 8 (2021): 744839. https://doi.org/10.3389/fmed.2021.744839

[6] Martin, Tracey A., Lin Ye, Andrew J. Sanders, Jane Lane, and Wen G. Jiang. "Cancer invasion and metastasis: molecular and cellular perspective." In *Madame Curie Bioscience Database [Internet]*. Landes Bioscience, 2013.

[7] Sosnowski, Roman, Marcin Kuligowski, Olga Kuczkiewicz, Katarzyna Moskal, Jan Karol Wolski, Marc A. Bjurlin, James S. Wysock, Piotr Pęczkowski, Chris Protzel, and Tomasz Demkow. "Primary penile cancer organ sparing treatment." *Central European Journal of Urology* 69, no. 4 (2016): 377. https://doi.org/10.5173/ceju.2016.890

[8] Alzubaidi, Laith, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions." *Journal of big Data* 8 (2021): 1-74. https://doi.org/10.1186/s40537-021-00444-8

[9] Waddingham, William, Stella AV Nieuwenburg, Sean Carlson, Manuel Rodriguez-Justo, Manon Spaander, Ernst J. Kuipers, Marnix Jansen, David G. Graham, and Matthew Banks. "Recent advances in the detection and management of early gastric cancer and its precursors." *Frontline Gastroenterology* (2020). https://doi.org/10.1136/flgastro-2018-101089

[10] Xiong, Zheng, Yuxin Cui, Zhonghao Liu, Yong Zhao, Ming Hu, and Jianjun Hu. "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation." *Computational Materials Science* 171 (2020): 109203. https://doi.org/10.1016/j.commatsci.2019.109203

[11] Shu, Xiaoling, and Yiwan Ye. "Knowledge Discovery: Methods from data mining and machine learning." *Social Science Research* 110 (2023): 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

[12] B. Ganger and F. Pérez, "JupyterLab," *jupyter*, 2022.

[13] Singh, Jaspreet, Sarada Prasad Pradhan, Mahendra Singh, and Bingxiang Yuan. "Modified block shape characterization method for classification of fractured rock: A python-based GUI tool." *Computers & Geosciences* 164 (2022): 105125. https://doi.org/10.1016/j.cageo.2022.105125

[14] D. P. Yin *et al.*, *Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython*. 2001.

[15] Vanneschi, Leonardo, and Sara Silva. "Support Vector Machines." In *Lectures on Intelligent Systems*, pp. 271-281. https://doi.org/10.1007/978-3-031-17922-8_10 Cham: Springer International Publishing, 2023.

[16] Cristianini, Nello, and Bernhard Scholkopf. "Support vector machines and kernel methods: the new generation of learning machines." *Ai Magazine* 23, no. 3 (2002): 31-31.

[17] Hidayati, Nur, and Arief Hermawan. "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation." *Journal of Engineering and Applied Technology* 2, no. 2 (2021): 86-91. https://doi.org/10.21831/jeatech.v2i2.42777

[18] Kenyhercz, Michael W., and Nicholas V. Passalacqua. "Missing data imputation methods and their performance with biodistance analyses." In *Biological Distance Analysis*, pp. 181-194. Academic Press, 2016. https://doi.org/10.1016/B978-0-12-801966-5.00009-3

[19] Yan, Kang, Song Jinling, Bian Mingming, Feng Haipeng, and Mohamed Salama. "Red tide monitoring method in coastal waters of Hebei Province based on decision tree classification." *Applied Mathematics and Nonlinear Sciences* 7, no. 1 (2022): 43-60. https://doi.org/10.2478/amns.2022.1.00051

[20] Myles, Anthony J., Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. "An introduction to decision tree modeling." *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, no. 6 (2004): 275-285. https://doi.org/10.1002/cem.873

[21] Hoffman, Julien IE. *Biostatistics for medical and biomedical practitioners*. Academic press, 2015.